Performance Evaluation 148 (2021) 102199

Contents lists available at ScienceDirect

Performance Evaluation

journal homepage: www.elsevier.com/locate/peva

Modelling user experience of adaptive streaming video over fixed capacity links☆

Åke Arvidsson [a,*,1], Milosh Ivanovich [b], Paul Fitzpatrick [b]

[a] Kristianstad University, Kristianstad, Sweden
[b] Telstra Corp, Melbourne, Australia

ARTICLE INFO

Article history:
Received 4 April 2020
Received in revised form 20 November 2020
Accepted 22 February 2021
Available online 28 February 2021

Keywords:
Adaptive streaming video
User QoE metrics
Video quality metrics
Proportion of time with reduced video quality
Proportion of videos with reduced quality

ABSTRACT

Streaming video continues to experience unprecedented growth. This underscores the need to identify user-centric performance measures and models that will allow operators to satisfy requirements for cost-effective network dimensioning delivered with an acceptable level of user experience. This paper presents an analysis of two novel metrics in the context of fixed capacity links: (i) the average proportion of a video's playing time during which the quality is reduced and (ii) the average proportion of videos which experience reduced quality at least once during their playing time, based on an $M/M/\infty$ system. Our analysis is shown to hold for the more general $M/G/\infty$ system for metric (i), but not for (ii) and simulation studies show an unexpected form of sensitivity of metric (ii) to the flow duration distribution, contrary to the norm of increasing variance causing worse performance. At typical operational loads these new metrics provide a more sensitive and information rich guide for understanding how user experience degrades, than the widely used average throughput metric does. We further show that only the combined use of this existing and our new metrics can provide a holistic perspective on overall user performance.

1. Introduction

Streaming video has experienced unprecedented growth in both fixed and wireless networks over the last few years [1,2] and this development is expected to continue. Cisco [1] predicts annual video growth rates 2016–2021 of 27% and 55% for fixed and mobile access respectively whereas Ericsson [2] predicts an annual video growth rate for mobiles access 2017–2023 of 50%. The strong growth means that video has become the dominant source of internet traffic and that its dominance will increase. Cisco [1] reports that the fraction of consumer internet traffic that relates to video traffic was 72% in 2016 and predicts that it will reach 81% in 2021 and Ericsson [2] predicts that video will account for 75% of all mobile data traffic in 2023. The vast majority of this is made up of Adaptive BitRate (ABR) video from streaming sites such as Netflix, YouTube and Amazon Video [3].

Cost-effective delivery of this type of traffic with an acceptable user Quality of Experience (QoE) is a profound challenge for service providers because of the stresses it places on access and core network infrastructure alike. In the case of

☆ The authors acknowledge Telstra and Ericsson for the support of this work.
* Corresponding author.
E-mail addresses: ake.arvidsson@hkr.se (Å. Arvidsson), milosh.ivanovich@team.telstra.com (M. Ivanovich), paul.g.fitzpatrick@team.telstra.com (P. Fitzpatrick).
[1] Formerly with Ericsson AB, Stockholm, Sweden.

https://doi.org/10.1016/j.peva.2021.102199

the latter, it is critical to be able to optimally dimension the domestic and international "peering links" between a service provider and the various content providers which supply the video traffic. Although content caching within service provider networks is one technique that is often used, there is always the need to have "right sized" direct links for non-cached content or when cache capacity is exhausted. In this context, the development and application of best practice dimensioning methods provide two tangible benefits to operators: (i) customer satisfaction and loyalty directly flowing from winning in the competitive environment where content providers publicise their ranking of operators' video performance (see, *e.g.*, Netflix [4] which publishes such rankings regularly) and (ii) cost minimisation from avoidance of over-dimensioned links.

This raises the practical problem of characterising the behaviour of a large number of ABR video flows which share a high capacity traffic aggregation link in the context of an Internet Protocol (IP)-based core network not amenable to classical CAC, call admission control. The ability of ABR video to adapt by scaling its quality (*i.e.* resolution) in line with the available bitrate of the end to end path [5], motivates the need for a model which predicts how many ABR videos can share such a bottleneck before their QoE degrades to unacceptable levels.

In this paper we focus on fixed capacity links carrying IP traffic and develop a model for this scenario based on two key user-centric performance metrics: (i) the average proportion of a video's playing time during which the quality is reduced and (ii) the average proportion of videos which experience reduced quality at least once during their playing time. For the purpose of our current model, "reduced quality" is initially defined as a video experiencing at least one period during its playing time where it is not at the highest possible resolution (and hence highest encoded bitrate). Later, this definition is extended and generalised to include the ability to quantify the performance impact of any specified degree of rate reduction. It should be noted however that various definitions of "reduced quality" can be used, and this is a particularly active area of research due to the challenges imposed by the practice of encrypting ABR video traffic. Various methods have been developed to deal with these challenges [6], but this is beyond the scope of the present work.

Instead, our focus in the development of the proposed model and metrics has been to tackle the challenge of identifying user experience metrics that are related to network performance without being overly complicated by needing to be modelled at the user device, as would be the case for metrics like "time to start", "buffer stalls" and "buffer stall duration".

Bitrate is an example of a metric that can be modelled at the network level. However as shown in Section 3.2, on its own it cannot capture the user experience dimension as well as when complemented by our two proposed metrics. Furthermore, our metrics are shown to have a greater sensitivity than bitrate to changes in link parameters and this helps operators better understand the trade-offs between network cost and user experience. The application of our model to bottleneck links with variable capacity (*e.g.* a wireless network) remains an area for further study.

2. System model

We consider a link that carries a variable number of ABR video flows delivered over an IP network, and we are interested in the extent to which the bitrate of these flows is reduced from their highest encoded bitrate, due to congestion on the link and the reaction of the ABR mechanism to that congestion. In this context a traffic flow is the transmission of an ABR video. To this end, we model the system as an M/M/∞ system where the number of customers $k \geq 0$ in the system represents the number of flows on the link and we let $K \geq 0$ represent the number of flows that can be supported before rates are reduced due to congestion.

In formal terms, denote the arrival inter-arrival time of video flows by $1/\lambda$ seconds, the average duration of video flows by $1/\mu$ seconds, the highest encoded bitrate of a video flow by c bits/second and the link capacity by C bits/second. We then have that the system load (*i.e.*, average number of flows in progress) $\rho = \lambda/\mu$ and that the critical level of congestion (*i.e.*, the level above which rate reduction kicks in) is $K = \lfloor C/c \rfloor$.

We are interested in two performance metrics, *viz.* the fraction of time a flow is subject to rate reduction and the fraction of flows which are subject to rate reduction

2.1. Fraction of time a flow is subject to rate reduction

To calculate this metric we consider a tagged flow, *i.e.* with duration t which is subject to rate reduction during a time b and write

$$b = rt \tag{1}$$

where r denotes the fraction of the duration during which rate reduction applies. Taking the averages in (1) we get

$$\mathrm{E}[b] = \mathrm{E}[rt]$$

By definition r is a fraction and t is the time. That is, b and t are clearly dependent (we know, e.g., that $b \leq t$) but while b and t both are related to the duration of a tagged flow, r is related to the state of the system during the sojourn time of that flow. The tagged flow can be seen as a random observer of the fraction of time spent in states above state K, i.e., the probability that the state of the Markov chain exceeds K, and the state of the Markov chain is independent of the observation period, hence the observed ratio r is independent of the observation time t. It is thus the underlying system load ρ, and not the duration t of the tagged flow, that drives $r = b/t$. Significantly, this intuitive understanding of the

independence of r and t is confirmed experimentally via simulation and shown in Fig. 6. Applying the independence we get

$$\mathrm{E}[b] = \mathrm{E}[rt] = \mathrm{E}[r]\,\mathrm{E}[t]$$

which after some trivial algebra gives

$$\mathrm{E}[r] = \frac{\mathrm{E}[b]}{\mathrm{E}[t]}.$$

But we also know that by the definition of r

$$\mathrm{E}[r] = \mathrm{E}\left[\frac{b}{t}\right]$$

hence we see that

$$\mathrm{E}\left[\frac{b}{t}\right] = \frac{\mathrm{E}[b]}{\mathrm{E}[t]}.$$

This is an important property, because $\mathrm{E}[t]$ is known and we can derive $\mathrm{E}[b]$ whereas deriving $\mathrm{E}[b/t]$ directly is a very complex task.

To derive $B = \mathrm{E}[b]$ we consider a flow arriving to state k and let $B_k = \mathrm{E}[b|k]$ denote the conditional expectation of the time during which rate reduction applies. Using the memoryless property of the Markov model we may now write

$$B_k = \begin{cases} \text{For } k < K\text{:} \\ \frac{k\mu}{\lambda+(k+1)\mu} B_{k-1} + \frac{\lambda}{\lambda+(k+1)\mu} B_{k+1} + 0 \\ \text{For } k \geq K\text{:} \\ \frac{k\mu}{\lambda+(k+1)\mu} B_{k-1} + \frac{\lambda}{\lambda+(k+1)\mu} B_{k+1} + \frac{1}{\lambda+(k+1)\mu} \end{cases} \tag{2}$$

To see this, note that a tagged flow that arrives to state k will change the state to $k+1$ and on the average accumulate a congestion time 0 if $k < K$ and $1/(\lambda + (k+1)\mu)$ if $k \geq K$, the third terms in Eq. (2), after which three things can happen, *viz.*

1. The tagged flow departs. In this case it will not accumulate more congestion time hence there are no corresponding terms in Eq. (2).
2. Another flow departs. In this case the state changes to k and the tagged flow will, because of the renewal properties of the Markov model, on the average accumulate the same congestion time B_{k-1} as a flow that arrives to state $k-1$ (and thus changes the state to k). This case corresponds to the first terms in Eq. (2) where the first factor is the probability of another flow departing and the second factor is the expected, additional accumulated time.
3. Another flow arrives. In this case the state changes to $k+2$ and the tagged flow will, because of the renewal properties of the Markov model, on the average accumulate the same time B_{k+1} as a flow that arrives to state $k+1$ (and thus changes the state to $k+2$). This case corresponds to the second terms in Eq. (2) where the first factor is the probability of another flow arriving and the second factor is the expected, additional accumulated time.

Multiplying both sides in Eq. (2) by $(\lambda + (k+1)\mu)\,p_k$, where p_k is the probability that there are k flows in the system, and summing the left hand side and the right hand sides over all k-values yields

$$\sum_{k=0}^{\infty} p_k\,(\lambda + (k+1)\mu)\,B_k = \sum_{k=0}^{\infty} p_k\,\lambda\,B_{k+1} + \sum_{k=1}^{\infty} p_k\,k\mu\,B_{k-1} + \sum_{k=K}^{\infty} p_k, \tag{3}$$

where the first sum on the right hand side can be rewritten as

$$\sum_{k=0}^{\infty} p_k\,\lambda\,B_{k+1} = \sum_{k=1}^{\infty} p_{k-1}\,\lambda\,B_k = \sum_{k=1}^{\infty} p_k\,k\mu\,B_k = \sum_{k=0}^{\infty} p_k\,k\mu\,B_k,$$

by using that $p_k = \rho^k/k!\;e^{-\rho}$ and that $\rho = \lambda/\mu$, and the second sum on the right hand side can be rewritten as

$$\sum_{k=1}^{\infty} p_k\,k\mu\,B_{k-1} = \sum_{k=1}^{\infty} p_{k-1}\,\lambda\,B_{k-1} = \sum_{k=0}^{\infty} p_k\,\lambda\,B_k,$$

again by using that $p_k = \rho^k/k!\;e^{-\rho}$ and that $\rho = \lambda/\mu$, after which we may simplify Eq. (3) by removing terms that cancel out

$$\mu \sum_{k=0}^{\infty} p_k\,B_k = \sum_{k=K}^{\infty} p_k$$

in which we may identify the sum on the left hand side as the unconditional expectation B of B_k

$$\sum_{k=0}^{\infty} p_k B_k \equiv B = \frac{1}{\mu} \sum_{k=K}^{\infty} p_k$$

hence we finally obtain

$$\mathrm{E}[r] = \frac{B}{\mathrm{E}[t]} = \sum_{k=K}^{\infty} p_k = 1 - \sum_{k=0}^{K-1} p_k$$

where we have used the fact that by definition $\mathrm{E}[t] = 1/\mu$.

This result can be generalised to an M/G/∞ queueing system by applying the following two principles. Firstly, both [7] and [8] show two important properties of Little's Law: (i) its application is independent of the arrival and departure processes distributions and (ii) it can be applied to any part of a queueing system provided the appropriate mean arrival rate, mean number of flows in the system and mean time in the system are respected. Secondly, [7] and [8] show that the steady state probability of k customers in an M/G/∞ queueing system is identical to that of an M/M/∞ queueing system, $p_k = \rho^k/k!\, e^{-\rho}$ where $\rho = \lambda/\mu$.

Applying Little's Law to that part of the $M/G/\infty$ system comprising the congested states, let the mean number of flows in the congested states be N_{con} and noting that the arrival rate of flows is λ. This results in

$$B = \frac{N_{\mathrm{con}}}{\lambda}. \tag{4}$$

The mean number of flows in the congested states, N_{con}, can be found by applying the appropriate relationships for the M/G/∞ queue ($p_k = \rho^k/k!\, e^{-\rho}$ and $\rho = \lambda/\mu$).

$$N_{\mathrm{con}} = \sum_{k=K+1}^{\infty} k p_k = \sum_{k=K+1}^{\infty} \rho p_{k-1} = \rho \sum_{k=K}^{\infty} p_k. \tag{5}$$

Combining (4) and (5) leads to

$$B = \frac{N_{\mathrm{con}}}{\lambda} = \frac{\rho}{\lambda} \sum_{k=K}^{\infty} p_k = \frac{1}{\mu} \sum_{k=K}^{\infty} p_k$$

and we obtain as shown above

$$\mathrm{E}[r] = \frac{B}{\mathrm{E}[t]} = B\mu = \sum_{k=K}^{\infty} p_k = 1 - \sum_{k=0}^{K-1} p_k.$$

where we again have used the fact that $\mathrm{E}[t] = 1/\mu$.

2.2. Fraction of flows subject to rate reduction

To begin we note that flows which arrive to state $k : k < K$ *may* be subject to rate reduction whereas flows that arrive to state $k : k \geq K$ *will* be subject to rate reduction. Let a be the probability that a tagged flow will be subject to rate reduction, which is therefore

$$a = \sum_{k=0}^{K-1} p_k \mathrm{Pr}\left\{ \begin{array}{c} \text{Tagged flow arrives in state } k : k < K \text{ and} \\ \text{visits states higher than } K \text{ prior to completion} \end{array} \right\} + \sum_{k=K}^{\infty} p_k.$$

For states $k = 0, \ldots, K-1$, the Pr{Tagged flow arrives in a state $k : k < K$ and visits states higher than K prior to completion} is essentially an absorption problem where the tagged flow arrives in a state $k : k < K$ and experiences rate reduction when absorbed into state $K+1$. Let this probability be a_k for which, similar to Eq. (2), the following relationships hold for the a_k terms

$$\begin{aligned} a_0 &= \frac{\lambda}{\lambda + \mu} a_1 \\ a_1 &= \frac{\lambda}{\lambda + 2\mu} a_2 + \frac{\mu}{\lambda + 2\mu} a_0 \\ &\vdots \\ a_k &= \frac{\lambda}{\lambda + (k+1)\mu} a_{k+1} + \frac{k\mu}{\lambda + (k+1)\mu} a_{k-1} \\ &\vdots \end{aligned} \tag{6}$$

$$a_{K-2} = \frac{\lambda}{\lambda + (K-1)\mu} a_{K-1} + \frac{(K-2)\mu}{\lambda + (K-1)\mu} a_{K-3}$$
$$a_{K-1} = \frac{\lambda}{\lambda + K\mu} + \frac{(K-1)\mu}{\lambda + K\mu} a_{K-2},$$

where the different terms refer to different events. The arrival of another flow is represented by the first terms, the departure of another flow but the tagged one is represented by the second terms, and the departure of the tagged flow itself is not represented as it vanishes since the flow ended before any rate reduction took place.

To obtain a we uncondition on k,

$$a = \sum_{k=0}^{K-1} p_k a_k + \sum_{k=K}^{\infty} p_k, \tag{7}$$

but we note that the set of equations (6) does not lend itself to the same simple procedure as the corresponding set of equations (2). This is because the time t is memoryless and cumulative, which means that Eq. (2) is infinite, whereas absorption a is not, which means that Eq. (6) is limited.

From this observation, and after multiplying both sides in Eq. (6) by $\lambda + (k+1)\mu$ and applying some simple algebra, we thus instead obtain the following set of K linear equations for the K conditional probabilities $a_0, \ldots, a_{K-1}$,

$$\begin{aligned}
0 &= (\lambda + \mu)\, a_0 - \lambda\, a_1 \\
0 &= (\lambda + 2\mu)\, a_1 - \lambda\, a_2 - \mu\, a_0 \\
&\vdots \\
0 &= (\lambda + (k+1)\mu)\, a_k - \lambda\, a_{k+1} - k\mu\, a_{k-1} \\
&\vdots \\
0 &= (\lambda + (K-1)\mu)\, a_{K-2} - \lambda\, a_{K-1} - (K-2)\mu\, a_{K-3} \\
\lambda &= (\lambda + K\mu)\, a_{K-1} - (K-1)\mu\, a_{K-2},
\end{aligned} \tag{8}$$

which can be solved directly or by a recursive procedure the details of which, however, are left for possible further work.

We also remark that, contrary to the previous metric, this metric is sensitive to the distribution of the flow durations.

2.3. Generalised metrics

Our two metrics $\mathrm{E}[r]$ and a both capture the extent to which flows experience *any* rate reduction but not by *how much*. The calculations are, however, valid for any choice of "critical level" K hence we can analyse performance in much more detail by varying K.

To this end we extend the definition of K to $K(\xi) = \lfloor C/((1-\xi)c) \rfloor$ where $\xi = [0, 1)$ denotes the *degree of rate reduction*. Using this extended definition of K, we can use the same mathematical treatment to study two scenarios which are equivalent numerically but have different real-world meaning.

To see this, first consider the original critical level of no reduction and note that $K(0) = \lfloor C/c \rfloor = K$, *i.e.* the same as above. Then consider a new critical level of, *e.g.*, 20% reduction and note that $K(0.2) = C/(0.8c) = 1.25C/c = 1.25K$. In this $\xi = 0.2$ example the interpretation could either be:

1. A larger link of capacity $(C/0.8) = 1.25C$ which is carrying flows of the same rate "c", and where the analysis based on $K(\xi)$ can determine for this larger link the fraction of time flows experience some rate reduction compared to the flow baseline rate "c".

2. A link of the same capacity C carrying flows which were originally rate "c" but due to congestion have been reduced to rate "$0.8c$". Here the analysis based on $K(\xi)$ will determine the fraction of time flows on a capacity C link will experience a rate reduction of at least $\xi = 0.2$ compared to "c".

Although both interpretations are numerically equivalent, it is the latter which is more general and thus more useful to us. With it, we can thus calculate $\mathrm{E}[r(\xi)]$ and determine not only the fraction of time with *some* rate reduction, but also the fraction of time with rate reduction *larger than any level* ξ. Similarly we can calculate $a(\xi)$ to determine not only the fraction of flows which are subject to *some* rate reduction, but also the fraction of flows with rate reductions *larger than any level* ξ.

3. Results

3.1. Parameters

The selection of values for K depends on typical values for the video bitrate c and the link capacity C through the relationship $K = \lfloor C/c \rfloor$. This is complicated by the dependence of c on the video codec capability and also on any choice of target average video rate. While there are standard video codecs, the most popular codecs in use are determined by the

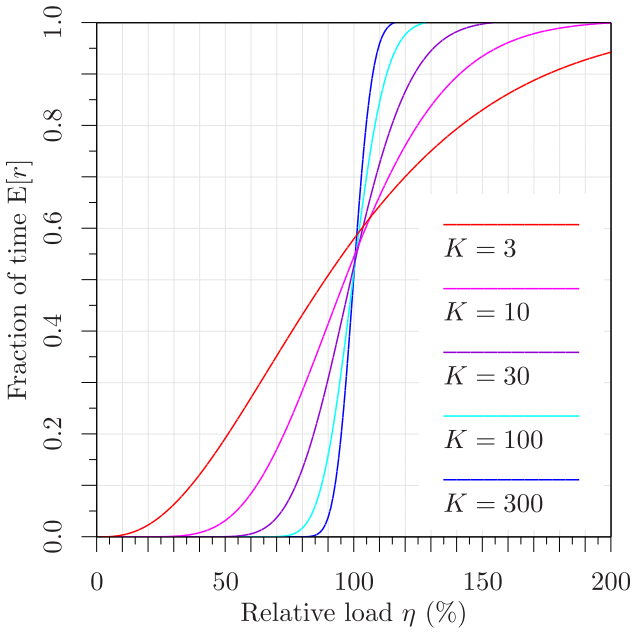

Fig. 1. Expected fraction of time a flow is subject to congestion $\mathrm{E}[r]$ as a function of relative link load η for different link sizes.

most popular video content providers. Examples of representative codec data rates can be found in [9] and [10]. Further, there is a trade-off in the average video rate depending on the choice of ρ relative to C. Examples of typical average video data rates achieved in practice can be found in reports by Netflix [4] and YouTube [11].

In this study a wide range of conditions are modelled by using values of $3 \leq K \leq 300$, as these range across a suitable mix of realistic scenarios that allow for the trade-offs to be demonstrated. Values for the relative system load $\eta = \rho/K$ of $0\% \leq \eta \leq 200\%$ also allow a broad view of system performance. It is seen that both metrics increase with increasing load and that the increase becomes more distinct the larger the system. We note that both observations are expected (*cf. e.g.*, the Erlang-B function).

3.2. Performance

Figs. 1 and 2 respectively depict $\mathrm{E}[r]$, the expected fraction of time during which a flow experiences rate reduction, and a, the probability that a flow is subject to rate reduction as functions of the relative system load η for different system sizes.

An example of the generalised metrics for a link with a nominal capacity of $K = \lfloor C/c \rfloor = 30$ flows is shown in Figs. 3 and 4. It is seen that the leftmost points of the curves $\xi = 0$ agree with Figs. 1 and 2 respectively and that the fractions drop as higher rate reduction thresholds are applied.

These results illustrate how the presented model provides a practical capability for ensuring that user quality of experience, can be taken into account during network dimensioning. In other words, computationally efficient analysis of practical engineering trade-offs becomes possible, whereby minimal network capacity can be found while still satisfying thresholds for one or more user video performance parameters.

As a simple illustrative example, in the context of Figs. 1 and 2 consider setting a particular composite threshold target of $E[r] < 0.05$ and $a < 0.1$, noting that we may want to set the first target slightly more "tightly" to ensure that however many users are impacted by congestion, the impact of that congestion on an individual user session is not too extreme, *i.e.* 10% chance of a user session being impacted, but that impact being restricted to less than 5% of the sessions' average duration.

Furthermore, these composite thresholds can be substantially enriched using the metric generalisation approach developed in Section 2.3. The network operator can derive significant additional value from making these $\mathrm{E}[r]$ and a threshold targets dependent on levels of rate reduction as illustrated in the example presented in Figs. 3 and 4. This would allow the operator to make the dimensioning process better coupled to user quality of experience by considering issues like: (i) level of rate degradation within a flow which leads to severe enough video stalling and rebuffering to make content "unviewable", and (ii) the possibility of allowing some tolerable fraction of user flows which are rate degraded beyond the set threshold.

By simultaneously searching for the maximum loads allowed in Figs. 1 and 2 which still respect our composite threshold, and taking into account typical levels of offered traffic ρ, we would then be able to find appropriate system sizes

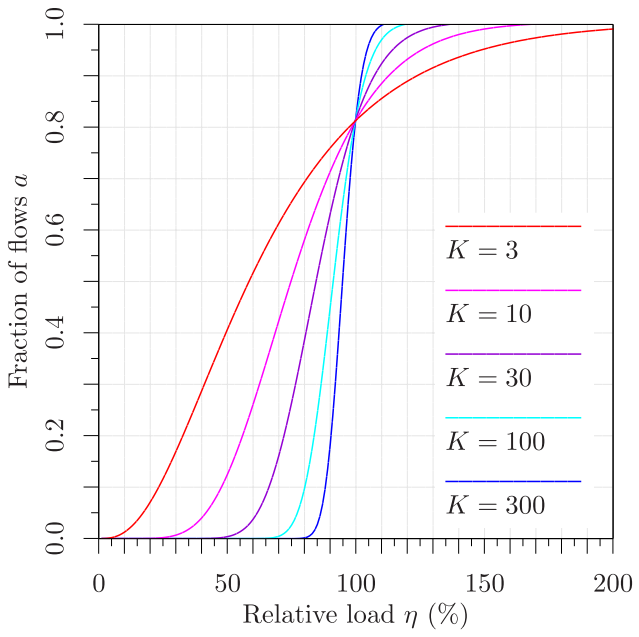

Fig. 2. Probability of a flow being subject to congestion a as a function of relative link load η for different link sizes.

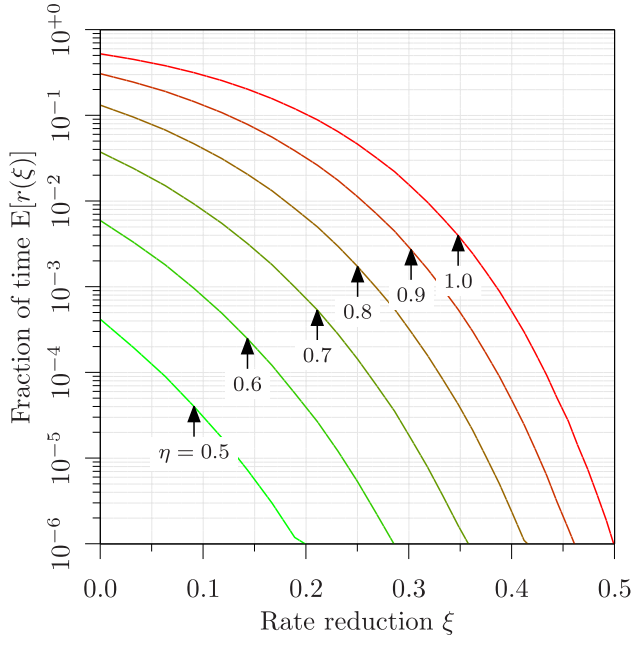

Fig. 3. Expected fraction of time flows are subject to rate reduction by at least ξ, $\mathrm{E}[r(\xi)]$, as a function of the rate reduction ξ for different relative link loads η as indicated.

which meet our criteria. For the present example and assuming hypothetically $\rho < 270$, we could choose our $K = 300$ contour noting that it would result in a relative load $\eta < 90\%$ and the composite threshold being met in both figures.

To illustrate the additional value of our metrics relative to more conventional metrics Fig. 5 shows the behaviour of the normalised average data rate ν (where $\nu = (\sum_{k=1}^{\infty} p'_k \times \min(C/k, c))/c$ and $p'_k = p_k / \sum_{k=1}^{\infty} p_k$) versus relative load η for the same system sizes. Over the range of values $\eta \leq 100\%$ the deviation of ν from 1 occurs at slightly higher loads than the deviation of $\mathrm{E}[r]$ and a from 0. Importantly, the rate of decrease in ν with increasing η is less than the rate of increasing $\mathrm{E}[r]$ and a with increasing η. This results in smaller changes in ν corresponding to larger changes in $\mathrm{E}[r]$ and a

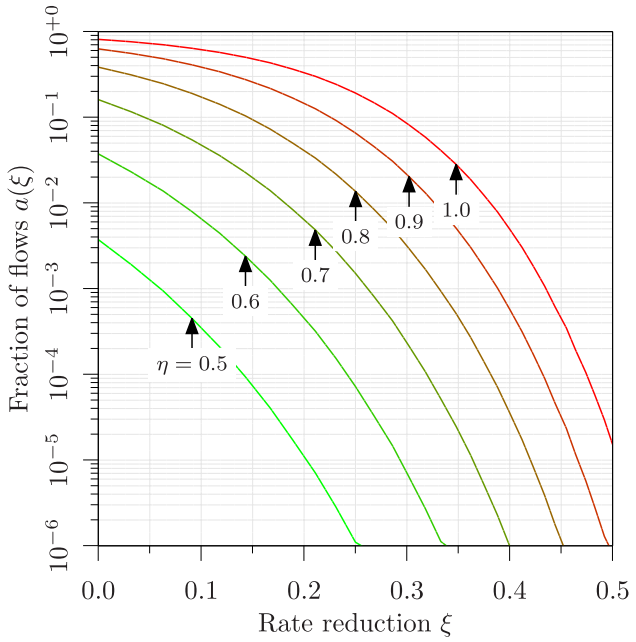

Fig. 4. Fraction of flows subject to rate reduction by at least ξ, $a(\xi)$, as a function of the rate reduction ξ for different relative link loads η as indicated.

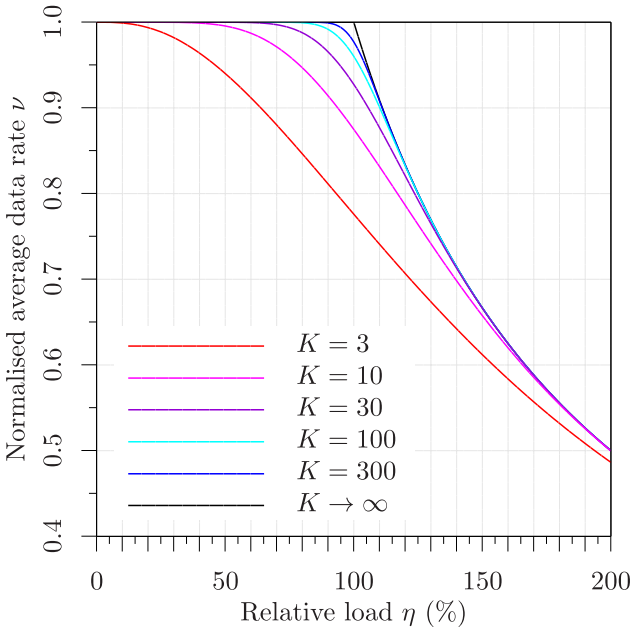

Fig. 5. Normalised average rate ν as a function of relative link load η for different link sizes.

for the same η change with the degree of change dependent on system size K. Here, at typical operational loads, it should be noted that ν only captures the magnitude of the rate reduction due to congestion. Our new metrics provide operators with a significantly broader view of the user experience with $a(\xi)$ and $\mathrm{E}[r(\xi)]$ at each single load and average data rate point.

Table 1
Example of the impact of the distribution of the duration for $K = 30$ and $\rho = 30$.

Duration distribution	E[r]		a	
	Sim.	Cal.	Sim.	Cal.
Negexp	0.5244	0.5243	0.8129	0.8129
Erlang-3	0.5245		0.8635	
Hyperexp-2	0.5241		0.7513	
Erlang-10	0.5242		0.8860	
Hyperexp-2	0.5244		0.7048	

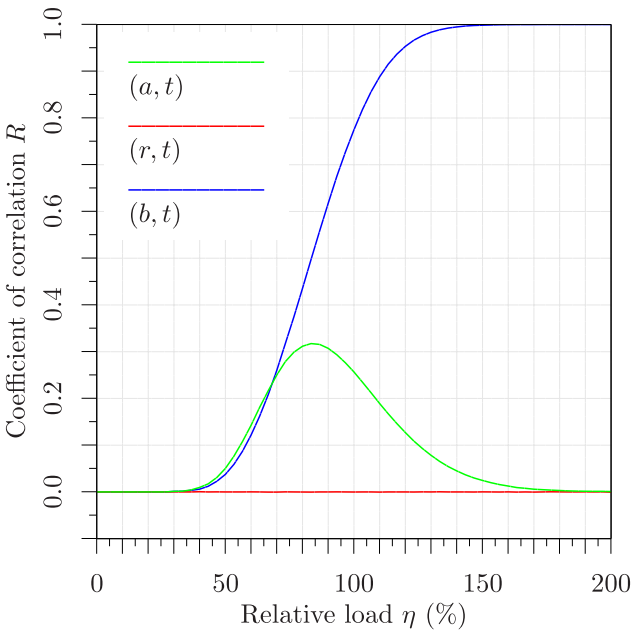

Fig. 6. Coefficient of the observed correlation R between a and t (a, t), between r and t (r, t) and between b and t (b, t) as a function of the relative load η for $K = 30$.

3.3. Sensitivity

Returning to the earlier discussion about the significance of the distribution of the duration, Table 1 provides a numerical example obtained by simulating 300 million arrivals to a system with a capacity of $K = 30$ subject to a load of $\rho = 30$ under five different duration distributions. The first case, the negative exponential distribution, corresponds to the theoretical model, the two Erlang-k distributions scale the coefficients of variation and skewness by $1/k$ and $1/\sqrt{k}$ respectively and the corresponding two Hyper-exponential-2 distributions were set to scale the two coefficients by k and $\sqrt{k}$ respectively.

The results confirm the fact that $\mathrm{E}[r]$ is insensitive to the distribution of the duration (except its mean) and that the same does not apply to a. In more detail we note that a tends to drop with increasing variance, *i.e.* that performance improves, and we remark that this behaviour differs from most performance metrics in queueing theory where increasing variance typically is associated with worse performance.

To understand the last observation we note that rate reduction can be seen as having two causes; *viz.*, first, congestion immediately on arrival and, second, no congestion on arrival but before departure. The first event only depends on the state probabilities and, since these are independent of the distribution of the service time, this event is equally likely no matter the distribution of the duration. The second event, however, depends on the duration and the shorter the duration the less the risk of remaining in the system long enough to experience congestion prior to departure. Noting that increasing the variance under a fixed mean results in gradually fewer flows with gradually longer durations and at the same time more flows with shorter durations, we conclude that the former group, which runs an increasing risk of experiencing congestion, shrinks while the later group, which runs a decreasing risk of experiencing congestion, grows.

This is illustrated in Fig. 6 which depicts the observed coefficient of correlation $R(a, t)$ between a and t and, for comparison, $R(r, t)$ between r and t and $R(b, t)$ between b and t for the same system as in Table 1. The curves are based on 60 points each of which represents 30 million arrivals.

It is seen that the correlation between a and t (green curve) is weak for very low loads and for very high loads, but peaks for medium loads. The poor correlation for low and high loads is explained by the importance of the state of the system when the flow arrives and this is independent of the flow itself (first cause above). The peak for medium loads is similarly explained by the fact that the load is low enough for many flows to arrive to an uncongested system but high enough for long lasting flows not to depart before the system becomes congested (second cause above).

To supplement the discussion we see that increasing load means a growing correlation between the congested time b and the total time t (blue curve) and we note that this can be explained by the intuitive observation that, as load increases, the congested time and total time of each arriving flow converge. We also note that, as expected, there is no correlation between the fraction of congested time r and the total time t (red curve).

Considering the earlier illustrative example, the implications of dependence of a with t can be seen in terms of the impact on different durations t for videos on user experience. Figs. 1 and 2 provide insight into the average user performance, but Table 1 highlights that mean metrics may not be sufficient to understand the full impact of duration distribution on user experience.

4. Related work

Much of the literature on ABR video, in terms of a general overview and detailed studies respectively, is primarily focused on improving the operation of ABR. See [12] as an example. From a network operator's perspective, a large fraction of ABR traffic appears as over the top (OTT) traffic and as such, optimisation of the ABR mechanism is outside their scope to impact other than through some form of limiting of the video throughput. However, operators need to adequately dimension the links that deliver the video to users in a way that accounts for its adaptive behaviour. Theirs is an interest in cost-effective delivery of ABR video that meets users' video quality expectations. From this perspective, models that capture the trade-offs between link capacity and video user experience are most useful.

The M/M/∞ queue has been used to model the performance of telecommunications systems and general queueing problems. Some examples of its application can be found in [13–17], and [18]. Relevant works in telecommunications are [13] (and its generalisation in [14] and [15]) and [16]. These papers, particularly [13] develop solutions for the system level metrics, the average time for a system exceeding a given capacity C (equivalent to the number of flows $k \geq K$ in this paper), and the average number of customers arriving during this congestion period (also referred to as the C-congestion period in [16]). The results from these papers are directly applicable to deriving the same system level metrics in this work because they share the M/M/∞ system, but only in respect of these metrics. They assume that "the required transmission of a single burst" (or flow) is fixed. Whereas in this work, the ABR system can reduce or increase the transmission rate per flow according to the congestion on the link. From these perspectives all of these works do not readily provide the solution for the two user focused performance metrics of interest, $\mathrm{E}[r]$ and a, nor do they address the user focused extension of K (or C in those papers) to $K(\xi)$. Further, the work in this paper also identifies that while $\mathrm{E}[r]$ is insensitive to the flow duration distribution, a is sensitive to this distribution.

The work by Bonald et al. in [19] is an important complementary contribution in this area, particularly because of its ability to capture the interaction of elastic and video streaming traffic in a cellular network. However, with regard to the video streaming aspect, it does have a different focus compared to our work. The authors' modelling assumptions of (i) dealing with a relatively low traffic aggregation level (*i.e.* a single cell in a cellular network); (ii) the ability to store the entire video file in the client's play-out buffer and (iii) the ability to download the video file's chunks at the maximum instantaneous cell throughput, mean that their model is dealing with video streaming at both a very different timescale and traffic aggregation level from a network topology perspective.

Our work is predicated on the assumption of a high traffic aggregation level (*i.e.* backbone part of a network) which is dominated by the longer timescales corresponding to video play-out durations (order of minutes). At this longer timescale the concept of buffering does not matter as much, because we are concerned about the overall video session durations rather than the chunks of which they are composed. In line with this perspective, the user and system performance metrics which we introduce focus on the big picture view considering the question of "how many users are allowed to share an aggregation link somewhere in the network backbone"? Note that at these timescales, knowledge of the fine-grained per-video-chunk buffering dynamics will not be overly useful in answering this fundamental question.

On the other hand, the work in [19] is predicated on the assumption of a lower-traffic aggregation point (a cell), where the shorter timescales associated with video chunk delivery do matter, especially given the desire to capture interaction with other traffic types (*e.g.* elastic traffic). As a result, it can be postulated that together, our approach and that of [19] holistically account for video streaming dynamics by describing it at two important levels of network topology (access and backbone).

5. Conclusions

This paper has investigated two simple and useful metrics for evaluating the user experience in shared fixed capacity ABR video links and developed analytic expressions for both. In addition a simulation evaluation of their sensitivity to the flow duration distribution has been undertaken. The two metrics of interest, the fraction of time a flow is subject to rate reduction $\mathrm{E}[r(\xi)]$ and the fraction of flows subject to rate reduction $a(\xi)$ provide alternatives to metrics like the average

rate and allow a more “user centric" view when performing link dimensioning. Furthermore, both metrics are able to quantify specific rather than arbitrary levels of rate reduction. This is a key requirement to assist operators in their quest to balance cost-effective design against user video experience.

From the perspective of the $\mathrm{E}[r(\xi)]$ metric, the M/G/∞ system model is shown to be applicable, suggesting performance that is independent of the flow duration distribution. However, surprisingly the same is shown to not hold in the case of the $a(\xi)$ metric, which was based on the more restrictive M/M/∞ model and where sensitivity to the flow duration distribution is indeed observed via simulations. What is more, this sensitivity is novel because performance improves with increasing variance in the duration. This result runs counter to other systems where increased variance is associated with worse performance. A further important contribution of this paper has been to demonstrate via Fig. 6 that r and t are not correlated due to independence, while both b and t and a and t show load-dependent levels of correlation.

It can also be observed that for relative loads $<$ 100%, the transition from low probability of poor performance to higher probability of poor performance for both of these metrics happens over relatively small changes in relative load. This is contrary to the change in the average data rate metric, which on face value is more gradual. So these metrics can provide a more “sensitive" view of the degradation in user experience over this load range. However, as we have shown, our user focused metrics can buttress the legacy metric to provide a more holistic perspective on overall user performance. The extension of our model to better quantify the extent and frequency of bitrate switching, and how this maps to user quality of experience including the relationship between rate reduction and device screen size, remains a topic for future work.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors wish to acknowledge the anonymous reviewers for the contribution that their comments have made to improving this paper. The authors acknowledge Telstra and Ericsson for the support of this work.

References

[1] Cisco, Cisco Visual Networking Index: Forecast and Methodology, 2016–2021, White paper, 2017, pp. 1–17, URL https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.pdf.
[2] Ericsson, Ericsson Mobility Report, White paper, 2017, pp. 1–32, URL https://www.ericsson.com/assets/local/mobility-report/documents/2017/ericsson-mobility-report-november-2017.pdf.
[3] J. D'Onfro, More than 70% of internet traffic during peak hours now comes from video and music streaming, 2015, Accessed 23/01/2019, URL https://www.businessinsider.com.au/sandvine-bandwidth-data-shows-70-of-internet-traffic-is-video-and-music-streaming-2015-12?r=US&IR=T.
[4] Netflix, Netflix ISP Speed Index, https://ispspeedindex.netflix.com/.
[5] T.-Y. Huang, N. Handigol, B. Heller, N. McKeown, R. Johari, Confused, Timid, and Unstable: Picking a Video Streaming Rate is Hard, in: Proceedings of the 2012 Internet Measurement Conference, 2012, pp. 225–238.
[6] S.C. Madanapalli, H.H. Gharakhieli, V. Sivaraman, Inferring Netflix User Experience from Broadband Network Measurement, in: Proceedings of the 3rd Network Traffic Measurement and Analysis Conference (TMA) 2019, pp. 41–48.
[7] D. Gross, C. Harris, Fundamentals of Queueing Theory, in: Wiley Series in Probability and Mathematical Statistics, Wiley, 1998.
[8] L. Kleinrock, Queueing Theory Volume 1: Theory, John Wiley and Sons, 1975.
[9] G. Sullivan, P. Topiwala, A. Luthra, The H.264/AVC Advanced Video Coding Standard: Overview and Introduction to the Fidelity Range Extensions, in: Proceedings of SPIE, Applications of Digital Image Processing XXVII, Vol. 5558, 2004, pp. 454–474.
[10] M. Řeřábek, P. Hanhart, P. Korshunov, T. Ebrahimi, Quality Evaluation of HEVC and VP9 Video Compression in Real-Time Applications, in: Proceedings of 7th International Workshop on Quality of Multimedia Experience, 2015, pp. 1–6.
[11] Google, Google Video Quality Report, https://www.google.com/get/videoqualityreport/.
[12] J.W. Kleinrouweler, S. Cabrero, R. van der Mier, P. Cesar, Modeling Stability and Bitrate of Network-Assisted HTTP Adaptive Streaming Players, in: Proceedings of 27th International Teletraffic Congress, 2017, pp. 177–184.
[13] F. Guillemin, A. Simonian, Transient Characteristics of an M/M/∞ System, Adv. Appl. Prob. 27 (3) (1995) 862–888.
[14] F. Guillemin, D. Pinchon, Continued Fraction Analaysis of the Duration of an Excursion in an M/M/∞ System, Adv. Appl. Prob. 35 (1) (1998) 165–183.
[15] D. Flajolet, F. Guillemin, The Formal Theory of the Birth-and-Death Processes, Lattice Path Combinatorics and Continued Fractions, Adv. Appl. Prob. 32 (3) (2000) 750–778.
[16] F. Roijers, M. Mandjes, H. van den Berg, Analysis of Congestion Periods of an M/M/∞-Queue, Perform. Eval. 64 (7) (2007) 737–754.
[17] M. Rumsewicz, P. Taylor, A Spot Welding Reliability Problem, J. Aust. Math. Soc. Ser. Appl. Math. 29 (3) (1988) 257–265.
[18] M. Ramakrishnan, D. Sier, P.G. Taylor, A two-time-scale model for hospital patient flow, IMA J. Manage. Math. 16 (3) (2005) 197–215.
[19] T. Bonald, S.E. Elayoubi, Y.-T. Lin, A Flow-Level Performance Model for Mobile Networks Carrying Adaptive Streaming Traffic, in: Proceeding of IEEE Global Communications Conference (GLOBECOM) 2015, 2015, pp. 1–7.

Åke Arvidsson obtained his M.Sc. and Ph.D. degrees in Electrical Engineering from Lund University, Sweden, in 1982 and 1990 respectively. He has worked with several consultancy companies and held various academic positions in Sweden and Australia and became full professor of teletraffic systems at Blekinge Institute of Technology in 1995. In 1998 he joined Ericsson and since 2018 he is full professor at Kristianstad University. His current research interests include performance analysis and optimisation of cellular networks, transport protocols and content distribution.

Milosh Ivanovich is Principal — Traffic Modelling and Optimisation within the Networks and I.T. division of Telstra and is an Honorary Research Fellow at Monash University. Milosh's interests lie in queueing theory, teletraffic modelling, performance analysis of wireless networks, adaptive flow control and enhancement of TCP/IP in hybrid fixed/wireless environments. Milosh obtained a B.E. (1st class Hons) in Electrical and Computer Systems Engineering (1995), a Master of Computing (1996) and a Ph.D. in Information Technology (1998), all at Monash University Australia. He is the author of several edited book chapters, a patent, and over 50 international journal and conference publications.

Paul Fitzpatrick received his B.E. in Electrical Engineering from Caulfield Institute of Technology, Melbourne, Australia in 1979 and his Ph.D. in Electrical Engineering from Swinburne University, Melbourne, Australia in 1997. Paul has over 40 years experience working in the telecommunications industry and academia. He is currently with the Modelling and Analytics group within Networks and I.T. at Telstra Corp., and an Adjunct Research Associate in the Department of Electrical and Computer System Engineering, Monash University. His current work covers teletraffic engineering, wireless network performance modelling and analysis, user experience modelling and evaluation and applications of data analysis.