User-level Fairness Delivered: Network Resource Allocation for Adaptive Video Streaming

Mu Mu, Steven Simpson, Arsham Farshad, Qiang Ni, Nicholas Race School of Computing and Communications, Lancaster University
LA1 4WA Lancaster, United Kingdom
Email: {m.mu,s.simpson,a.farshad,q.ni,n.race}@lancaster.ac.uk

***Abstract*—HTTP adaptive streaming (HAS) technology is becoming a popular vehicle for online video delivery. HAS applications often compete for network resources without any coordination between each other in a shared network. This leads to quality of experience (QoE) fluctuations and unfairness between end users. This paper introduces a user-level fairness model (UF) which exploits video quality, switching impact and cost efficiency as the fairness metrics to achieve user-level fairness in resource allocation. Experimental results demonstrate how this model is a foundation to orchestrate the resource consumption of HAS streams.**

I. Introduction

Globally, IP video traffic is estimated to be 79% of all consumer Internet traffic in 2018, up from 66% in 2013 [3]. HTTP adaptive streaming (HAS) protocols, especially the MPEG Dynamic Adaptive Streaming over HTTP (DASH), are becoming popular for online video delivery, thanks to their unique adaptation feature that allows dynamic selection of quality representations in the face of network fluctuation. Most HAS protocols adopt TCP as the transport layer protocol whose retransmission mechanism greatly mitigates the impact of network impairments, such as packet loss, to the delivered video quality. TCP aims to increase bandwidth utilization whilst avoiding congestion, which allows the application to fully exploit the available network resources. The adaptation between representations is often managed at the client side and is not specified by the DASH standard. Recent years have seen an increasing number of single-stream HAS optimization algorithms [10], [22], [14], [19] with the main objective being to maximize the quality of user experience of a single HAS stream by exploiting bandwidth estimation, client-side buffer management, and QoE measurement. However such optimization algorithms work on user clients independently without any coordination with other devices in the same network. This leads to QoE fluctuations and unfairness when multiple HAS streams on heterogeneous user devices compete for network resources. The conventional fairness models such as proportional fairness [12] are not suitable for HAS applications where video streams of different characteristics exhibit distinctive utilities. Additionally, crucial HAS QoE impact factors such as the representation switching impact are often neglected.

This paper presents a user-level fairness UF model to be exploited by either of the orchestration frameworks. The UF model incorporates video quality, switching impact and cost efficiency as the fairness metrics, and provides recommendations on resource allocation with respect to the particular balance to be achieved between the three fairness metrics. We carried out a number of experiments to verify the performance of the UF model against the conventional network provisioning mechanisms with different degrees of network fluctuations and various numbers of competing user clients. We also investigated how each dimension of fairness contributes to the overall user-level fairness as well as the impact of the UF model to the user perceived video quality for HAS streams. The results demonstrate the benefits of enabling user-level fairness as a service for future networks. The remainder of the paper is organized as follows. Section II explores background and related work in the field of network management for HAS streams. The design of the UF model is detailed in Section III. Section IV introduces the experiment set-up as well as the results. Discussions and conclusions are given in Section V and VI respectively.

978-1-4673-7113-1/15/$31.00

II. Background and Related work

Using MPEG-DASH as an example, HAS content is prepared in multiple *representations*, each of which is a version of the same content prepared using different encoding specifications (such as bitrate and resolution) according to predefined use scenarios. A number of associated representations formulate an *Adaptation Set*. Each representation is served with time-coded chunks addressable using URLs and retrievable via HTTP. By exploiting the switching points between chunks, "seamless" switching between representations becomes possible. Upon a playback request, the user client receives a *Media Presentation Description (MPD)*, the manifest file that specifies all the resources and structure information required for video retrieval and playback. A DASH client often starts from a representation that matches its screen resolution and at modest bitrates. Once the client detects an increase in available bandwidth, it can switch to a higher bitrate.

Most HAS protocols adopt TCP as the transport mechanism. TCP aims to increase bandwidth utilization whilst avoiding congestion so that HAS applications can fully exploit the network resources and potentially maximize the quality of the delivered content. However, this presents two challenges in delivering a good quality of user experience especially in a shared network environment with heterogeneous user devices. The first challenge is the network fluctuation caused by the

packet delivery scheme. A DASH client may continuously inflate its receiver window during TCP ON periods. This inadvertently forces the sender to burst as much traffic as possible on to the network, until either enough video chunks are buffered at the client (which then switches to OFF mode), or until the sender incurs TCP packet loss. This behavior causes extremely bursty traffic and TCP inefficiency, as connections are repeatedly restarted between ON and OFF periods, resulting in unstable video playback quality[1], [9]. The second challenge is the gap between network-level fairness and user-level fairness when network resources are shared between multiple HAS streams on different devices over heterogeneous networks. Conventionally, the goal of resource allocation in the network is to maximize the aggregate utility of all the users in the network subject to the capacity constraints of the network [12], [17]. One known implementation based on this theory is proportional fairness [12]. Proportional fairness performs well when all users follow the same utility. When users have different QoS needs, proportional fairness favors users who require lower rates to achieve high utility. Therefore, maximizing the combined utility does not necessary lead to fairness between users. In fact, research work in the field of video quality analysis shows that there is no linear correlation (video quality utility) between the bitrate of a video stream and its perceptual quality [2]. This suggests that an increase in bitrate could lead to a significant gain in resulting quality or very limited quality improvement that is barely noticeable to the end user. In order to optimize the efficiency of network resource allocation whilst maintaining a satisfactory level of user experience, it is essential to incorporate the video-quality utility of media streams.

Recent years have seen significant work put forward to improve the QoE of HAS (particularly MPEG-DASH) video distribution. One solution is to have some cross-layer interaction between TCP and HTTP in order to provide the streaming application with better metrics and to allow TCP to reach steady-state [9]. This would indeed improve TCP performance, but would not control the ON/OFF nature of DASH-style applications. Furthermore, it would not attain network-wide fairness across all devices. Tian and Liu [22] use throughput-prediction algorithms to attenuate video rate fluctuations. Mansy et al. [18] have shown that DASH's bursty nature leads to excessive queuing in the network (a phenomenon commonly referred to as bufferbloat [7]), and they proposed adjusting DASH's buffering behavior to keep the size of the client's receiver window low. FESTIVE [11] attempts to improve fairness, stability and efficiency using a DASH player with a stateful, delayed-bitrate update mechanism. Huang et al. introduce a buffer-based approach to rate adaptation to reduce the rebuffer rate in online HAS streaming [10]. A client-side rate-adaptation algorithm for HAS is introduced in [14]. Georgopoulos et al. incorporated OpenFlow and video quality utility as part of work towards network-wide QoE fairness [6].

Most of the aforementioned work focuses on optimizing the network efficiency or the quality of user experience of individual media streams. There is currently a lack of research addressing the user-level fairness of network resource provisioning in a multi-HAS-stream environment. With the increasing number of high-throughput HAS streams delivered in IP networks, quality assurance via over-provisioning becomes less feasible. It is crucial for the service providers to depart from simply providing best-effort networks to orchestrating the network resource consumption with a better understanding of the user-level requirements of user applications, especially the resource-intensive HAS applications. The utility and QoE metrics are defined to a class of application such as HAS streaming and not customized for a service provider. Hence, our proposition also does not conflict with the framework of network neutrality.

III. User-level fairness model

The ultimate goal of the user-level fairness UF model is to orchestrate network resource allocation between HAS streams to improve the QoE fairness. The model takes into account multiple fairness metrics such as video quality (fidelity of video frames), HAS adaptation impact as perceived by the end user, as well as the efficiency of media distribution in achieving user experience from the perspective of service providers. The underlying principle of the UF model is depicted in Figure 1.

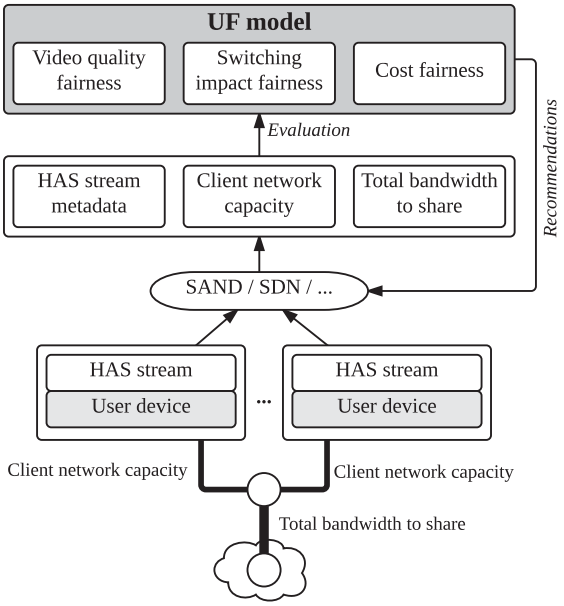


Fig. 1. User-level fairness model

The UF model incorporates three fairness metrics: video-quality fairness, switching-impact fairness and cost-efficiency fairness. The metrics address either the fairness between HAS streams using QoE measurements, or the fairness between network providers and resource consumers as a whole. The fairness metrics are evaluated using: 1) input parameters including context information regarding HAS streams such as current playback bitrate, resolution, etc., 2) current network capacity of user devices, and 3) total bandwidth to share between multiple devices. Both the network capacity and the total bandwidth are dynamic and can be affected by the change of link capacity or background traffic. The input parameters

can be derived using a network-level or application-level QoE framework. The actual extraction of input parameters is further discussed in our related work [4]. Once the UF model determines the best resource allocation strategy to warrant user-level fairness based on given information, recommendations can be forwarded to the network management functions for QoE-aware network management such as metering or QoE routing. Recent advances in networking, such as software-defined networking (SDN) and network-function virtualization (NFV), enable network-wide flexibility and programmability allowing comprehensive network and service functions to be deployed easily in an on-demand fashion.

A. Video quality fairness

In order to fairly share network resources between HAS streams with respect to the QoE, it is crucial to understand and model the impact of network impairments on the delivered video quality. A HAS application chooses an optimal resolution that best matches the native resolution of the playback device and dynamically selects a representation (of a certain bitrate) from the adaptation set of the same frame resolution according to available bandwidth. Based on the assumption that the same encoding scheme (e.g., Group of Picture structure, motion estimation schema, etc.) is employed, a higher encoding bit-rate results in less compression loss and therefore yields higher video quality pertaining to picture fidelity. A video-quality utility (*VQ*) function (a type of rate-distortion function) is often employed to capture the relationship between bitrate and video quality. It captures the notion of the law of diminishing returns [20] – a certain addition of resources to what one already has increased the total worth, but it contributes less and less to the increase if one has more of the resource already.

We adopt the utility functions derived in our previous work [6] as the foundation of the QoE modelling. Figure 2 shows the scatter plot and the fitted utility curves of the HAS video bitrate utility in the three named resolutions. The figure reflects the common understanding that more resources are required to deliver video of the same quality on higher resolutions. The utility plot quantifies such a relationship between video quality and bitrate.

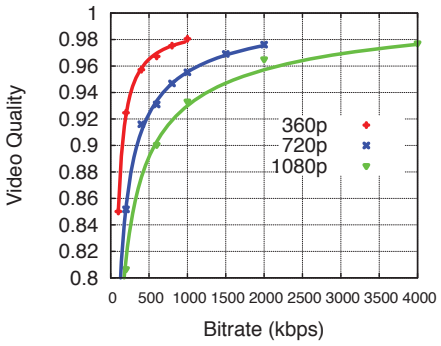


Fig. 2. Scatter plot and utility curves of bitrate utilities

$$Q_{res} = ar^b + c \tag{1}$$

$$Q_{720p} = -4.85r^{-0.647} + 1.011 \tag{2}$$

Equation 1 is the generic QoE utility function. r denotes the video bitrate and the Q is the video quality. A Q of 1 is the maximum possible video quality (when no compression or lossless compression is applied to the content). A utility function $\mathcal{U}_{res}(r)$ is often developed to capture the relative video fidelity ($Q_{res}(r)$) compared with Q_{res}^{MAX}. With the Q_{res}^{MAX} rescaled to be 1 as the reference, $Q_{res}(r)$ has the data range of (0,1]. a, b, and c are the coefficients that instantiate the utility function for certain video resolutions. For instance, Equation 2 is an instance of utility function for 720p videos. The utility function is of low complexity (suitable for real-time quality assessment) and yet offers high performance. Equation 2 shows significant correlation (R^2 of 0.9983 and $RMSE$ of 0.002923) to the observed experimental results in [6].

The standard utility function $\mathcal{U}_{res}(r)$ provides quality estimation based on the theoretically highest quality. In practice, content providers offer a finite set of representations in an adaptation set and the highest quality offered for the content (Q_{res}^{TOP}) is defined in a content manifest such as the MPD by the best representation ($r = r^{TOP}$) in the adaptation set. By adopting such a companion content descriptor provided by the content provider, we prevent the over-provisioning of resources beyond the capacity of HAS streams and therefore avoid HAS streams with higher top representations being penalized. To this end, we are able to further improve the video quality fairness measurement by incorporating both the modeling of user perception and the nature of media applications. We rescale $\mathcal{U}_{res}(r)$ so that the video quality reaches its maximum value 1 when the best representation is active (i.e., $Q_{res}^{TOP} = \mathcal{U}'_{res}(r^{TOP}) = 1$). Hence the adjusted video quality utility function is:

$$\mathcal{U}'_{res}(r) = \frac{\mathcal{U}_{res}(r)}{\mathcal{U}_{res}(r^{TOP})} \tag{3}$$

In practice, the maximum feasible quality of a stream is also limited by the network capacity at the user device. The network capacity then determines the highest bitrate feasible (r^{MAX}) for the corresponding media stream. For instance, a user may have only 2 Mbit/s network capacity using Wifi networks in the garden, though she subscribes to 50 Mbit/s broadband network over DSL. Hence, it is not necessary to provision more than 2 Mbit/s of network resource on the shared access network for this user. Network capacity is often determined by the link capacity and any background traffic on the same link. The video quality utility is therefore tuned to reflect the network resource constraint at a user device:

$$\mathcal{U}'_{res}(r) = \frac{\mathcal{U}_{res}(r)}{\mathcal{U}_{res}(r^{MAX})},\ if\ r^{MAX} < r^{TOP} \tag{4}$$

Using such video quality utility functions, we can then divide network resources between HAS media streams in a way that minimizes any discrepancy between the delivered video quality on all HAS streams, hence achieving the video quality fairness.

$$Q_1 = Q_2 = ... = Q_N \tag{5}$$

$$\mathcal{U}'_{res_1}(r_1) = \mathcal{U}'_{res_2}(r_2) = ... = \mathcal{U}'_{res_N}(r_N), \quad \text{with } r_1 + r_2 + ... + r_N = B \tag{6}$$

Overall, the optimal video quality fairness of a HAS media stream can be achieved mathematically using Equation 6 as influenced by the available bandwidth and the adaptation set given by the content provider. The fairness between media streams can be measured using Relative Standard Deviation (RSD) (Equation 8), obtained by multiplying the standard deviation s by 100 and dividing this product by the mean $\bar{Q}$. RSD captures not only the deviation but also the scale of the video quality difference. A small RSD means less difference between video quality perceived over related HAS streams, which also suggests better fairness at a user level.

$$s_{VQ} = \sqrt{\frac{1}{M-1}\sum_{j=1}^{M}(Q_j - \bar{Q})^2} \tag{7}$$

$$\Im^{VQ} = s_{VQ-RSD} = 100 \times \frac{s_{VQ}}{\bar{Q}} \tag{8}$$

B. Switching impact fairness

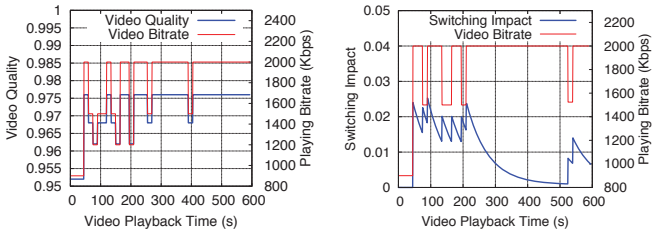
HAS media streams have the capability of switching between representations as the means to adapt to the available network resource. The purpose of switching can be increasing the bitrate to improve video quality or reducing the bitrate to avoid buffer underrun and playback stalling. However, the switching process itself may cause disturbance to the end user. The impact of quality switches (*SI*) is influenced by the *amplitude* and the *distribution* of switching events [5]. The amplitude is determined by the perception of video quality changes between representations. We define such quality change as $\Delta_{VQ} = |Q - Q'|$ with Q' as the projected video quality after the representation switch. A higher change of video quality leads to more severe perceptual impact at the time of switch. In a related work, Liu et al. observed that the impact caused by "increasing switch" is much smaller than "decreasing switch" of the same scale [15]. The modeling of this advanced feature requires further subjective experiments, which will be carried out in our future work. A crucial aspect when modeling the HAS switching impact is the *forgiveness effect*, which captures the psychological observations that, when followed by intact content, the impact of quality distortion degrades over time. The *forgiveness effect* related to video quality degradation was first studied and modeled by Seferidis et al. [21] and Hands [8]. One of the key findings from the user ratings is that the impact of quality distortion is reduced to nearly 70% after 20 seconds. We incorporate the *forgiveness effect* (Equation 9) in our model based on the generalized model introduced by Liu et al [16]. Equation 9 is a function of intensity of quality changes (Δ_{VQ}) and the duration of time since a switching event ($t - t_i$).

$$SI_i(t) = (\Delta_{VQ})\mathrm{e}^{-0.015(t-t_i)}, \quad t_i \text{ is the time of the quality switch } i \tag{9}$$

Using Equation 9, the initial switching impact will eventually diminish to a negligible value when $t - t_i$ is sufficiently large. We consider the QoE as the overall acceptability of a video session as perceived by human user. Therefore, we define 10% of initial switch impact as a residual influence that lasts for the user's entire viewing session. This means that the residual impact from multiple switching events will accumulate till the end of the viewing session. The impact function is updated as:

$$SI_i(t) = max((\Delta_{VQ})\mathrm{e}^{-0.015(t-t_i)}, 0.1\Delta_{VQ}) \quad t_i \text{ is the time of the quality switch } i \tag{10}$$

Figure 3(a) shows the video quality (VQ) for playing the test video with options to switch between different video bitrates in 720p video resolution. The figure demonstrates the non-linear mapping between the video bit-rate and the video quality. For instance, a switch between two very high bitrates shows less impact on the video quality compared with the same amount of bitrate change between representations of lower bitrates. Such QoE measurements are valuable for both single-stream quality optimization and QoE fairness between media streams. Switching impact accounts for the frequency and distribution of changes over the playing time. As demonstrated in Figure 3(b), high switching impact can be caused by high video quality variation or small, but temporally close, changes.



(a) Video quality (b) Switching impact

Fig. 3. Impact of HAS adaptation

Through switching impact measurement, we can evaluate a resource allocation solution or compare different solutions using the following switching impact fairness function based on RSD:

$$s_{SI} = \sqrt{\frac{1}{M-1}\sum_{j=1}^{M}(SI_j - \bar{SI})^2} \tag{11}$$

$$\Im^{SI} = s_{SI-RSD} = 100 \times \frac{s}{\bar{SI}} \tag{12}$$

During network fluctuations representation switching is inevitable on one or multiple HAS streams. Switching impact fairness captures the impact of switches throughout the entire life-cycle of a HAS stream, and balances such impact between related HAS streams. As an example, a relatively high RSD measure on switching impact suggests that one or more HAS streams have experienced more frequent or severe quality adaptation between representations. Using the SI fairness metric,

the UF model would be able to mitigate such imbalance and potentially protect the playback bitrate of certain HAS streams from further variations.

C. Cost efficiency fairness

Content consumers, especially those having invested in high-definition TV and broadband internet connections, expect on-demand movies or live football matches to be delivered at the highest possible quality. Distributing video content in high definition, high framerate, and high color depth with the companion multi-channel high quality audio tracks is only possible with a high degree of guaranteed network bandwidth. Such requirements place great challenges on network operators, particularly during prime time when a large amount of concurrent video streams must be supported by shared network resources. High throughput video streams can also overwhelm "vulnerable" segments of delivery networks and deteriorate packet delivery of other applications. It is in network operators' interests to assure a high degree of user satisfaction on HAS video streams whilst moderating any excessive utilization of network resources.

We define cost efficiency *CT* as a metric to capture the notion of fairness between content consumers and network operators. CT of a target network segment (usually a bottleneck) is quantified as *the required (or consumed) bandwidth per unit of total targeted (or delivered) video quality*. A high CT denotes low cost efficiency as it requires more bandwidth to deliver a unit of video quality. Given the bitrate of selected representations of related video streams and their adjusted utility functions $\mathcal{U}'$, CT can be quantified using Equation 13. Unlike video quality fairness and switching impact fairness, which capture the level of deviation in a quality metric between media streams, the cost-efficiency fairness is evaluated based on all related HAS media streams as a whole over the measured network segment.

$$\Im^{CT} = \frac{\sum_{i}^{N} r_i}{\sum_{i}^{N} U'_{res_i}(r_i)} \tag{13}$$

It is also possible to determine the most (theoretically) cost-effective bandwidth-provisioning solution(s) using Lagrange multipliers to find the minimal value(s) of Equation 13 subject to the constraint $\sum_{i}^{N} r_i \leq B$. However, a fairness model built entirely based on CT would most likely favor bitrates towards the lower end of the chart due to the nature of utility curves (Figure 2). Therefore, CT fairness should be in principle exploited in balance with at least a complementary metric such as video quality.

D. Fairness-aware resource allocation

Using video-quality fairness, switching-impact fairness, and cost-efficiency fairness as the user-level metrics, a QoE service can dynamically program specific segments of a network using platforms like SDN so that network resources can be provisioned fairly with respect to the user perception of video content, and cost efficiency of the network, to deliver good user experience. Incorporating the fairness metrics in production networks either as a network service or a QoE middleware poses a number of challenges.

Firstly, the adaptation sets of HAS streams comprise discrete and finite representations, hence the optimal solution to share available bandwidth between media streams cannot be derived directly from any continuous utility functions. Ultimately a decision is made from the many combinations of representations of each media streams. For the case that N HAS streams, each with M representations to adapt to, are present for bandwidth sharing, N^M combinations are available per fairness metric. This leads to the second challenge of computational complexity. Taking HAS streams of 10 representations per resolution as an example, the presence of 4 streams results in upto $10,000$ potential solutions for bandwidth sharing using a single fairness metric. Every new stream joining the network would increase the number of combinations tenfold, and the UF model is to be queried to provide a new provisioning solution whenever there is a major change (e.g., a new client joining) in the network. When multiple QoE metrics are incorporated, the complexity of the fairness model will also increase accordingly. For stateless metrics such as VQ and CT, which are independent from past status (e.g., bitrate), it is possible to employ techniques such as dynamic programming to improve runtime performance [13]. Stateful metrics like SI require historical information related to quality switches from the start of the video session, hence it is more difficult to reduce their runtime complexity.

In order to improve the feasibility of the UF model for live resource allocation optimization, we approach this challenge with an optimization method of three internal stages. At the first stage, the framework uses the continuous VQ utility functions to derive the theoretically optimal sharing of bandwidth which ensures that an identical degree of video quality is delivered on all HAS streams. The process also maximize the utilization of total available bandwidth to share. This is done by solving the following equation of adjusted utility functions:

$$\begin{aligned} \mathcal{U}'_{res_1}(r_1) = &\mathcal{U}'_{res_2}(r_2) = ... = \mathcal{U}'_{res_N}(r_N), \\ &with\ r_1 + r_2 + ... + r_N = BW \end{aligned} \tag{14}$$

Because VQ utility functions are monotonically increasing, Equation 14 gives at most one set of results: $\hat{R} = [\hat{r_1}, \hat{r_2}, ..., \hat{r_N}]$.

The second stage takes the optimal solution given by the first stage as the starting point, and conducts a bi-directional search of nearest representations (defined in MPD) of every optimal bitrate in $\hat{R}$. The search returns one or two playback rates for each $\hat{r}$:$[[r_1^l, r_1^h], [r_2^l, r_2^h], ..., [r_N^l, r_N^h]]$. $[r_i^l, r_i^h]$ are the bitrates of representations that best approximate the optimal rate of $\hat{r_i}$, with $r_i^l \leq \hat{r_i}$ and $r_i^h \geq \hat{r_i}$. In the cases that $\hat{r_i}$ is higher than the highest representations or lower than the lowest representations, only r_i^l or r_i^h will be available. The searching stage serves as a screening process that greatly reduces the complexity of resource allocation between N streams with M

levels of representations from M^N (exhaustive search) to a much more manageable candidate list C of a maximum of 2^N items.

The last stage of the optimization process evaluates the candidate list C using three fairness metrics: VQ, SI and CT, and identifies the candidate that achieves the best balance of fairness between all three metrics. The process begins by calculating video quality fairness $\Im_c^{VQ}$, switching impact fairness $\Im_c^{SI}$, and cost fairness $\Im_c^{CT}$ of all c in candidate list C. We then continue with a pooling process by combining all three measurements and deriving an overall rating for each c. Because the fairness metrics are in different scales, we rescale the fairness measurements using the maximum value of the same metric as the rescaling factor. For instance:

$$\ddot{\Im}_c^{VQ} = \frac{\Im_c^{VQ}}{max(\Im_C^{VQ})} \tag{15}$$

The rescaled fairness measurement $\ddot{\Im}_c$ has the scale of $[0, 1]$. Because a higher value in our fairness metrics denote lower fairness, $\ddot{\Im}_c = 1$ represents the worst solution from all candidates with respect to a given fairness. Any value between 0 and 1 quantifies the level of improvement a solution c achieves on a fairness metric compared with the worst solution. Using the rescaled fairness measurements, we then combine the three fairness measurements using the weighted-sum method:

$$\begin{aligned} \ddot{\Im}_c^{combined} =& w_c^{VQ} * \ddot{\Im}_c^{VQ} + w_c^{SI} * \ddot{\Im}_c^{SI} + w_c^{CT} * \ddot{\Im}_c^{CT} \\ & with\ w_c^{VQ} + w_c^{SI} + w_c^{CT} = 1 \end{aligned} \tag{16}$$

w_c is the weight coefficient for each fairness metric and it defines how fairness of video quality, switching impact and cost is balanced. We define the UF model by reaching an equal balance between three fairness metrics (i.e., $w_c^{VQ} = w_c^{SI} = w_c^{CT}$). This standard configuration help us investigate the impact of each fairness metric to the overall resource allocation solution. In practice, a QoE management framework may adopt a different balance between video quality fairness, switching impact fairness and cost fairness according to specific preferences. The candidate solution c which exhibits the minimum value of combined fairness $\ddot{\Im}_c^{combined}$ is considered to be the best option to achieve the overall user-level fairness.

IV. Experiments

In order to assess the effectiveness of the UF model under different network conditions, we use a purpose-built evaluation testbed (Figure 4). Using profiles specified by a tester (e.g., number of clients, frequency of network fluctuations, client link capacities, etc.), the test scripter function generates randomized network events for test manifests. A testbed function parses any given test manifest and simulates client arrival/departure and network fluctuation accordingly. The resource-allocation function encapsulates a number of APIs which allow the testbed to specify network status and metadata of HAS streams, and acquire solutions to optimize resource allocation between all relevant media streams. Session logs, which capture time-coded representation changes for all media streams, are also maintained for stateful metrics such as SI fairness. In order to study the characteristics of each fairness metric in achieving user-level fairness, we employed three additional fairness models, each exclusively uses one of the three fairness metrics (VQ, SI, and CT) to direct resource allocation. We also incorporate a baseline model, which resembles how network resources are provisioned through transport protocols without the help of an overarching orchestration framework. Together with the UF model, there are five fairness models (i.e., VQ, SI, CT, UF, and baseline) to be studied in the experiment. The testbed is therefore designed to carry out any experiment with five independent threads, each hosting one of the five fairness models. As a result, we are able to comparatively study the results given by each model based on identical test conditions.

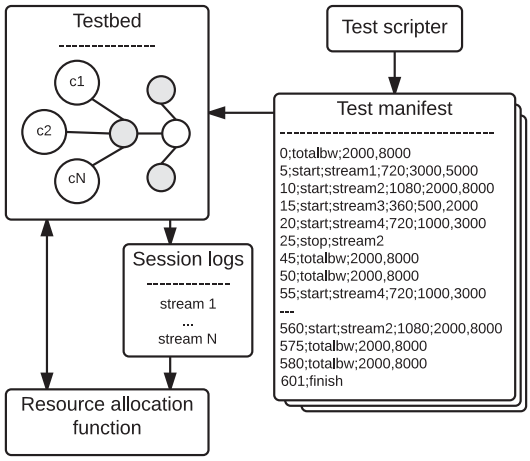


Fig. 4. Evaluation testbed

Tests are specified using a generic tree structure. A number of clients (c_1,...,c_N) connect to an aggregation node via corresponding links (l_1,...,l_N) and share the resource of an access network. One HAS stream is delivered to each of the clients. The first test is defined with the duration of 10 minutes and four clients (streams). The total available bandwidth accessible to video streams on the access network randomly fluctuates between 2Mb/s to 8 Mb/s as influenced by background traffic. The resolution of video streams delivered to the clients is randomly selected between 360p, 720p, or 1080p. The bitrates for each resolution are given in Table I. The available bandwidth on each client link also changes randomly between 500kb/s to 8Mb/s during the course of the test to simulate change of link capacity (such as in wireless networks) or background traffic. The total available access network bandwidth is shared between all video streams with respect to the resource available on client links.

For the first test, we compare how the baseline model and UF fairness model provision network resource differently in a dynamic environment. Figure 5(a) and Figure 5(b) show how network resource is shared between four video streams in the first 65 seconds of the test as instructed by the two different

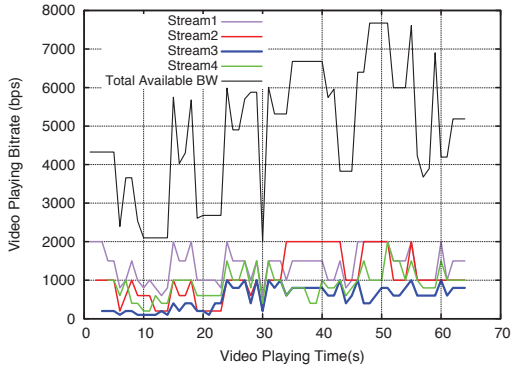
TABLE I. SET BITRATES FOR THREE VIDEO RESOLUTIONS

Resolution	Video Bitrate (kbps)
1080p	100, 200, 600, 1000, 2000, 4000, 6000, 8000
720p	100, 200, 400, 600, 800, 1000, 1500, 2000
360p	100, 200, 400, 600, 800, 1000

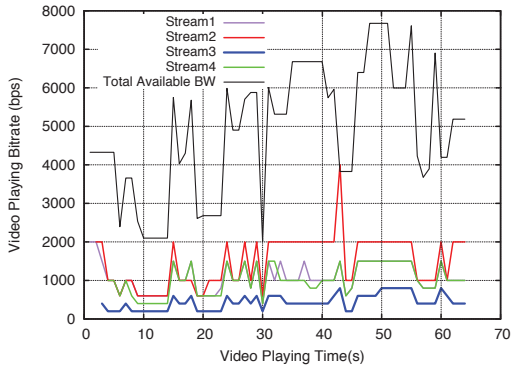
models. The resultant video quality of each video stream is given in Figure 6(a) and Figure 6(b) for baseline model and UF model respectively. The results clearly demonstrate the significant difference of the network provisioning strategy adopted by the user-level model compared with the conventional network-level baseline model. The baseline model allows video streams with more intensive requests at the transport layer to acquire more resources, leading to some video streams being heavily penalized (Figure 5(a)). Using the first 20 seconds of the test as an example, stream2, stream3 and stream4 all suffered from low video quality and severe quality fluctuation while the quality of stream1 remains high through the entire test (Figure 6(a)). This example demonstrates the gap between network-level and user-level fairness. Using the bespoke UF model, which takes advantage of three fairness metrics, the network-management element in the testbed is able to schedule the resource according to the QoE requirements and link status of every HAS stream (Figure 5(b)). As a result, network resources are dynamically provisioned in a way that similar video quality is maintained on all related media streams for the entire course of the experiment (Figure 6(b)). Furthermore, the UF model also avoided any severe video quality fluctuation thanks to its incorporation of switching-impact fairness.

In order to further investigate the performance of the UF fairness model and specifically how each individual fairness metric contributes to the user-level fairness, we defined a test manifest similar to the first test and enabled all five fairness models. Test manifests are defined with respect to a test scenario such as "busy wireless home network with a DSL broadband connection". It specifies the number of HAS streams, and the overall frequency at which the total shared bandwidth and the network capacity at user devices fluctuate. The exact timing and scale of the dynamics are purposely randomized and will only be instantiated at run time. Therefore every iteration of the test will generate a unique test configuration of a predefined scenario and hence test results. This also allows us to evaluate the consistency of the model. Exploiting such a feature of the testbed, we repeated the test 50 times. Figure 7(a) compares how five models perform in terms of video quality fairness. It reflects our previous observations in Figure 5 that the UF model significantly outperforms the baseline model (a lower value in fairness metric denotes better fairness). Between the VQ, SI and CT models, VQ (whose objective is to maximize the video-quality fairness exclusively without considering other fairness metrics) yields the best results, unsurprisingly. The SI and CT models compromise on video-quality fairness to balance switching impact or cost fairness but still greatly outperform the baseline model.

The evaluation based on switching impact fairness is shown in Figure 7(b). Similar to the conclusions in Figure 7(a), all user-level fairness models achieve better performance than the



(a) Resource allocation by baseline model



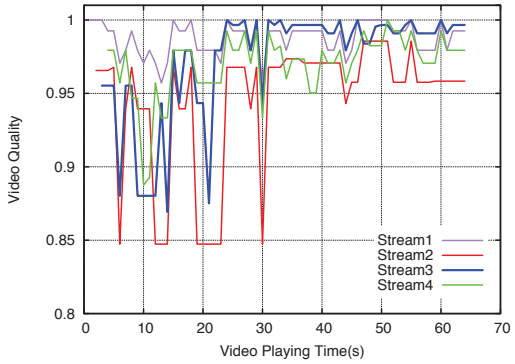
(b) Resource allocation by UF model

Fig. 5. Resource allocation of UF and baseline model

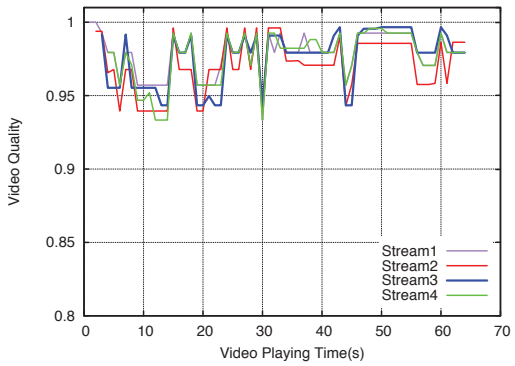
baseline model, while we can maximize the switching impact fairness using the SI model. The results delivered by the VQ and CT models are between SI and baseline. The results on cost-efficiency fairness are slightly different (Figure 7(c)) to the other two. In this case, the baseline model exhibits better performance in delivering cost-efficiency fairness compared with VQ, SI and UF, and is only beaten by the CT model. This is due to the fact that gaining a unit of video quality is easier when the bitrate of video is low according to the video quality utility function which resembles the law of diminishing returns. With more streams in the lower-bitrate and lower-quality ranges, the baseline model can be more cost effective in terms of consumed bitrate per unit of delivered quality, though the delivered video quality is still much lower than UF and other models as demonstrated in Figure 6(a).

Overall, video-quality fairness, switching-impact fairness and cost-efficiency fairness all exhibit their distinctive benefits to the overall user-level fairness. A model achieving the best on one fairness metric usually shows sub-optimal performance on the other two fairness metrics. Experimental results suggest that by combining the fairness models, it is possible to achieve a good balance on all fairness measurements.

In practice, a shared network can be very quiet or extremely busy. To study the consistency of the UF model during network fluctuations to various degrees, we specified new test manifests by manipulating the probability of a bandwidth change using

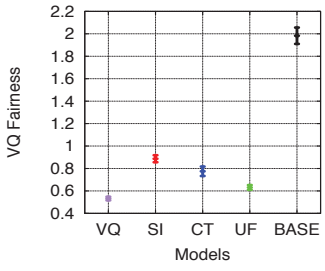


(a) Video quality delivered by baseline model

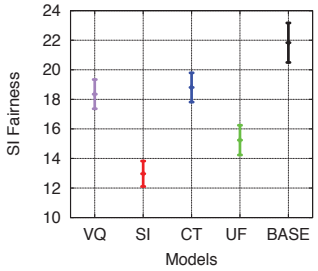


(b) Video quality delivered by UF model

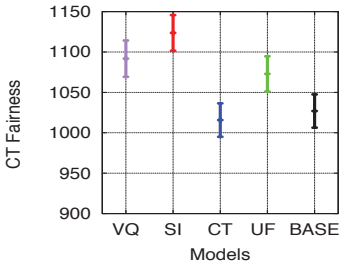
Fig. 6. Resultant video quality of UF and baseline model



(a) Video quality fairness measurement



(b) Switching impact fairness measurement

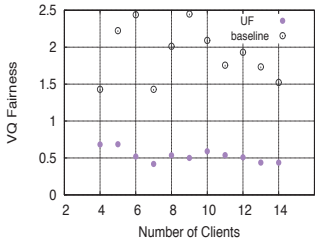


(c) Cost efficiency fairness measurement

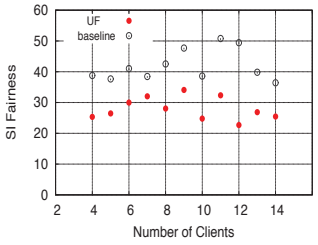
Fig. 7. Fairness measurements of resource allocation models

the test scripter. We generated a total of 400 10-minute-long tests with the number of bandwidth changes varying from around 20 (one change to the shared bandwidth every 30 seconds) to 120 (one change to the shared bandwidth every 5 seconds). Every change of the shared bandwidth leads to a reallocation process instructed by the resource-allocation function. Again, we use both the UF and baseline models for a comparative analysis. The increasing number of quality fluctuations is believed to have an impact on the UF model especially through its stateful SI metric, where every change of video quality is accounted for in the user experience.

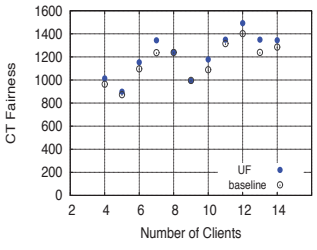
Figure 8 compares how the UF and baseline models deliver user experience on media streams using the three fairness measurements for networks of different characteristics. Each point on the figure projects the mean value of the corresponding fairness measure of the entire test. The UF model achieves its design objectives of delivering a good level of video quality fairness and switching impact fairness whilst maintaining the cost efficiency compared with the baseline model. The SI fairness measurements are more scattered between tests of fewer bandwidth changes than the tests of frequent bandwidth changes (Figure 8(b)). This is due to the fact that switching impact is a stateful metric that recognizes the dependency between consecutive changes of playback bitrate. Hence, tests with larger numbers of quality switches are more likely to statistically capture the performance of the a model on SI fairness. Furthermore, as defined in Equation 16, the UF model may be configured to balance between the three fairness metrics equally (as for the experiment), or to favour certain metric(s) with respect to a particular service strategy. For instance, a service provider may allow a level of discrepancy on video quality whilst giving more priority to maximizing the cost efficiency and minimizing the switching impact.



(a) Video quality fairness measurement



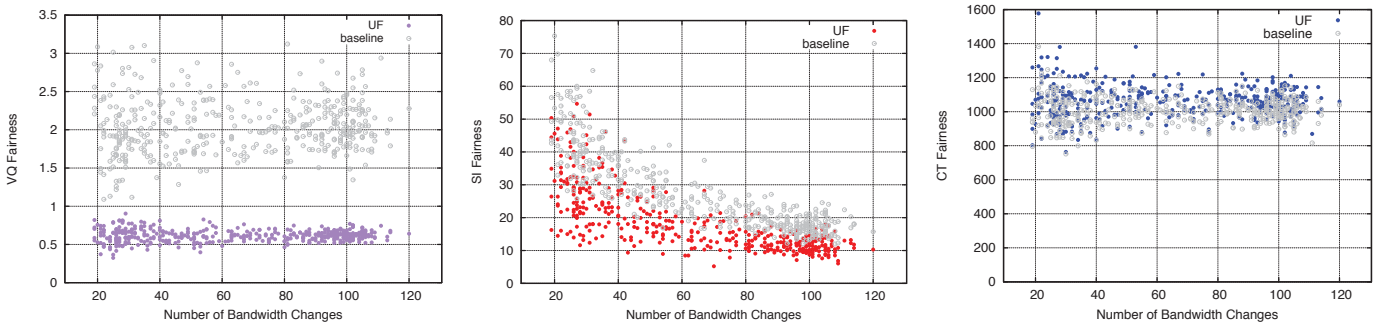
(b) Switching impact fairness measurement



(c) Cost efficiency fairness measurement

Fig. 9. Performance of UF model influenced by the number of clients

One important performance index of a resource allocation



(a) Video quality fairness measurement (b) Switching impact fairness measurement (c) Cost efficiency fairness measurement

Fig. 8. Performance of UF model influenced by the number of bandwidth fluctuations

algorithm is its scalability. We continue the experiments using test manifests that allows different numbers of user clients (media streams), ranging from 4 (used by previous tests) to 14. The total bandwidth is configured to be fluctuating between 2Mbps to 8Mbps for around 35 times during the course of the 10-minute-long test. This set-up helps us to investigate whether the UF model can perform with the same level of user-level fairness when more clients join the shared network and share the same pool of network resources. Figure 9 suggests that increasing the number of clients does not significantly impact the output of the UF model. There seems to be a trend of CT fairness reduction when the number of clients increases beyond 10. However, the fairness measurements still stay within the data range observed in Figure 8(c) where the number of clients is 4.

V. Discussions

In Section IV, we investigated the effectiveness of the UF model in achieving user-level fairness using a tree-like network topology. The UF model, designed to be topology-agnostic, can be exploited for different use scenarios where resources are shared between multiple entities. For more complex network structures, the UF model may be applied recursively or in a level of abstraction (e.g., resource provisioning between two sub-networks).

The goal of the UF model is to look beyond network-level metrics and maximize the fairness of resource allocation at a user level. In other words, the model keeps end users *equally happy* by provisioning network resources according to application and user requirements. We consider the UF model as our first step towards user-level fairness. There are still a number of challenges to be addressed. For instance, the ultimate user-level fairness may come at the cost of sacrificing the QoE of some HAS streams (such as stream1 in Figure 6(a) and Figure 6(b)). It is worth studying whether certain streams are overly penalized. In the worst case, achieving fairness may also result in all users being *equally unhappy*. Therefore, a balance has to be reached between user experience on individual HAS streams and the overall fairness between users.

VI. Conclusions

HAS is becoming a popular vehicle for online video delivery. The adaptiveness of HAS maximizes the utilization of network resources for better video quality and ensures smooth playback during network fluctuations. However, HAS applications often work independently without coordination between each other in the same network. This leads to QoE fluctuations and unfairness between end users. This paper introduces a UF model, which facilitates the orchestration of network resource allocation to achieve user-level fairness. The UF model incorporates video quality, switching impact, and cost efficiency as the fairness metrics. The performance and scalability of the model is evaluated through a number of experiments. Future work will look into employing the UF model and piloting in-network and transparent QoE management using technologies such as software defined networking.

Acknowledgement

The work presented is supported by the European Commission within FP7 project FI-Content2 (grant 603662) and by the UK EPSRC project TOUCAN (grant EP/L020009/1).

References

[1] S. Akhshabi, A. Begen, and C. Dovrolis. An Experimental Evaluation of Rate-adaptation Algorithms in Adaptive Streaming over HTTP. In *Proc. 2nd annual ACM Conference on Multimedia Systems*, MMSys '11, pages 157–168, 2011.

[2] G. Cermak, M. Pinson, and S. Wolf. The relationship among video quality, screen resolution, and bit rate. *Broadcasting, IEEE Transactions on*, 57(2):258–262, 2011.

[3] Cisco. Cisco visual networking index: Forecast and methodology, 20132018. *http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.pdf*, 2014.

[4] A. Farshad, P. Georgopoulos, M. Broadbent, M. Mu, and N. Race. Leveraging sdn to provide an in-network qoe measurement framework. In *IEEE INFOCOM 2015 Workshop on Communication & Networking Techniques for Contemporary Video*. IEEE, 2015.

[5] M.-N. Garcia, F. D. Simone, S. Tavakoli, N. Staelens, S. Egger, K. Brunnstrm, and A. Raake. Quality of experience and http adaptive streaming: a review of subjective studies. *Proceedings of the 6th International Workshop on Quality of Multimedia Experience (QoMEX)*, 2014.

[6] P. Georgopoulos, Y. Elkhatib, M. Broadbent, M. Mu, and N. Race. Towards network-wide qoe fairness using openflow-assisted adaptive video streaming. In *Proceedings of the 2013 ACM SIGCOMM workshop on Future human-centric multimedia networking*, pages 15–20. ACM, 2013.

[7] J. Gettys and K. Nichols. Bufferbloat: Dark Buffers in the Internet. *ACM Queue*, 9(11):40–54, Nov. 2011.

[8] D. Hands. Temporal characterization of forgiveness effect. *Electronics Letters*, 37, 2002.

[9] T.-Y. Huang, N. Handigol, B. Heller, N. McKeown, and R. Johari. Confused, Rimid, and Unstable: Picking a Video Streaming Rate is Hard. In *Proc. ACM IMC*, pages 225–238, 2012.

[10] T.-Y. Huang, R. Johari, N. McKeown, M. Trunnell, and M. Watson. A buffer-based approach to rate adaptation: Evidence from a large video streaming service. In *Proceedings of the 2014 ACM conference on SIGCOMM*, pages 187–198. ACM, 2014.

[11] J. Jiang, V. Sekar, and H. Zhang. Improving Fairness, Efficiency, and Stability in HTTP-based Adaptive Video Streaming with FESTIVE. In *Proc. ACM CoNEXT*, pages 97–108, 2012.

[12] F. P. Kelly, A. K. Maulloo, and D. K. Tan. Rate control for communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research society*, pages 237–252, 1998.

[13] Z. Li, A. C. Begen, J. Gahm, Y. Shan, B. Osler, and D. Oran. Streaming video over http with consistent quality. In *Proceedings of the 5th ACM Multimedia Systems Conference*, pages 248–258. ACM, 2014.

[14] Z. Li, X. Zhu, J. Gahm, R. Pan, H. Hu, A. Begen, and D. Oran. Probe and adapt: Rate adaptation for http video streaming at scale. 32(4):719–733, 2014.

[15] Y. Liu, S. Dey, D. Gillies, F. Ulupinar, and M. Luby. User Experience Modeling for DASH Video. In *Packet Video Workshop (PV), 2013 20th International*, pages 1–8. IEEE, 2013.

[16] Z. Liu, Y. Shen, K. W. Ross, S. S. Panwar, and Y. Wang. Layerp2p: Using layered video chunks in p2p live streaming. *IEEE Transactions on Multimedia*, 11(7):1340–1352, 2009.

[17] S. H. Low and D. E. Lapsley. Optimization flow controli: basic algorithm and convergence. *IEEE/ACM Transactions on Networking (TON)*, 7(6):861–874, 1999.

[18] A. Mansy, B. Ver Steeg, and M. Ammar. SABRE: A Client based Technique for Mitigating the Buffer Bloat Effect of Adaptive Video Flows. In *Proc. 3rd annual ACM Conference on Multimedia Systems*, MMSys '12. ACM, 2012.

[19] R. K. Mok, X. Luo, E. W. Chan, and R. K. Chang. Qdash: a qoe-aware dash system. In *Proceedings of the 3rd Multimedia Systems Conference*, pages 11–22. ACM, 2012.

[20] D. O'Neill, E. Akuiyibo, S. Boyd, and A. J. Goldsmith. Optimizing adaptive modulation in wireless networks via multi-period network utility maximization. In *Communications (ICC), 2010 IEEE International Conference on*, pages 1–5. IEEE, 2010.

[21] V. Seferidis, M. Ghanbari, and D. Pearson. Forgiveness effect in subjective assessment of packet video. *Electronics Letters*, 28(21):2013–2014, 1992.

[22] G. Tian and Y. Liu. Towards agile and smooth video adaptation in dynamic http streaming. In *Proceedings of the 8th international conference on Emerging networking experiments and technologies*, pages 109–120. ACM, 2012.