This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2504732, IEEE Transactions on Circuits and Systems for Video Technology

Improving QoE and Fairness in HTTP Adaptive Streaming over LTE Network

Sergio Cicalò, Nesrine Changuel, Velio Tralli, Bessem Sayadi, Frédéric Faucheux, Sylvaine Kerboeuf

***Abstract*—HTTP adaptive streaming (HAS) has emerged as the main technology for video streaming applications. Multiple HAS video clients sharing the same wireless channel may experience different video qualities, as well as, different play-out buffer levels, as a result of both different video content complexities and different channel conditions. This causes unfairness in the end-user quality of experience (QoE). In this paper, we propose a quality-fair adaptive streaming solution with fair buffer (QFAS-FB) to deliver fair video quality and to achieve asymptotically fair play-out buffer levels among HAS clients competing for the same wireless resources in an LTE cell. In the QFAS-FB framework the share of radio resources is optimized according to video content characteristics, play-out buffer levels and channel conditions. The proposed solution is compared with other state-of-the-art strategies and the numerical results show that it significantly improves the quality fairness among heterogeneous HAS users, it reduces the video quality variations, and improves the fairness among the user's play-out buffers.**

***Index Terms*—HTTP adaptive streaming, QoE, LTE, fairness.**

I. Introduction

Current trend is toward an explosion of the mobile video traffic due to the increasing demand from a variety of multimedia-friendly portable devices such as tablets and smart phones. According to [1], streaming video applications over wireless networks will generate three-quarter of the mobile data traffic with the latter increasing of nearly ten times between 2014 and 2019. Although the latest enhancement of the 3GPP mobile standard, *i.e.*, the Long Term Evolution advanced (LTE-Advanced) [2], substantially improves end-user throughput and reduces user plane latency, the provisioning of enhanced and fair Quality of Experience (QoE) to multiple multimedia users still remains a challenge. In fact, besides a higher probability of traffic congestion in the cellular networks, mobile users will still experience different channel conditions with limited link capability and large throughput fluctuations due to the time-varying nature of the wireless interface.

HTTP adaptive streaming (HAS) [3] has emerged as the prominent technology to deliver video streams over internet. In addition to the commercial implementations, *i.e.*, Microsoft smooth streaming (MSS) [4], Apple HTTP live streaming [5] and Adobe HTTP dynamic streaming [6], HAS has been recently standardized by 3GPP [7] and MPEG, and denoted as dynamic adaptive streaming over HTTP (DASH) [8]. In HAS-based technologies, the video content is encoded at multiple bit-rates, also called profiles, which may consist of different temporal, spatial and quality resolutions. For each profile the video is segmented in several chunks, whose durations generally range between 2 and 10 seconds. At the end of profiles encoding, or periodically during encoding, the server generates a manifest file, which contains synthetic information describing the available profiles of each chunk. The client, after receiving a chunk, requests the subsequent chunk by selecting one of its available profiles according to the play-out buffer status and current downloading rate, thus enabling adaptive streaming. A comprehensive review of the MPEG-DASH standard for multimedia streaming over the internet can be found in [9].

So far, HAS principle has targeted an end-to-end optimization between the client and the server. However, in multi-user cellular systems the achievable data-rate depends on the UEs channel conditions, which may be widely heterogeneous. Hence, QoS-aware channel-dependent optimization is the primary key-tool to improve the fairness among HAS UEs. Mobile networks offer capability for controlling the UE rate thanks to the radio scheduler at the evolved node B (eNB). Resource allocation is realized upon the UE's radio channel condition and network utilization. LTE [10] supports different types of services including web browsing, video streaming, VoIP, online gaming, real-time video, etc., with standardized quality class indicators (QCI) [11]. Each QCI defines a set of requirements for quality of service (QoS) bearers, *e.g.*, maximum tolerable delay, packet loss rate and/or guaranteed bit-rate (GBR). A GBR bearer allows to define a minimum bit-rate and a maximum bit-rate (MBR) to be allocated to a particular UE.

Nevertheless, if the optimization is performed only in term of QoS, *e.g.*, by trying to provide the same data-rate to the UEs [12], the QoE may become significantly unfair. In fact, the UEs requesting low-motion videos, *e.g.*, interviews or news, require less data-rate to achieve an excellent QoE compared to UEs streaming high-motion videos, *e.g.*, sport or music events. We should keep in mind that video quality does not depend on the encoding rate only, but it also depends on the complexity of the video scenes [13]–[15]. Moreover, the QoE can be significantly degraded if stalling occurs during the video play-out due to a re-buffering event. In particular, the probability of stalling may increase for the user equipments (UEs) coming up into the network when the cell load is high, since they may have less possibility to build an adequate play-out buffer with

S. Cicalò and V. Tralli are with the Engineering Department of the University of Ferrara (ENDIF), CNIT, Via Saragat 1 - Italy. N. Changuel, B. Sayadi, F. Faucheux and S. Kerboeuf are with Alcatel-Lucent Bell Labs, Route de Villejust, Nozay, France.

respect to the already present UEs.

In order to meet QoE requirements we bring content-aware intelligence into the mobile network. The goal is to provide a fair video quality to different video flows and to reduce video quality fluctuations, by also allowing each client to fairly build an adequate play-out buffer. When multiple videos are delivered through the same eNB, this can be done through content- and buffer-aware resource allocation [13], [16]–[19].

In this paper, we propose a quality-fair adaptive video streaming solution with fair buffer level (QFAS-FB) to provide comparable QoE to HAS clients competing for the same radio resources in an LTE cell. The proposed solution allows to derive optimized GBR values to each HAS UE in order to achieve the following objectives: (i) maximize the overall video quality under quality fairness constraint in presence of different video complexities and channel conditions, and (ii) control the play-out buffer at the clients, *i.e.*, achieve an adequate fair amount of video playback time available in the buffer, even when UEs request videos at different time instants.

The QFAS-FB solution requires a MANE (media-aware network element) which primarily acts as a pre-scheduler on top of the LTE radio scheduler. The MANE derives the GBR values of each UE by solving an optimization problem aimed at maximizing the aggregate video utility under minimum and maximum rate constraints, available resources, and quality-fair and buffer-fair constraints across multiple video clients. The resulting solutions are then transparently provided to the LTE radio scheduler and used as GBR and MBR values. Two strategies are proposed in this paper: (i) the *look-ahead* strategy, where the quality level is decided by the client which reacts to the optimized GBR computed by the MANE, and (ii) the *overwrite* strategy, where the quality level of each chunk is decided by the MANE without client influence.

Numerical evaluations resulting from extensive and detailed *ns2* simulations with different UEs' starting time show that both *look-ahead* and *overwrite* QFAS-FB solutions provide significant improvement to the quality received by the end-users demanding more complex video scenes, even when they are experiencing bad channel conditions, while the degradation of the quality perceived by clients receiving low-complexity video chunks remains tolerable, when compared to other state-of-the-art frameworks. The *overwrite* strategy significantly reduces the video quality fluctuations, compared to the *look-ahead* strategy, which already outperforms the benchmarks. Moreover, our frameworks lead to an overall improvement of the buffer control by dynamically balancing the buffer levels at the clients. The overall QoE fairness is thus well improved compared to other state of the art framework.

The QFAS-FB framework proposed in this paper is built on the preliminary contribution in [19], where a basic *look-ahead* QFAS solution without buffer control has been investigated. The main novelty of this paper with respect to [19] is the design of a joint quality- and buffer-fair adaptive streaming solution, which improves the QoE fairness among the end-users. The asymptotic buffer fairness optimality of the QFAS-FB solution is also proved here in a rigorous way. Moreover, compared to [19], this paper specifically addresses the design of both the *look-ahead* strategy and the novel *overwrite* strategy to allow the short-term multi-user optimization of the QoE also in presence of asynchronous chunk requests.

This paper is organized as follows. Next section provides an overview of the literature on HAS. Section III introduces the system model. Section IV illustrates the *look-ahead* and the *overwrite* adaptation processes by introducing the main QFAS-FB optimization problem. In Section V the basic QFAS optimization problem is formulated and solved. Section VI details the extended QFAS-FB solution with buffer control. Section VII presents some final remarks. The performance of the proposed framework is evaluated in Section VIII and conclusions are drawn in Section IX.

II. Related Work

Early researches on HAS have focused on the optimization of a single server-client link, by improving the rate decision algorithm (RDA) of state-of-the-art commercial clients according to the trade-off among number of re-buffering events, video quality, quality oscillations and video play-out deadline [20]–[25]. However, a user-driven approach is generally suboptimal in a system where multiple HAS clients compete for the same resources.

In fact, the Authors in [26] have shown that, in a constant-bandwidth multi-user scenario, three major issues, *i.e.*, efficiency, stability and bandwidth estimation accuracy, have to be addressed. Efficiency and stability issues result when clients do not fully exploit the available resources, and perform needless bit-rate switches. The third issue is due to the fact that users might fail to precisely estimate the bandwidth in presence of periodic requests of video chunks, which produce ON-OFF intervals. In fact, when no limitation on the allocated resources is taken into account, competing players with non-overlapping ON-OFF intervals may overestimate their share of bandwidth. The Authors in [27] analyzed the main causes of these problems and proposed a novel fair, efficient, and stable adaptive (FESTIVE) RDA. The framework combines an optimized bandwidth estimator based on the harmonic mean of the past measured throughput samples, an improved profile selector that allows the RDA to converge to a stable profile, and a randomized chunk scheduler, which increases the accuracy in the estimation of the bandwidth. Nevertheless, the large window used to estimate the throughput, which improves the stability of the quality levels selection, could not cope with uneven large throughput drops typical of wireless channel with mobility. To improve the stability-responsiveness trade-off, the Authors in [28] proposed PANDA, a probe and adapt client RDA, which results in better bandwidth utilization, with respect to FESTIVE. The work in [29] extended the PANDA approach to also consider the content of the video by showing that content-aware adaptation schemes achieve better QoE when compared to conventional PANDA scheme. However, without an efficient in-network optimization, the clients may still experience abrupt QoE changes, especially in mobility scenarios. This motivated research to investigate HAS multi-client in-network optimization for enhancing service capacity, buffer stability and video quality [12], [17], [18], [30].

A re-buffering aware gradient algorithm (RAGA) to constraint the re-buffering probability in a multi-client wire-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2504732, IEEE Transactions on Circuits and Systems for Video Technology

less scenario has been proposed in [30]. It takes advantage of the periodic report of the buffer level standardized by DASH [7] by introducing a further constraint on the classical proportional-fair scheduling problem, thereby allowing a significant reduction of the buffer outage events. The Authors in [12] proposed an efficient method, named adaptive GBR (AGBR), to optimally and adaptively set up the GBR of each video flow in a LTE network with heterogeneous traffic. The approach is intended to achieve a level of fairness among the video flows, according to device characteristics, *e.g.*, screen size, while preventing starvation of other data flows. A similar framework was presented in [31], which also aims at improving stability and resource utilization, but without considering heterogeneous traffics. In these frameworks the definition of the utilities is not content-aware and may not lead to the best possible quality fairness among the video flows.

While several contributions have proposed enhanced content-aware transmission techniques for RTP/UDP video transmission, *e.g.*, [13], [16], [32], few works have recently investigated content-aware multi-user HAS delivery optimization [13], [15], [33]–[36]. A QoE-based HAS delivery framework over wireless networks has been recently proposed in [35]. According to the proposed QoE-continuum model, it derives a greedy solution for the maximization of both cumulative video quality and playback smoothness of each UE subject to wireless resource constraints. The resulting optimized profile level is used to overwrite the chunk requests. In [37] the Authors present a simple but asymptotically optimal QoE-driven network optimization for HAS video adaptation (NOVA). The aim is to jointly optimize the network resource allocation and the distributed RDAs at the client under a general wireless system model, in order to achieve the best users' QoE. Numerical results show that it achieves significant gain in terms of the average QoE compared to traditional proportional-fair resource allocation methods and state-of-the-art RDAs. The Authors in [15] proposed a content-aware multi-user HAS video delivery framework in LTE network. Similarly to here, a MANE (named proxy) is in charge of selecting the streaming rate required by each client in order to maximize the aggregate video utilities under resource constraint and may act in both the *overwrite* and *look-ahead* approaches. The approaches have been extended to also include buffer level optimization, in which the clients that have a high buffer level may select a profile rate that is larger than the actual streaming rate. Specifically the rate increment is proposed to be equal to the ratio between the buffer level and the chunk duration. They show that the proposed strategies can provide significant improvements with respect to content-blind optimization strategies. However, QoE fairness is not considered. Hence, UEs having bad channel conditions and/or small buffer level experience lower QoE with respect to luckier UEs, while a similar QoE should be provided to UEs subscribing the same service. In our approach, we aim to provide the same video quality by also dynamically prioritizing UEs having a low buffer level with respect to UEs having already an adequate buffer. More specifically, the novelty of our proposed approach is to tackle the problem in terms of buffer fairness among the users. Moreover, due to the problem formulation for long-term time basis in [15], the peculiarities of the *look-ahead* delivery strategy of HAS technology, where the video quality changes according to the content of each chunk, are not taken into account. Here we specifically design the *look-ahead* strategy to allow the client to follow the utility variations between consecutive chunks.

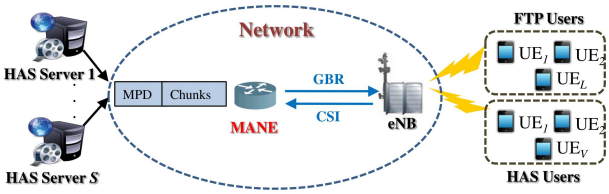


Fig. 1. System architecture.

All of the aforementioned works, as well as the present paper, consider single-layer encoding, *i.e.*, AVC/H.264 standard, to encode each profile, due to its high encoding efficiency. However, several contributions have shown the benefits of using scalable video coding (SVC) standard [38], [39]. The loss in coding efficiency of SVC is partially recovered thanks to a reduction of the storage requirement at both the server and the content distribution network, and thanks to the possibility of further optimizing the delivery strategy. The Authors in [34] proposed a quality-based optimization in which the eNB is in charge to jointly optimizing radio resource allocation and proactively overwrite the profile requested according to a cross-layer solution. In their approach, clients with good channel and buffer conditions may request enhancement layers of already downloaded chunks before the chunks playback deadline, thereby increasing the overall perceived QoE. However, the proposed cross-layer strategy requires out-of-standard radio level optimization and additional processing power to the eNB, whose resources are generally scarce. Here we specifically exploit the QoS differentiation allowed by the LTE standard: the eNB is transparently and dynamically instructed about the GBR and the MBR to be used for each HAS client, by keeping unchanged the standard scheduling and radio resource allocation functionalities.

III. System Model

As depicted in Fig. 1, we consider an LTE wireless access network comprising an eNB serving a total of K UEs, subdivided in V HAS UEs indexed by the set $\mathcal{V} = \{1, \ldots, V\}$, and L data UEs. We assume that the eNB sets up one dedicated GBR bearer for each HAS UE, whereas the data UEs are considered as non-GBR.

One or more HAS servers, connected to the eNB through high-speed backbone network, encode video sequences at multiple bit-rates and, after segmentation, generate for each of them a manifest file, also named media presentation descriptor (MPD). In HAS, each bit-rate defines a profile. We define $\mathcal{R}_k = \{r_{k1}, \ldots, r_{kN}\}$ as the set of N available rate profiles listed in the MPD of video k, each representing a different video quality at constant frame rate and resolution. In order to allow for content-aware optimization, we assume that the HAS servers generate synthetic video quality information for each chunk and inserts it in the MPD.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2504732, IEEE Transactions on Circuits and Systems for Video Technology

Notation	Description	Notation	Description
$\mathcal{V}$, V	Set and total number of HAS UEs	Π	Amount of resources for HAS UEs
w_k	Inverse of the k-th UE achievable rate	R_k	GBR of UE k
U_k	Utility of HAS UE k	$\Delta(x,y)$	Utility difference metric
Y_k, Z_k	Min. and max. rate of video k	$\mathcal{R}_k$	Set of profile rates of UE k
T_s	Scheduling time interval	T_c	Chunk duration
$c_k[n]$	Last chunk requested by UE k at nT_s	$\tilde{r}_k[n]$	Profile rate of chunk $c_k[n]$ send to UE k
B_k	Buffer fullness of client k	$\delta B_k(c_j)$	Buffer fullness difference between UE j and UE k
Acronym	**Full name**	**Acronym**	**Full name**
HAS	HTTP Adaptive Streaming	QFAS-FB	Quality-Fair Adaptive Streaming with Fair-Buffer
UE	User Equipment	QCI	Quality Class Indicator
MANE	Media-Aware Network Element	GBR	Guaranteed Bit-Rate
MBR	Maximum Bit-Rate	SNR	Signal-to-Noise Ratio
RDA	Rate Decision Algorithm	MPD	Media Presentation Description

TABLE I
LIST OF MOST USED SYMBOLS AND ACRONYMS

We consider the following parametric rate-utility model, which is used to describe the utility U_k, in terms of video quality, of requesting a particular profile with rate r_k:

$$U_k(r_k) = f_U\left(r_k; \mathbf{a}_k\right). \tag{1}$$

Here, $\mathbf{a}_k \in \mathcal{A} \subset \mathbb{R}^{N_a}$ is a time-varying and content-dependent vector of N_a parameters, which represents the synthetic quality information. For all values of $\mathbf{a}_k$, $f_U\left(r, \mathbf{a}_k\right)$ is assumed to be a continuous, invertible and strictly increasing function of r[1]. The model (1) may represent the relationship between the peak signal-to-noise ratio (PSNR), the Structural SIMilarity (SSIM) index [40], or any other strictly increasing quality metric, and the encoding rate [41]. Due to the high correlation with subjective tests in assessing the perceived video quality [42][43], in our numerical evaluation we consider the SSIM video quality metric. To model the dependency between the utility (here SSIM) and the rate, we consider a logarithmic utility function, *i.e.*,

$$f_U\left(R_k; \mathbf{a}_k\right) = a_1 \log(a_2 R_k + a_3), \quad R_k \in [Y_k, Z_k] \tag{2}$$

where Y_k, Z_k are the minimum and maximum rates of video k and the parameters a_1, a_2, a_3 depend on the spatial and temporal complexity of each chunk and are derived through curve-fitting over the actual discrete empirical points. We use the non-linear least square trust-region curve-fitting algorithm, whose convergence is generally achieved with an average number of iterations and function evaluations approximately equal to 100 and 300, respectively.[2]

A MANE, suitably located close to the eNB, is able to intercept and process the MPD requested by each HAS client in order to get rate and quality information. It may be connected to several eNB, and may be built upon the content distributed networks which are widely deployed in multiple locations, often over multiple backbones. The MANE primarily acts as a pre-scheduler. Based on the information available in the MPDs of all clients, it sets up the rate values, *i.e.*, GBRs, that the eNB will try to guarantee to the clients, in order to globally optimize the video streaming QoE of all HAS UEs. The MBR is set equal to the GBR, in order (i) to allow each client to precisely estimate its throughput [28] and (ii) to limit the HAS UEs resource with respect to those reserved to the data UEs in agreement with a resource sharing constraint.

The eNB allocates the available resources by using a general proportional-fair scheduler with minimum bit-rate, *i.e.*, GBR, and MBR constraints for each UE. Following the approach in [12] and similar to [44], we consider a simplified air interface model where the achievable rate (averaged over short term channel variations) for each UE is estimated according to the average channel state information (CSI) of its link. If γ_k is the average signal-to-noise ratio (SNR) experienced by UE k, the average rate per unit bandwidth is estimated as $\log_2(1+\gamma_k)$, by using the Shannon formula. The eNB sends the CSI to the MANE only when the average channel conditions of the UE significantly change. This information exchange requires a negligible burden compared to the typical signaling among the LTE entities.

IV. FAIR QOE ADAPTATION

As already mentioned, each HAS client is generally in charge of selecting the most suitable bit-rate profile according to its own RDA, given the throughput experienced during previous chunk transmissions. In this case, the most intuitive optimization strategy for the MANE is to operate according to a *look-ahead* strategy, *i.e.*, to act as pre-scheduler that dynamically sets up the feasible guaranteed minimum and maximum rates to transparently help the client in selecting optimized profile in the subsequent chunk requests. Nevertheless, the client's RDA is generally designed to avoid unnecessary quality fluctuation and to heuristically stabilize the buffer in best-effort wireless network [12], [24], [27]. Hence it may not react promptly to variations of the measured throughput resulting from a dynamic optimization. To overcome this limitation, besides the *look-ahead* strategy, we also design a full proactive strategy, which overwrites the users' chunk requests.

[1]Although f is a continuous function, the set of admissible values of r_k is discrete with finite cardinality.

[2]The amount of extra-information and the computational complexity required to generate the set of parameters $\boldsymbol{a}_k$ are negligible compared to the amount of data usually inserted in the MPD (chunk URLs, duration, size, etc.) and the encoding complexity of the profiles, respectively.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2504732, IEEE Transactions on Circuits and Systems for Video Technology

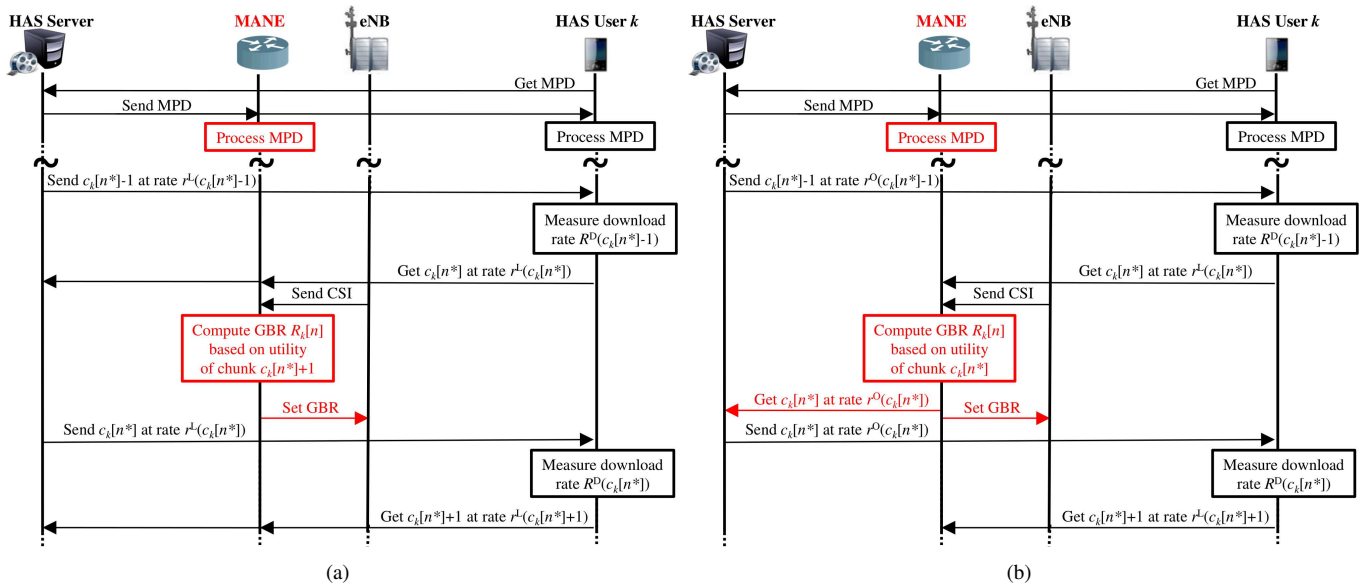


Fig. 2. Proposed approaches for a single HAS video delivery optimization: *look-ahead* strategy (a), and *overwrite* strategy (b). In the figure n^*T_s is the time instant at which chunk c_k is requested.

A. The Adaptation Process

Let us define T_s as the pre-scheduling time interval. The role of the MANE is to set up the optimized continuous GBRs $R_k[n], \forall k \in \mathcal{V}$, $n \in \mathbb{N}$, at every time instant nT_s in order to provide a fair short-term QoE to the HAS users. Due to the time-varying conditions of the network and the different streaming starting times, the chunk requests of the clients may occur at different time instants and are not synchronized with pre-scheduling time scale. In order to handle the multi-client asynchronous operations within the framework of the short-term multi-user adaptation, we build our analysis on the following assumptions:

a) long-term wireless channel conditions and average capacity change slowly in the time scale of chunk transmissions[3]
b) rate-utility functions change slowly between consecutive chunks of each video program
c) T_s is much smaller than the chunk transmission time, so that the chunk requests can be considered to be exactly aligned with pre-scheduling time instants[4]

Based on these assumptions, we define $c_k[n]$ as the index, at the time instant nT_s, of the last chunk requested by user k (as an example, if chunk 3 is requested at the discrete time 297 and chunk 4 is requested at the time 352, then $c_k[300] = 3$, $c_k[351] = 3$, $c_k[352] = 4$). In the following we will detail the two different adaptation strategies in an ideal scenario, according to what is shown in Fig. 2.

1) Look-ahead strategy: The behavior of the RDA at the client is described by assuming that the wireless network delivers each chunk to the respective client with a rate that exactly follows the GBR set up by the MANE. Hence, we further assume that once a chunk, *i.e.*, chunk $c_k - 1$ in Fig. 2(a), has been fully received by client k, the RDA of the client completes the measurement of the chunk download rate $R^D(c_k - 1)$ and requests the chunk c_k with a profile rate

$$r_k^L(c_k) = \max_{r \in \mathcal{R}_k(c_k), r \le R_k^D(c_k-1)} r. \quad (3)$$

where $\mathcal{R}_k(c_k)$ is the set of available rate profiles for chunk c_k. If this chunk request occurs at the time instant n^*T_s, the MANE, based on the current CSI, updates the scheduling rate $R_k[n^*]$ of the client k. The GBR values are updated only for the UEs that actually requested a chunk in the n^*-th pre-scheduling epoch. The GBR set up will be maintained, *i.e.*, $R_k[n] = R_k[n^*]$, for $n > n^*$, until a new chunk request occurs.

According to the assumptions stated in Sect. IV-A, when the channel variations are very slow in time, at the next chunk request of client k the measured download rate will be $R_k^D(c_k) = R_k[n^*]$, *i.e.*, the GBR value $R_k[n^*]$ used for transmission of chunk c_k will be the basis for the profile rate request of chunk $c_k + 1$. Hence, as highlighted in Fig. 2(a), in order to avoid mismatch, the GBR value $R_k[n^*]$ will be computed based on the video utility of the chunk $c_k + 1$.[5]

2) Overwrite strategy: In the *overwrite* strategy, the MANE is still in charge of intercepting the request of chunk c_k by the client k, occurring at time instant n^*T_s, and to update the prescribed GBR $R_k[n^*]$, but, as shown in Fig. 2(b), before forwarding it to the server, the profile rate request is overwritten by the new rate

$$r_k^O(c_k) = \max_{r \in \mathcal{R}_k(c_k), r \le R_k[n^*]} r. \quad (4)$$

The server sends the chunk c_k at the rate $r^O(c_k)$. The MANE, as in the *look-ahead* approach, collects the CSI and updates

[3]As also illustrated in Sect. III, note that the air interface is modeled through the average SNR of the links, which is a long-term channel description.

[4]In practical applications the pre-scheduling interval is in the order of milliseconds, which is usually much less than the chunk download time.

[5]If we use the video utility of chunks c_k to evaluate $R_k[n^*]$, then the k-th UE will request chunk $c_k + 1$ with a profile that the RDA selects based on the utility of chunk c_k, but the utilities may vary between consecutive chunks.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2504732, IEEE Transactions on Circuits and Systems for Video Technology

the GBR value $R_k[n^*]$ when it intercepts the chunk request, but differently from the previous strategy, $R_k[n^*]$ is computed based on the video utility of the chunk c_k, since the RDA at the client has no actual impact in the chunk request.

B. Multi-user Optimization: Problem Formulation

According to the selected strategy, our goal is to design the sequence of optimized continuous GBRs $R_k[n], \forall k \in \mathcal{V}$, which maximize the aggregate video quality and jointly allow the users to experience comparable video quality and to achieve similar buffer level. This is done by exploiting at each time n the CSI, the video utility, *i.e.*, the content information in the MPD of each chunk $c_k[n]$ (or $c_k[n]+1$ in the *look-ahead* strategy), and the buffer information. While the first objective can be achieved in the short term, the second objective will be reached in the long-term, especially considering that UEs may have very large buffer level gaps at the time they are requesting their first chunk.

In order to unify the analysis for both approaches, we introduce the indicator $I_{\mathrm{S}} = 1$ for the *look-ahead* strategy and $I_{\mathrm{S}} = 0$ for the *overwrite* strategy and we will use it to define, according to (3) and (4), the discrete profile rate requested for the chunk $c_k[n] + I_{\mathrm{S}}$, which depends on the scheduling rate at time n. It is given by

$$\tilde{r}_k[n] = r_k(c_k[n] + I_{\mathrm{S}}) = \max_{r \le R_k[n], r \in \mathcal{R}_k(c_k[n]+I_{\mathrm{S}})} r. \quad (5)$$

Since $\tilde{r}_k[n] \le R_k[n]$, the difference between them is the amount of rate available to build the buffer of client k. To this end, let us define $B_k(c_k)$ as the playback time of the video content still in the buffer when the client k requests the chunk c_k. The value of $B_k(c_k)$ is derived by exploiting the periodic buffer status feedback, which is included in the the DASH standard [7]. We also refer to it as buffer fullness. Given the chunk duration T_{c}, the scheduling rate $R_k[n]$ is implicitly related to the buffer fullness samples, *i.e.*,

$$B_k(c_k[n]+1) = B_k(c_k[n]) - \frac{\tilde{r}_k[n] T_{\mathrm{c}}}{R_k[n]} + T_{\mathrm{c}}, \quad (6)$$

where $\tilde{r}_k[n]T_{\mathrm{c}}$ is the size of chunk $c_k[n]$ and $\tilde{r}_k[n]T_{\mathrm{c}}/R_k[n]$ is its download time. The relationship holds if $R_k[n]$ is kept constant during the transmission of chunk $c_k[n]$.

One challenge for the design of the scheduling GBRs $R_k[n]$ for all the UEs is to optimize the use of the multi-user wireless channel at each time n, despite the GBRs are required to be asynchronously updated only at the times of chunk requests. Under the assumptions stated in Sect. IV-A, the approach proposed in this paper is to first evaluate the set of unconstrained scheduling rates $\hat{R}_k[n]$ that jointly optimize the video quality and the buffer level of the users and the use of the wireless channel at time n. Then, the rate $\hat{R}_k[n]$ will be used to update the GBR value $R_k[n]$ at the time instants at which UE k requests the chunks. We motivate this approach by considering that when the channel and video complexity variations are slow with respect to chunk transmission intervals, the actual GBRs $R_k[n]$ resulting from the asynchronous rate update will approximately follow the optimized rates $\hat{R}_k[n]$. The main benefit of this design choice is the possibility of building an analytical framework for optimization, adaptation and buffer control. The proposed approach will be validated by the numerical results obtained through simulations.

To derive the unconstrained scheduling rates $\hat{R}_k[n]$ we want to maximize the aggregate video utilities while minimizing the utility difference among multiple video at each chunk request, subject to minimum and maximum utility constraints, wireless resource constraints and asymptotic buffer fairness constraints. By keeping in mind the relationships in eq. (5) and eq. (6), the optimization problem can be stated as follows[6]

$$\max \sum_{k \in \mathcal{V}} U_k(\tilde{r}_k[n]) \quad (7a)$$

$$\min \sum_{i \in \mathcal{V}} \sum_{j > i} \Delta\big(U_i(\tilde{r}_i[n]), U_j(\tilde{r}_j[n])\big) \quad (7b)$$

$$s.t.\ U_k(Y_k[n]) \le U_k(\tilde{r}_k[n]]) \le U_k(Z_k[n]), \forall k \in \mathcal{V} \quad (7c)$$

$$\sum_{k \in \mathcal{V}} w_k[n] R_k[n] \le \Pi[n] \quad (7d)$$

$$\lim_{n \to \infty} \delta B_k(c_j[n]) = 0, \forall k, j \in \mathcal{V}, k \ne j \quad (7e)$$

where $w_k[n] = [\log_2(1 + \gamma_k[n])]^{-1}$ is the inverse of the rate per unit bandwidth depending on the average SNIR $\gamma_k[n]$ experienced by UE k. The value of $\Pi[n]$ defines the average amount of resources dedicated to the HAS UEs, which is dynamically computed as in [12] at every scheduling interval based on number of UEs and scaling factors.

The utility-fairness metric in (7b) is defined as:

$$\Delta(U_i, U_j) = \begin{cases} 0 & \text{if } U_i = f_U(Y_i; \mathbf{a}_i) \wedge U_j < U_i \\ 0 & \text{if } U_i = f_U(Z_i; \mathbf{a}_i) \wedge U_j > U_i \\ |U_i - U_j| & \text{otherwise.} \end{cases} \quad (8)$$

where $\wedge$ denotes the AND operator. This metric was introduced in [13] in terms of video distortion and extends the simple fairness metric $|U_i - U_j|$ to the case where U_i, U_j are constrained to their minimum and maximum values. In fact, in presence of utility constraints, if a video achieves its maximum utility, it is reasonable to use the available resources to increase the utilities of other videos. On the other hand, in a case of scarce resources, if decreasing the rate of the i-th video is not possible since its minimum utility value has been already reached, it is necessary to decrease the rate of the other videos, at the price of decreasing the related utility.

Eq. (7e) is the fairness condition for buffer fullness, given by the buffer fullness difference between the UE j and the UE $k \ne j$, *i.e.*,

$$\delta B_k(c_j) = \min_{i=0,1} \big| B_j(c_j) - B_k(c_k[n^*(c_j)] + i) \big| \quad (9)$$

where $n^*(c_j)$ is the time instant at which the request of chunk c_j occurs. The $\min$ operator ensures that the evaluation of the buffer fullness difference is carried out between aligned chunk requests in the asymptotic condition.

We should remark that the asynchronous update of the GBRs, occurring at the UEs' chunk request, only affects the constraint (7d), which may not instantaneously hold. In this

[6]Although we use $U_k(\tilde{r}_k[n])$ for the sake of conciseness, we remark that U_k also depends on time index n through the set of parameters $\boldsymbol{a}_k$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2504732, IEEE Transactions on Circuits and Systems for Video Technology

case, a part of radio resources is drained from, or released to, best effort data UEs.

The problem in (7) is a mixed integer non-linear problem (MINLP), thus NP-hard, and has multiple objectives. Moreover, part of the optimization has a short-term time basis, whereas the buffer fullness constraint in (7e) holds on the long-term. As the optimal solution is computationally prohibitive, we propose to tackle the problem in two steps. The first step aims at finding a solution for short-term quality fairness without constraints on the buffer. The second step addresses buffer fullness fairness constraint in the long term, by keeping unchanged the utilities achieved in the first step.

We address the first step in the next section by deriving an adaptive quality-fair solution, without buffer constraints. In Section VI, we then handle buffer fairness by deriving a quality-fair and buffer-fair solution that is proved to achieve long-term buffer fairness among the clients.

V. Quality-Fair Adaptive Streaming (QFAS): Problem and Solution

In this section we aim at deriving scheduling rates in order to maximize the overall video quality while minimizing the quality difference among multiple videos, but without considering buffer constraints. Since the resulting problem (7a)-(7d) is handled in the short-term, in the following we will drop the time index n for the sake of clarity.

Due to the relationship in (5) between R_k and $\tilde{r}_k$, the problem (7a)-(7d) is still MINLP. In order to derive low-complexity sub-optimal solution, we consider a continuous (relaxed) domain for r_k and for the utility U_k by using $r_k = R_k$. Accordingly, as shown in [45], the resulting relaxed version of the multi-objective problem (7a)-(7d) can be rewritten as a convex single-objective maximization problem where the objective in (7b) is eliminated and replaced by an equality constraint. The basic QFAS optimization problem is then stated as follows:

$$\max_{R_k, k\in\mathcal{V}} \sum_{k\in\mathcal{V}} U_k(R_k) \tag{10a}$$

$$s.t.\ U_k(Y_k) \le U_k(R_k) \le U_k(Z_k), \forall k \in \mathcal{V} \tag{10b}$$

$$\sum_{k\in\mathcal{V}} w_k R_k \le \Pi \tag{10c}$$

$$\Delta\big(U_i(R_i), U_j(R_j)\big) = 0 \quad \forall i,j \in \mathcal{V}, i \ne j \tag{10d}$$

The optimization problem in (10) admits a feasible solution under the condition $\sum_{k\in\mathcal{V}} w_k Y_k \le \Pi$. By considering the trivial condition $\sum_{k\in\mathcal{V}} w_k Z_k \ge \Pi$ as true, it has been proved in [13] that the problem (10) collapses in a constraint-satisfaction problem where the objective is achieved by fulfilling constraint (10c) with an equality constraint, *i.e.*, $\sum_{k\in\mathcal{V}} w_k R_k = \Pi$. The solution of problem (10), denoted with $\tilde{R}_k$, can be derived by relaxing constraint (10b) with two boolean variables and by applying the procedure with quadratic complexity described in Algorithm 1.

More specifically, we define $y_k, z_k \in \{0,1\}$, $k \in \mathcal{V}$, with $(y_k, z_k) \ne (0,0)$, the binary variables that indicate whether (1) or not (0) the two constraints $R_k \ge Y_k$ and $R_k \le Z_k$, respectively, are satisfied. We also define the function

$$\Gamma(\boldsymbol{y}, \boldsymbol{z}, U) = \sum_{k\in\mathcal{V}} y_k z_k w_k f_U^{-1}(U; \mathbf{a}_k) - \Pi(\boldsymbol{y}, \boldsymbol{z}) \tag{11}$$

where

$$\Pi(\boldsymbol{y}, \boldsymbol{z}) = \Pi - \sum_{k\in\mathcal{V}} w_k \big[(1-y_k)Y_k + (1-z_k)Z_k\big], \tag{12}$$

and f_U^{-1} is the inverse function of f_U. Since $f_U(r; \mathbf{a}_k)$ is a continuous and strictly increasing function of r, $f_U^{-1}(U; \mathbf{a}_k)$ is continuous and strictly increasing function of U. Algorithm 1 is finally derived by following the methods applied to the problem presented in [13], which has a similar structure. At each iteration in the inner loop (lines 9-18), it evaluates the fair utility value that allows to fulfill the available resources for all the HAS UEs not violating the constraints in (10b). Then, it iteratively checks whether or not the utility solutions of these UEs violate constraints in (10b). If this happens for one HAS UEs, the algorithm assigns the relative minimum (lines 12-17) or maximum (lines 19-23) utility to this specific UE and re-evaluates the fair utilities for all the other UEs.

It is worth to emphasize that, when the solutions $\tilde{R}_k$ at time n are known, the evaluation of $\tilde{r}_k[n]$ with (5) has the meaning of finding the nearest discrete values of rates that replace the relaxed continuous solutions. We finally note that the expected gap in terms of utility between the proposed sub-optimal and the optimal solutions is bounded with high probability by the sum of the utility differences between consecutive user profiles, which in turn depend on both the number of available profiles and their rate values. In fact, the larger is the number of available profiles, the better is the accuracy given by the continuous rate relaxation of the problem, considering that the utility difference between consecutive profiles decreases.

Algorithm 1 Pseudo code to solve problem (10)

1: **if** $\sum_{k\in\mathcal{V}} w_k Y_k \ge \Pi$ **then**
2: report infeasibility
3: **else if** $\sum_{k\in\mathcal{V}} w_k Z_k \le \Pi$ **then**
4: Set $\tilde{R}_k = Z_k$, $\forall k \in \mathcal{V}$
5: **else**
6: $z_k = 1$, $\forall\ k \in \mathcal{V}$;
7: **repeat**
8: $cond_{\mathrm{Z}}$ = false; $y_k = 1$, $\forall\ k \in \mathcal{V}$;
9: **repeat**
10: $cond_{\mathrm{Y}}$ = false;
11: Compute $\tilde{U} : \Gamma(\boldsymbol{y}, \boldsymbol{z}, \tilde{U}) = 0$;
12: **for all** $k \in \mathcal{V} : z_k y_k = 1$ **do**
13: $\tilde{R}_k = f_U^{-1}(\tilde{U}; \mathbf{a}_k)$;
14: **if** $\tilde{R}_k < Y_k$ **then**
15: $R_k = Y_k$; $y_k = 0$; $cond_{\mathrm{Y}}$ = true;
16: **end if**
17: **end for**
18: **until** $cond_{\mathrm{Y}}$ is false
19: **for all** $k \in \mathcal{V} : y_k z_k = 1$ **do**
20: **if** $\tilde{R}_k > Z_k$ **then**
21: $\tilde{R}_k = Z_k$; $z_k = 0$; $cond_{\mathrm{Z}}$ = true;
22: **end if**
23: **end for**
24: **until** $cond_{\mathrm{Z}}$ is false
25: **end if**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2504732, IEEE Transactions on Circuits and Systems for Video Technology

VI. QFAS WITH FAIR-BUFFER (QFAS-BF): PROBLEM AND SOLUTION

In section V we have derived content-aware continuous rate solution which enforces quality-fair video delivery. If this solution were directly used at the eNB to update the GBR values of the users, the following main fairness issues could arise concerning the buffer level, *i.e.*,

- In the case that the transmission rate of one user results to be equal to the discrete profile rate at each chunk request (*e.g.*, equal to the minimum rate $\tilde{R}_k[n] = Y_k[n], \forall n$), the client will have no possibility at all to build its own buffer.
- Users coming up into the system at different times when traffic load conditions are severe may have less possibility to build their own buffer with respect to the already present users.

In other words, by using basic QFAS solutions, we have no control on the buffer level among the users. This may lead to buffer starvation in practical scenarios where, due to the random channel variability, the actual download rate of the chunks may differ from the download rate estimated by the client or the MANE. provided by the eNB.

We overcome these issues by reshaping the transmission rate values $\tilde{R}_k[n]$ in order to satisfy the buffer fairness constraint (7e) of the general problem (7), without changing the discrete quality-fair rates achieved in the first step of optimization. More specifically, given the discrete profile rate solutions $\tilde{r}_k[n]$, evaluated as function of $\tilde{R}_k[n]$ as

$$\tilde{r}_k[n] = \max_{r \le \tilde{R}_k[n], r \in \mathcal{R}_k(c_k[n]+I_S)} r, \tag{13}$$

and the buffer fullness of each client, we derive a new set of unconstrained scheduling rates $\hat{R}_k[n]$ that allows to satisfy the constraints (13), (7d), and (7e) of the main problem, *i.e.*,

$$\hat{R}_k[n] > \tilde{r}_k[n], \forall k \in \mathcal{K} \tag{14}$$

$$\Pi[n] = \sum_{k \in \mathcal{V}} w_k[n] \hat{R}_k[n] \tag{15}$$

$$\lim_{n \to \infty} \delta B_k(c_j[n]) = 0, \forall k, j \in \mathcal{V}, k \neq j \tag{16}$$

The eq. (14) readily holds when suitable functions $\xi_k(\mathcal{B}[n]) > 0$ of a given set of buffer fullness values $\mathcal{B}[n] = \{b_1[n], \ldots, b_V[n]\}$ are considered, such that

$$\hat{R}_k[n] = \tilde{r}_k[n]\big[1 + \xi_k(\mathcal{B}[n])\big]. \tag{17}$$

With this choice we rewrite eq. (15) as

$$\sum_{k \in \mathcal{V}} w_k[n] \tilde{r}_k[n] (1 + \xi_k(\mathcal{B}[n])) = \Pi[n] \tag{18}$$

which is always satisfied by introducing a new unconstrained function $\eta(b_k[n]), k \in \mathcal{V}$, to have

$$\xi_k(\mathcal{B}[n]) = \frac{\Pi[n] - \sum_{i \in \mathcal{V}} w_i[n] \tilde{r}_i[n]}{\sum_{i \in \mathcal{V}} w_i[n] \tilde{r}_i[n] \eta(b_i[n])} \eta(b_k[n]). \tag{19}$$

This can be proved through simple algebra manipulations. To obtain new scheduling rates through (17) and (19), we finally need to determine the function $\eta(b_k[n])$ and the set $\mathcal{B}[n]$, that allows to satisfy condition (16). We have the following lemma.

Lemma 1: If at least one of the discrete rates $\tilde{r}_k[n]$ is strictly smaller than the corresponding QFAS continuous rate solution of Algorithm 1, *i.e.*,

$$\exists\, s \in \mathcal{V} : \tilde{r}_s[n] < \tilde{R}_s[n], \forall n \tag{20}$$

then the function

$$\eta(b_k[n]) = \frac{1}{b_k[n]} \begin{cases} 1 & \text{if } b_k[n] = \max_{v \in \mathcal{V}} b_v[n] \\ 1 + \epsilon & \text{otherwise} \end{cases} \tag{21}$$

with $b_k[n] = B_k(c_k^*[n]), \forall k$, where

$$c_k^*[n] = c_k[n^*(c_1)] + \underset{i=0,1}{\operatorname{argmin}} \left| B_1(c_1[n]) - B_k(c_k[n^*(c_1)] + i) \right| \tag{22}$$

and $0 < \epsilon \ll 1$ is an arbitrary constant value close to zero, satisfies all the conditions (14), (15) and (16).

Proof: See Appendix.

By combining the result of Lemma 1 together with (17) and (19), we finally obtain

$$\hat{R}_k[n] = \tilde{r}_k[n] \left(1 + \frac{1}{\mu_k B_k(c_k^*[n])} \frac{\Pi[n] - \sum_{i \in \mathcal{V}} w_i[n] \tilde{r}_i[n]}{\sum_{i \in \mathcal{V}} \frac{w_i[n] \tilde{r}_i[n]}{\mu_i B_i(c_i^*[n])}} \right). \tag{23}$$

where $\mu_k = 1$ if $B_k(c_k^*[n]) = \max_{v \in \mathcal{V}} B_v(c_v^*[n])$, $\mu_k = 1/(1+\epsilon)$ otherwise. Lemma 1 implies that if the UEs experience different values of buffer fullness at a given time, due to, *e.g.*, different start-up time, the selection of the GBR as in eq. (23) allows each UE to asymptotically achieve a fair buffer fullness value. This is done by selecting a transmission rate that is larger than the QFAS discrete rate of a factor inversely proportional to the buffer fullness of each client.

We remark that in the *look-ahead* approach the rate solution $\hat{R}_k[n] > \tilde{r}_k[n]$, should however not exceed the nearest discrete rate $r'_k[n] > \tilde{r}_k[n]$. Otherwise, the k-th client may select the profile rate $r'_k[n]$, given $\hat{R}_k[n] \geq r'_k[n]$, thus changing the discrete quality-fair rates resulting from the first step of optimization. A solution to the problem in (17)-(19) with the additional constraint $\hat{R}_k[n] < r'_k[n]$ can be derived with algorithmic methods similar to the ones proposed in the outer loop of Algorithm 1. However, we omit the discussion of this solution since the solution in (23) satisfies the constraint with high probability, given a reasonably large ratio between consecutive discrete rate profiles. In our numerical evaluation the constraint was never violated in all the scenario considered.

VII. FINAL REMARKS

The proposed QFAS-FB framework can be extended to consider the following important aspects of practical applications:

1) When the rate-utility functions change significantly between consecutive chunks of each video program, contradicting assumption *b* in Sect. IV-A, uncontrolled quality variations may occur by using the proposed framework. In order to limit the quality fluctuations (as in [35], [37]), the problem in (7) can be slightly modified by extending constraint (7c) as follows

$$\begin{cases} U_k(r_k[n]) \leq \min(U(Z_k), U(r_k^*[n-1]) + \delta^{\text{TH}}) \\ U_k(r_k[n]) \geq \max(U(Y_k), U(r_k^*[n-1]) - \delta^{\text{TH}}) \end{cases} \tag{24}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2504732, IEEE Transactions on Circuits and Systems for Video Technology

Cell layout	Single hexagonal cell
Cell Range	2 km
Carrier Frequency	700 MHz
Path Loss	COST 231-Hata model
Channel Model	ITU A Pedestrian model [47]
Shadowing	Log-normal
User Speed	3 km/h
System Bandwidth	10 MHz
UL/DL duplexing	FDD
Pre-scheduling time interval	100 ms

TABLE II
SIMULATION PARAMETERS

Sequence	Spatial Compl.	Temporal Compl.	Description
clip	Medium	Medium	A music video clip
spiderman	Medium	High	*Spiderman* movie
sport	Very High	Very High	Canoe competition
interview	Low	Very Low	An interview
bunny	Medium	Medium	*Big Buck Bunny* movie
home	High	High	*Home* documentary movie

TABLE III
TEST VIDEO SEQUENCES : COMPLEXITY (COMPL.) AND DESCRIPTION

where δ^{TH} is the maximum quality variation between consecutive chunks and $r_k^*[n]$ is the discrete rate solution at time n. This allows to keep the quality variations among consecutive chunks within a proper range.[7]

2) The QFAS-FB framework allows each client to infinitely increase the buffer level until the maximum buffer capacity is reached. However, in practical scenarios most of the users interrupt the play-back before the end of the full video. Therefore, the higher the maximum buffer level, the larger the waste of bandwidth for downloading data that will be discarded. To overcome this problem, the proposed framework can be easily extended to limit the transmission rate $\hat{R}_k[n]$ to exactly the value of the quality-fair discrete rate $\tilde{r}_k[n]$, when a suitably defined buffer fullness target B^{T} is reached by client k. This can be easily done by setting $b_k[n] = \infty$, if $b_k[n] \geq B^{\mathrm{T}}$, without compromising the asymptotic optimality of the proposed framework. In this case, the buffer fullness can be stabilized to B^{T} and the remaining resources can be exploited by the other HAS/data users.

VIII. NUMERICAL RESULTS

The numerical results are obtained through Monte Carlo simulations on a *ns2*-platform which includes HAS servers and clients, LTE radio interface and radio resource management, as well as the different protocol layers (TCP/IP, PDCP and RLC). The UEs are uniformly distributed in a cell with an average SNR ranging from 2 to 25 dB. We consider $V = 6$ HAS UEs and $L = 24$ FTP UEs, and a log-normal shadowing with a standard deviation of 8 dB, if not otherwise specified. The main system parameters are listed in Table II. The simulation time of each drop is set to 360 seconds and UEs are randomly activated during the first 60 seconds. The sequences are extracted from real time programs with 4CIF resolution and frame rate equal to 30 fps. Each of them is encoded by the DASH Encoder [46] with 10 profiles having rates ranging from 150 kbps to 5 Mbps. Chunk duration T_{c} is set to 2 seconds.

The RDA engine is based on the Microsoft Smooth Streaming player [4], which aims at heuristically stabilizing the buffer fullness to 20 seconds. The RDA may select an higher (lower) profile rate if the buffer fullness is larger (smaller) than a suitable threshold, *e.g.*, 30 (10) seconds, and the slope of the buffer evolution is high, *i.e.*, the buffer level is quickly changing. The maximum buffer threshold under which the RDA requests a chunk immediately after the previous chunk is downloaded is set to 40 seconds. The so-called "panic" buffer threshold, which causes the RDA to necessarily request the lowest profile is set to 5 seconds. The first chunk is requested at the lowest profile. Each client starts to play-out the video after the first chunk is received.

As mentioned in Sect. III, we consider the SSIM metric to assess the perceived video quality. The SSIM is a value between -1 and 1. According to [48] SSIM values larger than 0.94 correspond to a good visual quality, *i.e.*, mean opinion score (MOS) equal to 4, whereas SSIM values ranging between 0.86 and 0.94 result in a MOS of 3, *i.e.*, fair quality. SSIM values lower than 0.86 indicate poor quality. The validation results of the model in eq. (2) show almost perfect correlation to empirical points with a Pearson coefficient always higher than 0.99 for each chunk of the considered video sequences.

We compare three QFAS-based proposed strategies, *i.e.*, (i) the basic *look-ahead* QFAS strategy (QFAS-LH) where the GBRs of the clients are evaluated as the outcome rates of Algorithm 1 without the buffer control, (ii) the *look-ahead* QFAS with fair buffer strategy (QFAS-FB-LH) where the GBR is evaluated according to eq. (23), and (iii) the *overwrite* QFAS with buffer fair strategy (QFAS-FB-OW), with the two following approaches: (i) best effort (BE), where all UEs are non-GBR with QCI equal to 9 [11]; (ii) AGBR approach proposed in [12] where the GBR values are updated every 2 seconds for each HAS UE (non-GBR UEs have QCI equal to 9 and HAS UEs have QCI equal to 4).

In order to have a fair comparison between AGBR and QFAS-based solutions, the amount of available resources Π dedicated to HAS UEs is dynamically updated every pre-scheduling time instant, according to the on-line implementation proposed in [12], *i.e.*,

$$\Pi[n] = \pi \frac{V[n]}{V[n] + L[n]} + (1 - \pi)\Pi[n-1] \qquad (25)$$

where $V[n]$ and $L[n]$ are the number of active HAS and data UEs, respectively, at time nT_{s}. The parameter $\pi = 10^{-3}$ is selected to slowly adapt the required resource partitioning to the number of UEs change. In order to have a clear understanding of the results, it is important to remind that the behavior of the system in the *look-ahead* strategy depends on both QFAS-based algorithm and RDA at the client. If we focus on the buffer, the main task of the QFAS-FB-LH algorithm is to allow buffer growth to all clients, since the buffer level is

[7] In the results that will be presented in Sect. VIII, we have used this extension with $\delta^{\mathrm{TH}} = 0.05$. However, in all our experiments, the constraint in (24) has never been violated.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2504732, IEEE Transactions on Circuits and Systems for Video Technology

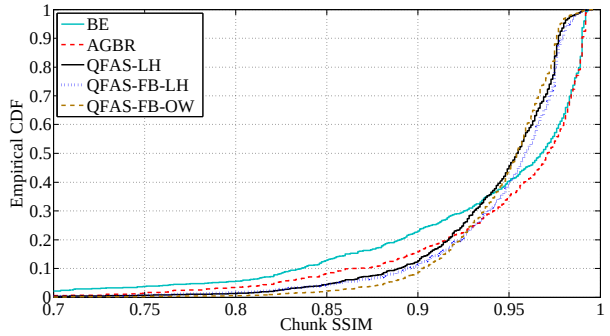


Fig. 3. Empirical CDF of the chunk-by-chunk SSIM at the clients

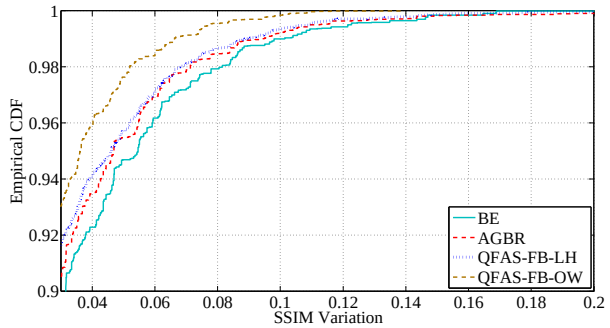


Fig. 4. Empirical CDF of the chunk-by-chunk SSIM variation δ_{SSIM} at the clients

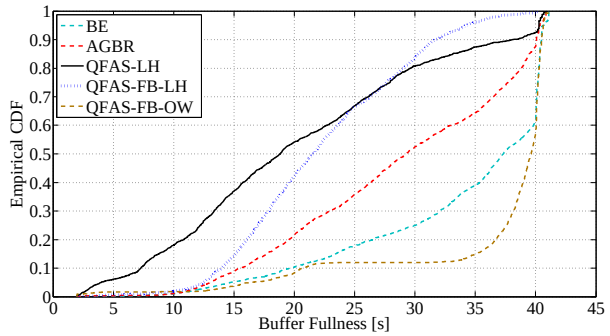


Fig. 5. Empirical CDF of the buffer fullness evaluated at each chunk request of the clients

controlled by the RDA. In the *overwrite* strategy, QFAS-FB algorithm is also able to provide buffer fairness, since it has full control of the video streaming optimization.

Fig. 3 shows the empirical cumulative distribution function (CDF) of the chunk-by-chunk SSIM at the clients for each strategy. We can note how AGBR approach allows to strictly increase the received quality with respect to BE approach, mainly thanks to the increasing of the average rate provided to HAS clients. The QFAS-based approaches perform similarly in terms of received SSIM by improving the received quality of approximately the 40th and the 30th percentile of the received chunks, with respect to BE and AGBR approaches, respectively. Moreover, in the 10 % of the cases the QFAS-FB-OW strategy increases the video quality from 0.85 to 0.9 with respect to content-blind optimization strategies, whereas high-SSIM users lose no more than 0.03. Nevertheless, such

FTP UEs	HAS UEs	BE	AGBR	QFAS-LH	QFAS-FB-LH	QFAS-FB-OW
24	6	0.935	0.947	0.941	0.945	0.946
	12	0.925	0.942	0.938	0.939	0.940
	18	0.918	0.935	0.926	0.928	0.931
12	6	0.958	0.967	0.967	0.967	0.970
36		0.914	0.934	0.930	0.931	0.935

TABLE IV
AVERAGE SSIM AT THE CLIENTS

FTP UEs	HAS UEs	BE	AGBR	QFAS-LH	QFAS-FB-LH	QFAS-FB-OW
24	6	0.068	0.054	0.029	0.027	0.017
	12	0.072	0.060	0.032	0.032	0.021
	18	0.073	0.064	0.038	0.036	0.029
12	6	0.041	0.032	0.017	0.017	0.009
36		0.086	0.065	0.039	0.036	0.031

TABLE V
STANDARD DEVIATION OF THE SSIM AT THE CLIENTS

quality degradation is hardly be perceived by the end-user, as many contributions in literature have shown, *e.g.*, [49].

Table IV and V report the SSIM averaged over all clients and simulations, and the related standard deviation for different numbers of HAS and FTP users. When the number of HAS users increases and the number of FTP users is fixed, the average portion of resources dedicated to each single HAS user decreases (see eq. (25)), leading to a general worsening of the mean and the standard deviation of the video quality. The average SSIM of the QFAS-based strategies is approximately equal to that of AGBR approach, but with a significant improvement in the quality fairness among clients. The standard deviation of the utilities, *i.e.*, the amount of quality variation perceived by each client for each chunk, is reduced up to about four times. The QoE fairness improvement of the proposed strategies is also evident in Table VI where the average Jain index [50] of the SSIM is reported. The Jain index evaluates the fairness of a set of V values and ranges from $1/V$ (worst fairness) to 1 (maximum fairness).

The improvements achievable by the QFAS-FB-OW strategy are more noticeable in Fig. 4 where the stability of the considered approaches, *i.e.*, the ability to avoid large quality fluctuation at the clients, is investigated by evaluating the normalized SSIM variation between consecutive chunks, *i.e.*, $|U_k(c_k[n]+1) - U_k(c_k[n])|/U_k(c_k[n])$. To this purpose, the empirical CDF reported in Fig. 4 has a range limited to the high percentile values where fluctuations on the SSIM are larger than 0.04. The possibility to overwrite the chunk request allows the QFAS-FB-OW strategy to achieve the best performance, by limiting the quality fluctuations generated by the RDA at the clients, to a value smaller than 0.03 in 95 % of the cases. It is interesting to note that also QFAS-FB-LH (QFAS-LH is omitted for sake of clarity), strictly overtakes AGBR and BE strategies.

The behavior of the buffer at the clients resulting from the investigated strategies is analyzed in Fig. 5 and Table VI, where the empirical CDF of the buffer fullness samples $B_k[n]$, *i.e.*, the amount of video time still available to play-out in the buffer at each chunk request, and the average Jain index of the buffer fullness levels computed every 2 seconds of simulation, are reported. The basic QFAS-LH approach provides the worst results in terms of buffer stability and fairness, especially

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2504732, IEEE Transactions on Circuits and Systems for Video Technology

for clients requesting most of the times the chunks with the lowest profile rate. Compared to BE and AGBR, QFAS-FB-LH significantly increases the buffer fairness at the clients as the average Jain index moves from approximately 0.90 to 0.94. Nevertheless, it still performs worse with respect to the QFAS-FB-OW strategy, which provides almost perfect buffer fairness, with an average Jain index approximately equal to 0.99.

This behaviour appears more clear in Fig. 6 and Fig. 7, where the time evolution of both chunk-by-chunk SSIM and buffer fullness at the clients are reported for one drop of simulation, respectively. Here, we consider a more challenging scenario, where the standard deviation of the log-normal shadowing is set to 12 dB. It can be clearly noted in Fig. 6a that a fair-rate approach, *i.e.*, AGBR, provides unfair video quality by significantly impairing the QoE of the clients streaming high-motion videos, *e.g.*, "sport" and "home", with respect to other clients playing low-motion movies, *e.g.*, "interview". In the QFAS-LH strategy without control of the buffer (see Fig. 7b), the buffer of "interview" is always close to the panic threshold and the user experiences several rebuffering events, *i.e.*, video playback interruptions. In this specific case the rebuffering percentage is equal to 7 % with a maximum rebuffering time of 3.5 seconds. The QFAS-FB-LH strategy counteracts this issue by allowing all the clients to build their buffer. As a result, the video playbacks are never interrupted. However, the buffer stability, as well as the quality improvements, are still not remarkable. This is also due to the fact that the state-of-the-art RDA at the client generates quality fluctuations. This issue is definitely overcome by the QFAS-FB-OW strategy, which provides both video fair quality and buffer stability.

	BE	AGBR	QFAS-LH	QFAS-FB-LH	QFAS-FB-OW
BF	0.904	0.902	0.815	0.939	0.989
SSIM	0.881	0.914	0.987	0.990	0.999

TABLE VI
AVERAGE JAIN INDEX OF THE BUFFER FULLNESS (BF) AND THE QUALITY (SSIM) AT THE CLIENTS FOR $V = 6$ AND $L = 24$

Finally, in Fig. 8 we report the empirical CDF of the average FTP UE throughput. We note that, although in the QFAS-based approaches the GBRs are updated asynchronously, thereby draining some FTP resources in the short-term, the average performance of the FTP UEs is not degraded with respect to AGBR scenario, where the resource constraint (7d) always holds.

IX. CONCLUSIONS

In this paper, we have proposed the QFAS-FB framework to optimize QoE and fairness for HAS video delivery in LTE networks, according to two strategies, named *look-ahead* and *overwrite* strategies. By adding intelligence in the network, *i.e.*, through the use of a MANE, the proposed approaches are able to control the rate provided to each HAS user in order to obtain fair video quality among multiple HAS clients and to asymptotically achieve a fair buffer fullness level. This is achieved even when HAS users are requesting programs with significant differences in video complexity and

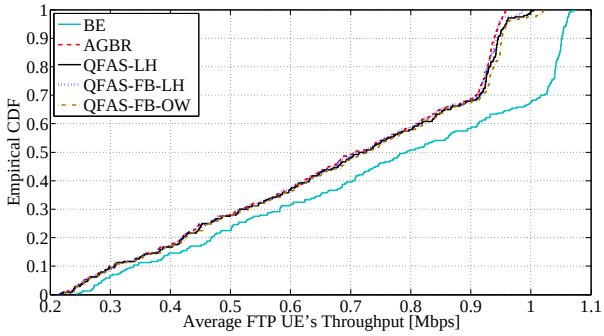


Fig. 8. Empirical CDF of the average FTP UE's throughput.

buffer level, and are experiencing different channel conditions. Numerical results have shown that, compared to other state-of-the-art frameworks, our proposed QFAS-FB solution provides a significant improvement of the overall quality delivered to users demanding complex videos at the expense of a tolerable degradation of the other low-complexity videos. Moreover, The QFAS-FB *overwrite* strategy is able to significantly reduce the quality fluctuations at the clients by also providing the best results in terms of buffer fairness, when compared to other content-agnostic HAS solutions. The proposed framework has been built on the assumption that HAS clients request a profile rate not larger than the transmission rate, thereby enforcing video quality and buffer stability. However, the QoE might be improved if HAS clients with a large buffer level were allowed to occasionally increase their profile rate. An interesting topic for further research is to jointly address the multi-client quality fairness and the trade-off among video quality, buffer stability and quality variations, following the ideas in [37].

APPENDIX
PROOF OF LEMMA 1

For the sake of conciseness, we use here symbols n_k and m_k to denote the two time instants at which the request of chunk $c_k^*[n]$ occurs, and the download of chunk $c_k^*[n]$ ends, respectively. It is straightforward to prove that $\delta B_k(c_1[n]) \xrightarrow{n\to\infty} 0, \forall k \in \mathcal{V} \setminus \{1\}$ is a sufficient condition for (16). Therefore, after defining $b_k[n_k] = B_k(c_k^*[n])$ and $b_k[m_k] = B_k(c_k^*[n]+1)$, we rewrite it as

$$\delta b_k[n_k] = |b_1[n_1] - b_k[n_k]| \xrightarrow{n_1\to\infty} 0. \tag{26}$$

We prove that the choice of $\eta(b_k[n_k]) = 1/(\mu_k b_k[n_k])$ satisfies (26), if condition (20) holds. From eq. (6) and eq. (17) we have

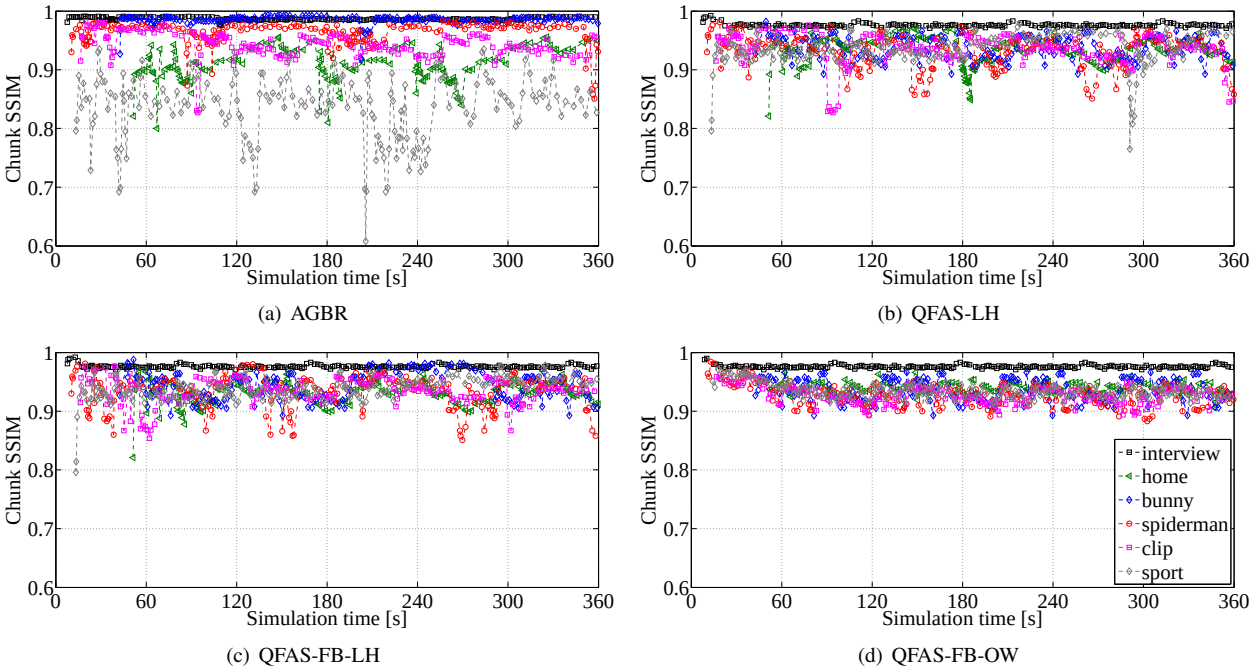
$$\hat{R}_k[n] = \tilde{r}_k[n]\left(1 + \frac{1}{\mu_k b_k[n_k]} g(\mathcal{B}[n])\right) = \frac{\tilde{r}_k[n] T_c}{T_c - \Delta b_k[m_k]} \tag{27}$$

where $\Delta b_k[m_k] = b[m_k] - b[n_k]$ is the buffer increment at time m_k, and

$$g(\mathcal{B}[n]) = \frac{\Pi[n] - \sum_{i\in\mathcal{V}} w_i[n]\tilde{r}_i[n]}{\sum_{i\in\mathcal{V}} \frac{w_i[n]\tilde{r}_i[n]}{\mu_i b_i[n_i]}}. \tag{28}$$

From the assumption in (20) we have $\Pi[n] > \sum_{i\in\mathcal{V}} w_i[n]\tilde{r}_i[n]$, thus $g(\mathcal{B}[n]) > 0$ readily leads to

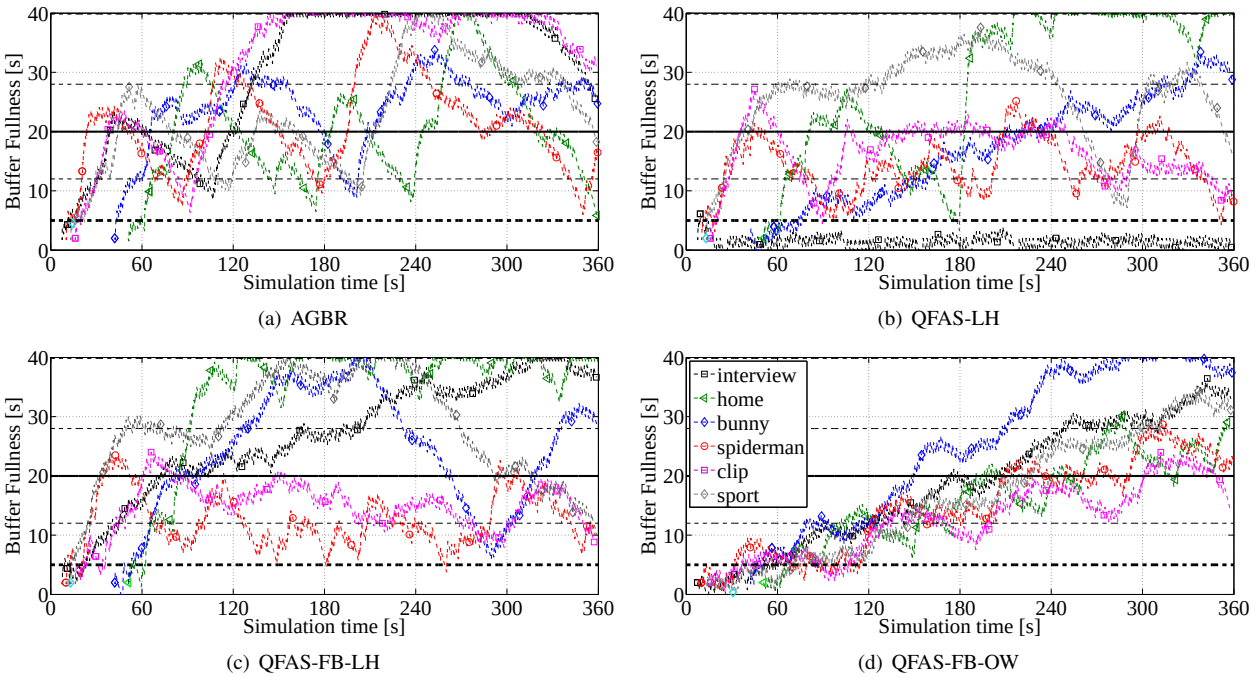
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2504732, IEEE Transactions on Circuits and Systems for Video Technology



(a) AGBR (b) QFAS-LH

(c) QFAS-FB-LH (d) QFAS-FB-OW

Fig. 6. An example (one drop of simulation) of the chunk-by-chunk SSIM at the clients over time for the GBR-aware strategies.



(a) AGBR (b) QFAS-LH

(c) QFAS-FB-LH (d) QFAS-FB-OW

Fig. 7. An example (one drop of simulation) of the evolution of the buffer fullness at the clients for the GBR-aware strategies.

$\Delta b_k[m_k] > 0$. After some simple algebra manipulation of (27) we obtain:

$$\Delta b_k[m_k] = T_c \frac{g(\mathcal{B}[n])}{g(\mathcal{B}[n]) + \mu_k b_k[n_k]} > 0 \qquad (29)$$

which, as expected, leads to $\Delta b_k[m_k] < T_c$. Without loss of generality, we assume $b_1[n_1] = \max_{v \in \mathcal{V}} b_v[n_v]$ and we consider the non-trivial case $b_k[n_k] < b_1[n_1], \forall k \in \mathcal{V} \setminus \{1\}$ leading to $\mu_1 = 1, \mu_k = 1/(1+\epsilon)$. We next prove that the following three conditions hold, *i.e.*,

$$|\delta b_k[m_k]| \leq |\delta b_k[n_k]|, \forall k \in \mathcal{V} \setminus \{1\}, \forall n \qquad (30)$$

$$\delta b_k[n_k] \delta b_k[m_k] \geq 0, \forall k \in \mathcal{V} \setminus \{1\}, \forall n \qquad (31)$$

where the equalities hold if and only if $b_1[n_1] = b_k[n_k], \forall k$, and

$$\lim_{n_1 \to \infty} \delta b_k[m_k] - \delta b_k[n_k] = 0 \Leftrightarrow \lim_{n_1 \to \infty} \delta b_k[n_k] = 0. \qquad (32)$$

thereby proving the lemma. All conditions are verified by noting that

$$\delta b_k[m_k] - \delta b_k[n_k] = \Delta b_1[m_1] - \Delta b_k[m_k] \tag{33a}$$

$$= T_c \frac{g(\mathcal{B}[n])}{g(\mathcal{B}[n]) + b_1[n_1]} - T_c \frac{g(\mathcal{B}[n])}{g(\mathcal{B}[n]) + \mu_k b_k[n_k]} \tag{33b}$$

$$= T_c \frac{g(\mathcal{B}[n])(\mu_k b_k[n_k] - b_1[n_1])}{(g(\mathcal{B}[n]) + b_1[n_1])(g(\mathcal{B}[n]) + \mu_k b_k[n_k])} \tag{33c}$$

$$< - T_c \frac{g(\mathcal{B}[n])\delta b_k[n_k]}{(g(\mathcal{B}[n]) + b_1[n_1])(g(\mathcal{B}[n]) + \mu_k b_k[n_k])}. \tag{33d}$$

To summarize, we have:

$$\delta b_k[m_k] < \delta b_k[n_k](1 - h(\mathcal{B}[n])) \tag{34}$$

where $h(\mathcal{B}[n]) = \frac{g(\mathcal{B}[n])T_c}{(g(\mathcal{B}[n])+b_1[n_1])(g(\mathcal{B}[n])+\mu_k b_k[n_k])} > 0$. Moreover, $b_1[n_1] + \mu_k b_k[n_k] > T_c$ easily follows from condition $b_k[n_k] > T_c, \forall k$, which implies $h(\mathcal{B}[n]) < 1$. By combining this last inequality with (34), both conditions (30) and (31) are readily proved. From (34), it is also evident that $(\delta b_k[m_k] - \delta b_k[n_k]) < \delta b_k[n_k] h(\mathcal{B}[n])$ satisfies (32) thanks to the strict lower bound of $h(\mathcal{B}[n])$, thus ending the proof.

REFERENCES

[1] "Cisco visual networking index: Global mobile data traffic forecast update, 2014-2019," Cisco, Tech. Rep., 2014.

[2] H. Holma and A. Toskala, *LTE Advanced: 3GPP Solution for IMT-Advanced*. Wiley, 2012.

[3] T. Stockhammer, "Dynamic adaptive streaming over HTTP design principles and standards," *in Proc. ACM conference on Multimedia systems*, pp. 133–144, 2011.

[4] A. Zambelli. Smooth streaming technical overview microsoft corporation, 2009. [Online]. Available: http://www.iis.net/learn/media/on-demand-smooth-streaming/smooth-streaming-technical-overview

[5] Apple HTTP live streaming. [Online]. Available: http://tools.ietf.org/html/draft-pantos-http-live-streaming-07

[6] Adobe HTTP adaptive streaming. [Online]. Available: http://www.adobe.com/products/hds-dynamic-streaming.html

[7] 3GPP, "LTE; transparent end-to-end packet-swiched streaming service (PSS); progressive download and dynamic adaptive streaming over HTTP (3GP-DASH)," TS 26.247 v10.7.0, 2012.

[8] "International Standards Organization/International Electrotechnical Commission (ISO/IEC), 23009-1:2012 Information Technology Dynamic Adaptive Streaming over HTTP (DASH) Part 1: Media Presentation Description and Segment Formats," 2000.

[9] I. Sodagar, "The MPEG-DASH standard for multimedia streaming over the internet," *IEEE MultiMedia*, vol. 18, no. 4, pp. 62–67, April 2011.

[10] S. Sesia, I. Toufik, and M. Baker, *LTE, The UMTS Long Term Evolution: From theory to practice*. Wiley, 2009.

[11] 3GPP, "Policy and charging control architecture," TS 23.203, v10.7.0, 2012.

[12] D. De Vleeschauwer, H. Viswanathan, A. Beck, S. Benno, G. Li, and R. Miller, "Optimization of HTTP adaptive streaming over mobile cellular networks," in *Proc. IEEE INFOCOM*, 2013, pp. 898–997.

[13] S. Cicalò and V. Tralli, "Distortion-fair cross-layer resource allocation for scalable video transmission in OFDMA wireless networks," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 848–863, April 2014.

[14] N. Changuel, B. Sayadi, and M. Kieffer, "Control of multiple remote servers for quality-fair delivery of multimedia contents," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 4, pp. 746–759, April 2014.

[15] A. El Essaili, D. Schroeder, E. Steinbach, D. Staehle, and M. Shehada, "QoE-based traffic and resource management for adaptive HTTP video delivery in LTE," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 6, pp. 988–1001, June 2015.

[16] H. Park and M. van der Schaar, "Fairness strategies for wireless resource allocation among autonomous multimedia users," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 2, pp. 297–309, Feb 2010.

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

[18] N. Bouten, S. Latre, J. Famaey, W. Van Leekwijck, and F. De Turck, "In-network quality optimization for adaptive video streaming services," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2281–2293, Dec 2014.

[19] S. Cicalò, N. Changuel, R. Miller, B. Sayadi, and V. Tralli, "Quality-fair HTTP adaptive streaming over LTE network," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 714–718.

[20] S. Mehrotra and W. Zhao, "Rate-distortion optimized client side rate control for adaptive media streaming," in *Proc. IEEE International Workshop on Multimedia Signal Processing (MMSP)*, Oct 2009, pp. 1–6.

[21] D. Jarnikov and T. Özelebi, "Client intelligence for adaptive streaming solutions," *Signal Processing: Image Communication*, vol. 26, no. 7, pp. 378–389, 2011.

[22] C. Liu, I. Bouazizi, and M. Gabbouj, "Rate adaptation for adaptive HTTP streaming," in *Proc. 2nd Annual ACM Conf. on Multimedia Systems (MMSys '11)*. New York, USA: ACM, 2011, pp. 169–174.

[23] C. Müller, S. Lederer, and C. Timmerer, "An evaluation of dynamic adaptive streaming over HTTP in vehicular environments," in *Proc. of the 4th Workshop on Mobile Video (MoVid)*. ACM, 2012, pp. 37–42.

[24] Y. Huang, R. Johari, and N. McKeown, "Downton abbey without the hiccups: buffer-based rate adaptation for HTTP video streaming," in *Proc. ACM SIGCOMM Workshop on Future Human-centric Multimedia Networking (FhMN '13)*. ACM, 2013, pp. 9–14.

[25] X. Zhu, Z. Li, R. Pan, J. Gahm, and H. Hu, "Fixing multi-client oscillations in HTTP-based adaptive streaming: A control theoretic approach," in *Proc. IEEE 15th Int. Workshop on Multimedia Signal Processing (MMSP)*, Sept 2013, pp. 230–235.

[26] S. Akhshabi, L. Anantakrishnan, A. C. Begen, and C. Dovrolis, "What happens when HTTP adaptive streaming players compete for bandwidth?" in *Proc. 22nd Int. Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV '12)*. New York, USA: ACM, 2012, pp. 9–14.

[27] J. Jiang, V. Sekar, and H. Zhang, "Improving fairness, efficiency, and stability in HTTP-based adaptive video streaming with FESTIVE," in *Proc. 8th Int. Conf. on Emerging Networking Experiments and Technologies (CoNEXT)*. New York, USA: ACM, 2012, pp. 97–108.

[28] 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 J. Sel. Areas Commun.*, vol. 32, no. 4, pp. 719–733, 2014.

[29] S. Hu, L. Sun, C. Gui, E. Jammeh, and I.-H. Mkwawa, "Content-aware adaptation scheme for QoE optimized DASH applications," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec 2014, pp. 1336–1341.

[30] V. Ramamurthi and O. Oyman, "Video-QoE aware radio resource allocation for HTTP adaptive streaming," in *Proc. IEEE Int. Conf. on Commun. (ICC)*, June 2014, pp. 1076–1081.

[31] J. Chen, R. Mahindra, M. A. Khojastepour, S. Rangarajan, and M. Chiang, "A scheduling framework for adaptive video delivery over cellular networks," in *Proc. 19th Annual Int. Conf. on Mobile Computing; Networking (MobiCom '13)*, 2013.

[32] Y. P. Fallah, H. Mansour, S. Khan, P. Nasiopoulos, and H. Alnuweiri, "A link adaptation scheme for efficient transmission of H.264 scalable video over multirate WLANs," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 7, pp. 875–887, July 2008.

[33] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hossfeld, and P. Tran-Gia, "A survey on quality of experience of HTTP adaptive streaming," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 469–492, March 2015.

[34] M. Zhao, X. Gong, J. Liang, W. Wang, X. Que, and S. Cheng, "QoE-driven cross-layer optimization for wireless dynamic adaptive streaming of scalable videos over HTTP," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 3, pp. 451–465, March 2015.

[35] Z. Yan, J. Xue, and C. W. Chen, "QoE continuum driven HTTP adaptive streaming over multi-client wireless networks," in *Proc. IEEE Int. Conf. on Multimedia and Expo (ICME)*, July 2014, pp. 1–6.

[36] X. Liu and A. Men, "QoE-aware traffic shaping for HTTP adaptive streaming," *International Journal of Multimedia & Ubiquitous Engineering*, vol. 9, no. 2, 2014.

[37] V. Joseph and G. de Veciana, "NOVA: QoE-driven optimization of DASH-based video delivery in networks," in *Proc. IEEE INFOCOM*, April 2014, pp. 82–90.

[38] Y. Sanchez, T. Schierl, C. Hellge, T. Wiegand, D. Hong, D. De Vleeschauwer, W. Van Leekwijck, and Y. Le Louédec, "Efficient HTTP-based streaming using scalable video coding," *Signal Processing: Image Communication*, vol. 27, no. 4, pp. 329–342, 2012.

[39] J. Famaey, S. Latre, N. Bouten, W. Van de Meerssche, B. De Vleeschauwer, W. Van Leekwijck, and F. De Turck, "On the merits of SVC-based HTTP adaptive streaming," in *Proc. IFIP/IEEE*

Int. Symposium on Integrated Network Management (IM 2013), 2013, May 2013, pp. 419–426.

[40] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[41] K. Seshadrinathan, R. Soundararajan, A. Bovik, and L. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427 – 1441, June 2010.

[42] H. Kim and S. Choi, "QoE assessment model for multimedia streaming services using QoS parameters," *Multimedia Tools and Applications*, vol. 72, no. 3, pp. 2163–2175, 2014.

[43] K. Seshadrinathan, R. Soundararajan, A. Bovik, and L. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, June 2010.

[44] A. E. Essaili, D. Schroeder, D. Staehle, M. Shehada, W. Kellerer, and E. Steinbach, "Quality-of-experience driven adaptive HTTP media delivery," in *Proc. IEEE Int. Conf. on Commun. (ICC)*, Budapest, Hungary, Jun 2013.

[45] S. Cicalò, A. Haseeb, and V. Tralli, "Fairness-oriented multi-stream rate adaptation using scalable video coding," *Signal Processing: Image Communication*, vol. 27, no. 8, pp. 800–813, 2012.

[46] S. Lederer, C. Müller, and C. Timmerer, "Dynamic adaptive streaming over HTTP dataset," in *Proc. 3rd Multimedia Systems Conference (MMSys '12)*. New York, USA: ACM, 2012, pp. 89–94.

[47] 3GPP, "User equipment (UE) conformance specification, radio transmission and reception. part 1: Conformance testing," TS 36.521-1, v11.0.1, 2013.

[48] M. Shehada, S. Thakolsri, Z. Despotovic, and W. Kellerer, "QoE-based cross-layer optimization for video delivery in long term evolution mobile networks," in *Proc. 14th Int. Symposium on Wireless Personal Multimedia Communications (WPMC)*, Oct 2011, pp. 1–5.

[49] T. Zinner, O. Hohlfeld, O. Abboud, and T. Hossfeld, "Impact of frame rate and resolution on objective QoE metrics," in *Proc. 2nd Int. Workshop on Quality of Multimedia Experience (QoMEX)*, June 2010.

[50] R. K. Jain, D. M. W. Chiu, and W. R. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," Digital Equipment Corp., Maynard, MA, USA, DEC-301, Tech. Rep., 1984.

Sergio Cicalò (S'10-M'14) received the B.S. Degree, the M.Sc. degree and the Ph.D. Degree from the University of Ferrara, Italy, in 2006, in 2010, and in 2014, respectively, all in Communication Engineering. He is currently a Research Engineer at the Engineering Department of the University of Ferrara. He authored 10+ peer-reviewed publications in international Journal and Conference Proceedings and serves as a reviewer for IEEE Transactions/Journals and Conferences. His research interests are in the wide area of wireless communication and video delivery systems.

Nesrine Changuel is Program manager at Microsoft in Stockholm managing projects on video streaming for Skype and Lync. Previously, she worked as a Research Engineer in Alcatel Lucent Bell Labs. She received the B.S. degree in 2006 and M.S./Eng. degree in electrical engineering in 2008 from the Ecole Nationale Supérieure d'Electronique et de Radio-électricité de Grenoble, France, and the Ph.D. degree in Control and Signal Processing in 2011 from the Paris-Sud University, Orsay, France. She authored over 20 publications in journal and conference proceedings and serves as a regular reviewer for several peer review journals and conferences, and has six patent applications in the area of video coding and delivery.

Velio Tralli (S'93-M'94-SM'05) received the Dr.Ing. degree in electronic engineering (cum laude) and the Ph.D. degree in electronic engineering and computer science from the University of Bologna, Italy, in 1989 and 1993, respectively. From 1994 to 1999, he was a Researcher of the National Research Council (CNR) at CSITE, University of Bologna. In 1999, he joined the Engineering Department of the University of Ferrara, Italy where he is currently an Associate Professor. His research interests are within the areas of digital transmission and coding, and wireless communications, with emphasis on radio resource management, cross-layer design and optimization, OFDM and multiantenna systems, and cooperative communications. He participated in several national and European research projects addressing short-range communications systems, wireless sensor networks, 3G-4G wireless networks, wireless video communications. He published more than one hundred papers in refereed journals, including Transactions of the IEEE, and international conferences. He is a Senior Member of IEEE where he serves as a reviewer for Transactions/Journals and Conferences, and as a TPC member for several international conferences. He is an associate Editor for the European Transactions on Emerging Technologies. He also served as a Co-Chair for the Wireless Communication Symposium of ICC2006 and for the Communication Theory Symposium of ICC2013.

Bessem Sayadi received a Telecommunication Engineering degree in 1999 from SUPCOM Tunis and both M.Sc. (2000) and Ph.D. (2003) degrees in Control and Signal processing from Supélec, Paris-Sud University, with highest distinction. He was a postdoctoral fellow for two years in the National Centre for Scientific Research (CNRS), and worked for Orange Labs as senior researcher from 2005 to 2006. Since 2006, Dr Sayadi is a Senior Researcher in Wireless Technology Program at Alcatel-Lucent Bell Labs. His research expertise covers wireless networks, ranging from physical and MAC layer design, scheduling, broadcasting to transport protocols, video coding and delivery and Future Internet architectures. He has participated in many EU projects. He was the project coordinator of FP7 MEDIEVAL project. Now, he is leading Alcatel-Lucent Bell-Labs France's team in 5G-PPP/5G-NORMA Project. He has authored over 60 publications in journal and conference proceedings and serves as a regular reviewer for several technical journals and conferences. He holds 17 patents and has more than twenty five patent applications pending in the area of video coding and wireless communications.

Frédéric Faucheux is a researcher in the End2End Mobile Network and Services project, part of the Wireless Program at Alcatel-Lucent Bell Labs in Villarceaux, France. He graduated from Télécom SudParis Engineering school in Evry. After joining Alcatel-Lucent, he worked on the development of GPRS and UMTS software before focusing on video topics when entering the research and innovation department of Alcatel-Lucent (Bell Labs). His current research interests are media streaming technologies and network functions virtualization for 5G.

Sylvaine Kerboeuf is a senior researcher in the Wireless Program at Alcatel-Lucent Bell Labs in Villarceaux, France. She received an M.S. degree in physics (1991) and a Ph.D. in solid state physics from Paris Sud University, Orsay France (1994). After her Ph.D. in the superconductivity field at Centre National d'Etude des Télécommunications (CNET) at France Telecom, she joined Alcatel's Research and Innovation department and worked for several years on research projects in optoelectronics. In 2004, she joined a project working on radio access networks and focusing on fourth generation (4G) discontinuous networks and on caching technology. Her current research interests include end-to-end video delivery over wireless networks, video transport protocols, video quality of experience optimization and software defined networking for 5G.