QoE-Aware Sustainable Throughput for Energy-Efficient Video Streaming

Markus Fiedler, Adrian Popescu, Yong Yao
Blekinge Institute of Technology, Karlskrona, Sweden
{markus.fiedler|adrian.popescu|yong.yao}@bth.se

***Abstract*—This work motivates and details the concept of QoE-aware sustainable throughput in the area of video streaming. Sustainable throughput serves as a mean to compare video streaming solutions in terms of Quality of Experience (QoE) and energy efficiency (EE). It builds upon the QoE Provisioning-Delivery Hysteresis (PDH) and denotes the maximal throughput at which QoE deteriorations can be kept below a quantifiable level, which in turn allows to compare the EE of different video streaming solutions on QoE-fair grounds. In this work, we particularly focus on delivery problems stemming from outage-prone links, as they are typical for mobile systems. Well-adapted to the nature of the video-associated data streams and disturbances, a stochastic fluid flow model is used that allows for straightforward calculation of sustainable throughput values. We also discuss the application of the sustainable throughput for comparisons among different streaming solutions and their offered QoE and EE, respectively.**

***Index Terms*—Quality of Experience, Quality of Service, Stochastic Fluid Flow Model, Energy Efficiency**

I. Introduction

Today, the users of video streaming services may use a plethora of wired, wireless and mobile options for data transfer. Every underlying link technology has own resource offerings, benefits and drawbacks. The higher the user-exploitable speed is, the higher the bearer frequency becomes, and thus, the coverage area becomes smaller due to damping and shadowing effects. Furthermore, the intensity of errors grows with physical distance, data rate, interference caused by surroundings and other users, and also due to multi-path propagation. Consequently, depending on the communication scenario (fixed; wireless; mobile; ad-hoc), different link technologies have different resource facilities and associated specifications (*e.g.* bit rate offered to the user) as well as means to cope with adversary conditions (*e.g.* forward error correction or automatic repeat-request mechanisms). Thus, depending on technology and operator, users may be provided with different offered rates and error patterns (losses; freezes; outages), different data plans as well as differences in energy consumption. The user's Quality of Experience (QoE), which is defined as the "degree of delight or annoyance" [1], is influenced by the above-mentioned parameters. At the same time, energy consumption and battery lifetime have shown to be of particular importance for user perception [2]. Thus, we are facing the challenge of evaluating the offerings of different wireless technologies in a way that considers quality, cost and energy consumption properly into account.

For instance, it was observed in mobile systems that even moderate video bit rates, yet far below the advised capacity of the mobile channel, caused significant QoE disturbances. Reducing the picture size of the video, which means lowering the video bit rate, yielded disturbance-free operations that was – in spite of the small video picture size – more appreciated than a large picture size with frequent freezes [3]. Obviously, users trade larger and more detailed pictures for the fluent appearance of the video, which seems to be more important. These observations indicate the presence of a throughput threshold, whose transgression triggers QoE problems to become apparent. Exceeding the particular threshold means entering a region of unstable QoE, which is not feasible for any quality-and-cost comparison between different links and operators. Thus, in order to evaluate the properties and to optimize the usage of the link, it is necessary to determine the particular throughput threshold for different technologies and/or operators.

We therefore suggest the concept of *QoE-aware sustainable throughput* as being the bit rate threshold value to enable comparisons of video streaming solutions in terms of QoE and energy consumption. QoE-aware sustainable throughput builds upon the QoE Provisoning-Delivery Hysteresis [4] and denotes the maximal throughput at which QoE problems can still be kept at a desired level. Based on previous experience, we are particularly interested in delivery problems stemming from outage-prone links, as they appear in mobile systems. In order to support such performance evaluation, we contribute a stochastic fluid flow model (FFM), which is well-suited to model the streaming video scenario at hand, including its disturbances [3], [5]. This model has the appealing feature that it allows for straightforward calculations of sustainable throughput values. We identify the key parameters for sustainable throughput, and discuss their impacts. Finally, we give suggestions on how to use sustainable throughput for the comparison of different mobile video streaming solutions [6].

In the remainder of this paper, Section II reviews related work; Section III introduces the Provisioning-Delivery Hysteresis; and Section IV provides the formal definition and develops the fluid flow model of QoE-aware sustainable throughput. Section V presents and discusses relevant examples, focusing on the impact of key parameters. Section VI highlights the importance of sustainable throughput for optimizing energy efficiency. Section VII concludes the paper and points out directions for future work.

II. RELATED WORK

In the networking domain, the notion of *sustainable throughput* was introduced in reference [7] as achievable throughput in a multi-hop WLAN scenario. Reference [8] considered the sustainable throughput to be the maximal throughput in a WLAN scenario to assure stability. Reference [9] reported the so-called reliable throughput values for mobile channels with focus on automatic repeat-request (ARQ). The latter works have in common that throughput values are obtained from traffic modeling through analytic models, similar to our work.

Regarding QoE of video delivery, the impacts of freezes and outages on QoE have been studied rather extensively, *e.g.* in [5], [10], [11]. The latter work used an M/M/1 model for QoE-related buffer dimensioning purposes, which is complementary to our fluid flow model-based approach to QoE-related input flow restriction. Reference [12] describes a monitoring approach that reveals whether the video throughput trespassed its sustainable value. This is achieved by reconstructing buffer filling levels and deriving risks for video stalling from traffic measurements along the video delivery path. In contrary to our work, no direct calculation of sustainable throughput is performed.

The optimization of QoE for video delivery has attracted a lot of attention, mostly in the context of adaptive video streaming [13], which is however not the focus of our work. Reference [14] suggests to adapt the video resolution to mobile device capabilities in a QoE-aware manner, yielding considerable savings in file sizes.

None of the works mentioned so far focused explicitly on energy or power efficiency. Energy savings have now become a prominent feature when implementing and deploying efficient video streaming over Internet [15]. Reference [16] investigates the energy consumed by the smartphone to record and stream video contents. In [17], the authors suggested a new power model to estimate the energy consumed by routers and switches with high traffic flows. The experimental results were also reported to show the feasibility and effectiveness of the suggested model. Based on this work, the authors of [18] reported the modeling of energy consumption for a single video flow traversing the mobile network. Clearly, the focus of these studies is on energy modeling for the network elements only.

During the recent years, the areas of QoE and energy efficiency have met [2] and developed jointly [19]. In [20], the authors suggest a new metric called *QoE perceived per user per Watt* (QoEW). This metric is useful given the difficulties to measure and to model the QoE-EE trade-offs for different particular scenarios and network architectures. Recently, dependencies between QoE models and energy savings on terminals have been established [19], [21]. A recent research effort is the CELTIC PLUS project CONVINcE, which aims at minimizing the energy consumption in IP-based video distribution networks combined with the best possible QoE obtained at the terminal [6], [22].

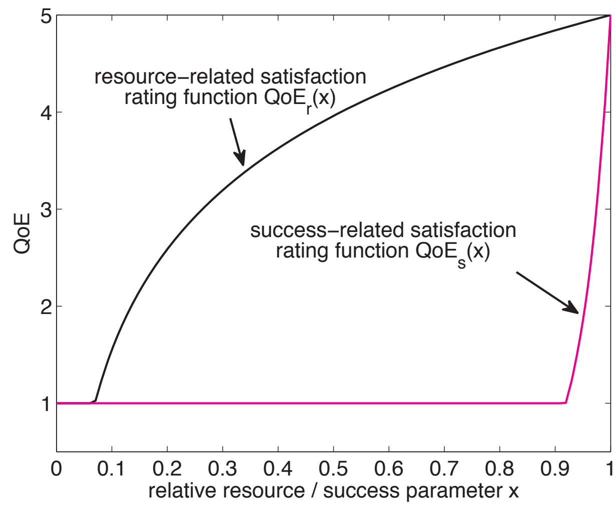

Figure 1: Original illustration of the QoE Provisioning-Delivery Hysteresis with resource-related provisioning branch and success-related delivery branch [4].

Subsequently, the ultimate motivation of our work comes down to using the sustainable throughput to numerically bridge the relationship between QoE and energy consumption, thus paving the way for QoE-aware optimization for energy-efficient video streaming. To the best of our knowledge, our approach is new in the sense that it allows for a straight-forward calculation and application of QoE-aware sustainable throughput through the combination of QoE, teletraffic and power modeling approaches.

III. PROVISIONING-DELIVERY HYSTERESIS

A recent definition [1] denotes QoE as the "degree of delight or annoyance". The goal is to offer as good quality as possible, while avoiding as much annoyance as possible. In fact, these two sides of QoE (delight and annoyance) coincide with the Provisioning-Delivery Hysteresis (PDH), originally formulated in [4] and investigated for the first time for video in [23]. The PDH basically states that impacts on QoE are pretty much different to whether QoE is built up by provisioning measures (contributing to delight), or torn down by delivery issues (contributing to annoyance). An illustration of the PDH is provided in Figure 1 [4], [23].

The (upper) resource-related branch is obtained by dimensioning, and can thus be used to control quality. Typically, this so-called provisioning branch of the PDH follows a logarithmic relationship according to the Weber-Fechner-Law [24]

$$QoE_{\mathrm{P}} \sim log(R) \tag{1}$$

which is a shape that is known in the domain of utility functions [25]. Due to its concave shape, it is well-suited for optimizations. This means that the growth of QoE over additional resources is initially rather intense, but it is getting smaller with higher amount of resources

$$\frac{\mathrm{d}QoE_{\mathrm{P}}}{\mathrm{d}R} \sim \frac{1}{R}\,. \tag{2}$$

This also means that a multiplicative increase of the throughput R yields an additive QoE improvement, *e.g.* any doubling of resources leads to the same absolute QoE improvement, until QoE becomes saturated at R^*.

Assume now that the system under study is provisioned such that the optimal level QoE^* is met at a resource spending of R^*. We call the *goodput ratio* as being the ratio of the available throughput R and optimal (or desirable) throughput R^*

$$x = \frac{R}{R^*} \leq 1\,. \tag{3}$$

A goodput ratio less than 1 can be implemented in a controlled manner for instance by choosing a lower video resolution, which comes along with reduced quality $QoE < QoE^*$ according to (1).

The other branch, the so-called delivery branch, is related to success or failure of the service. It is observed in the context of quality distortion, such as data loss of rate

$$L = R^* - R\,. \tag{4}$$

In contrast to provisioning, the delivery may not be controllable. With (3), the goodput ratio becomes

$$x = 1 - \frac{L}{R^*}\,. \tag{5}$$

Typically, the delivery branch of the QoE shows a rather steep, exponential-type decrease with growing L:

$$QoE_{\mathrm{D}} \sim \exp(-L)\,. \tag{6}$$

The derivative reads

$$\frac{\mathrm{d}QoE_{\mathrm{D}}}{\mathrm{d}L} \sim -\exp(-L)\,. \tag{7}$$

The QoE-QoS relationship according to (6) and (7) is commonly known as *IQX Hypothesis* [26]. It expresses that – comparable to atomic and human memory decay processes [27] – the decay rate of QoE_{D} depends just on the actual level of QoE_{D}. A further implication is that an additive increase of the disturbance (*i.e.* decrease of the rate) yields a multiplicative decrease of the QoE.

It can be proven [4] that, for the same amount of resources R, the provisioning branch supersedes the delivery branch, *i.e.*

$$QoE_{\mathrm{P}}(R) \geq QoE_{\mathrm{D}}(R)\,. \tag{8}$$

Thus, riding the PDH provisioning branch by controlling the amounts of consumed resources, *e.g.* through adaptations of coding, resolutions and data rates [13], [14] is preferable to being exposed to the PDH delivery branch with its mainly uncontrollable sidekicks on QoE.

IV. QoE-Aware Sustainable Throughput

A. Definition

As indicated above, it is important to provide resources according to the provisioning branch with its implicit logarithmic, optimization-friendly QoE-throughput relationship. The larger the target throughput $R_{\mathrm{target}} \leq R^*$ gets, the better $QoE_{\mathrm{P}}(R_{\mathrm{target}})$ becomes. However, in many communication systems, disturbances (such as loss and jitter) tend to increase in frequency and size as the resource load increases, particularly as R_{target} approaches the capacity limits of the corresponding sub-system. As R_{target} grows, there is an increased risk that the delivery branch $QoE_{\mathrm{D}}(R_{\mathrm{target}})$ gets invoked, with $QoE_{\mathrm{D}}(R_{\mathrm{target}}) < QoE_{\mathrm{P}}(R_{\mathrm{target}})$ according to (8). There exists the risk of a rather quick decay of $QoE_{\mathrm{P}}(R_{\mathrm{target}})$ with growing disturbances triggered by too generous R_{target} values.

On this background, we define the *QoE-aware sustainable throughput (ST)* R_{s} as being the maximal value of the throughput R_{target} that keeps the QoE disturbances due to delivery problems at an acceptable minimum:

$$R_{\mathrm{s}} = \max\{R_{\mathrm{target}}|_{QoE_{\mathrm{P}}(R_{\mathrm{target}})-QoE_{\mathrm{D}}(R_{\mathrm{target}})<\delta_{QoE}}\}\,. \tag{9}$$

So far, we consider this throughput as a constant bit rate (CBR) value, which can be used (1) as an upper bound for a variable bit rate (VBR) source, or (2) during quasi-stationary throughput phases. The subsequently presented model (see Section IV-A) actually allows for VBR video traffic, *cf.* Section VI-B, which opens up for a subsequent generalization of the sustainable throughput concept.

We now establish the link between QoE_{D}, QoE_{P} and the underlying resource QoS parameters. We use a simple QoE-QoS model that is described in [19], [28]:

$$QoE_{\mathrm{D}}(R_{\mathrm{target}}) = 4.59\exp(-3.44\,p_{\mathrm{freeze}}(R_{\mathrm{target}}))\,. \tag{10}$$

It approximates QoE as function of the probability p_{freeze} that video freezes occur, which depends on R_{target} and is weighted with the sensitivity factor -3.44 in the argument of the exponential function. The latter is weighted with factor 4.59 out of 5, which denotes the maximal value of the modeled user ratings. We associate the provisioning curve with the case that no disturbances appear, i.e. $p_{\mathrm{freeze}} = 0$, which yields

$$QoE_{\mathrm{P}}(R_{\mathrm{target}}) = 4.59\,. \tag{11}$$

Figure 2 illustrates how an increase of p_{freeze} affects QoE_{D}. Let us consider the maximal freeze probability $p_{\mathrm{freeze,max}}$ as the value that ensures an upper bound on the impact of delivery problems on user perception. Then

$$\begin{aligned}\delta_{QoE} &= QoE_{\mathrm{P}} - QoE_{\mathrm{D}}(p_{\mathrm{freeze,max}}) \\ &= 4.59\,(1-\exp(-3.44\,p_{\mathrm{freeze,max}}))\,.\end{aligned} \tag{12}$$

This yields

$$p_{\mathrm{freeze,max}} = -\frac{\ln(1-\delta_{QoE}/4.59)}{3.44}\,. \tag{13}$$

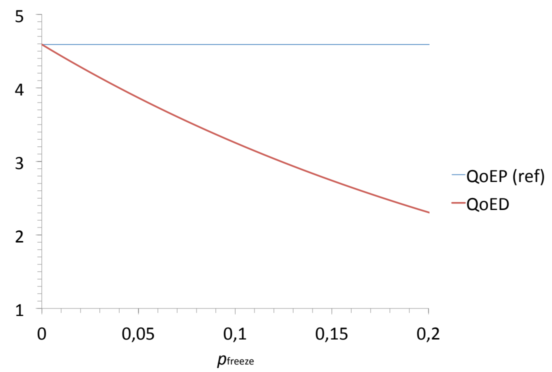

Figure 2: Relationship between freeze probability and QoE_{D} according to (10), with QoE_{P} as reference value.

Close to the origin, the following linear dependencies can be used with good approximation, due to the properties of the exponential function (10, 12):

$$\delta_{QoE} \simeq 15.8\, p_{\text{freeze,max}} \quad \text{for} \quad p_{\text{freeze,max}} \ll 1\,; \tag{14}$$

$$p_{\text{freeze,max}} \simeq 0.064\, \delta_{QoE} \quad \text{for} \quad \delta_{QoE} \ll 1\,. \tag{15}$$

The challenge consists in modeling the interdependencies between throughput R and freeze probability p_{freeze} such that the sustainable throughput at a tolerable QoE deviation δ_{QoE} according to (9) can be obtained. The maximal throughput that does not allow the freeze probability to exceed a given bound $p_{\text{freeze,max}}$ has to be determined.

B. Stochastic Fluid Flow Model

As mentioned above, we are looking for the sustainable throughput R_{s} as being the maximal CBR that allows for keeping the freeze probability and the QoE degradation below desired limits. We focus on an outage-prone VBR Gilbert-Elliott type channel with exponential on-off durations of a good "on" state (with nominal capacity) and a bad "off" state (with vanishing capacity), as it applies to mobile conditions [3], [5], [29].

The channel can be perceived as being off due to reasons like coverage issues, scheduling actions, and repair efforts of symbol errors by the underlying ARQ mechanism [9], to ensure the highest-possible degree of completeness and order of packets in the corresponding stream. The buffer used for interconnecting the CBR inlet with the VBR outlet is necessary for scheduling and repair. Moreover, it allows for limiting potential loss due to the discrepancy of CBR inlet and zero outlet. We distinguish the following phases:

1) During normal operations, the channel is on, and the video traffic can pass it without any hinder, *i.e.* the inlet and outlet rates match.
2) Once the channel turns off, *e.g.* due to outages provoked by the ARQ mechanism, the video traffic is buffered at its full CBR, while the outlet rate is zero.
3) Once the buffer is full, loss happens also at the full CBR, *i.e.* gaps emerge in the video stream. At the receiver side, these gaps are perceived in a similar way as freezes [28]. We assume a playout buffer of at least the same size so that no additional QoE disturbances occur due to delayed traffic.
4) Once the channel turns back on, the buffer empties at a constant rate $C > R_{\text{s}}$ until it becomes empty, which means that the outlet rate goes back from C to R_{s}.

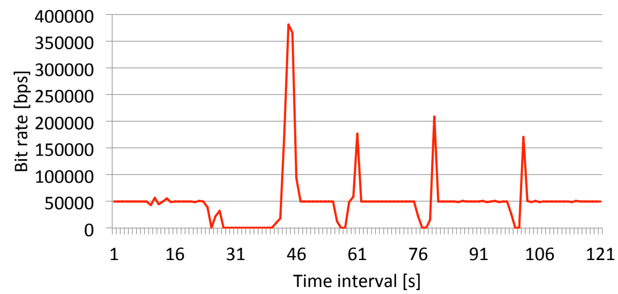

Figure 3: Throughput trace at the outlet of a mobile channel, with a UDP-based CBR stream sent from a moving car [5].

Figure 3 shows the result of these phases as seen by the receiver behind the outlet of the buffer. Observe that the capacity C can be much larger than the CBR inlet rate; in the example in Figure 3, $C/R = 384\,\text{kbps}/50\,\text{kbps} = 7.68$.

For this purpose, we use a performance modeling approach that works well for such bursty cases [30] and in particular for variable capacity scenarios such as mobile channels [5], [29]. The corresponding model parameters are as follows:

- CBR inlet rate $R \simeq \text{const.}$, to be upper-bounded by the sustainable throughput R_{s};
- VBR on/off outlet rate, toggling between *capacity* C and bit rate zero, with
- exponentially distributed on time with average $\mathbf{E}T_{\text{on}}$;
- exponentially distributed off time with average $\mathbf{E}T_{\text{off}}$; and
- buffer size K.

From the fluid flow model analysis [29], [31], [32], we obtain

$$p_{\text{loss}} = \mathbf{FFM}(R, C, \mathbf{E}T_{\text{on}}, \mathbf{E}T_{\text{off}}, K)\,. \tag{16}$$

Due to data and time unit scaling properties of the fluid flow model, the dependencies can be simplified to

$$p_{\text{loss}} = \mathbf{FFM}(\rho, \alpha, \kappa)\,. \tag{17}$$

with the following parameters:

1) *throughput factor*, relating the CBR at the inlet to the maximal capacity:
$$\rho = \frac{R}{C}\,; \tag{18}$$
2) *on-factor*, denoting the nominal availability of the channel:
$$\alpha = \frac{\mathbf{E}T_{\text{on}}}{\mathbf{E}T_{\text{on}} + \mathbf{E}T_{\text{off}}}\,; \text{ and} \tag{19}$$

3) *full buffer depletion time factor*, expressing the buffer size as multiples of the amount that could be transported by the channel during an average off time:

$$\kappa = \frac{K}{C\mathbf{E}T_{\text{off}}}\,. \tag{20}$$

From the above description of phase 3, we can observe that the loss probability (17) is given by the joint probability that the channel is off and the buffer is full:

$$p_{\text{loss}} = f(\rho,\alpha,\kappa)\,\exp\left(\kappa\,\frac{\rho-\alpha}{\alpha\,\rho\,(1-\rho)}\right) \tag{21}$$

The first term $f(\rho,\alpha,\kappa)$ is a rather voluminous expression [32], and it is omitted for the sake of brevity.

In order to obtain the *sustainable throughput factor*

$$\rho_{\text{s}} = R_{\text{s}}/C = \mathbf{FFM^{-1}}(p_{\text{loss,max}}|\alpha,\kappa)\,, \tag{22}$$

we need to solve the equation (21) for ρ. Due to the transcendental nature of (21), this is merely possible in a numerical way, *e.g.* using a branch-and-bound algorithm. However, the pre-factor $f(\rho,\alpha,\kappa)$ in (21) can be approximated by the probability for the off state, $(1-\alpha)$, which yields

$$\begin{aligned}\check{p}_{\text{loss}} &= (1-\alpha)\,\exp\left(\kappa\,\frac{\rho-\alpha}{\alpha\,\rho\,(1-\rho)}\right) \\ &\geq p_{\text{loss}}\,.\end{aligned} \tag{23}$$

We obtain the closed-form approximation of the sustainable throughput factor as

$$\begin{aligned}\check{\rho}_{\text{s}} &= \frac{\alpha-\beta(\alpha)-\sqrt{(\alpha-\beta(\alpha))^2+4\alpha^2\beta(\alpha)}}{2\alpha} \\ &\geq \rho_{\text{s}}\end{aligned} \tag{24}$$

with

$$\beta(\alpha) = \frac{\kappa}{\ln(p_{\text{loss,max}})-\ln(1-\alpha)}\,. \tag{25}$$

The tightness of this bound is examined in Section V. The structure of Equation (24) is similar to the corresponding formula reported in [7].

V. Numerical Results

We now report a study focusing on the key parameters α and κ.

- We focus on on-factors α in the realistic range between 90 % and 99 %.
- The buffer sizes are given in multiples of the depletion capacity and the average off time through the parameter κ according to (20). For the sake of convenience, we define small buffers via a full buffer depletion time of less than one average off time, i.e. $\kappa < 1$. Conversely, large buffers have a full buffer depletion time of more than one average off time, i.e. $\kappa > 1$.

In both cases, the values for $\kappa = 1$ are included for the sake of comparison.

We choose an acceptable QoE disturbance of $\delta_{QoE} = 0.1$ which, according to (13) and the corresponding simplification (15), yields $p_{\text{freeze,max}} \simeq 0.0064$. The reference values $\rho_{\text{s}} = \mathbf{FFM^{-1}}(p_{\text{freeze,max}}(\delta_{QoE})|\alpha,\kappa)$ are calculated numerically via branching-and-bounding of (17), while the closed-form approximation is obtained from (24).

A. Small buffers

From the Table I, we can see that the bound $\check{\rho}_{\text{s}}$ is rather tight and found to be on the safe side, *i.e.* the sustainable throughput factor is not overestimated by its closed-form approximation (24). Thus, it is well motivated to use the closed formula (24) instead of the reference values that require numerical iterations, which has been done for Figure 4. For $\alpha = 0.99$, the smallest buffer size considered (yielding full buffer depletion within 1/10 of the average off time) allows for a sustainable throughput in the order of 20 % of the depletion capacity, which is a realistic value as seen from lab measurements. As the on-factor α sinks, so does the corresponding sustainable throughput value. That means that the buffer's role is to compensate to some extent for a sinking on-factor. There is an almost linear dependency of κ until saturation takes place. The sustainable throughput is upper-bounded by αC, which avoids overload.

Another kind-of-extreme case is obtained when passing the full buffer depletion time factor to the limit $\kappa \to 0$, which means that the disturbance model turns into a Gilbert-Elliott loss model. When there is no chance to buffer any traffic during the off state, loss happens as soon as the channel is off. The emerging simple relationship $\alpha + p_{\text{loss}} = 1$ implies the availability-related constraint $\alpha \geq 0.9936$ for $p_{\text{loss}} \leq 0.0064$.

B. Large buffers

From the Table I, we observe that a larger buffer drives the sustainable throughput factor towards its upper limit α. However, such large buffers are of rather theoretical interest. In practice, they would imply too large jitter buffers at the receiver, challenging the patience of users that have to wait until the video finally can start [10]. Looking at the other sustainable throughput coefficient values, we observe a somewhat decreased tightness of the bound on the way towards reaching saturation. The reason for this is found in suppressing the impact of the limited buffer size on the loss probability. Actually, Equation (24) assumes an unlimited buffer for the exponential buffer term. Yet, the errors remain within the order of 5 %, which again motivates the use of the closed-form approximation (24) that allows to calculate the sustainable throughput from easily measurable parameters.

Equation (24) also forms the basis for Figure 5. It is clearly visible that, for $\alpha = 0.99$, the sustainable throughput is close to saturation, while the other two on-factors show – similarly to $\alpha = 0.99$ in the short buffer case – first a somewhat linear interdependency to κ and then a trend towards saturation $\rho \to \alpha$. However, the smaller α, the less pronounced these two behaviors become.

C. Discussion

As the buffer size grows, the sustainable throughput values get close to the corresponding capacities. The reasons for this

α	ρ	$\kappa = 0.1$	$\kappa = 0.25$	$\kappa = 0.5$	$\kappa = 1$
0.99	ρ_s	0.224	0.559	0.984	0.989
	$\breve{\rho}_s$	0.223	0.553	0.941	0.982
0.95	ρ_s	0.049	0.121	0.241	0.473
	$\breve{\rho}_s$	0.049	0.121	0.239	0.464
0.90	ρ_s	0.036	0.091	0.179	0.349
	$\breve{\rho}_s$	0.036	0.090	0.178	0.343

Table I: Sustainable throughput factors for different on-factors and small full buffer depletion time factors, $p_{\text{loss,max}} = 0.0064$.

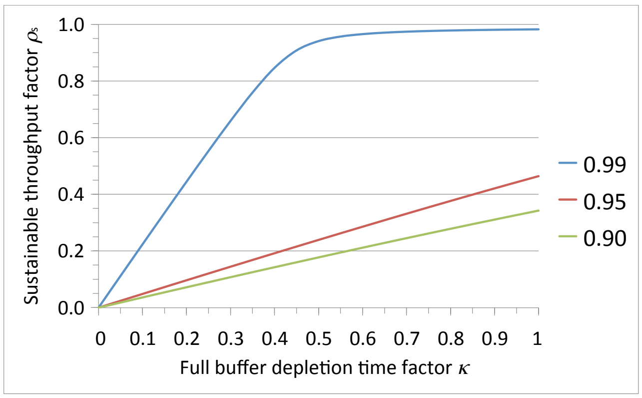

Figure 4: Sustainable throughput factors (24) versus full buffer depletion time factor for different on-factors in the case of small buffers, $p_{\text{loss,max}} = 0.0064$.

α	ρ	$\kappa = 1$	$\kappa = 2$	$\kappa = 4$	$\kappa = 10$
0.99	ρ_s	0.989	0.990	0.990	0.990
	$\breve{\rho}_s$	0.982	0.987	0.989	0.990
0.95	ρ_s	0.473	0.828	0.938	0.949
	$\breve{\rho}_s$	0.464	0.785	0.910	0.939
0.90	ρ_s	0.349	0.634	0.842	0.896
	$\breve{\rho}_s$	0.343	0.604	0.802	0.872

Table II: Sustainable throughput factors for different on-factors and large full buffer depletion time factors, $p_{\text{loss,max}} = 0.0064$.

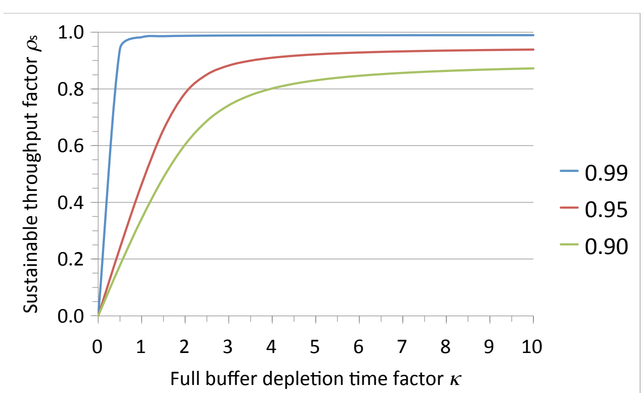

Figure 5: Sustainable throughput factors (24) versus full buffer depletion time factor for different on-factors in the case of large buffers, $p_{\text{loss,max}} = 0.0064$.

are partly found in the low off probability of the channel $(1 - \alpha)$, partly in the increasing capability of the buffer to take care of the overshooting amount of traffic.

However, from the point of view of the source, the buffer "looks bigger" due to the fact that $R_s < C$. Considering the duration T_s of CBR traffic at a rate of R_s that the buffer can take before it overflows, we observe

$$K = \kappa\, C\, \mathbf{E}T_{\text{off}} = R_s\, T_s \Rightarrow T_s = \kappa \frac{C}{R_s}\, \mathbf{E}T_{\text{off}} \qquad (26)$$

In particular when R_s is small, this translation needs to be taken into account in order not to arrive at wrong impressions about the capability of the system to buffer the traffic during an off phase. The following example demonstrates this in detail.

D. A Real-World Example

The example from the drive test reported in Section IV-B has the following parameters: $\mathbf{E}T_{\text{off}} = 3.24\,\text{s}$ and $\mathbf{E}T_{\text{on}} = 200\,\text{s}$, yielding $\alpha = 0.984$; and a full buffer depletion time of approximately 2.2 s, seen from the disappearance of the large spike in Figure 3. From this, we can estimate $\kappa = 0.68$ and, for $p_{\text{loss}} = 0.0064$, we obtain $\breve{\rho}_s = 0.74$ from (24), and can thus estimate $R_s \simeq 0.74 \cdot 384\,\text{kbps} = 284\,\text{kbps}$. This is much higher than the exploited throughput of $R = 50\,\text{kbps} = 0.13\,C$.

Indeed, we faced a loss ratio of 6 out of 60000 packets due to outages. Assuming the corresponding value $p_{\text{loss}} = 0.0001$, we obtain $\breve{\rho}_s = 0.13$ from (24), which matches our offered load. As expected, staying far below the originally calculated sustainable throughput value (factor 5.7) also yields a much better loss performance (factor 64) than originally anticipated.

However, from the point of view of the source, we obtain $T_s = 16.9\,\text{s}$, which is more or less the duration of the long outage before the large spike in Figure 3.

VI. Application to Energy Efficiency

This section presents a set of use cases for the sustainable throughput as well as some possibilities for further extensions of this concept. They are so far of theoretical nature and will be complemented with concrete numerical values in future work.

A. Comparative Analysis

As outlined in [6], we aim at the comparison of different video streaming solutions in terms of QoE and EE, captured by the QoEW measure [20]

$$QoEW(R) = \frac{QoE_{\text{P}}(R)}{P(R)}\,, \qquad (27)$$

expressed in user opinion scores per Watt.

Just using a specific video coded at a specific rate for EE considerations does not reveal the whole picture; an unnecessarily low rate is expected not to exploit the full QoE provisioning potential, while an excessive rate may entail significant disturbances. In both cases, the QoE as such is suboptimal, and PE comparisons are potentially based on unfair grounds. It remains to be investigated where $QoEW(R)$ reaches its optimum, which then can be compared amongst different systems and configurations.

1) Optimization: To this end, consider the following set of relationships:

$$QoE_{\mathrm{P}}(R) = a_Q + b_Q \ln(R)\,; \tag{28}$$
$$P(R) = a_P + b_P R\,. \tag{29}$$

For the power P, a linear dependency on R is assumed [17]. Henceforth, we consider the normalized rate $R' = R/R_{\mathrm{s}}$ and the normalized expression

$$QoEW(R') = c\,\frac{1 + q\,\ln(R')}{1 + p\,R'}\,. \tag{30}$$

The necessary conditions for maximal QoEW are

$$\frac{\mathrm{d}}{\mathrm{d}R'}\,QoEW(R') = 0 \text{ and} \tag{31}$$
$$\frac{\mathrm{d}^2}{\mathrm{d}R'^2}\,QoEW(R') < 0\,. \tag{32}$$

Equation (31) implies $pq(1 - \ln(R')) = p - q/R'$. Such a transcendental condition is not straightforward to handle. However, the specific case $R' = 1$ and $p = q$ fulfills both (31) and (32) meaning that the maximal QoEW is obtained just at the sustainable throughput $R = R_{\mathrm{s}}$.

2) A Trivial Case: It has been observed that b_P is rather small; for switches and routers, [17] reports results in the order of $b_P/a_P \simeq 0.05/10\,$Gbps. This rather flat gradient motivates us to investigate the trivial case $b_P \to 0$, *i.e.* $p \to 0$, which yields

$$\frac{\mathrm{d}}{\mathrm{d}R'}\,QoEW(R') \to c\,\frac{q}{R'}\,. \tag{33}$$

No maximum according to (31,32) is obtained, but the first derivative (33) is positive for any $R' > 0$, which means that the optimal value of QoEW (*i.e.* the maximal EE) is obtained at $R' = 1$, *i.e.* for the sustainable throughput $R = R_{\mathrm{s}}$.

3) Comparison Between Systems: If different video delivery systems are to be compared in terms of their EE, their $QoE_{\mathrm{P}}(R_{\mathrm{s}})$ values should be taken into account, which captures both optimal QoE and optimal EE. One should be aware that using the same QoE for different systems might be disadvantageous for systems that would allow for reaching higher QoE values if their sustainable throughput was exploited. Indeed, as outlined above, systems running at $R < R_{\mathrm{s}}$ are not likely to run in the most energy-efficient way.

B. Potential Extensions

This introductory work opens up for a non-exhaustive list of extensions, as follows:

1) The QoE model may be attached to different parameters, e.g. the average freeze time. As long as these parameters can be linked to the underlying fluid flow model, they can be taken into account.
2) The VBR model of the channel can be refined, considering the ARQ behavior at hand [9].
3) The CBR characteristic at the inlet can be exchanged to a VBR characteristic with two or more states, potentially modulating each other in a multiplicative fashion [33]. As long as the modulating processes are of Markov type, the corresponding combinations can be addressed. However, in most cases, closed-form solutions with close approximation behavior cannot be expected. Yet, asymptotic solutions with potentially loose bounding behavior should be available [30].
4) Additional traffic can be considered to compete with the (video) stream of interest, either in a best-effort manner or by using priorities. Again, as long as the modulating processes are Markovian, the resulting system can be analyzed.

Due to the inherent modeling flexibility of the Markov-modulated stochastic fluid flow model, the above items can be addressed in any order, as demand arises.

VII. Conclusions

In this work, we motivated and described the concept of QoE-aware sustainable throughput as a means to tune the load of an outage-prone system such that QoE provisioning can be maximized without deteriorating it. To this end, we reviewed the QoE Provisioning-Delivery-Hysteresis and used it for deriving the corresponding definition of sustainable throughput. Next, we derived mathematical descriptions of sustainable throughput, including reference values and a tight closed-form approximation. Through a set of numerical results, we were able to judge the tightness of the bounds provided by the approximation, and found one particular closed-form solution to be rather tight to the reference values. This means that we are now able to calculate the sustainable throughput for the outage-prone links in closed form, given that we know the QoE and interrelated QoS target as well as the parameters of the system at hand. We also provided the reasoning for using the sustainable throughput as basis for comparisons of systems with regards to both QoE and power efficiency.

Future work is about the inclusion of more elaborate QoE, video and channel models as well as more complex traffic conditions with interfering traffic and potential traffic prioritization, all of them includable in the extensible and flexible fluid flow model analysis. Furthermore, sustainable throughput values will be determined through measurements and user experiments, and will be compared to other proposed formulas as far as applicable. We will definitely use the concept for comparing video streaming solutions in terms of QoE and EE.

Acknowledgment

This work was triggered and is supported by the European CELTIC PLUS project CONVINcE [22]. It was partially funded by Finland, France, Sweden and Turkey.

References

[1] Qualinet White Paper on Definitions of Quality of Experience (2012). European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Patrick Le Callet, Sebastian Möller and Andrew Perkis, eds., Lausanne, Switzerland, Version 1.2, March 2013. Available at http://www.qualinet.eu (last seen 2016-05-31).

[2] S. Ickin *et al.*, "Factors influencing quality of experience of commonly-used mobile applications," IEEE Communications Magazine, Special Issue on QoE Management in Emerging Multimedia Services, April 2012.

[3] M. Fiedler, "On the limited potential of buffers to improve Quality of Experience," in Proc. 2014 IEEE PERCOM Workshops, Budapest, Hungary, March 2014.

[4] M. Fiedler and T. Hoßfeld, "Quality of Experience-related differential equations and provisioning-delivery hysteresis," in Proc. 21st ITC Specialists Seminar on Multimedia Applications – Traffic, Performance and QoE, Miyazaki. Japan, Mar. 2010.

[5] M. Fiedler, J. Shaikh, and V.J.D. Elepe, "Exponential on-off traffic models for Quality of Experience and Quality of Service assessment," Praxis der Kommunikationstechnik (PIK) 37(4):297–304, 2014.

[6] R. Monnier, A. Popescu, and R. Ljung, "CONVINcE: Towards performance optimization in video distribution networks," in Proc. 19th ICIN Conference, Paris, March 2016.

[7] P.C. Ng and S.C. Liew, "Throughput analysis of IEEE802.11 multi-hop ad hoc networks," IEEE/ACM Transactions on Networking 15(2):309–322, Apr. 2007.

[8] Y.J. Zhang, S.C. Liew, and D.R. Chen, "Sustainable throughput of wireless LANs with multipacket reception capability under bounded delay-moment requirements," IEEE Transactions on Mobile Computing 9(9):1223–1240, Sept. 2010.

[9] S. Akin and M. Fidler, "Backlog and delay reasoning in HARQ systems.," in Proc. ITC 2015, Ghent, Belgium, Sept. 2015.

[10] T. Hoßfeld *et al.*, "Initial delays vs. interruptions: Between the devil and the deep blue sea," in Proc. 2012 Int. Workshop on Quality of Multimedia Experience (QoMEX), Yarra Valley, Australia, Jul. 2012.

[11] T. Hoßfeld, C. Moldovan, and C. Schwartz, "To each according to his needs: Dimensioning video buffer for specific user profiles and behavior." in Proc. IFIP/IEEE International Symposium on Integrated Network Management (IM), 2015.

[12] M. Eckert, T. Knoll, and F. Schlegel, "Advanced MOS calculation for network based QoE estimation of TCP streamed video services," in Proc. 7th Int. Conf. on Signal Processing and Communication Systems (ICSPCS), Carrara, VIC, Dec. 2013, pp. 1–9.

[13] S. Petrangeli, J. Famaey, M. Claeys, S. Latré, and F. De Turck, "QoE-driven rate adaptation heuristic for fair adaptive video streaming," ACM Transactions on Multimedia Computing, Communications and Applications (ACM TOMM), Nov. 2015.

[14] A. Devlic, P. Kamaraju, P. Lungaro, Z. Segall, and K. Tollmar, "QoE-aware optimization for video delivery and storage", in Proc. IEEE Int. Symp. on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), Boston, MA, June 2015.

[15] A. Popescu, "Greening of IP-Based Video Distribution Networks: Developments and Challenges," IEEE 10th International Conference on Communications, Bucharest, May 2014.

[16] S.V. Rajaraman, M. Siekkinen, and M.A. Hoque, "Energy Consumption Anatomy of Live Video Streaming from a Smartphone," IEEE 25th International Symposium on Personal, Indoor and Mobile Radio Communications, Washington DC, 2014.

[17] A. Vishwanath, K. Hinton, R.W.A. Ayre, and R.S. Tucker, "Modeling Energy Consumption in High-Capacity Routers and Switches," IEEE Journal on Selected Areas in Communications 32(8):1524–1532, 2014.

[18] Y. Yao and A. Popescu, "On Energy Consumption in Mobile Multimedia Networks with OpenFlow Switch," The 11th International Conference on Communications, Bucharest, Romania, June 2016.

[19] S. Ickin, "Quality of Experience on smartphones: Network, Application, and Energy Perspectives," Ph.D. thesis, Blekinge Institute of Technology, Karlskrona, Sweden, 2015.

[20] X. Zhang, J. Zhang, Y. Huang, and W. Wang, "On the study of fundamental trade-offs between QoE and energy efficiency in wireless networks", Transactions on Emerging Telecommunications Technologies, 24:259–265, 2013.

[21] L.G.M. Ballesteros *et al.*, "Energy Saving Approaches for Video Streaming on Smartphone based on QoE Modeling," in Proc. 13th Annual IEEE Consumer Communications & Networking Conference (CCNC), Las Vegas, USA, Jan. 2016.

[22] European CELTIC-PLUS project CONVINcE: "Consumption OptimizatioN in Video NEtworks," https://www.celticplus.eu/project-convince/ (last seen 2016-05-31).

[23] T. Zinner, T. Hoßfeld, T.N. Minhas, and M. Fiedler, "Controlled vs. Uncontrolled Degradations of QoE: The Provisioning-Delivery Hysteresis in Case of Video," in Proc. EuroITV 2010 Workshop: Quality of Experience for Multimedia Content Sharing, Tampere, Finland, 2010.

[24] P. Reichl, S. Egger, R. Schatz, and A. D'Alconzo, "The Logarithmic Nature of QoE and the Role of the Weber-Fechner Law in QoE Assessment," in Proc. IEEE ICC 2010, Cape Town, South Africa, May 2010, pp 1–5.

[25] J. Norstad, "An Introduction to Utility Theory," available at http://www.norstad.org/finance/util.pdf (last seen 2016-05-31).

[26] M. Fiedler, T. Hossfeld, and P. Tran-Gia, "A generic quantitative relationship between Quality of Experience and Quality of Service," IEEE Network, 24(2):36–41, 2010.

[27] T. Hoßfeld *et al.*, "The Memory Effect and Its Implications on Web QoE Modeling." in Proc. ITC 2011, San Francisco, USA, Sept. 2011.

[28] S. Ickin and M. Fiedler, "Quality of Experience on smartphones: Network, Application, and Energy Perspectives," in Proc. 19th ICIN Conference, Paris, March 2016.

[29] M. Fiedler and U. Krieger, "The impact of varying channel capacity on the quality of advanced data services in PCS networks," in Proc. 12th ITC Specialists Seminar, Lillehammmar, 2000.

[30] D. Anick, D. Mitra, and M. Sondhi, "Stochastic theory of a data-handling system with multiple sources," The Bell System Technical Journal 61(8):1871–1894, Oct. 1982.

[31] R. Tucker, "Accurate method for analysis of a packet-speech multiplexer with limited delay," IEEE Transactions on Communications 36(4):479–483, Apr. 1988.

[32] M. Fiedler, "Erforderliche Kapazität beim Multiplexen von ATM-Verbindungen," (in German, English title: "Required capacity for multiplexing ATM connections") Ph.D. thesis, Universität des Saarlandes, Saarbrcken, 1998.

[33] P. Carlsson, "Multi-Timescale Modelling of Ethernet Traffic," Tekn.Lic. thesis, Blekinge Institute of Technology, Karlskrona, Sweden, 2003.

APPENDIX

Table III: Main Abbreviations and Parameters

Notion	Definition
α	(Channel) On-factor
a_Q, b_Q, q	Approximation coefficients for the QoE model
a_P, b_P, p	Approximation coefficients for the power model
c	Constant
C	Capacity, *i.e.* the constant rate of emptying the buffer
CBR	Constant Bit Rate
δ_{QoE}	QoE tolerance between provisioning and delivery for calculating the sustainable throughput
EE	Energy Efficiency
$\mathbf{E}T_{\text{on}}$	Average channel on time
$\mathbf{E}T_{\text{off}}$	Average channel off time
FFM	Fluid Flow Model analysis algorithm, exact
K	Buffer size
κ	Full buffer depletion time factor
L	Data loss rate
PDH	Provisioning-Delivery Hysteresis
p_{freeze}	Freeze probability
p_{loss}	Loss probability
$\check{p}_{\text{loss}}$	Dito, approximation
R	Throughput
R'	Normalized throughput
R_{s}	Sustainable throughput
ρ	Throughput factor
ρ_{s}	Sustainable throughput factor
$\check{\rho}_{\text{s}}$	Dito, approximation
QoE	Quality of Experience
QoE_{P}	QoE, provisioning branch
QoE_{D}	QoE, delivery branch
$QoEW$	QoE perceived per user per Watt
QoS	Quality of Service
T_{s}	Buffer fill-time at CBR
VBR	Variable Bit Rate