To Each According to his Needs: Dimensioning Video Buffer for Specific User Profiles and Behavior

Tobias Hoßfeld
University of Duisburg-Essen
Modeling of Adaptive Systems
Essen, Germany
Email: tobias.hossfeld@uni-due.de

Christian Moldovan
University of Duisburg-Essen
Modeling of Adaptive Systems
Essen, Germany
Email: christian.moldovan@uni-due.de

Christian Schwartz
University of Würzburg
Chair of Communication Networks
Würzburg, Germany
Email: christian.schwartz@uni-wuerzburg.de

***Abstract*—Today's video streaming platforms offer videos in a variety of quality settings in order to attract as many users as possible. But even though a sufficiently dimensioned network can not always be provided for the best experience, users are asking for high QoE. Users consume the content of a video streaming platform in different ways, while video delivery platforms currently do not account for these scenarios and thus ensure at best mediocre QoE. In this paper, we develop a queuing model and provide a mean-value analysis to investigate the impact of user profiles on the QoE of HTTP Video Streaming for typical user scenarios. Our results show that the user profile and particularly the scenario have to be respected when dimensioning the buffer. Further, we present recommendations on how to adapt player parameters in order to optimize the QoE for individual users profiles and viewing habits. The provided model leads to relevant insights that are required to build a system that guarantees each user the best attainable QoE.**

I. Introduction

There are many ways to improve the QoE in Hypertext Transfer Protocol (HTTP) video streaming. However, these improvements, such as adaptive video streaming, often come with a harsh trade-off like a reduction of the videos resolution. Nonetheless, a user might insist on watching a certain video in HD even though his network does not provide enough bandwidth to stream the whole video without stalling. A different way to optimize the QoE is dimensioning the size of the video buffer. However, QoE varies for each user and users perceive problems differently [1]. E.g. some users prefer many short stalling events while others prefer fewer but longer stalling events. In addition, users consume videos in different ways. Some users may want to watch the entire video. In contrast, other users may search for certain content by watching only a short part of a video if it does not match the content they are searching for. Measurement studies indicated that 60 % of the videos requested were watched for less than 20 % of their total duration [2].

For these behaviors, different things are necessary in order to provide a good QoE. We know that the QoE depends on the number and length of stalling events. Further, we know that initial delay has an impact on the QoE. All these variables are dependent on the buffer size. Therefore, QoE can be controlled by dimensioning the buffer size. A larger buffer means that stalling events are less but longer and that the initial delay is larger. The key questions investigated in this paper are as follows.

- Do we need to know the QoE preferences of the user in order to optimize QoE? E.g. whether a user prefers few long stalling events over many short stalling events?
- Do we need to know the users behavior or the usage scenario? E.g. video browsing or whether a user only watches a few seconds of a video?
- Do we need to know the video characteristics and the current network situation?

We approached these questions by modeling the video buffer for HTTP Video Streaming as $M/M/1$ queue with pq-policy. The model provides the status of the buffer at any point in time for a given network bandwidth and the video bit rate. We conduct a mean-value analysis of the queueing model that returns stalling frequency and stalling ratio. Using these results, we can calculate the QoE of a video for any given user. To answer the questions above, we extend and combine existing QoE models in such a way that different user profiles can be specified and analyzed. We introduce profile parameters that describe a user's general sensitivity to stalling and whether he prefers many short stalling events or few long stalling events. In the analysis, we consider the three major video streaming usage scenarios, (1) 'watch-later', (2) default video streaming, and (3) video browsing. To the best of our knowledge, this is the first study that addresses the questions above and investigates individual user preferences for HTTP video streaming QoE. As a further contribution, recommendations are provided how to take into account the results in practice. The answers and results are an important step for realizing QoE-centric management for HTTP video streaming.

The remainder of the paper is structured as follows. Section II gives a short background on HTTP video streaming and reviews approaches for improving its QoE with a focus on video buffer dimensioning. In Section III an $M/M/1$ queueing model with pq-policy is presented for analyzing the video buffer. An extended QoE model based on YouTube QoE is described in Section IV which allows to analyze individual user profiles. Section V presents analytic results for different usage scenarios and infers recommendations for QoE improvement in practice. The paper is finally concluded in Section VI.

II. Background and Related Work

The increasing popularity of video streaming has driven intensive research activities on how to optimize the video delivery to the end user concerning QoE. In particular, HTTP streaming is deployed by large video service delivery platforms like YouTube or Netflix and represents the major video delivery solution, especially for video-on-demand. HTTP video streaming is a combination of download and concurrent playback. Video data is transmitted to the client via HTTP and stored in an application buffer. After the download of a sufficient amount of data p (which is in the order of a few video seconds, e.g. for YouTube [3]), the video play out starts at the client. As soon as the video buffer falls below a certain threshold q, the video stalls [3]. We refer to this threshold policy as pq-policy and model the video buffer at the client side by a queueing model with pq-policy.

Due to the reliable transmission over TCP, no video degradations can be observed, but resource problems in the network or at the video server manifest as initial delays or as interruptions and stalling of the video during play out.

Many different solutions are proposed in literature how to overcome those QoE degradations for HTTP video streaming. The solutions can be differentiated among others into (A) network-based solutions overcoming networking problems or resource limitations, (B) adaptive streaming approaches lowering the network demands at the cost of lower video quality, (C) buffer-based solutions, optimizing/adapting the video buffer sizes.

A. Solution Approaches for Improved HTTP Streaming QoE

Various resource management mechanisms to improve QoE for YouTube have been proposed in literature, e.g. in Wifi mesh networks [4], e.g. using Software-Defined Networking (SDN) [5]. SDN enhances the interaction between networks and applications and allows a more dynamic and demand-based allocation of network resources.To overcome resource limitations in the content delivery infrastructure, [6] proposes client-based local caching, P2P-based distribution, and proxy caching which reduces network traffic significantly and can therefore avoid QoE degradations.

Recently, big service providers like YouTube rely on HTTP adaptive streaming (HAS) which adapts the video to the current network conditions. The video adaptation may be realized by changing the frame rate, resolution, or quantization of the video. Although the adaptation results in lower quality, the major benefits compared to classical HTTP video streaming is the reduction of stalling events. [7] surveys QoE for HTTP adaptive streaming and gives an overview of the recent developments. Besides improved quality adaptation mechanisms like [8], other approaches aim for example at optimizing the segmentation of the videos [9] .

B. Analysis of Video Buffer Adaptation and Dimensioning

Subjective studies showed that users prefer initial delays instead of stalling events [10]. An analytical framework for the dimensioning of appropriate video buffers for TCP streaming shows that the initial buffering delay and the size of the buffer should be as small as possible, yet large enough to avoid buffer underflows [11]. A concrete approach [12] determines the optimal, i.e. minimal, initial delay at the client. During this time, the the video buffer is filled such that no stalling occurs.

Two buffer size adaptation policies are proposed by [13] which are evaluated by means of a fluid model in terms of freezing probability. [14] evaluates the impact of network dynamics and QoS provision on user's video quality. An analytical framework models the playback buffer at the receiver as a $G/G/1$ queue, however no pq-policy is considered. Further, video quality is considered in terms of the start-up delay or fluency of video playback. Based on that, adaptive playout buffer management schemes are proposed.

So far, no queueing system with pq-policy is applied to analyze QoE for HTTP video streaming and to dimension video buffers accordingly. We go even one step further and consider individual user profiles by utilizing a parametrized QoE model. In queueing theory, the related threshold policy is denoted as N-policy introduced by [15] with $N = p$ and $q = 0$; the server stops whenever the system becomes empty and resumes service when the number of waiting customers in the system (i.e. the video buffer in our case) reaches a threshold value N; in contrast to the transient phase in the steady state, q has no influence on the performance, see Section III-B.

Various researchers analyzed the N-policy. [16] derives the stationary joint distribution of queue length and the server's status for the $GI/M/1$. [17] obtains the steady state probability distribution of the number of customers in a finite system for the $M/GI/1$ system with N-policy. A transient solution of the $M/M/1$ queue under pq-policy is derived by [18].

C. Perception of Individual HTTP Video Streaming Users

Results from queueing theory may be applied to dimensioning the video buffer for HTTP streaming in order to optimize QoE. However, the approaches mentioned above are either considering QoS parameters only or they apply QoE models based on mean opinion scores (MOS) of subjects. However, differences in how QoE degradations are observed by individual users are not considered. Therefore, this work is an important step in QoE-centric management of HTTP video streaming. An analytical model is developed (Section III) which allows to investigate individual user profiles based on this parametrized QoE model (Section IV).

Most user studies on HTTP video streaming quantify and report QoE in terms of MOS, e.g. [3]. However, there is a diversity in user perception which is eliminated by the process of averaging subjective ratings. A relationship between the MOS and the second moment of the user ratings is formulated as SOS hypothesis and a standard deviation for particular MOS values is observed up to 0.8 for video QoE [1]. Thus, user perceptions may fluctuate between good and poor quality under the same conditions. [1] observes different user types, denoted as 'hectic', 'regular', 'insensitive' depending on their sensitivity to QoE degradations. In this paper, we extend existing YouTube QoE models by user profile parameters which allows to investigate such diverging user perceptions.

III. System model

We provide a system model for video playback, in order to study the stalling behavior of HTTP video streaming. We consider the playback of a video consisting of multiple frames. The frames are downloaded in-order and arrive at the client with rate λ while the playback time is given by the video framerate μ, resulting in an offered load of $a = {}^{\lambda}/_{\mu}$. Here, a quantifies the available network bandwidth normalized by the video framerate.

In order to reduce the number of stalling events during playback, the video player uses a playback buffer. Video playback stops, if less than q frames are currently available for playback and is only resumed if the buffer contains $p = q + d$ frames. The normalized buffer size d^* (in video seconds) relates the buffer size d (in frames) to the video framerate μ, i.e. $d^* = {}^{d}/_{\mu}$.

Next, we introduce metrics used to evaluate the influence of the playback buffer parameter selection. The relative amount of time spent in stalling compared to the total duration of the playback process including stalling is given by the stalling ratio R and the number of stalling events normalized by the video length N^*. For the case of finite videos we furthermore consider the stalling duration L which gives the sum of times spent in stalling states during the complete video playback.

To derive the key performance metrics, we model the system as a $M/M/1/\infty$ queueing model with pq-policy in Section III-A. A mean value analysis allows then to investigate the impact of system parameters in the steady state (Section III-B) but also in the transient phase for the analysis of short (finite) videos and user aborts (Section III-C).

A. M/M/1 Queue with pq-Policy

The state of the video playback is characterized by the tuple (i, j), where $i \in \{0, 1\}$ is the playback state of the client, i.e. the video is not played back if i is 0 and the video is played back if i is 1 and $j \geq 0$ gives the number of unplayed frames currently available at the client and . Furthermore, we give the probability of the playback being in state (i, j) as $x(i, j)$. We obtain the following equilibrium state equations.

$$\begin{aligned}
\lambda x(0,0) &= 0 \\
\lambda x(0,i) &= \lambda x(0,i-1) && i \in [1,q) \\
\lambda x(0,q) &= \lambda x(0,q-1) + \mu x(1,q+1) \\
\lambda x(0,i) &= \lambda x(0,i-1) && i \in (q,p) \\
(\lambda+\mu) x(1,q+1) &= \mu x(1,q+2) \\
(\lambda+\mu) x(1,i) &= \lambda x(1,i-1) && i \in (q+1,p) \\
&\quad + \mu x(1,i+1) \\
(\lambda+\mu) x(1,p) &= \lambda (x(0,p-1) + x(1,p-1)) \\
&\quad + \mu x(1,p+1) \\
(\lambda+\mu) x(1,i) &= \lambda x(1,i-1) + \mu x(1,i+1) && i \in (p,+\infty)
\end{aligned}$$

Due to space constraints, no proof is given for the state probabilities which can be obtained analogously to [16].

$$\begin{aligned}
x(0,i) &= 0 && i \in [0,q) \\
x(0,i) &= \tfrac{1-a}{d} && i \in [q,p) \\
x(1,i) &= \tfrac{a(1-a^{i-q})}{d} && i \in (q,p] \\
x(1,i) &= \tfrac{a^{j-p+1}(1-a^{d})}{d} && i \in (p,+\infty]
\end{aligned}$$

From this we obtain the stalling ratio R as the probability of being in a stalling state, i.e.

$$R = \sum_{i=0}^{p-1} x(0,i) = 1 - a \,. \tag{1}$$

B. Mean Value Analysis of Steady State

A mean value analysis of the $M/M/1/\infty$ queueing model with pq-policy is now conducted which can be derived by considering Figure 1, in order to obtain the number of stalling events N^* and the stalling ratio R.

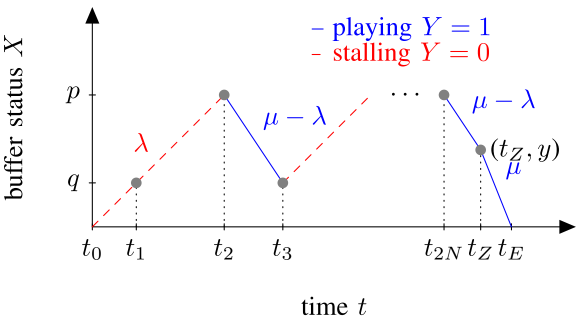

Fig. 1. Video buffer status evolving over time with constant video bitrate and network bandwidth for a finite video of duration T and Z frames.

The initial download begins at t_0 and new frames arrive with rate λ at the client. The number of frames in the buffer exceeds q the first time at t_1. At time t_2, the threshold of p is reached for the first time and playback begins. While the download of new frames continues with rate λ, frames are played out with rate μ, resulting in a buffer change with rate $\lambda - \mu$. Thus, the number of buffered frames reaches q at time t_3. This process repeats which results in an alternating chain of stalling and playback phases.

In this analysis we consider the steady state, i.e. especially neglecting the time $t_1 - t_0$. First, we consider the time required for the buffer to fill from q frames to p frames, i.e. obtaining d frames while no playback is occurring. This time depicts the average duration L of a single stalling event. In Figure 1 this is given as the time between t_1 and t_2, and we get

$$L = t_2 - t_1 = \tfrac{p-q}{\lambda} = {}^{d}/_{\lambda} = {}^{d^*}/_{a} \,. \tag{2}$$

The average stalling length L only depends on the actual buffer size d and the network bandwidth λ. Next, we consider the time required for the number of frames in the buffer to decrease from p to q, i.e. the time between t_2 and t_3, $t_3 - t_2 = \frac{d}{\mu - \lambda}$. Combining these two equations we get the time between two stalling event as $t_3 - t_1 = (t_3 - t_2) + (t_2 - t_1) = \frac{\mu d}{(\mu - \lambda)\lambda}$.

The stalling ratio R follows as

$$R = \tfrac{t_2 - t_1}{t_3 - t_1} = 1 - a \,, \tag{3}$$

yielding the same result as in Eq.(1) in Section III-A.

Finally, we can obtain the number of of stalling events normalized by video duration by analysing the busy periods of the system. Here, the mean idle period is given by $L = {}^{d}/_{\lambda}$.

For the mean busy period B it holds $\frac{B}{B+L} = 1 - R = a$, which yields $B = \frac{a}{1-a} \frac{\lambda}{d}$ and the normalized number of stalls,

$$N^* = \tfrac{1}{B} = \tfrac{\mu - \lambda}{d} = \tfrac{1-a}{d^*} \,. \tag{4}$$

Eq.(4) can also be derived by considering $N^* = \frac{1}{t_3 - t_2}$. While N^* relates the stalls to the video duration, the stalling frequency F denotes the number of stalls per time. It holds $F = \frac{1}{t_3 - t_1} = aN^*$ which is also equal to $F = x(0, p-1)\lambda$ to change the player with the state probability $x(0, p-1)$ and the network arrival rate λ. However, from an end user's point of view, the metric N^* but not F is of importance.

Beside the network bandwidth λ[1] and the video bitrate μ, the number N^* of stalling events depends only on the video buffer size $d = p - q$, but not on the concrete values of p and q in the steady state.

C. Mean Value Analysis of Finite Videos and User Aborts

As we will see later in Section V, the steady state analysis is sufficient to dimension the buffer. However, in practice, playback is finite, either because the video is of finite length T, or because a user aborts playback after a number of T seconds. This behaviour is shown in Figure 1.

We do not consider the time until the initial playback, i.e. the time between t_0 and t_2 as stalling, since it has a much lower impact on the perceived quality than stalling [19]. First, we consider the case where the user plays back the complete video. Given the network bandwidth λ and a video of Z Frames, the required download time for the complete video is $t_Z = {}^{Z}\!/_{\lambda}$. Within t_Z there are N phases of stalling and playback and each phase is of duration $t_3 - t_1$.

$$N = \left\lfloor \frac{t_Z - t_1 + t_0}{t_3 - t_1} \right\rfloor \tag{5}$$

Next, we consider the case where the user aborts playback of the video after T seconds of video have been watched. Here, the number of stalling phases is given as

$$N = \lfloor T/(t_3 - t_2) \rfloor \ , \tag{6}$$

rounding down as we do not consider the initial delay before playback as stalling. Again, we can obtain the number of stalling events normalized by video length as $N^* = {}^{N}\!/_{T}$.

IV. YouTube QoE Model

A. Stalling QoE Model Q_1

The QoE of HTTP streaming depends mainly on the actual number of stalling events N for a video of duration T and the average length L of a single stalling event. A QoE model combining both key influence factors into a single equation $f(L, N)$ is provided in [3] and found to follow the IQX hypothesis [20] describing an exponential relationship between the influence factors and QoE. In particular, the model function returns mean opinion scores (MOS) on a 5-point absolute category rating scale with 1 indicating the lowest QoE and 5 the highest QoE.

$$f(L, N) = 3.5e^{-(0.15L+0.19)N} + 1.50 \tag{7}$$

Due to well known rating scale effects, the model in Eq.(7) has a lower bound of 1.50, as users avoid the extremities of the scale called “saturation effect”, see e.g. [21]. In contrast, if the video is not stalling, no degradation is observed and users rate the impact of stalling as 'imperceptible', i.e. a value of 5. It has to be noted that the model function in Eq.(7) is based on subjective user studies with videos of duration up to $T = 30\,\text{s}$. For other video durations, the normalized number $N^* = {}^{N}\!/_{T}$ of stalling events has to be considered which requires to adapt the parameters $\alpha = 0.15$ and $\beta = 0.19$ in Eq.(7), respectively.

As the goal of our investigation is the analysis of the impact of different user profiles, we parametrize the function in Eq.(7) with α and β and conduct a parameter study on their impact. For the sake of simplicity, we normalize the QoE value to be in the range $[0; 1]$. As a result, we arrive at Eq.(8) as parametrized QoE model Q_1 to quantify the impact of stalling on QoE for different user profiles expressed by α and β. Thereby, the parameter α adjusts the sensitivity of the user to the stalling duration $L \cdot N^*$, while β quantifies the sensitivity of the user to the actual number of stalling events, i.e. the video interruptions. Therefore, we will also use the term 'duration parameter' and 'interruption parameter' for α and β, respectively.

$$Q_1(L, N^*) = e^{-(\alpha L + \beta)N^*} \tag{8}$$

The model function Q_1 in Eq.(8) has the same form as Eq.(7) and follows the IQX hypothesis, but allows to investigate different user profiles. For example, some users may suffer stronger from interruptions which is then adjusted by a higher value of β. Thus, a user profile is expressed by α and β.

B. Initial Delay QoE Model Q_2

Another impairment on HTTP streaming QoE are initial delays before the video playout start. The impact of initial delays T_0 is modeled by the following function g and the model parameters are obtained from subjective tests [10].

$$g(T_0) = -0.963\,\log10(T_0 + 5.381) + 5 \tag{9}$$

The results in [10] show that the impact of the initial delay is independent of the video duration which was either 30 s or 60 s in the user tests. Further, it was observed that users have a clear preference of initial delays instead of stalling and that service interruptions have to be avoided in any case, even at costs of increased initial delays for filling up the video buffers.

For the sake of simplicity, we normalize the function in Eq.(9) yielding to the QoE model Q_2 for initial delays T_0, such that Q_2 returns values in $[0; 1]$ and that $Q2(0) = 1$. The user profile is parametrized with γ determining the impact of initial delays. The constant $c = 5.381$ is taken from Eq.(9) defining the shape of the curve. Since the logarithm is not bounded, only positive values are considered to ensure $Q_2(T_0) \in [0; 1]$.

$$Q_2(T_0) = -\gamma\,\log10\,(T_0 + c) + \gamma\text{log10}\,(c) + 1 \tag{10}$$

C. Combined QoE Model Q

For dimensioning the video buffers, we are interested in a QoE model which considers both, the impairments due to stalling and due to initial delays of the video playout. However, to the best of our knowledge no combined model exists so far which has been validated by proper subjective user studies. Therefore, we suggest the following model Q. Since the impact of stalling events clearly dominates the user perception [3], [10], we consider the following rationale for the combined QoE model. A user facing an initial delay T_0 experiences a QoE value of $Q_2(T_0)$. If additional stalling events occur, this will lower the QoE further. Thus, $Q_2(T_0)$ is the upper bound of QoE. For N^* stalling events with an average length L, the QoE will be further decreased by $Q_1(L, N^*)$.

ITU-T Recommendation P.1201 proposes an additive QoE model for non-adaptive HTTP streaming which is referred to as buffer-related perceptual indicator in the Appendix III [22].

[1] For the sake of readability, we use the term 'network bandwidth λ' instead of 'network bandwidth in terms of frames' or 'network frame arrival rate'.

This model follows the same rationale above, start from the maximum QoE value which is $1 = Q(0,0,0)$, subtract the degradation $1 - Q_2(T_0)$ stemming from initial delay, and from stalling $1 - Q_1(L, N^*)$.

Then, we arrive at the following additive QoE model Q used in the analysis.

$$\begin{aligned} Q(T_0, L, N^*) &= 1 - (1 - Q_1(L, N^*)) - (1 - Q_2(T_0)) \\ &= Q_1(L, N^*) + Q_2(T_0) - 1 \qquad (11) \end{aligned}$$

V. QoE Study for Typical User Scenarios

Typical usage scenarios of video streaming services reflect the following user behaviors: A. *Watch Later*, B. *Regular Video Streaming*, C. *Video Browsing*. For these scenarios, the video buffer size d^* is optimized concerning QoE and the impact of the user profile (α, β) is analyzed for realistic ($\alpha = 0.15, \beta = 0.2$) and extreme values ($\alpha \in \{0.05, 0.45\}, \beta \in \{0.05, 0.8\}$). The difference between the steady state (Section III-B) and the finite case (Section III-B) is 0.2 points on a 5-point MOS scale for 30 s videos. For the *Watch Later* and *Regular Video Streaming* scenario, we assume longer video durations and can use the steady state results. In contrast, for *Video Browsing* short viewing times of 10 s require the finite case results. For the sake of readability we transformed the QoE value linearly to be in the range $[0; 1]$.

A. Watch later Scenario

In the 'watch later' scenario, a user requests a video, but the user does not expect that the video playout starts immediately. This may be the case for example when the user wants to watch an HD movie even though the network bandwidth is low. During that initial delay, the user may do something else, e.g. opening another web page in a parallel tab in the browser, or getting some snacks in the kitchen. Thus, QoE is not affected by initial delays and we only need to consider Q_1 in Eq.(8). In the steady state, it is $L = {}^{d}/_{\lambda}$ and $N^* = \frac{\mu-\lambda}{d}$ and we obtain the following QoE relation in Eq.(12).

$$Q_1(L, N^*) = e^{(\mu-\lambda)({}^{\alpha}/_{\lambda}+{}^{\beta}/_{d})} = e^{-\alpha\frac{1-a}{a}-\beta\frac{1-a}{d^*}} \qquad (12)$$

Since the QoE function in Eq.(12) is strictly monotonically increasing with the buffer size d^*, the optimum is achieved for $Q_+ = \lim_{d^*\to\infty} Q_1(L, N^*) = e^{-\alpha\frac{1-a}{a}}$. Thus, the QoE value only depends on the parameter α in the limit. To see for which buffer size we are close to the optimum, we consider the relative difference $\frac{Q_+ - Q_1(L,N^*)}{Q_+}$ when it is less than $\Omega = 5\,\%$. This is true for $d^* > -\beta\frac{1-a}{\log(1-\Omega)}$.

For $\beta \in \{0.05, 0.2\}$, a small buffer size of $d^* > 4$ s is already sufficient to be close to the optimum Q_+ for any offered network condition a. For users extremely sensitive to stalling ($\beta = 0.8$) buffer sizes up to 15 s are required. However, a buffer of 4 s is sufficient for a relative difference to the optimum of 20 %. In general, the larger the buffer size the better the QoE is in this scenario. In practice, a buffer size of 4 s is a good choice.

B. Default Video Streaming Scenario

In the case of normal streaming, the user wants to watch a video immediately. In contrast to the 'watch later' scenario, the initial delay now impacts the QoE according to Eq.(11). Figure 2 shows QoE depending on the buffer size for the streaming scenario and different user profiles in a network situation $a = 0.5$ leading to a stalling ratio $R = 0.5$. Now, QoE optima exist for finite buffer size, if the impact of the initial delay is taken into consideration. We notice that α does increase the QoE but has no significant impact on the optimal buffer size. In contrast, for different β we observe different optima for the buffer size. Therefore, we can ignore α when optimizing the buffer size in regard to the QoE. A buffer size less than 0.5 s results in a severe loss of QoE for all users. A buffer size of 2-4 s offers a good QoE for the average user and any sensitive user. Increasing the buffer size further decreases the QoE. In practice, QoE is only marginally improved if the user profile is known (see resulting optimal QoE values for different β values in Figure 2).

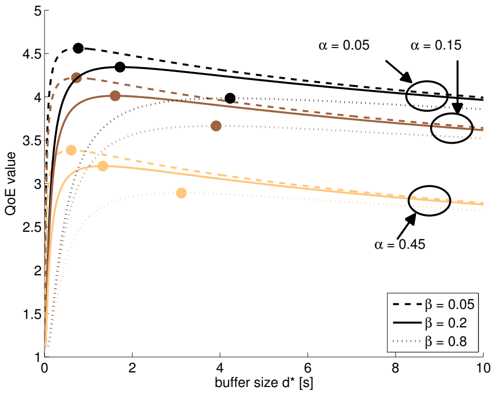

Fig. 2. Dimensioning of buffer size in the **Streaming Scenario** for available network bandwidth of $a = 0.5$. Maxima marked as dots mainly depend on β.

C. Video Browsing Scenario

In the case of Video Browsing, the user watches a video for a short period of time. This includes cases such as, viewing a short video completely, viewing a short part of a long video or skipping ahead in a video frequently (thus watching multiple short parts of a video). Since we know from the previous section that α and β have a marginal impact on the optimal QoE, we consider only the default parameters $\alpha = 0.15$ and $\beta = 0.2$ in the following. However, for video browsing, the impact of the initial delay may be more important for the user. Therefore, we consider $\gamma = 0.3$ corresponding to Eq.(9) as well as a delay sensitive user $\gamma = 0.6$.

In Figure 3, the impact of the buffer size on the QoE is depicted for the case that the video is aborted after the first 10 s. Multiple local QoE maxima exist independently of γ, which appear when the number of stalls changes. The results for the steady state are also included. We observe that the steady state represents a worst case buffer dimensioning, but there is little difference between steady state and the finite case. However, for larger buffer sizes, the difference between the local maxima and the steady state increases. Nevertheless, in those cases, the initial delay exceeds tens of seconds. So this scenario can not be described as realistic video browsing.

In general, if the exact viewing length of a video was known (e.g. short videos will be watched completely), the

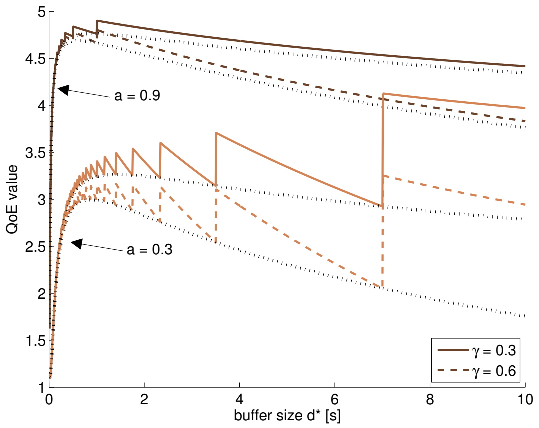

Fig. 3. Dimensioning of buffers for **Video Browsing** users with varying QoE sensitivity to initial delays ($\gamma = 0.3, 0.6$) in two network situations ($a = 0.9, 0.3$). Users abort the video after 10 s. Steady state results i.e. for long videos are indicated by dotted lines

buffer size could be set so that the QoE lies at a local maximum which is independent of γ. However, this method can result in a severe loss of QoE (depending on γ) if the user aborts earlier. In practice, a buffer size of 1-2 s is recommended for video browsing. If the buffer size is set too large, γ determines again the actual QoE loss.

VI. Conclusions and Future Work

Optimizing QoE of HTTP streaming can be realized by proper dimensioning of video buffers. The question arises whether the QoE optimization needs to take into account the individual user profile or preferences as well as the usage scenario. Therefore, we extended existing YouTube QoE models by user profile parameters and analyzed such diverging user profiles based on a queueing model of the video buffer.

The results of a mean-value analysis show that the optimal video buffers require the actual user profile (α, β) parametrizing the sensitivity to stalling, while the initial delay sensitivity parameter γ can be ignored. In practice, recommendations for the buffer size allows to neglect the actual user profile. However, those practical recommendation then need to consider the concrete usage scenario. In particular, it has to be differentiated if the user is video browsing (buffer size 1-2 s) or watching video immediately or later (buffer size 4 s). It has to be noted that for the practical recommendations the network and video characteristics are not relevant. In order to identify the scenario, we recommend adding options to the player or the website, the player is embedded in, in order to give the user a direct choice on how to consume the traffic. Proper user interfaces in the video client, e.g. specific gestures for mobile phones, may indicate browsing behavior. These insights are an important step for QoE-centric management of video delivery in the Internet. Future work needs to address subjective user studies for video browsing and for combined QoE models.

Acknowledgment

This work was partly funded by Deutsche Forschungsgemeinschaft (DFG) under grants HO 4770/1-1 and TR257/31-1. The authors alone are responsible for the content.

References

[1] T. Hoßfeld, R. Schatz, and S. Egger, "SOS: The MOS is not enough!" in *Third International Workshop on Quality of Multimedia Experience (QoMEX 2011)*. Mechelen, Belgium: IEEE, 2011, pp. 131–136.

[2] A. Finamore, M. Mellia, M. M. Munafò, R. Torres, and S. G. Rao, "Youtube everywhere: Impact of device and infrastructure synergies on user experience," in *2011 IMC*, Berlin, Germany, 2011.

[3] T. Hoßfeld, R. Schatz, E. Biersack, and L. Plissonneau, "Internet video delivery in youtube: from traffic measurements to quality of experience," in *Data Traffic Monitoring and Analysis*. Springer, 2013, pp. 264–301.

[4] F. Wamser, D. Hock, M. Seufert, B. Staehle, R. Pries, and P. Tran-Gia, "Using Buffered Playtime for QoE-Oriented Resource Management of YouTube Video Streaming," *Transactions on Emerging Telecommunications Technologies*, vol. 24, Apr. 2013.

[5] T. Zinner, M. Jarschel, A. Blenk, F. Wamser, and W. Kellerer, "Dynamic application-aware resource management using software-defined networking: Implementation prospects and challenges," in *IEEE Network Operations and Management Symposium (NOMS)*, 2014, pp. 1–6.

[6] M. Zink, K. Suh, Y. Gu, and J. Kurose, "Watch global, cache local: Youtube network traffic at a campus network: measurements and implications," in *Electronic Imaging 2008*, 2008.

[7] 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 Communications Surveys Tutorials*, vol. PP, no. 99, pp. 1–1, 2014.

[8] C. Sieber, T. Hoßfeld, T. Zinner, P. Tran-Gia, and C. Timmerer, "Implementation and User-centric Comparison of a Novel Adaptation Logic for DASH with SVC," in *IFIP/IEEE QCMan 2013*, Ghent, Belgium, May 2013.

[9] J. Lievens, S. M. Satti, N. Deligiannis, P. Schelkens, and A. Munteanu, "Optimized segmentation of h. 264/avc video for http adaptive streaming," in *IFIP/IEEE QCMan 2013*, Ghent, Belgium, 2013.

[10] T. Hoßfeld, S. Egger, R. Schatz, M. Fiedler, K. Masuch, and C. Lorentzen, "Initial Delay vs. Interruptions: Between the Devil and the Deep Blue Sea," in *QoMEX 2012*, Yarra Valley, Australia, Jul. 2012.

[11] J. Yan, W. Mühlbauer, and B. Plattner, "An analytical model for streaming over tcp," in *Smart Spaces and Next Generation Wired/Wireless Networking*. Springer, 2011, pp. 370–381.

[12] T. Hoßfeld, F. Liers, R. Schatz, B. Staehle, D. Staehle, T. Volkert, and F. Wamser, "Quality of Experience Management for YouTube: Clouds, FoG and the AquareYoum," *PIK Journal*, vol. 35, Aug. 2012.

[13] M. Fiedler, "On the limited potential of buffers to improve quality of experience," in *6th Int. Workshop on Information Quality and Quality of Service for Pervasive Computing*, March 2014, pp. 419–424.

[14] T. H. Luan, L. X. Cai, and X. Shen, "Impact of network dynamics on user's video quality: analytical framework and qos provision," *IEEE Transactions on Multimedia*, vol. 12, no. 1, pp. 64–78, 2010.

[15] M. Yadin and P. Naor, "Queueing systems with a removable service station," *OR*, pp. 393–405, 1963.

[16] Z. G. Zhang and N. Tian, "The N threshold policy for the $GI/M/1$ queue," *Operations Research Letters*, vol. 32, no. 1, pp. 77–84, 2004.

[17] K.-H. Wang and J.-C. Ke, "A recursive method to the optimal control of an $m/g/1$ queueing system with finite capacity and infinite capacity," *Applied Mathematical Modelling*, vol. 24, no. 12, pp. 899–914, 2000.

[18] W. Böhm and S. G. Mohanty, "The transient solution of $M/M/1$ queues under (M, N) policy. A combinatorial approach," *Journal of Statistical Planning and Inference*, vol. 34, pp. 23–33, 1993.

[19] M.-N. Garcia, D. Dytko, and A. Raake, "Quality impact due to initial loading, stalling, and video bitrate in progressive download video services," in *QoMEX 2014*, Singapore, Sep. 2014, pp. 125–130.

[20] M. Fiedler, T. Hoßfeld, and P. Tran-Gia, "A generic quantitative relationship between quality of experience and quality of service," *IEEE Network*, vol. 24, no. 2, pp. 36–41, 2010.

[21] S. Möller, *Assessment and prediction of speech quality in telecommunications*. Springer, 2000, vol. ISBN: 0792378946.

[22] ITU-T P.1201, *Parametric non-intrusive assessment of audiovisual media streaming quality. Amendment 2: New Appendix III – Use of ITU-T P.1201 for non-adaptive, progressive download type media streaming*, International Telecommunications Union, Dec. 2013.