QoE-based Traffic and Resource Management for Adaptive HTTP Video Delivery in LTE

Ali El Essaili, *Student Member, IEEE,* Damien Schroeder, *Student Member, IEEE,* Eckehard Steinbach, *Senior Member, IEEE,* Dirk Staehle, and Mohammed Shehada, *Member, IEEE*

***Abstract*—There is a growing interest in over-the-top (OTT) dynamic adaptive streaming over HTTP (DASH) services. In mobile DASH, a client controls the streaming rate and the base station in the mobile network decides on the resource allocation. Different from the majority of previous works which focus on client-based rate adaptation mechanisms, this paper investigates the mobile network potential for enhancing the user Quality-of-Experience (QoE) in multi-user OTT DASH. Specifically, we first present proactive and reactive QoE optimization approaches for adapting the adaptive HTTP video delivery in an LTE network. We then show, using subjective experiments, that by taking a proactive role in determining the transmission and streaming rates, the network operator can provide a better video quality and a fairer QoE across the streaming users. Furthermore, we consider the playout buffer time of the clients and propose a novel playout buffer-dependent approach that determines for each client the streaming rate for future video segments according to its buffer time and the achievable QoE under current radio conditions. In addition, we show that by jointly solving for the streaming and transmission rates, the wireless network resources are more efficiently allocated among the users and substantial gains in the user perceived video quality can be achieved.**

***Index Terms*—adaptive HTTP streaming, DASH, QoE, video transport, LTE, resource allocation.**

I. Introduction

The lion share of mobile traffic is dominated by streaming of video and audio, all delivered over the top (OTT) [1]. More specifically, TCP/IP based streaming is dominant over traditional RTP/UDP based streaming. Indeed, YouTube alone accounts for 27% of the mobile downlink traffic in North America at peak hours [1]. This represents a pragmatic shift in multimedia streaming to ensure higher transmission reliability. In this context, Dynamic Adaptive Streaming over HTTP (DASH [2]) is emerging as the new standard for mobile multimedia streaming which utilizes TCP/IP and offers intra session rate adaptation capable to deal with the variability of wireless networks. It mitigates the playout interruptions and initial buffering delays, encountered in progressive download, used for example in YouTube.

Mobile video is identified as a main reason for congestion in mobile networks [3]. The proliferation of powerful mobile devices and resource-demanding multimedia applications is outpacing the capacity enhancements in next generation wireless networks. It is crucial for mobile network operators to guarantee a high Quality of Experience (QoE), in particular for adaptive HTTP video streaming. The Quality of Service (QoS) control mechanisms for DASH support in mobile networks are based on mapping functions that translate application-specific information into QoS parameters (e.g., QoS Class Identifiers (QCIs), guaranteed bit-rate (GBR)) [4] [5]. This allows the network to establish bearers with correct characteristics for DASH users [6] [7] but does not reflect the dynamic and variable characteristics of the video contents during transmission. Moreover, the majority of video streaming users use non-GBR bearers [3]. Consequently, a need arises for user-centric approaches to dynamically adapt the adaptive HTTP streaming (segment granularity) and which complement the QoS mechanisms in the mobile network. As a result, recent standardization activities have focused on introducing optimizing modules in the radio access network and core network which allow to cache and adapt the multimedia transmission in the mobile cell [3] [8] (Section VII-C).

So far, DASH has been mainly studied from an end-to-end client-server perspective [9] [10]. In DASH, a video stream is encoded at different representations which are accessed through a standard HTTP server. A client uses HTTP requests to download the representation that matches its transmission capacity. The base station in the mobile network determines the resource allocation but acts as a black box in the DASH server-client system. This brings along several challenges. First, the DASH client adapts its video quality to the resources allocated by the wireless scheduler. The scheduler, however, is typically not content-aware and assigns resources based on the channel conditions without considering the characteristics of the transported content. Second, the clients are selfish and make decisions irrespective of other clients sharing the network resources. Indeed, recent studies show that the DASH client behaviour results in network under-utilization, fluctuating and unfair throughput allocation [11] [12]. Huang et al. [13] shows that inaccurate throughput estimation at the client can lead to a variable and lower video quality.

To address these issues, a line of literature focuses on improving the rate control logic at the client [14] [15] [16] [17] or server [18] for stabilizing the client's behaviour and efficiently using the network resources. All of these approaches, however, are specific for a single client. Prior work on multi-user resource allocation for adaptive HTTP video delivery has shown that the streaming performance can be

A preliminary version of this work has been published at the IEEE International Conference on Communications (ICC), June 2013.

A. El Essaili, D. Schroeder, and E. Steinbach are with the Institute for Media Technology, Technische Universität München (TUM), Munich, Germany (e-mail: elessaili@tum.de; damien.schroeder@tum.de; eckehard.steinbach@tum.de).

D. Staehle and M. Shehada are with DOCOMO Communications Laboratories Europe GmbH, Munich, Germany (e-mail: staehle@docomolab-euro.com; shehada@docomolab-euro.com).

improved by jointly optimizing the network resources among multiple clients [19] [20] [21]. The investigated approaches can be classified as reactive, i.e., they focus on optimizing the resource allocation to meet an objective criteria and the clients react to the assigned network resources.

This paper addresses the following challenging questions: can a mobile network operator optimize the adaptive HTTP video delivery by exploiting its knowledge on the cell load and radio conditions in a mobile network? What are the benefits of in-network traffic and resource management in the context of adaptive HTTP streaming? It bridges the gap between client-based and network-based optimization approaches by jointly optimizing the multi-user network resource allocation and the streaming rate of the DASH clients. Moreover, our objective is to proactively adapt the adaptive HTTP mobile video delivery by considering the radio conditions, content characteristics, and playout buffer levels of the clients. More specifically, the proposed optimization involves two processes: a QoE optimization in the mobile network to determine the target transmission rate for each user, and a proxy-based method to match the streaming rate of each user to the QoE optimization result. Furthermore, the proposed rate adaptation approach at the proxy makes use of the available multiple bitrate encodings within the adaptive HTTP content and thus requires no further processing of the video content. This makes our approach particularly suitable for OTT streaming services where the DASH server lies outside the operators' network.

This paper builds on our preliminary work in [22]. The main added contributions are: first, we perform subjective tests to assess the performance of our QoE-based adaptive HTTP streaming system. Then, we extend our system from [22] in two directions. We first consider the playout buffer level feedback from the clients and propose a novel QoE-driven buffer-aware approach for selecting the representation rates at the proxy. Moreover, we jointly solve for the transmission and streaming rates of the mobile users. Compared to [22], this allows a mobile operator to further optimize the network resource allocation based on the buffer demands of the users, which leads to an overall improvement in the mobile users' QoE.

The rest of the paper is organised as follows. In the next section we review related work. In Section III, we present our QoE-based optimization approach for adaptive HTTP streaming and outline our system model. In Section IV, we introduce the playout buffer-aware QoE adaptation approach. The joint streaming and transmission rate allocation problem is then formulated in Section V. Section VI describes our subjective test methodology and results. In Section VII, the simulation results are presented and practical deployment scenarios are discussed. Finally, Section VIII concludes the paper.

II. Related work

A. DASH Overview

In contrast to UDP streaming which is push-based, DASH is a pull-based client-driven streaming protocol. The client sends HTTP requests to retrieve media content that matches its available throughput using the standard HTTP protocol.

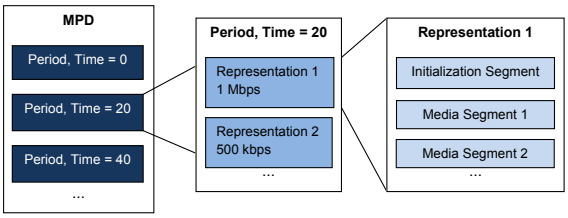

Fig. 1. Media Presentation Description (MPD) for adaptive HTTP streaming [2].

The server splits the media content into segments that can be independently decoded at the client. The protocol defines a media presentation description (MPD) for communication between an HTTP server and a streaming client (Figure 1). Each MPD is composed of one or more presentation periods. A period can include multiple representations of the same video content which correspond to different encoding characteristics (bit-rate, resolution, codec, etc.). Each representation consists of an initialization segment which provides the client with the metadata that describes the content and one or more media segments. A client can seamlessly switch between the different segments during the streaming session by adaptively adjusting the streaming rate to its estimated transmission capacity.

B. Background on QoE-based Cross-layer optimization

Cross-layer optimization approaches for resource allocation have been considered to improve the quality of service by exchanging information across the different protocol layers (e.g., [23]). Conventional cross-layer optimization (CLO) adapts the instantaneous transmission parameters on a short timescale which is not optimal from a multimedia quality perspective [24]. Application-driven cross-layer optimization, on the other hand, directly maximizes an application-specific objective function while using abstracted models for the link layer and physical layer [25] [26].

QoE-based resource allocation in wireless networks has been proposed for traditional RTP/UDP streaming (e.g., [27]). The objective of the QoE-based optimization is to find a long-term resource allocation which maximizes the overall utility based on the application and channel conditions of the users in the cell. In-network content adaptation (e.g., transcoding [28]) is then used to shape the transmitted video streams. This, however, is costly in terms of computational resources and induces additional delays. Meanwhile, DASH provides inherent adaptivity by encoding the same content at multiple bit-rates which simplifies the rate adaptation compared to RTP/UDP based optimizations.

C. Related work on adaptive HTTP video delivery

DASH gives the control of the streaming rate to the client but the rate adaptation strategies are not specified in the standard [2]. As a result, a plethora of research proposes client-based rate adaptation approaches to enhance the user perception in adaptive HTTP media delivery. Oyman et al. [29] evaluates the end-to-end QoE in adaptive HTTP streaming over LTE. A client-driven adaptation algorithm that aims at

minimizing the rebuffering events is considered. Liu et al. [30] proposes a client-driven algorithm for determining the streaming rate and additionally managing the cached segments at a proxy cache. Huang et al. [31] determines the streaming rate by only considering the buffer information. Recently, much work studied the combination of rate, utility and buffer information. Li et al. [32] considers a finite horizon with constant bandwidth and formulates an optimization problem to determine the bit-rates for a set of segments by considering the quality of the segments and the buffer information. In [33], audio-visual metadata (rate-quality information) is added to the MPD as an extension to the *Subset* element of DASH. Each client computes the optimal rates individually that maximize its audio-visual quality. [34] describes different client-driven adaptation approaches which consider both the bit-rate and the quality of the DASH content. [35] proposes a client-driven adaptation algorithm based on the available TCP throughput and buffered media time. All these approaches, however, optimize the HTTP streaming of a single client without further considering the influence on other DASH users sharing the same network resources.

Meanwhile, there is a growing interest in exploiting the multi-user adaptive HTTP streaming scenario. Recent studies of available adaptive HTTP streaming clients show inconsistent and unfair behavior when two clients are competing for resources on a shared link [11]. With this objective, [14] proposes a client-driven adaptation algorithm for adaptive HTTP video streaming with the goal of achieving fairness, efficiency and stability among multiple clients. The algorithm, however, is not QoE-aware and is reactive to observed network conditions. In [19], network traffic management for adaptive HTTP video delivery across multiple clients is considered. The target bit-rate is determined by the network based on available throughput estimates of all users. The authors in [20] conclude that a simple rate shaping policy in a residential gateway can improve the adaptive HTTP experience among two competing clients. [36] proposes a rate adaptation algorithm for optimizing the adaptive HTTP streaming across multiple wireless clients. The approach, however, does not consider the individual content characteristics of the different clients and aims at stabilizing the user throughput. Also, different from our work, the authors of [36] propose to transcode the DASH stream, similar to typical RTP/UDP based optimizations (e.g., [27]), which is costly and may react too late. Recently, [37] studied multi-user streaming in a home networking scenario. They show that by jointly optimizing the DASH rates of multiple users a fair video quality can be achieved. Different from our approach, the authors in [37] do not consider the transmission rates of the users in their optimization.

The buffered media time feedback has been also considered for optimizing the resource allocation in the mobile network. The majority of prior work focuses on minimizing the stalling events (e.g., [38], [39]). [38] proposes a traffic prioritization approach at the scheduler in the mobile network that relies on playout buffer level feedback from YouTube videos. [39] studies the multiplexing of multiple variable bit-rate videos over a time-varying wireless channel with the goal of minimizing the number of playout stalls. [40] [21] study adaptive HTTP streaming in LTE networks with the aim of minimizing the number of interruptions. However, experimental results have shown that mature adaptive clients eliminate playout interruptions in real scenarios [41]. Indeed, recent studies show that it is important to address the multi-level rate switches as a measure of user dissatisfaction in adaptive HTTP streaming [42] [43] [44].

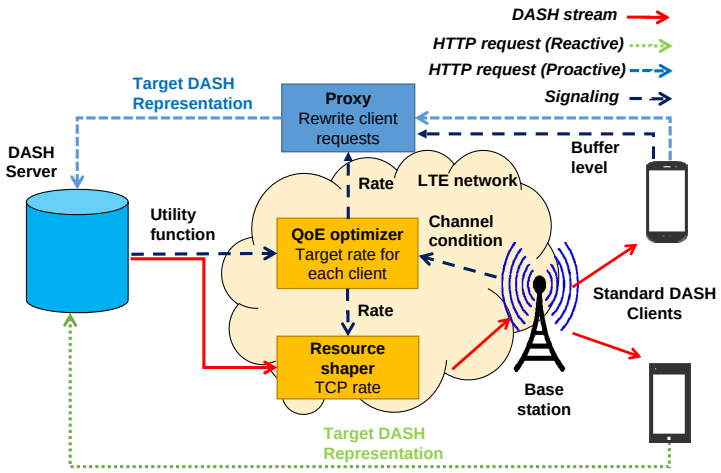

Fig. 2. Illustration of proactive and reactive optimization approaches for adaptive HTTP video delivery.

III. QoE-based Adaptive HTTP Streaming System

This paper proposes a proactive approach for optimizing the multi-user adaptive HTTP video delivery in mobile networks (Figure 2). At the DASH server, the utility information of each content is first extracted and added to the MPD. At the base station or close to it, a QoE optimizer collects utility and channel information about the different clients. Our application and radio layer models are presented in Sections III-A and III-B, respectively. The QoE optimizer then determines the target transmission rates using utility maximization, as described in Section III-C. The target rates are signaled to a proxy and a resource shaper for adapting the streaming and transmission rates of the DASH clients, respectively. To this end, we distinguish between proactive and reactive approaches for exploiting the QoE optimization result, which are described in Section III-D.

A. Application model

We express the user satisfaction or QoE for real-time video streaming on a Mean Opinion Score (MOS) scale [45]. The utility function U for video streaming is defined in [27] as a function of the application data rate R by:

$$U = f(R), f : R \rightarrow MOS \quad (1)$$

We apply a simple linear mapping between the peak signal-to-noise ratio (PSNR) and the MOS [46]. MOS can take on any value between 1.0 (30 dB) and 4.5 (42 dB), which represent the worst and best QoE, respectively. We also compare our results with more complex mappings in Section VI.

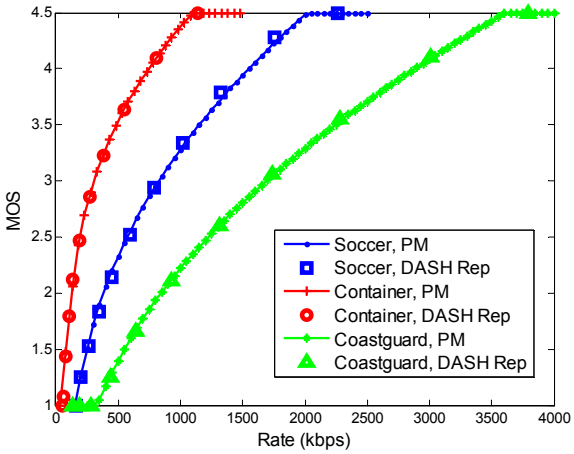

Fig. 3. Utility curves using the actual DASH representations and by fitting using the parametric model from [47].

The utility information is provided in the form of MOS-Rate pairs to the optimizer. For our optimization, the parametric model (PM) from [47] which requires three MOS-Rate pairs to model the performance of a video is considered. Figure 3 shows the utility curves for three different video sequences which correspond to the discrete MOS values of the actual DASH representations and the interpolated values using the PM from [47].

B. Radio model

We consider a long-term radio layer model with optimization cycles in the order of seconds. Our objective is to determine the resource share (i.e., physical resource blocks (PRBs) in LTE) of each client in each optimization round. This allows us to integrate our QoE-based optimization on top of the state-of-the art schedulers for LTE without the need to modify the scheduling mechanisms already deployed.

We use the link layer model originally proposed in [48]. It defines the data rate R_k for user k as a function of its resource share α_k and its maximum achievable rate $R_{max,k}$ if all the PRBs are allocated exclusively to user k.

$$R_k = g_k(\alpha_k) = \alpha_k R_{max,k} \quad 0 \leq \alpha_k \leq 1, \forall k \tag{2}$$

In each optimization round, a new $R_{max,k}$ is determined for each client based on its average channel statistics in the previous second. The channel information of the users is available to the base station in a downlink streaming scenario [49]. We use the link layer model from the 3GPP LTE recommendations [50] to determine the achievable throughput per PRB for a given Signal-to-Noise ratio (γ). The model from [50] approximates the throughput T in the downlink, after link adaptation and hybrid automatic repeat request, by an implementation loss $\beta = 0.6$ compared to the Shannon capacity. As baseline parameters, it further defines a γ_{min} of -10 dB, a γ_{max} of 23 dB and a maximum throughput T_{max} of 4.4 bps/Hz.

$$T = \begin{cases} 0 & \text{for } \gamma < \gamma_{min} \\ \beta \log_2(1+\gamma) & \text{for } \gamma_{min} \leq \gamma < \gamma_{max} \\ T_{max} & \text{for } \gamma \geq \gamma_{max} \end{cases} \tag{3}$$

C. QoE-based resource allocation

The objective of the QoE-based resource allocation is to determine the transmission rates of all clients that maximize the overall user satisfaction. In this work, we use the objective function as proposed in [51]. The optimization problem for K clients is given by:

$$\underset{(\alpha_1,\ldots,\alpha_K)}{\arg\max} \sum_{k=1}^{K} U_k(g_k(\alpha_k)) - P_k \tag{4}$$

$$subject\ to \quad \sum_{k=1}^{K} \alpha_k = 1, \quad g_k(\alpha_k) \geq R_{min,k} \tag{5}$$

$$where\ P_k = \min(0, |U_k(g_k(\alpha_k))_t - U_k(g_k(\alpha_k))_{t-1}| - \xi_{th})$$

where (4) determines the network resource share of each user that maximizes the sum of utilities. It additionally penalizes the temporal fluctuations of the video quality which are perceivable by the users. Specifically, it adds a penalty term P_k if the quality change between two successive optimization rounds (denoted by t and $t-1$) exceeds a just noticeable difference (JND) threshold ξ_{th}. In this work, we consider an average JND threshold of $\xi_{th} = 0.23$ MOS for all users which has been derived using subjective tests [51]. (5) constrains on the available resources and defines a minimum rate that should be allocated to each user (e.g., lowest representation).

Each α_k value corresponds to the fraction of total PRBs assigned to user k in each optimization round. A gradient-based greedy algorithm, similar to the work in [52], is used to determine the values of α_k. More specifically, the algorithm searches for the set of α_k values that maximizes (4). For arbitrary small $\alpha_k \rightarrow 0$, the algorithm can choose from a continuous set of rates for each user. Once the set of resource shares α_k is determined, the transmission rates R_k are returned by the QoE optimizer. In [27], it is shown that the greedy algorithm has low computational complexity and is scalable for a large number of users.

D. Enforcement of video quality adaptation

Knowing the target transmission rate of each user as described in (4)-(5), the objective is to adapt the application to the data rates supported at the lower layers. Specifically, two different paradigms for determining the streaming or representation rate of each user k, denoted by Q_k, given its target transmission rate are considered:

- *Proactive optimization*: We consider a proxy (e.g., at the edge of the wireless network) which intercepts the client HTTP requests and rewrites them according to the feedback from the QoE optimizer. In the proactive approach (Figure 2), the target rate of each client is signaled to a resource shaper and the proxy server. The

resource shaper limits the TCP throughput of each client. In addition, the main role of the proxy is to rewrite the client HTTP requests on the fly and forward them to the DASH server. The proxy operation is transparent to the DASH server and the DASH clients, i.e., no explicit communication between the proxy and the clients is required. Hence, each client will decode and play an optimized representation for its requested segment. Please note that we assume that the client will be able deal with a mismatch between the requested and the incoming rate as long as it is a valid representation. This is motivated by the observation that the actual bit-rate will differ from the rate defined in the MPD given the variable bit-rate nature of video coding.

Different approaches for determining the streaming rate are investigated in this paper. First, a buffer unaware approach, referred to as **QoE-Proxy**, is considered. Specifically, it rewrites the client requests to the closest lower representation given the rate feedback from the QoE optimizer (i.e., $Q_k \leq R_k$). In Section IV, alternative strategies which additionally consider the buffered media time of the clients are studied.

- *Reactive optimization*: Alternatively, a mobile operator can adapt the network resource allocation without interfering with the client decisions. In the reactive approach, referred to as **QoE-Reactive**, each client gets a TCP throughput equal to the target rate determined by the QoE optimizer (again enforced by the resource shaper in Figure 2). The representation rate, however, is only determined by the media streaming client which reacts to the throughput changes.

In both cases, a standard unmodified DASH client is used. Furthermore, both approaches are optimized for OTT DASH delivery and do not require to access or decode the transported video content. The approaches only differ in how the QoE-based resource allocation result is exploited for dynamic rate adaptation for overall QoE optimization.

IV. Playout Buffer-Aware Video Quality Adaptation

So far, the QoE-based proxy adaptation scheme presented in Section III optimizes the adaptive HTTP video delivery by considering the channel and content characteristics of the streaming users. This section investigates whether the playout buffer information can be utilized to further improve the QoE. The buffer level is defined as the length of buffered media time at the client which can be either estimated at the proxy or is directly reported by the client. We consider that the proxy is gathering QoE feedback from the client [53] [7]. One metric is the HTTP request/response transactions which includes the times when the requests are made and when they are received at the client. The proxy can use the information about the received segment times to estimate the buffer level assuming continuous playback at the client. In addition, the DASH specification defines a buffer level metric where the client reports the playout duration for which media data is available. Specifically, a QoE configuration is established between the server and the client that indicates the reporting frequency. Given that our algorithm relies on a long-term periodic feedback, the reporting interval can be configured according to the DASH specification [7].

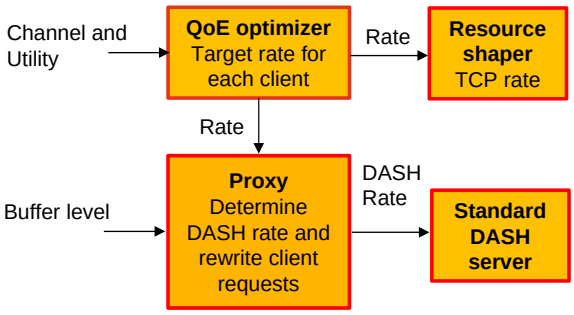

Fig. 4. Playout buffer-aware video quality enforcement at the proxy.

More specifically, the buffer levels of the mobile clients are considered at the proxy to enhance the streaming rate selection of the individual clients. The QoE optimizer first determines the target rate for each client as in (4)-(5) and provides the rates to the resource shaper and the proxy (Figure 4). The proxy then considers both the returned rate and the buffer level of each client when deciding on the representation rate. The transmission rate and buffer level are updated periodically (e.g., each 1 second). Specifically, two approaches are studied: 1) the first one is to solve for the highest representation that considers the throughput and buffer time of each client without any additional constraints. 2) The second approach aims at smoothing the playout video quality by considering playout buffer-aware quality bounds for selecting the representation rate of each client.

A. Maximum DASH rate selection

Given the available transmission capacity feedback from the QoE optimizer, the proxy solves for the highest representation for each user, given its current buffer level. This approach is similar to the QoE-Proxy scheme in (III-D) and additionally allows to stream at a higher rate than the current transmission capacity if there is enough buffered media time at the client. We define $OptQ(R_k, B_k) = U_k(Q_k)$ where the objective of user k is:

$$\arg\max_{Q_k} \quad U_k(Q_k) \tag{6}$$

$$subject\ to \quad Q_k \leq R_k(1 + \frac{B_k}{T_{seg}}) \tag{7}$$

$$where\ B_k = \text{buffer level (s)}, T_{seg} = \text{segment size (s)}$$

where R_k is the transmission rate as determined by the QoE optimizer, Q_k is the representation rate that maximizes the objective function in (6) and that satisfies the continuous playout constraint at the client (7). If the current buffer level at the client $B_k > 0$, then the user can stream at a representation rate which is higher than its transmission capacity and less than $R_k(1 + \frac{B_k}{T_{seg}})$, where T_{seg} is the segment duration. This

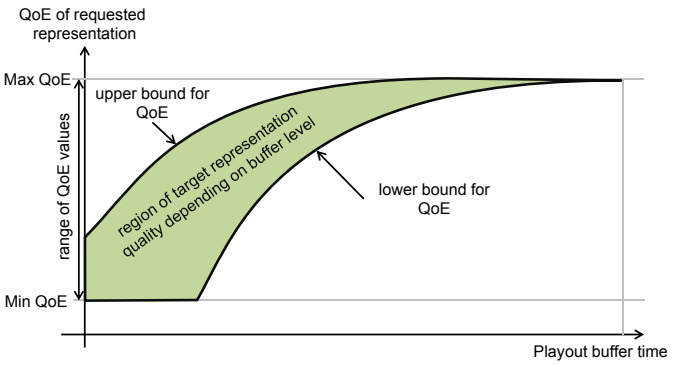

Fig. 5. Illustration of the playout buffer-dependent quality bounds approach for selecting the representation rate. The quality of requested representation is constrained to a target region depending on the current buffer level.

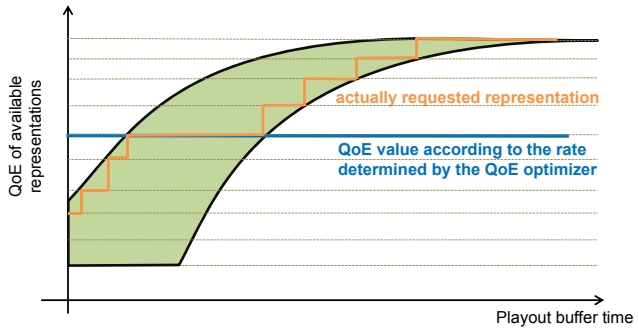

Fig. 6. The representation is selected based on the buffer level and achievable QoE according to the instantaneous transmission capacity feedback from the QoE optimizer.

represents the highest possible representation that does not violate the buffer underflow constraint.

B. Playout buffer-dependent quality bounds

Instead of solving for the highest representation rate (6)-(7), we foresee the benefit of allowing clients to build-up buffer that can later compensate for dynamic channel variations. The main idea is to introduce upper and lower quality bounds which constrain on the representation quality as a function of the playout buffer time at the client (Figure 5). To select the target representation, both the current buffer level and the achievable QoE according to the instantaneous transmission capacity feedback from the QoE optimizer are considered. This is illustrated in Figure 6 which shows the available representation qualities and the actually selected representations as a function of the playout buffer time. Users with few segments in their buffer can transmit at a lower representation rate in order to build up their buffers. Also, users with enough buffer can switch to a representation quality which is higher than the one at the instantaneous transmission rate. The overall objective is to improve the user experience by observing the buffer demands and the decision impact on the perceived user video quality. The optimization is realized in two steps:

1) We refer to ζ_k as the achievable MOS at the current transmission rate R_k determined by the QoE optimizer. That is, $\zeta_k = U_k(Q_k)$, where $Q_k \leq R_k$.

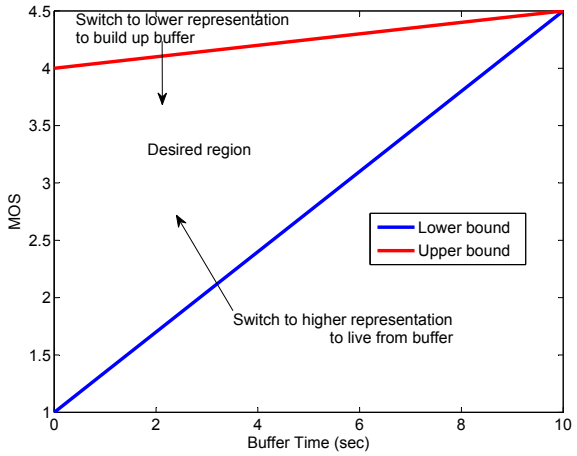

Fig. 7. Example of buffer-aware boundary conditions for selecting the representation rate used in this paper. Values for (T_L, a) equal to (1, 0.35) and values of (T_H, b) equal to (4, 0.05) are considered.

2) We define Upper bound (UB) and lower bound (LB) thresholds for finding the representation rate given the current buffer level B_k and ζ_k. For simplicity, in this work, linear thresholds are considered (cf. (11)). An example of these boundary conditions is given in Figure 7. The values of a and b determine how fast the users will deplete and build up their buffers, respectively. The values of T_L and T_H represent the lower and upper QoE bounds, when the buffer level of the client is equal to zero. We define $OptQ(R_k, B_k) = U_k(Q_k)$ where the DASH rate selection problem for user k is:

$$\arg\max_{Q_k} \quad U_k(Q_k) \quad s.t. \tag{8}$$

$$U_k(Q_k) \leq \min\left(UB, \zeta_k\right) \text{ if } \zeta_k \geq UB \tag{9}$$

$$U_k(Q_k) \geq \max\left(LB, \zeta_k\right) \text{ otherwise} \tag{10}$$

$$UB = T_H + B_k \cdot b, LB = T_L + B_k \cdot a \tag{11}$$

where (8) is the objective function for user k, (9) and (10) constrain on the desired region of representation quality.

In this work, the value of T_L is set to 1 on the MOS scale. That is, for users with empty buffers, the requested representation rate should not exceed the available transmission capacity. Also, the value of T_H is set to 4 on the MOS scale, so that the representation quality of the users with favorable conditions is not much degraded. For determining the a and b values (0.35 and 0.05, respectively), the assumption is that users with enough buffered segments (10 segments in this case) can request the highest quality representation. Please note that the upper and lower bound thresholds have been selected to control the QoE in a reasonable way. Nevertheless, other threshold values and in general more complex quality bound definitions could be considered.

V. Joint Playout Buffer-Aware Resource Allocation and Video Quality Adaptation

The objective is to optimize the network resource allocation based on the content and channel characteristics and the

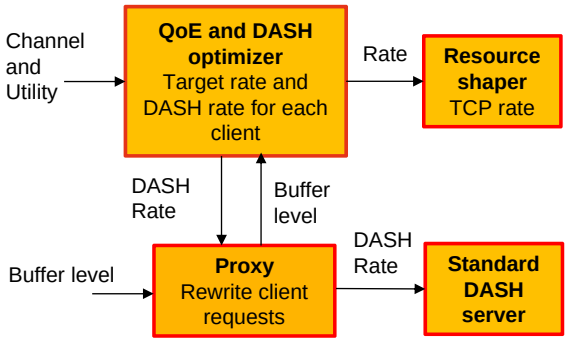

Fig. 8. Playout buffer-aware resource allocation and video quality adaptation.

client playout buffer levels. In this case, the buffer levels are additionally signaled to the QoE optimizer and considered in the multi-user resource allocation. We see a big potential in redistributing the network resources while considering the buffer information. This allows, for instance, to take some physical resources from a user without degrading the video quality of future representations if it has buffered enough segments and assigning these resources to another user with low buffer level which permits it to stream at a higher quality. Specifically, different from Sections III-C and IV, where the resource allocation and the enforcement of streaming rates were determined separately, we now jointly solve both rate allocation problems. In this case, the proxy first provides the buffer level of each client to the QoE optimizer. The optimizer then solves for the transmission rate and the representation rate of each client. The resulting rates are signaled back to the resource shaper and the proxy, respectively. This is illustrated in Figure 8.

Thus, the objective of the joint optimization is:

$$\underset{(\alpha_1,...,\alpha_K)}{\arg\max} \quad \sum_{k=1}^{K} OptQ(g_k(\alpha_k), B_k) - P_k, \tag{12}$$

$$s.t. \quad \sum_{k=1}^{K} \alpha_k = 1 \tag{13}$$

where $OptQ(g_k(\alpha_k), B_k)$ is the utility value of user k, as determined in (6), respectively (8), for a given network resource share α_k (R_k via (2)) and buffer level B_k.

To solve this problem, a greedy algorithm is considered (Algorithm 1). Let $((R_1)^0, ..., (R_K)^0)$ be the initial rate allocation vector and $((Q_1)^0, ..., (Q_K)^0)$ the corresponding DASH rates as determined in (6)-(7), respectively (8)-(10). At each iteration m, we search for the users i and j, where increasing the transmission rate $(R_i^+)^m$ and decreasing $(R_j^-)^m$ results in the maximum increase in the objective function (12). The corresponding representation rates $(Q_i)^m$ and $(Q_j)^m$ are updated. The above procedure is repeated until no further improvement in (12) is possible.

At each iteration, the objective function in (12) is maximized by determining the sensitivity of users i and j to gaining or losing a certain resource proportion, while keeping the optimization variables of the other users fixed, until convergence

Algorithm 1 Greedy algorithm based on [52]

Input: Number of users K, buffer level B_k, maximum number of iterations M, minimum utility improvement ΔU_{min}, iteration step size Δ_α

Output: Allocation $(\alpha_1, ..., \alpha_K)$

procedure GREEDY ALGORITHM
  Initialize $(\alpha_1, ..., \alpha_K)$ in a Round Robin way;
  Set iteration index $m = 0$;
  Compute $((R_1)^0, ..., (R_K)^0)$ using (2)
  Compute $((Q_1)^0, ..., (Q_K)^0)$ using (6)-(7) (respectively (8)-(10))
  Define $U_k = OptQ(g_k(\alpha_k), B_k) - P_k$, $k = 1...K$
  while $m < M$ **do**
    $i = \arg\max_{k,k=1...K} \{\Delta U_k | \alpha_k \leftarrow \alpha_k + \Delta\alpha\}$;
    $j = \arg\min_{k,k=1...K} \{\Delta U_k | \alpha_k \leftarrow \alpha_k - \Delta\alpha\}$;
    $\alpha_i = \alpha_i + \Delta_\alpha$; Update $(R_i)^m$ and $(Q_i)^m$;
    $\alpha_j = \alpha j - \Delta_\alpha$; Update $(R_j)^m$ and $(Q_j)^m$;
    $\Delta U_{inc,m} = \Delta U_i - \Delta U_j$;
    if $\Delta U_{inc,m} < \Delta U_{min}$ **then**
      break;
    end if
    $m = m + 1$;
  end while
  Output allocation $(\alpha_1, ..., \alpha_K)$;
end procedure

is achieved. The decision on the representation rate Q_k of user k for a given transmission rate R_k is done locally independent of the other users. This holds for both video quality adaptation schemes in (6)-(7) and (8)-(10).

VI. Subjective Quality Assessment

A. Experimental Setup

In order to assess how the different schemes impact the video perception, a subjective evaluation is first performed with human subjects. For our subjective tests, we have chosen the Microsoft Smooth Streaming client as a stable adaptive HTTP client. In our preliminary work in [22], we also considered the DASH-enabled VLC client by Mueller et al. [54]. Without loss of generality, we believe that our results are applicable to any DASH system. Moreover, the clients remains unmodified in our experiments. In DASH, the segments are per definition independently decodable. Consequently, the clients are able to decode and play the redirected segments in our proxy approach.

More specifically, two scenarios each with 8 adaptive HTTP streaming users in one LTE cell are simulated. A matlab-based LTE simulator [55] is considered to find the transmission rates of the users when applying the QoE-based resource allocation in (4)-(5). Furthermore, we use a standard HTTP server and emulate the wireless network. In other words, a resource shaper is placed between the server and the clients that limits the data rates per client to the output of the QoE optimizer. In our experiments, the Dummynet software [56] is used which allows enforcing bit-rate limitations at the TCP level. The experimental parameters are presented in Table I.

TABLE I
EXPERIMENTAL SETUP

Application Parameters	
Video codec	H.264/AVC, CIF, 30fps
Client software	MS Smooth Streaming
Number of representations	11
Quantization parameter	20...40
Segment size	2 sec
LTE Parameters	
Carrier frequency	2 GHz
System bandwidth	5 MHz
Number of PRBs	25
SNR averaging cycle	2 sec
Link layer model	[50]
Channel model	Urban macrocell
Shadowing	disabled

Each user is downstreaming a different video with specific rate-distortion characteristics (*soccer, ice, bus, coastguard, foreman, akiyo, container, harbour*). In the first scenario (scenario 1), users move in the cell with a speed of 30 km/h. The second scenario (scenario 2) presents more rapidly changing channel conditions as all users move with a speed of 120 km/h. Moreover, the **QoE-Proxy** and the **QoE-Reactive** approaches are assessed. These methods are also compared to a non-optimized scheme (**Non-Opt**) where the PRBs are equally shared among the users, and the streaming rate is dynamically decided by the client. As a result, there are 6 cases to evaluate.

B. Test Methodology

The subjective test is conducted using the SAMVIQ method [57], which is specifically designed for assessing multimedia applications. A 10 seconds long sequence extracted from the simulated 60 seconds is presented to the test subjects (viewers) for each scenario and each optimization method. More precisely, seconds 30 to 40 from the sequences are extracted. This allows avoiding the typical poor quality start-up phase of the adaptive streaming client (here: Microsoft Smooth Streaming). Additionally, 10 seconds of videos are a recommended duration for conducting subjective tests [57]. Besides the evaluated 6 cases, viewers also rate a reference sequence (best possible quality), a hidden reference and a poor quality sequence. The quality anchors are recommended to stabilize the subjective results [57]. Viewers rate the videos on a continuous scale from 0 to 100. After the screening procedure described in [57], the data of 20 test subjects was verified to be valid. For each sequence, the average rating over the 20 viewers is taken and a differential quality score (DMOS) [58] value is then computed by:

$$DMOS = \overline{R}(sequence) - \overline{R}(hidden\ reference) + 100 \quad (14)$$

where $\overline{R}$ is the average rating over the 20 viewers. The DMOS value is used in our data analysis as the subjective quality rating.

C. Test results

Figure 9 shows the mean DMOS over the 8 users in the cell. Additionally, the boxplot illustrates the distribution of the DMOS values for the 8 users.

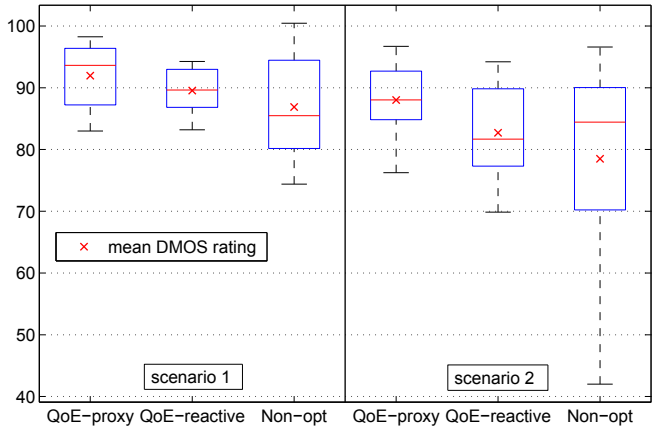

Fig. 9. DMOS of 8 users for 30 km/h (scenario 1) and 120 km/h (scenario 2). The error bars represent the worst and best user ratings. The boxplot shows the median, $25th$ and the $75th$ percentiles.

The results can be first interpreted horizontally by looking at the mean DMOS of the eight DASH users. We observe that for both scenarios, QoE-Proxy achieves the best mean DMOS, that is, provides the best mean QoE for the users. The QoE-Reactive approach achieves a better mean DMOS than the non-optimized approach. That is, both QoE-based optimization approaches improve the mean QoE compared to a non-optimized approach. In the more dynamic scenario (scenario 2), the gains in term of mean QoE are more pronounced compared to the less dynamic scenario (scenario 1).

The results can be further interpreted vertically by looking at the individual DMOS of the DASH users. The boxplot indicates a larger spread of the users' DMOS for the non-optimized scheme. The Non-Opt scheme provides the same resource share regardless of the spatiotemporal complexity of the video content which results in low DMOS for the resource-demanding videos. This is again more visible in the more dynamic scenario. On the contrary, both QoE-based approaches present a lower variance, i.e., the fairness is improved by using a QoE-based method. These subjective results confirm the experimental results in our preliminary work in [22].

Furthermore, no playout interruptions are observed at the client during our experiments. This is the main advantage of our proxy approach as it explicitly considers the available resources in the cell and hence avoids overloading the cell.

D. Assessment of the prediction model

We assess the applicability of the linear PSNR/MOS model that we use for the optimization problem. Therefore, a post-analysis is conducted to measure the correlation between the subjective results and different objective quality metrics. Specifically, we compare the linear mapping with two non-linear metrics based on the PSNR, namely the PSNR based Video Quality Metric (VQM_P) proposed in [59] and the STVQM proposed in [60]. The Pearson correlation between the subjective ratings and the predicted ratings for the 8 videos is presented in Table II. The predicted MOS values from the

TABLE II
PEARSON CORRELATION FOR EACH VIDEO

Metric	soccer	ice	bus	coastguard	foreman	akiyo	container	harbour	**mean**
Linear	0.9618	0.4762	0.8241	0.9489	0.9787	0.9115	0.9826	0.8959	**0.8725**
VQM_P	0.9866	0.8537	0.8241	0.9466	0.9976	0.9077	0.9872	0.9262	**0.9287**
STVQM	0.9697	0.7792	0.8241	0.9412	0.9988	0.9401	0.9884	0.9172	**0.9198**

linear model are converted on the rating scale as described in Appendix I of [61].

The main difference between the metrics is observed for the *ice* video, where the linear model achieves a low Pearson correlation. This is due mostly to the low resource-demanding characteristic of this video, which in our scenario leads to a very high perceived quality for all cases. The predicted ratings all being in a small range, the Pearson correlation is more sensitive to a variation of the subjective ratings. However, the mean Pearson correlation over the 8 videos is 0.8725 for the linear model. This is close to the highest mean Pearson correlation achieved by the VQM_P (0.9287). This shows that although the used linear is very simple, it performs almost as well as a more complex non-linear video quality metric in the case of an optimization over multiple videos.

VII. SIMULATION RESULTS

A. Simulation setup

In this section, we examine the performance of the proxy scheme and the buffer aware approaches introduced in Sections IV and V. We simulate a resource-constrained LTE cell with 8 clients streaming adaptive HTTP video content from a DASH server. Each content is encoded into 11 representations by varying the encoder's quantization parameter. We assume that the client requests a new segment immediately after the previous segment is downloaded. Each client provides periodic feedback (each 1 sec) on its playout buffer level. Please note that our optimization approach is independent of how the client requests the segments. The reported buffer level is considered at the proxy as a measure of the buffer fullness at the client and is incorporated into the optimization problem. The assumption of the segment requesting algorithm at the client is used for the simulations. Other requesting algorithms can be used without changing the proposed optimization approach.

A Matlab-based LTE simulator [55] is used to generate different mobility patterns and find the transmission rates of the users. The representation and transmission rates for each client are determined at a time scale of 1 sec as well. Moreover, the following parameters are defined for Algorithm 1: $\Delta_\alpha = 0.004$, $\Delta u_{min} = 0.0005$ and $M = 1000$. The simulation parameters are further illustrated in Table III.

Figure 10 shows the individual channel traces and the mean SNR of the 8 users. In each simulation run, the requested content and the channel condition of each user are shuffled. Each content is limited to 20 segments. During the simulation, a user periodically requests a new content type after it has downloaded the previous one. A pool of 12 videos with different content characteristics is considered.

In our simulations, we compare the following schemes:

- **QoE-Proxy**: This represents the buffer unaware proxy approach introduced in Section III and is used as a baseline for comparison with the buffer-aware schemes.
- **QoE-MR**: This corresponds to the maximum DASH rate selection approach in Section IV-A.
- **QoE-QB**: This represents the playout buffer-dependent quality bounds approach in Section IV-B.
- **QoE-MR-Joint**: This corresponds to the joint resource allocation and quality adaptation approach in Section V. The utilities are determined according to the maximum DASH rate selection approach in (6).
- **QoE-QB-Joint**: This also represents the joint resource allocation and quality adaptation approach in Section V. The utilities are determined according to the playout buffer-dependent quality bounds approach in (8).

B. Results

Figures 11 and 12 show the mean MOS and the mean buffer level of all users for the different schemes. Results are averaged over 50 simulation runs. The QoE-Proxy scheme always selects the highest possible representation below the available throughput. Consequently, the buffer level at the client will increase over time. Please note that buffer overflow is not considered here and it is assumed that clients have enough buffer depth. The QoE-MR scheme utilizes the buffer feedback to request a higher quality representation. Nevertheless, similar to the QoE-Proxy scheme, it adapts to the instantaneous rate. The QoE-QB provides a smoother mean MOS compared to the two other schemes. It builds up buffer at the client by requesting at a lower representation rate than the available throughput and then utilizes the buffer to smooth the playout curve over time. It also leads to a reasonably higher minimum mean MOS in the cell. (3.08 for QoE-QB and 2.79 for QoE-Proxy and QoE-MR schemes).

The QoE-MR-Joint and the QoE-QB-Joint refer to the case when joint throughput and DASH rate optimization is considered. The QoE-MR-Joint allows for a higher quality level compared to the QoE-MR scheme but it still runs the client buffers to their limits and ends in a fluctuating mean MOS over time. The QoE-QB-Joint, on the other hand, provides the best overall video quality. The gain comes from the building/depleting of the client buffers and the time multiplexing of the representation rates and the transmission rates of the clients. For instance, between 200 and 250 sec, users build up enough buffer which is used afterwards to stream at a higher quality when the channel deteriorates. In the QoE-QB scheme, however, the transmission rate only depends on the content and the channel properties of the different clients. This means that users with good channels will get a high transmission rate, irrespective of their buffer level. As a result, the gap between

TABLE III
SIMULATION PARAMETERS

Application parameters	
Video codec	H.264 AVC, CIF, 30 fps
Application type	Adaptive HTTP streaming
Number of representations	11
Quantization parameter	20...40
Segment size	1 sec
Pre-buffering time	1 segment
LTE parameters	
Carrier frequency	2 GHz
System bandwidth	5 MHz
Number of PRBs	25
Bandwidth per PRB	180 KHz
SNR averaging cycle	1 sec
Link layer model	[50]
Channel model	Urban macrocell
Shadowing standard deviation	8 dB
Correlation distance of Shadowing	50 m
Simulation parameters	
Number of users	8
Number of video sequences	12 videos
Simulation runs	50
Simulation time	300 sec

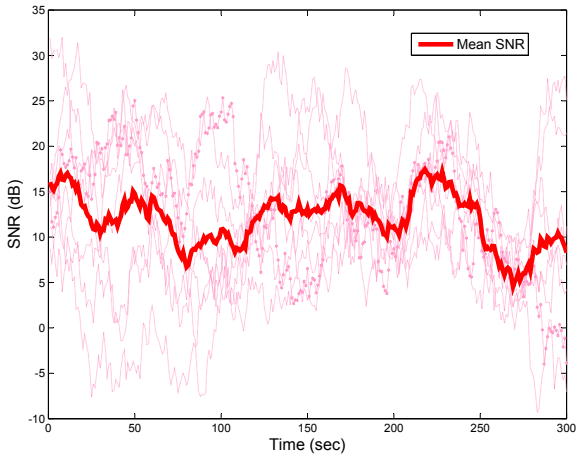

Fig. 10. Mean SNR of all users as a function of time. SNRs of individual users are shuffled at each simulation run.

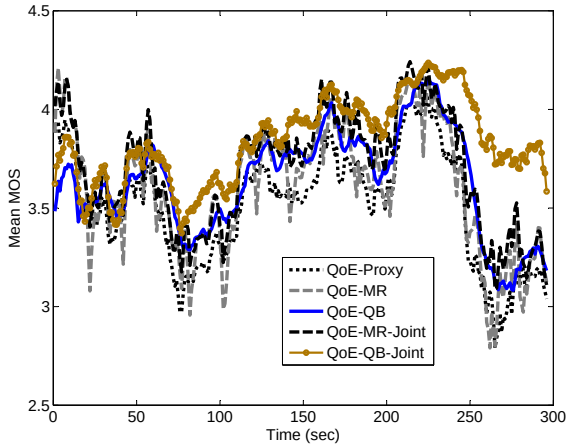

Fig. 11. Mean MOS of 8 users as a function of simulation time averaged over 50 simulation runs.

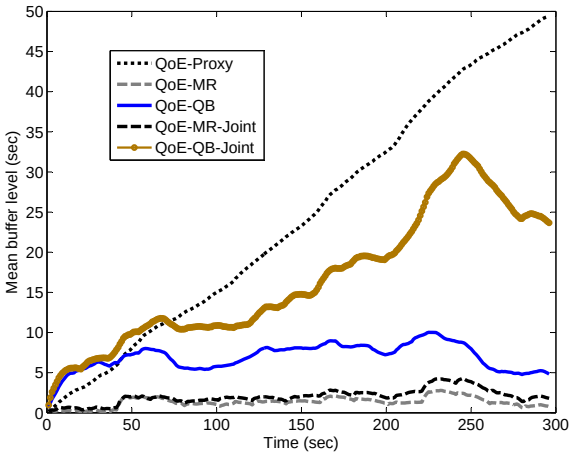

Fig. 12. Mean buffer level of 8 users as a function of simulation time averaged over 50 simulation runs.

the representation and transmission rates is reduced, overall less buffer is built up and the perceived quality will drop faster compared to the joint optimization scheme.

Figure 13 describes the mean MOS gain compared to the buffer unaware QoE-Proxy approach. The QoE-MR and QoE-QB, which use the same resource allocation approach as the QoE-Proxy, can result in a higher MOS by utilizing the buffer feedback when selecting the DASH representation. The joint optimization schemes can further improve the MOS level. Particularly, the proposed QoE-QB-Joint scheme results in up to 0.6 improvement on the MOS scale.

We additionally analyze the impact of the different schemes on the perceived video quality of the individual users. In Figure 14, we consider the quality switches which exceed one representation level as a measure of the temporal unsmoothness. The QoE-Proxy scheme optimizes the resources to minimize temporal quality fluctuations between two optimization rounds and leads to very few unsmooth quality switches. Similarly, the QoE-QB scheme smoothly increases and decreases the representation rate based on its boundary conditions and results in similar performance compared to the QoE-Proxy approach. The QoE-MR scheme, however, always adapts the representation rate to the instantaneous buffer level and the available transmission rate resulting in more abrupt quality switches. Finally, in the joint optimization function, a penalty term is used to penalize quality switches across two successive rounds. Subsequently, the quality switches for the QoE-MR-Joint and QoE-QB-Joint schemes are reduced.

Figures 15 and 16 show the mean MOS and the mean buffer level of all users for the different schemes when a buffer level constraint is considered. Specifically, a client will stop requesting new segments if the current buffer level exceeds 100 sec. The results show similar gains to Figures 11 and 12 which did not consider buffer accumulation.

We additionally compare the performance of our schemes with the buffer-based rate adaptation approach by Huang et al. [31]. We have used the same rate map as in [31] for the performance evaluation. Specifically, the rate adaptation algorithm by Huang et al. [31] is considered as an alternative

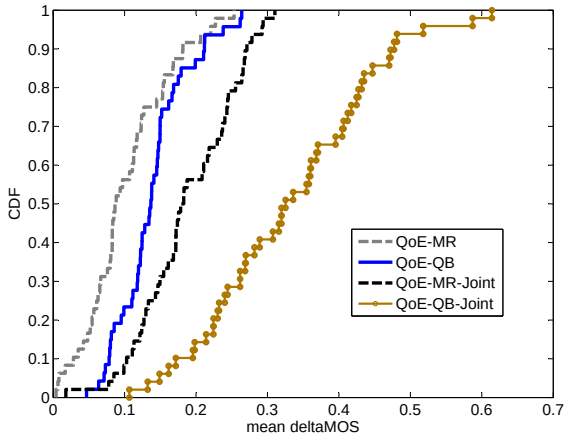

Fig. 13. CDF of mean MOS gain compared to QoE-Proxy scheme for 8 users. Results correspond to 50 simulation runs, 300 sec each.

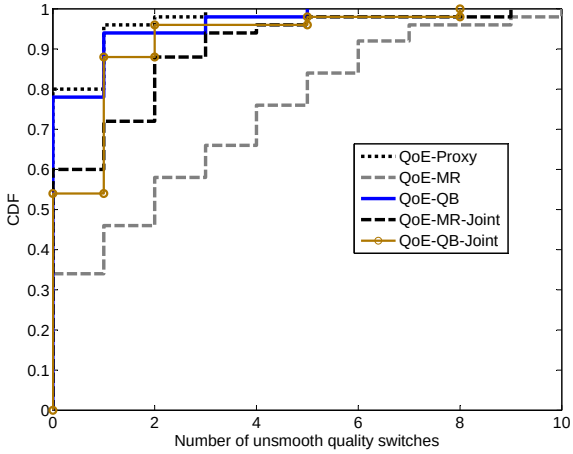

Fig. 14. CDF of mean number of non-smooth (jump more than 1 representation level) quality switches for each user. Simulation time is 300 sec. Results are averaged over 50 simulation runs.

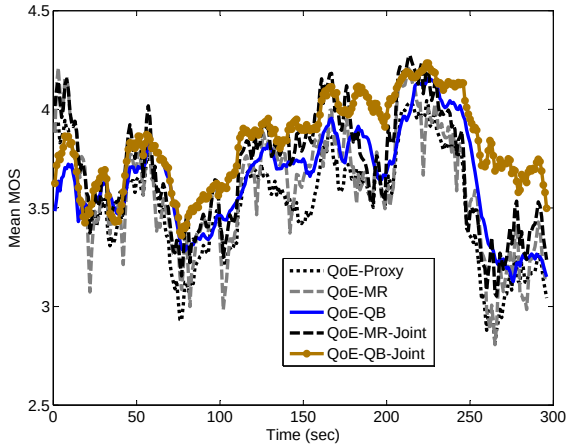

Fig. 15. Mean MOS of 8 users as a function of simulation time averaged over 50 simulation runs. A buffer level constraint of 100 sec is considered at the client.

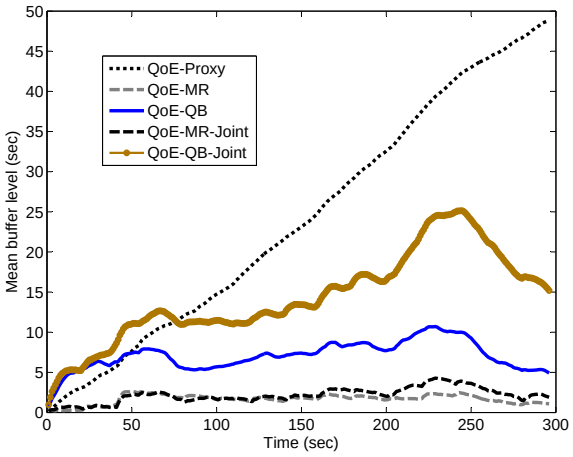

Fig. 16. Mean buffer level of 8 users as a function of simulation time averaged over 50 simulation runs. A buffer level constraint of 100 sec is considered at the client.

approach for determining the streaming rates of the clients while applying similar QoE optimized resource allocation in the network. This allows for a side-by-side comparison (shown in Figure 17) with our playout buffer-aware approach. First, we notice that the QoE-Huang scheme performs similar to the QoE-QB approach. Meanwhile, the QoE-Huang-Joint scheme shows that additional QoE gains can be achieved by jointly determining the streaming and transmission rates of the clients. This validates the importance of performing joint resource allocation and DASH rate adaptation. Furthermore, we observe that QoE-QB-Joint provides higher MOS improvements compared to the QoE-Huang-Joint scheme as it explicitly takes the QoE of the DASH users into account while redistributing the network resources.

C. Discussion

This study investigates the potential gains that can be achieved if QoE-based traffic and resource management is considered for adaptive HTTP streaming scenarios. The focus of this paper is on the radio access network (RAN) where the limited wireless resources are shared among multiple DASH clients. RAN user plane congestion is one of the main bottlenecks for mobile network operators and this is mainly due to video transmission [3]. The presented approaches show different dimensions of QoE gains when the content, channel and buffer time information of the users is available to a central controller in the mobile network.

The need for proactive in-network optimization presented in this paper is important for mobile network operators for different reasons. First, the QoS parameters (e.g., GBR) are used for admitting and charging the users when they join the network. These parameters, however, are not representative for video streaming applications where the video characteristics vary over time. Thus, the derived parameters may not truly represent the actual load during video transmission. In this paper, we propose to dynamically adapt the adaptive HTTP video transmission at a segment basis. Second, the scheduling mechanisms in LTE networks are performed on a short-time scale (order of ms). While it is also possible to incorporate the content characteristics into the short-term scheduling decisions, the benefits are rather questionable. Adaptive HTTP

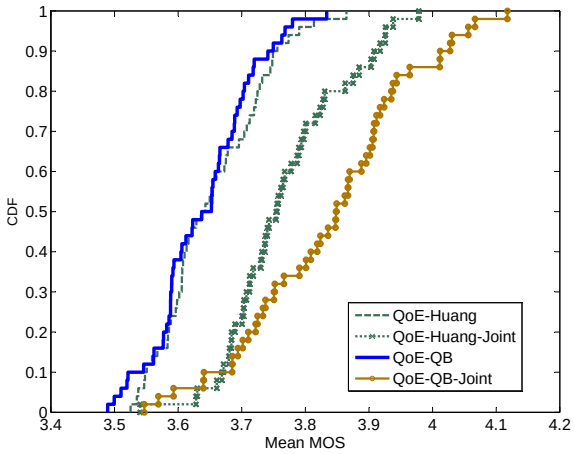

Fig. 17. CDF of mean MOS for 8 users. Results correspond to 50 simulation runs, 300 sec each. A buffer level constraint of 100 sec is considered at the client.

streaming segments are in the order of seconds. Our approach adapts the streaming and transmission rates on a long-term scale and can be integrated on top of the LTE scheduler. Third, the majority of streaming users use nonGBR bearers. Our objective is to optimize the system performance in loaded scenarios by admitting more users instead of blocking them. At the same time, by exploring the content information the quality of some users can be degraded in a controlled way while maximizing the average user satisfaction in the cell.

One of the key issues for deploying the proposed solutions in the current 3GPP LTE network is the location of the optimizer and the proxy: in the RAN close to the eNodeB, in the CN (Core Network), e.g., together with the PGW (Packet Data Network Gateway) or TDF (Traffic Detection Function), or outside the 3GPP network, e.g., as AF (Application Function). Placing the functionality in the RAN has the advantage that a precise cell view and channel information is available such that the optimizer may run on small time scales. The challenge will be to bring application and content information to the RAN. Placing the functionality into the CN or even outside the 3GPP network has the advantage that the application information, e.g., provided by TDF or ADC (Application Detection and Control) at the PGW is available. The challenge will then be to bring an up-to-date cell view and radio conditions to the CN which will lead to a significant signaling overhead.

Currently we can observe the following trends: (1) vendors offer multimedia optimization platforms that specially focus on caching, pacing, and transcoding video traffic transported via HTTP (progressive download). (2) The NGMN (Next Generation Mobile Network) Alliance started the Mobile Content Delivery Optimization (MCDO) project to standardize these multimedia optimization platforms and is about to initiate a work item in 3GPP [8]. MCDO describes in its use cases the options to place the multimedia optimization platform either in the RAN or in the CN. (3) The 3GPP work item UPCON (User Plane Congestion Management) is working on solutions for user plane congestion management. The proposed solutions [3] include (a) signaling congestion information from RAN to CN and (b) bringing application information to the RAN by packet marking.

This paper considers default bearers which represent the majority of mobile data traffic [3]. Moreover, the focus here is on managing the resource allocation on a long-term basis without interfering with the existing scheduling mechanisms in the RAN. The main challenge, however, for deploying the proposed solution in an LTE network is the availability of content-specific information. The UPCON solutions show the trend for better information exchange between RAN and CN. Most interesting is the architecture proposal raised in NGMN MCDO. Having a multimedia optimization platform located in the RAN will allow a mobile operator to deploy the QoE optimizer. The missing component, i.e., video specific utility information, may be embedded in the MPD by assigning the *qualityRanking* attribute for each representation [7] or added to the DASH segments [62]. The DASH standard also allows to extend the MPD to provide quality information within a program period as detailed in [33]. Alternatively, quality metadata can be downloaded from a specific utility server that provides utility information for popular videos, i.e., for those videos that are stored in the local cache.

VIII. Conclusion

This paper explores the improvements that QoE-based traffic and resource management in the mobile network can offer in the context of multi-user adaptive HTTP streaming. Compared to RTP/UDP based streaming, adaptive HTTP simplifies the rate adaptation process by providing multiple bit-rate encodings for the same content. Inspired by this a proactive QoE-based approach that rewrites the client HTTP requests at a proxy to the result of an overall network utility maximization is proposed. Subjective tests show a perceivable enhancement in video quality compared to reactive QoE optimization, that only adapts the throughput, and to non-optimized OTT DASH. In addition, our results indicate a fairer user experience, up to 35% MOS increase for the worst-case user, compared to non-optimized OTT DASH.

Moreover, the main contribution of this paper is the joint optimization of the transmission and representation rates of the mobile DASH users taking into account their buffer levels. Trading off the resources among the users allows a mobile network operator to allocate higher throughput for those running at a low buffer level. At the same time, reducing the data rates of users with enough buffered media time still allows them to request high quality representations. This leads to an additional mean gain of 0.3 on the MOS scale to the 0.35 gain that the proxy approach achieves compared to standard DASH with end-to-end adaptation.

References

[1] Sandvine, "Global internet phenomena report," Tech. Rep., 2013.
[2] T. Stockhammer, "Dynamic adaptive streaming over HTTP - standards and design principles," *Proc. MMSys 2011, California, USA*, Feb. 2011.
[3] 3GPP TR 23.705, "System Enhancements for User Plane Congestion Management," Tech. Rep., Aug. 2013.
[4] 3GPP TS 29.213 v12.4.0, "Policy and charging control signalling flows and Quality of Service (QoS) parameter mapping," Tech. Rep., June 2014.

[5] 3GPP TS 29.214 v12.3.0, "Policy and charging control over Rx reference point," Tech. Rep., March 2014.

[6] 3GPP TR 26.938 v1.6.0, "Improved Support for Dynamic Adaptive Streaming over HTTP in 3GPP," Tech. Rep., Jan. 2014.

[7] 3GPP 3GPP TS 26.247 v12.3.0, "Transparent end-to-end packet switched streaming service (PSS); Progressive download and dynamic adaptive streaming over HTTP (3GP-DASH)," Tech. Rep., June 2014.

[8] NGMN Alliance, Mobile Contents Delivery Optimization, "Deliverable 1. Use Cases, Draft version," Sept. 2013.

[9] A. Begen, T. Akgul, and M. Baugher, "Watching video over the web: Part 2: Applications, standardization, and open issues," *IEEE Internet Computing*, vol. 15, no. 3, pp. 59–63, May-June 2011.

[10] K. Ma, R. Bartos, S. Bhatia, and R. Nair, "Mobile video delivery with HTTP," *IEEE Comm. Magazine*, vol. 49, no. 4, pp. 166–175, Apr. 2011.

[11] S. Akhshabi, A. C. Begen, and C. Dovrolis, "An experimental evaluation of rate-adaptation algorithms in adaptive streaming over HTTP," *Proc. MMSys 2011, California, USA*, Feb. 2011.

[12] S. Akhshabi, L. Anantakrishnan, A. C. Begen, and C. Dovrolis, "What happens when HTTP adaptive streaming players compete for bandwidth?" in *Proceedings NOSSDAV '12, Toronto, Canada*, June 2012.

[13] T.-Y. Huang, N. Handigol, B. Heller, N. McKeown, and R. Johari, "Confused, timid, and unstable: Picking a video streaming rate is hard," in *Proceedings of the 2012 ACM Conference on Internet Measurement Conference*, Nov. 2012.

[14] J. Jiang, V. Sekar, and H. Zhang, "Improving fairness, efficiency, and stability in HTTP-based adaptive video streaming with festive," in *Proceedings CoNEXT '12, Nice, France*, Dec. 2012.

[15] L. De Cicco, V. Caldaralo, V. Palmisano, and S. Mascolo, "ELASTIC: A client-side controller for dynamic adaptive streaming over HTTP (DASH)," in *International Packet Video Workshop (PV) 2013, San Jose, CA, USA*, Dec 2013.

[16] Z. Li, X. Zhu, J. Gahm, R. Pan, H. Hu, A. C. Begen, and D. Oran, "Probe and adapt: Rate adaptation for HTTP video streaming at scale," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 4, pp. 719–733, April 2014.

[17] C. Mueller, S. Lederer, and C. Timmerer, "A proxy effect analyis and fair adatpation algorithm for multiple competing dynamic adaptive streaming over HTTP clients," in *IEEE Visual Communications and Image Processing (VCIP), San Diego, CA, USA*, Nov. 2012.

[18] S. Akhshabi, L. Anantakrishnan, C. Dovrolis, and A. C. Begen, "Server-based traffic shaping for stabilizing oscillating adaptive streaming players," in *Proceeding NOSSDAV '13 Oslo, Norway*, Feb. 2013.

[19] K. Ma and R. Bartos, "HTTP live streaming bandwidth management using intelligent segment selection," *Proc. IEEE Globecom 2011, Texas, USA*, Dec. 2011.

[20] R. Houdaille and S. Gouache, "Shaping HTTP adaptive streams for a better user experience," *Proc. MMSys 2012, North Carolina, USA*, Feb. 2012.

[21] T. Wirth, Y. Sánchez, B. Holfeld, and T. Schierl, "Advanced downlink LTE radio resource management for HTTP-streaming," in *Proceedings ACM MM '12, Nara, Japan*, Oct. 2012.

[22] A. El Essaili, D. Schroeder, D. Staehle, M. Shehada, W. Kellerer, and E. Steinbach, "Quality-of-experience driven adaptive HTTP media delivery," in *IEEE International Conference on Communications (ICC 2013)*, Budapest, Hungary, June 2013.

[23] R. Berry and E. Yeh, "Cross-layer wireless resource allocation," *IEEE Signal Processing Magazine*, vol. 21, no. 5, pp. 59–68, Sept. 2004.

[24] M. van der Schaar and N. Sai Shankar, "Cross-layer wireless multimedia transmission: challenges, principles, and new paradigms," *IEEE Wireless Communications*, vol. 12, no. 4, pp. 50–58, Aug. 2005.

[25] M. Ivrlac and J. Nossek, "Cross layer design - an equivalence class approach," *Proc. IEEE ISSSE '04, Linz, Austria*, Aug. 2004.

[26] S. Khan, Y. Peng, E. Steinbach, M. Sgroi, and W. Kellerer, "Application-driven cross-layer optimization for video streaming over wireless networks," *IEEE Communications Magazine*, vol. 44, no. 1, pp. 122–130, Jan. 2006.

[27] S. Thakolsri, S. Khan, E. Steinbach, and W. Kellerer, "QoE-driven cross-layer optimization for high speed downlink packet access," *Journal of Communications*, vol. 4, no. 9, pp. 669–680, Oct. 2009.

[28] Z. Lei and N. D. Georganas, "Adaptive video transcoding and streaming over wireless channels," *Journal of Systems and Software*, vol. 75, no. 3, pp. 253–270, Mar. 2005.

[29] O. Oyman and S. Singh, "Quality of experience for HTTP adaptive streaming services," *IEEE Communications Magazine*, vol. 50, no. 4, pp. 20–27, Apr. 2012.

[30] C. Liu, M. M. Hannuksela, and M. Gabbouj, "Client-driven joint cache management and rate adaptation for dynamic adaptive streaming over http," *International Journal of Digital Multimedia Broadcasting*, vol. Article ID 471683, 2013.

[31] T.-Y. Huang, R. Johari, and N. McKeown, "Downton abbey without the hiccups: Buffer-based rate adaptation for HTTP video streaming," in *Proceedings of the 2013 ACM SIGCOMM Workshop on Future Human-centric Multimedia Networking*, ser. FhMN '13, August 2013.

[32] Z. Li, A. C. Begen, J. Gahm, Y. Shan, B. Osler, and D. Oran, "Streaming video over HTTP with consistent quality," in *Proceedings of the 5th ACM Multimedia Systems Conference*, ser. MMSys '14, March 2014.

[33] T. Thang, Q. Ho, J. Kang, and A. Pham, "Adaptive streaming of audiovisual content using MPEG DASH," *IEEE Trans. on Consumer Electronics*, vol. 58, no. 1, pp. 78–85, Feb. 2012.

[34] S. Zhang, Y. Xu, P. Di, A. Giladi, C. Ai, and X. Wang, "Quality driven streaming using MPEG-DASH," *IEEE Communications Letters*, vol. 8, no. 2, pp. 34 –38, March 2013.

[35] G. Tian and Y. Liu, "Towards agile and smooth video adaptation in dynamic HTTP streaming," in *Proceedings of the 8th international conference on Emerging networking experiments and technologies, CoNEXT '12, Nice, France*, Dec. 2012.

[36] W. Pu, Z. Zou, and C. Chen, "Video adaptation proxy for wireless dynamic adaptive streaming over HTTP," *Proc. Packet Video Workshop 2012, Munich, Germany*, May 2012.

[37] P. Georgopoulos, Y. Elkhatib, M. Broadbent, M. Mu, and N. Race, "Towards network-wide QoE fairness using openflow-assisted adaptive video streaming," in *Proceedings of the 2013 ACM SIGCOMM Workshop on Future Human-centric Multimedia Networking*, Aug. 2013.

[38] F. Wamser, D. Staehle, J. Prokopec, A. Maeder, and P. Tran-Gia, "Utilizing buffered YouTube playtime for QoE-oriented scheduling in OFDMA networks," in *International Teletraffic Congress (ITC 24), Krakow, Poland*, Sept. 2012.

[39] P. Dutta, A. Seetharam, V. Arya, M. Chetlur, S. Kalyanaraman, and J. Kurose, "On managing quality of experience of multiple video streams in wireless networks," in *Proceedings IEEE INFOCOM, Orlando, USA*, March 2012.

[40] S. Singh, O. Oyman, A. Papathanassiou, D. Chatterjee, and J. Andrews, "Video capacity and QoE enhancements over LTE," in *IEEE International Conference on Communications (ICC), Ottawa, Canada*, June 2012.

[41] C. Mueller, S. Lederer, and C. Timmerer, "An evaluation of dynamic adaptive streaming over HTTP in vehicular environments," in *Proc. ACM MoVid '12, Chapel Hill, NC, USA*, Feb. 2012.

[42] B. Krogfoss, A. Agrawal, and L. Sofman, "Analytical method for objective scoring of HTTP Adaptive Streaming (HAS)," in *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), Soeul, Korea*, June 2012.

[43] R. K. Mok, E. W. Chan, X. Luo, and R. K. Chang, "Inferring the QoE of HTTP video streaming from user-viewing activities," in *Proceedings ACM SIGCOMM W-MUST, Toronto, Ontario, Canada*, Aug. 2011.

[44] R. K. P. Mok, X. Luo, E. W. W. Chan, and R. K. C. Chang, "QDASH: a QoE-aware DASH system," in *Proceedings MMSys '12, Chapel Hill, North Carolina, USA*, Feb. 2012.

[45] ITU, *Methods for subjective determination of transmission quality (ITU-T Recommendation P.800)*, International Telecommunication Union, Aug. 1996.

[46] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment."

[47] L. Choi, M. Ivrlac, E. Steinbach, and J. Nossek, "Sequence-level methods for distortion-rate behavior of compressed video," *Proc. IEEE ICIP'05, Genova, Italy*, Sept. 2005.

[48] A. Saul, S. Khan, G. Auer, W. Kellerer, and E. Steinbach, "Cross-layer optimization with model-based parameter exchange," *IEEE International Conference on Communications, ICC 2007, Glasgow, Scotland*, June 2007.

[49] F. Capozzi, G. Piro, L. Grieco, G. Boggia, and P. Camarda, "Downlink packet scheduling in LTE cellular networks: Key design issues and a survey," pp. 1–23, 2012.

[50] 3GPP TR36.942, "Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) system scenarios," 3rd Generation Partnership Project (3GPP), Tech. Rep., Jan 2011.

[51] S. Thakolsri, W. Kellerer, and E. Steinbach, "QoE-based cross-layer optimization of wireless video with unperceivable temporal video quality fluctuation," *Proc. IEEE ICC'11, Kyoto, Japan*, June 2011.

[52] D. Jurca and P. Frossard, "Media flow rate allocation in multipath networks," *IEEE Transactions on Multimedia*, vol. 9, no. 6, pp. 1227 –1240, Oct. 2007.

[53] ISO/IEC 23009-1, "Information technology- Dynamic adaptive streaming over HTTP (DASH)- Part I: Media presentation description and segment formats," Tech. Rep., Apr. 2012.
[54] C. Mueller and C. Timmerer, "A VLC media player plugin enabling dynamic adaptive streaming over HTTP," *Proc. ACM Multimedia 2011, Arizona, USA*, Nov. 2011.
[55] A. El Essaili, E. Steinbach, D. Munarretto, S. Thakolsri, and W. Kellerer, "QoE-driven resource optimization for user generated video content in next generation mobile networks," *Proc. IEEE International Conference on Image Processing, ICIP 2011, Brussels, Belgium.*, Sept. 2011.
[56] L. Rizzo, "Dummynet: a simple approach to the evaluation of network protocols," *Proc. ACM SIGCOMM Computer Communication Review*, Jan. 1997.
[57] ITU-R, "Rec. BT.1788 Methodology for the subjective assessment of video quality in multimedia applications," 2007.
[58] ITU-T, "Rec. P.910 Subjective video quality assessment methods for multimedia applications," Apr. 2008.
[59] S. Wolf and M. Pinson, "Video quality measurement techniques," NTIA, Tech. Rep. TR-02-392, Jun. 2002.
[60] Y. Peng and E. Steinbach, "A novel full-reference video quality metric and its application to wireless video transmission," in *IEEE International Conference on Image Processing (ICIP)*, Brussels, Belgium, Sep 2011.
[61] ITU-T, "Rec. G.107 The E-model: a computational model for use in transmission planning," Dec. 2011.
[62] ISO/IEC 23001-10, "Carriage of Timed Metadata Metrics of Media in ISO Base Media File Format," Tech. Rep., Jan. 2014.