SVC-based Multi-user Streamloading for Wireless Networks

S. Amir Hosseini[1], Zheng Lu[2], Gustavo de Veciana[2], and Shivendra S. Panwar[1]

[1]Department of Electrical and Computer Engineering, NYU Tandon School of Engineering
[2]Department of Electrical and Computer Engineering, University of Texas at Austin

***Abstract*—In this paper, we present an approach for joint rate allocation and quality selection for a novel video streaming scheme called *streamloading*. Streamloading is a recently developed method for delivering high quality video without violating copyright enforced restrictions on content access for video streaming. In regular streaming services, content providers restrict the amount of viewable video that users can download prior to playback. This approach can cause inferior user experience due to bandwidth variations, especially in mobile networks with varying capacity. In streamloading, the video is encoded using Scalable Video Coding, and users are allowed to pre-fetch enhancement layers and store them on the device, while base layers are streamed in a near real-time fashion ensuring that buffering constraints on viewable content are met.**

We begin by formulating the offline problem of jointly optimizing rate allocation and quality selection for streamloading in a wireless network. This motivates our proposed online algorithms for joint scheduling at the base station and segment quality selection at receivers. The results indicate that streamloading outperforms state-of-the-art streaming schemes in terms of the number of additional streams we can admit for a given video quality. Furthermore, the quality adaptation mechanism of our proposed algorithm achieves a higher performance than baseline algorithms with no (or limited) video-centric optimization of the base station's allocation of resources, e.g., proportional fairness.

***Index Terms*—Streamloading, scalable video coding, rate allocation, quality selection**

I. Introduction

Mobile video streaming services continue to gain popularity among cellular data users. Currently, video traffic has the largest share of cellular data (55% at the end of 2014), and this trend is predicted to continue growing [1]. In order to efficiently meet this demand with the limited bandwidth resources of wireless networks, the use of video quality adaptation has gained enormous interest in industry.

Adaptive video transmission over HTTP has been standardized under the commercial name DASH [2], where the video is divided into segments, and multiple versions of each segment are encoded at different bit rates. When a segment is to be downloaded for viewing, a decision is made based on the conditions in the network, or on the state of the receiver download buffer, regarding which segment representation to retrieve. Other video delivery systems use scalable video coding (SVC), an extension of the H.264/AVC standard. In SVC, rather than encoding each segment into multiple versions with different bit rates, the video segments are encoded into several streams called layers. The base layer may be encoded as a low-quality video, while additional enhancement layers provide incremental improvements in quality. This delivery scheme offers additional flexibility over DASH, and opens up new options to improve video delivery and network efficiency.

In the context of copyrighted video streaming, content owners tend to provide *conditional access* to users in order to tightly control the content being watched, prevent illegal distribution of content, implement smart content pricing mechanisms, etc. One of the most widely used conditional access schemes is to limit the amount of viewable video that can be pre-fetched and stored on the end user device ahead of the playback. This limit is specified in the license agreements between content owner and content distributor, and varies from tens of seconds to a few minutes.

Based on this, we distinguish between two service models for video delivery in wireless networks. We refer to *streaming* as a service model where only a limited number of video segments can be delivered to the user ahead of playback. Streaming services are usually inexpensive and content providers tend to monetize the service by injecting advertisements during playback or through subscription models. Due to the limited buffering, streaming may suffer from degraded quality of service under varying channel conditions. Apart from this, streaming license agreements prohibit making copies or distribution of video content [3]. We refer to *downloading* as a service model in which the amount of buffered video is not limited. Unlike streaming, a persistent Internet connection is not necessary during playback and users can watch the downloaded video at any, possibly constrained, future time. Since users are allowed to store copies of purchased content on their devices [4], this service model is subject to additional licensing restrictions such as the duplication license, and is therefore offered at significantly higher prices, typically two orders of magnitude higher than streaming. It should be noted that even if the downloaded content is encrypted, it falls under this category. Despite the higher price, downloading offers higher video quality to the user and also improves the efficiency of the data transmission, since there are no buffering constraints.

Recently, a hybrid service model for video delivery was proposed called *streamloading* [5]. In this scheme, the video is encoded into several layers using SVC, and the base layer is streamed in real-time with limited buffering at the end user device; while enhancement layers may be downloaded ahead of time without buffering restrictions. Using this approach, the

video quality is improved because the receiver can take advantage of available excess bandwidth to download enhancement layers associated with future segments, thereby smoothing the effect of variations in link capacity. Consequently, users may enjoy video quality similar to that of a downloading service, while still being classified as a streaming service from the content providers point of view [6]. The latter stems from the fact that enhancement layers cannot be decoded and are therefore of no value, unless the respective base layer is available [7] [1].

The main contributions of this paper are as follows. We first formulate an optimization framework to study the joint base station rate allocations and segment quality selection problem for a multi-user setting. Leveraging previous work on optimizing DASH-based algorithms, we propose the first comprehensive solution for streamloading. Our results show that streamloading provides significant benefits, e.g., high video quality and low re-buffering time, suggesting that this service model has the potential of providing high Quality of Experience (QoE) while meeting the legal requirements of a streaming service.

The rest of the paper is organized as follows. Section II contains a summary of existing research on optimal adaptive video delivery in multi-user networks. In Section III, we provide a detailed description of the three service models mentioned above. The system characteristics, as well as the video quality model, are discussed in Section IV along with an offline optimization formulation for multi-user streamloading. Section V contains the proposed online RAte allocation and QUality sELection algorithm (RAQUEL), followed by a discussion of practical implications of our proposed scheme on real networks in Section VI. A thorough simulation analysis is presented in Section VII. Finally, Section VIII concludes the paper and briefly discusses potential future research opportunities.

II. Related Work

A great deal of research has focused on optimal resource allocation and quality adaptation for video delivery in wireless networks, see, e.g., [8]–[11] and references therein. A large portion of this research deals with the algorithms and performance of DASH-based video delivery. Bandwidth management for live streaming HTTP-based applications is studied in [9]. Several commercial adaptive video streaming services are compared in [10] in terms of bandwidth utilization, fairness, and bit rate stability. The authors conclude that all current services fail to satisfy one or more of these requirements, and claim that a randomized scheduling and state dependent rate adaptation approach outperforms currently used services. In [8], the authors consider the problem of optimal rate adaptation of DASH video transmission from multiple content distribution networks. In order to keep the bit rate stable, they propose to perform block level rate allocation, where multiple segments are grouped together and are transmitted at the same bit rate. The work in [11] formulates the problem of optimal delivery of DASH-based video to wireless users as a dynamic network utility (video quality) maximization problem with re-buffering and delivery cost constraints. Based on this formulation, they develop an online algorithm called NOVA, which they prove to be asymptotically optimal in stationary regimes.

As scalable video gains acceptance, particularly after its inclusion in the new H.265(HEVC) [12] and VP9 [13] codecs, more research work is being dedicated to optimizing SVC delivery, especially in wireless networks. Several works have investigated the benefits of using SVC over AVC in terms of caching efficiency and adaptation performance [14], as well as reduced congestion, especially at the video server end [15]. The problem of optimal rate allocation at the base station for SVC video streaming in a wireless multi-user scenario is investigated in [16]–[20]. In [19], the authors model the quality-rate trade-offs for SVC using a piece-wise linear function, and derive a rate allocation scheme for fading wireless channels. In a similar study, a multi-modal sigmoid approximation is used to model the quality-rate trade-off where a utility-proportional optimization flow control method is used to achieve convexity [18]. The mapping of SVC layers into DASH representations is studied in [17]. In the same work, an optimal scheme for rate allocation and quality stabilization is proposed for wireless Orthogonal Frequency Division Multiple Access (OFDMA) systems using the Lagrangian dual decomposition method. A heuristic segment selection approach is used in [16] to dynamically select quality layers solely based on buffer level, without considering the link bandwidth. In an attempt to decrease the re-buffering time of SVC delivery, [21] presents a priority scheme, in which the base layer segments are pre-fetched, and enhancement layers are subsequently downloaded in order to increase video quality. To our knowledge, no comprehensive solution for joint resource allocation and quality selection has been proposed for a multi-user setting, and in particular for streamloading. This is the gap we attempt to fill in this paper.

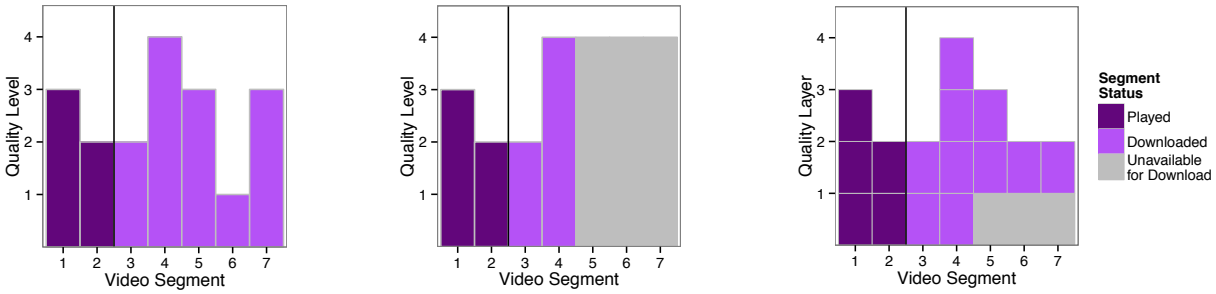
III. Service models

In this section, we define three different access schemes for wireless video delivery in terms of their associated restrictions on the amount of pre-fetched video content. A graphical illustration of these service models is shown in Figure 1.

Downloading: This model refers to the case where there is no restriction on the number of video segments that can be pre-fetched in advance. In this work, we consider a DASH-based delivery for this access scheme, in which the video segments are encoded at multiple quality levels. Upon delivery of each segment, the next segment is requested by the user. In Figure 1a, an example is shown for downloading a video sequence with four quality representations. It is important to note that here we are talking about short-term pre-fetching. It should not be confused with downloading videos overnight and viewing them later. Indeed, in such a scenario, the video could be delivered at the highest available quality.

Streaming: This service model differs from downloading in that no more than a certain number of segments can be

[1] Downloading encrypted video, which is not viewable without, for instance, an encrypted stream of keys, is still legally classified as downloading, and cannot be classified as a streaming service [6].



(a) Downloading **(b)** Streaming **(c)** Streamloading

Fig. 1: A sample snapshot of the end user buffer for downloading, streaming, and streamloading access models. The streaming and streamloading service models are shown with a buffer limit of two segments.

stored on the user device ahead of playback. We refer to this as a *buffer limit* from now on. Whenever this limit is reached, no further video segments can be received until some of the stored video is consumed. An example of streaming is shown in Figure 1b for a video sequence with four different representations and with a buffer limit of two segments.

Streamloading: As described in Section I, this service model uses SVC to encode each segment into multiple quality layers, and each layer can be transmitted independently. According to this scheme, base layer segments can only be pre-fetched up to a pre-defined buffer limit. However, the enhancement layer segments are not subject to such limitations. This scheme is illustrated in Figure 1c.

IV. Problem Formulation

In this section, we provide a detailed description of the multi-user streamloading problem and consider an offline optimization framework to explore what an ideal clairvoyant algorithm would aim to do. Table I shows the definition of all the notation used throughout the formulation.

A. System Model

For simplicity, we consider a wireless network consisting of a base station and a set of active mobile users $\mathcal{N}$, where $|\mathcal{N}| = N$. Time is assumed to be slotted with a slot duration of τ_{slot}. The users are viewing videos, each of which is divided into a sequence of segments of equal duration τ_{seg}. Each segment is encoded into one base layer and L enhancement layers. Our goal is to develop a scheme for joint base station scheduling and segment quality selection for "optimal" streamloading subject to a receiver base layer buffer limit of τ_{lim} seconds.

We let K denote an estimate for the number of time slots required to deliver the entire video to all users. In each time slot k, the base station allocates the rate $\mathbf{r}_k$ to all users, where $\mathbf{r}_k = (r_{i,k})_{i\in\mathcal{N}} \in \mathbb{R}_+^N$. The resource allocation is subject to time varying constraints determined by the link quality and achievable data rate. Therefore, the data rate that the base station can allocate to each user in time slot k is restricted to a possibly time varying convex rate region defined by $c_k(\mathbf{r}_k) \leq 0$, where c_k is assumed to be a real valued (continuous) convex function reflecting constraints on network resource allocation in slot k. We refer to this as the allocation constraint in slot k. This model encompasses a wide range of wireless systems [11].

TABLE I: Notation table

Variable	Definition
N	number of users
K	estimated number of total time slots
S	total number of video segments
L	total number of enhancement layers for each segment
τ_{slot}	length of time slot
τ_{seg}	length of video segment
τ_{lim}	length of buffer limit
$r_{i,k}$	rate allocated to user i in time slot k
$r^b_{i,k}$	rate allocated to user i in slot k to download base layer
$r^e_{i,k}$	rate allocated to user i in slot k to download enhancement layers
$x^b_{i,s,k}$	fraction of $r^b_{i,k}$ to download segment s
$x^e_{i,s,k}$	fraction of $r^e_{i,k}$ to download segment s
$q^l_{i,s}$	quality corresponding to delivering the first l enhancement layers for segment s to user i
β_i	estimated fraction of average re-buffering time to total download time for user i
$\overline{\beta}$	maximum value allowed for β_i $\forall i \in \mathcal{N}$
d_i	estimated time to download the entire video for user i
$c_k(\cdot)$	convex rate region in time slot k
$f_{i,s}(\cdot)$	quality-rate trade-off function for user i and segment s
$m^S_i(\cdot)$	average quality of entire video for user i
$v^S_i(\cdot)$	quality variation of entire video for user i
η	weight of variability in objective function
$\gamma_{i,k}$	number of time slots up to slot k that user i spent re-buffering
$S^b_{i,k}$	total number of base layers fully delivered to user i by slot k
$S^e_{i,k}$	total number of enhancement layers fully delivered to user i by slot k
$\mathcal{A}^b_{i,s,k}$	set of rates allocated to user i to fully download base layers by slot k
$\mathcal{A}^e_{i,s,k}$	set of rates allocated to user i to fully download enhancement layers by slot k

We denote by $q_{i,s}$, the perceived video quality achieved by user i for segment s. The quality of a segment in a scalable coded video increases with the number of successfully downloaded enhancement layers. Therefore, $q_{i,s}$ can take any value from the discrete set $\mathcal{Q}_{i,s} = \{q^0_{i,s}, \cdots, q^L_{i,s}\}$, where $q^l_{i,s}$ is the perceived quality obtained from delivering the first l

enhancement layers for segment s. The more enhancement layers the user downloads for a segment, the higher the required video data rate will be. We denote the average data rate associated with segment s downloaded by user i with quality level $q_{i,s}$ as $f_{i,s}(q_{i,s})$ in bits per second. Further, the segment data rate for a particular representation $f_{i,s}(.)$ is a convex function of the video quality, i.e., the relationship between quality and required video data is concave [22]. We refer to this convex function as the *quality-rate* trade-off. It should be noted that higher enhancement layers cannot be decoded if lower enhancement layers are not delivered.

B. Mathematical Formulation

The main objective of our streamloading problem is to maximize the overall video quality experienced by the users. We consider the average video quality along with the temporal quality variations as the key factors affecting the overall video quality. Therefore, the video quality of user i is calculated as $m_i^S(q_i) - \eta \mathrm{v}_i^S(q_i)$ where, $m_i^S(q_i) = \frac{\sum_{s=1}^{S} q_{i,s}}{S}$ represents the average video quality for user i after receiving the entire S segment long video, and $\mathrm{v}_i^S(q_i) = \sqrt{\frac{\sum_{s=1}^{S}(q_{i,s} - m_i^S(q_i))^2}{S}}$ is the quality variation for the same sequence of segments. The weight of quality variability on the overall video quality is determined by the constant η. A small value indicates that the quality depends less on the temporal variability and more on the average quality, and vice versa. The request of which quality representation to request next is sequentially made by the user. For each segment s, we define the quality vector $\mathbf{q}_i = (q_{i,s})_{s\in\mathcal{S}}$ as the sequence of requested quality levels by user i, where $\mathcal{S} = \{1, \cdots, S\}$. Hence, $\mathbf{q}_i$ is the decision variable of user i.

The resource allocation is performed at the base station and is subject to the channel capacity constraint as discussed in Section IV-A. Since in streamloading, base and enhancement layers can be scheduled and delivered separately, the base station scheduler has the capability to decide how to allocate rate not just among users, but also among base and enhancement layer segments of each user. Therefore, in order to differentiate the rates allocated to different layers, we denote the rate at slot k as $\mathbf{r}_k^b = (r_{i,k}^b)_{i\in\mathcal{N}}$ and $\mathbf{r}_k^e = (r_{i,k}^e)_{i\in\mathcal{N}}$ which correspond to the rate dedicated to base and enhancement layer segments, respectively. Hence, $\mathbf{r}_k^b$ and $\mathbf{r}_k^e$ are the decision variables of the base station at every time slot k.

Ideally, all users prefer to receive the full quality for all segments. However, because of the channel capacity constraint, increasing the load on the network causes delay in delivering video segments. This delay can result in *re-buffering*, which manifests itself to the user as a frozen video frame and causes major degradation to the QoE of video streaming. Therefore, quality selection mechanisms should limit re-buffering. In streamloading, since base and enhancement layer delivery is decoupled, the analysis of re-buffering is different than in single layered DASH as shown in Figure 2. In DASH video delivery, a segment is played back only if it has been fully delivered to the receiver. If the playback time reaches a segment which has not been completely delivered, re-buffering occurs and the playback stops until the delivery is complete as illustrated in Figure 2a. However, since in SVC, base layers can be decoded and played back with or without enhancement layers, the re-buffering time is solely determined by the time that the base layer buffer is empty. An example for re-buffering in streamloading is shown in Figure 2c. In order to limit the average re-buffering time of a client, we set an upper bound on the fraction of playback time that users experience re-buffering.

If we let d_i denote an estimate for the total time required to download S base layer segments for user i, we have:

$$d_i = \frac{\tau_{seg} \sum_{s=1}^{S} f_{i,s}(q_{i,s}^0)}{\frac{1}{K}\sum_{k=1}^{K} r_{i,k}^b}, \tag{1}$$

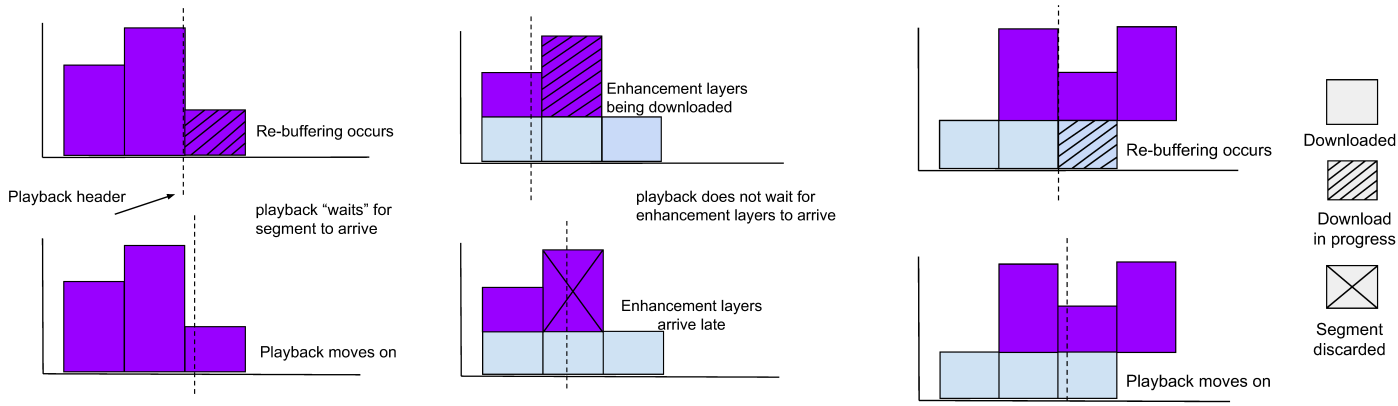
which is simply the total delivered base layer data over the average allocated rate to that user. For very large S in a stationary regime, the denominator in (1) gives an estimate for the average base layer download rate of user i over a long period. Consequently, one can estimate the fraction of time that user i is re-buffering, which we denote as β_i, as $\frac{d_i}{\tau_{seg} S} - 1$. We can rewrite β_i as follows:

$$\beta_i(\mathbf{q}_i, (r_{i,k}^b)_{k\in\mathcal{K}}) := \frac{\sum_{s=1}^{S} f_{i,s}(q_{i,s}^0)}{\frac{S}{K}\sum_{k=1}^{K} r_{i,k}^b} - 1, \tag{2}$$

where $(r_{i,k}^b)_{k\in\mathcal{K}}$ is the sequence of $r_{i,k}^b$ for all time slots and $\mathcal{K} = \{1, \cdots, K\}$. Since playback continuity depends only on the base layer segments, it is possible that enhancement layers are not fully delivered when the respective base layer is played. In other words, playback does not "wait" for enhancement layers to arrive, while it does so for base layers. In such scenarios, the waiting time for base layers is the re-buffering time. From now on, we refer to events where the enhancement layers arrive late as a *segment loss*, which is depicted in Figure 2b. Enhancement layers that are delivered after the base layer playback are discarded and do not contribute towards the aggregate video quality. Based on this discussion, any scheme designed for optimal streamloading should take into account both re-buffering and enhancement layer segment loss.

In addition to the constraints discussed above, there are other restrictions which further constrain the decision parameters for quality selection. For instance, users cannot request enhancement layers for segments that are already played back. Also, because of the base layer restrictions of streamloading, no base layer segments can be requested by users who have filled their buffer with base layer segments up to the buffer limit. All these constraints will be explained in detail throughout this section.

In order to formulate the streamloading problem mathematically, we present OPTSL, which incorporates all the discussed constraints in an offline optimization framework. In order to do that, we define a set of non-negative auxiliary variables $x_{i,s,k}^b$ and $x_{i,s,k}^e$, which indicate the respective fraction of $r_{i,k}^b$ and $r_{i,k}^e$ used to download base and enhancement layers of segment s in slot k, respectively. These auxiliary variables can be represented in vector form as $\mathbf{x}_{i,k}^b = (x_{i,s,k}^b)_{s\in\mathcal{S}}$ and $\mathbf{x}_{i,k}^e = (x_{i,s,k}^e)_{s\in\mathcal{S}}$.



(a) Re-buffering in DASH streaming

(b) Segment loss in streamloading

(c) Re-buffering in stream-loading

Fig. 2: This figure illustrates the different behavior between streamloading and conventional DASH based streaming in terms of re-buffering and segment late arrival. In each pair, the upper figure shows the buffer at a given time instant and the lower one shows the same buffer at a later point in time. The dashed vertical line indicates the location of the playback. It can be seen that in DASH streaming, the playback stops and re-buffers whenever the segment to be played back has not fully arrived. In streamloading, depending on whether the segment that has not been fully delivered for playback is a base, or enhancement layer, re-buffering, or segment loss, respectively, will occur.

We formulate the offline optimization problem OPTSL as follows:

$$\max_{\mathbf{r}_k^b,\mathbf{r}_k^e,\mathbf{q}_i,(\forall i\in\mathcal{N},k\in\mathcal{K})} \sum_{i\in\mathcal{N}} \left(m_i^S(q_i) - \eta \mathrm{v}_i^S(q_i)\right)$$

subject to

$$q_{i,s} \in \mathcal{Q}_{i,s},\ \forall s \in \mathcal{S}, i \in \mathcal{N} \quad (3)$$

$$x_{i,k,s}^b \geq 0,\ x_{i,k,s}^e \geq 0,\ \forall s \in \mathcal{S}, i \in \mathcal{N} \quad (4)$$

$$x_{i,k,s}^b = 0,\ x_{i,k,s}^e = 0\ \forall k, s \quad (5)$$

$$\text{s.t. } (k-\gamma_{i,k})\tau_{slot} > s\tau_{seg}, i \in \mathcal{N}$$

$$||\mathbf{x}_{i,k}^b||_1 \leq 1,\ ||\mathbf{x}_{i,k}^e||_1 \leq 1\ \forall k \in \mathcal{K}, i \in \mathcal{N} \quad (6)$$

$$c_k(\mathbf{r}_k) \leq 0,\ \forall k \in \mathcal{K} \quad (7)$$

$$r_{i,k}^b + r_{i,k}^e \leq r_{i,k},\ \forall k \in \mathcal{K}, i \in \mathcal{N} \quad (8)$$

$$\sum_{k=1}^{K} \tau_{slot} x_{i,k,s}^b r_{i,k}^b \geq \tau_{seg} f_{i,s}(q_{i,s}^0)\ \forall s \in \mathcal{S}, i \in \mathcal{N} \quad (9)$$

$$\sum_{k=1}^{K} \tau_{slot} x_{i,k,s}^e r_{i,k}^e \geq \tau_{seg}(f_{i,s}(q_{i,s}) - f_{i,s}(q_{i,s}^0)),$$

$$\forall s \in \mathcal{S}, i \in \mathcal{N} \quad (10)$$

$$\beta_i(\mathbf{q}_i, (r_i^b)_{1:K}) \leq \bar{\beta}, \forall k \in \mathcal{K}, s \in \mathcal{S}, i \in \mathcal{N} \quad (11)$$

$$\tau_{seg} S_{i,k}^b - (k-\gamma_{i,k})\tau_{slot} \leq \tau_{lim}, \forall k \in \mathcal{K}, i \in \mathcal{N} \quad (12)$$

$$\tau_{seg} S_{i,k}^e - (k-\gamma_{i,k})\tau_{slot} \geq 0, \forall k \in \mathcal{K}, i \in \mathcal{N}, \quad (13)$$

where $||\cdot||_1$ represents the ℓ^1 norm.

In OPTSL, we define $S_{i,k}^b$ and $S_{i,k}^e$ as the total number of base and enhancement layers, respectively, that are completely delivered to user i by slot k. The value of these two is derived as follows:

$$S_{i,k}^b = \sum_{s=1}^{S} \mathbb{1}_{\mathcal{A}_{i,s,k}^b}\left(r_{i,1}^b, \cdots, r_{i,k}^b\right),\ \forall k \in \mathcal{K}, i \in \mathcal{N} \quad (14)$$

$$S_{i,k}^e = \sum_{s=1}^{S} \mathbb{1}_{\mathcal{A}_{i,s,k}^e}\left(r_{i,1}^e, \cdots, r_{i,k}^e\right),\ \forall k \in \mathcal{K}, i \in \mathcal{N} \quad (15)$$

where $\mathcal{A}_{i,s,k}^b$ and $\mathcal{A}_{i,s,k}^e$ are the set of rates that if allocated to user i, will allow the user to fully download the base and enhancement layers of segment s by slot k, respectively. Hence, we define $\mathcal{A}_{i,s}^b$ and $\mathcal{A}_{i,s}^e$ as follows:

$$\mathcal{A}_{i,s,k}^b = \left\{ r_{i,1}^b, \cdots, r_{i,k}^b | \tau_{slot} \sum_{t=1}^{k} x_{i,t,s}^b r_{i,t}^b \geq \tau_{seg} f_{i,s}(q_{i,s}^0) \right\}, \quad (16)$$

$$\mathcal{A}_{i,s,k}^e = \Bigg\{ r_{i,1}^e, \cdots, r_{i,k}^e | \tau_{slot} \sum_{t=1}^{k} x_{i,t,s}^e r_{i,t}^e \geq \tau_{seg}(f_{i,s}(q_{i,s}) - f_{i,s}(q_{i,s}^0)) \Bigg\}. \quad (17)$$

Each time the allocated set of rates makes a full base or enhancement layer download possible, the total number of downloaded segments $S_{i,k}^b$ and $S_{i,k}^e$ are incremented.

Finally, $\gamma_{i,k}$ is the cumulative number of time slots up to slot k that user i has spent re-buffering. It is calculated in a manner similar to (16) and (17), as follows:

$$\gamma_{i,k} = \sum_{t=1}^{k} \mathbb{1}_{\{S_{i,t}^b \tau_{seg} < (t-\gamma_{i,t-1})\tau_{slot}\}}(S_{i,t}^b), \text{where } \gamma_{i,0} = 0$$

$$\forall k \in \mathcal{K}, i \in \mathcal{N} \quad (18)$$

The constraint shown in (8) ensures that the sum of base and enhancement layer rates does not exceed the total allocated rate in each slot. The causality constraint (5) ensures that segments are not downloaded if their respective playback time

has passed. Constraints (9) and (10) ensure that the total rate allocated for downloading a specific segment, regardless of the layer, should be at least equal to the size of the segment. The right hand side of (10) contains a decision variable indicating how many enhancement layers should be downloaded for each segment. The constraint should hold for any feasible choice of the number of enhancement layers. In (11), the upper limit on the fraction of time the user is re-buffering is set to $\bar{\beta}$. In other words, all segments need to be downloaded within $1+\bar{\beta}$ times the duration of the video. Therefore, $\bar{\beta}$ can take values greater than -1. However, the feasibility of the problem depends on the choice of $\bar{\beta}$, especially for negative values.

The buffer limitation on the base layer segments is captured in (12). Using this constraint, we ensure that at every time slot, the number of base layer segments currently stored in the buffer does not exceed the limit. The amount of buffered video at any given time slot is calculated as the total downloaded video duration minus the amount of time spent on playback. In order to avoid the occurrence of enhancement layer loss due to late arrivals, we introduce the *segment loss constraint* (13), which makes sure that the downloading header for the enhancement layers never falls behind the playback header.

The above problem jointly optimizes rate allocation over $\mathbf{r}_k^b$ and $\mathbf{r}_k^e$, and quality selection over $\mathbf{q}_i$, with respect to all given constraints. The feasibility depends on the choice of $c_k(\mathbf{r}_k)$, $\bar{\beta}$, and the quality rate trade-off functions. However, the solution to this problem is complex and requires channel state information for all future time slots, as well as the quality values of all future segments, thus it is not possible to implement it in practice. In order to overcome this problem, a sub-optimal online algorithm satisfying the constraints of OPTSL will be designed. The result of this algorithm is upper bounded by the optimal solution of OPTSL. In Section V, we propose this algorithm, called RAQUEL, that performs rate allocation and quality selection in an online fashion.

V. Online Algorithm RAQUEL

In this section, we present a simple online algorithm called RAQUEL that performs rate allocation and quality selection for multi-user streamloading in wireless networks. RAQUEL is based on an approach that was adopted for DASH video delivery in [11], [23]. The authors formulate the problem of DASH-based video delivery to a set of users in a wireless network in a similar setting. The formulation includes a subset of the constraints in OPTSL, namely the link capacity constraint (7) and the re-buffering constraint (11). Based on this formulation, an online algorithm is developed called NOVA, which is then proved to achieve optimality in a stationary regime. The fundamental idea in NOVA is the concept of *virtual buffer*, which estimates the Lagrange multipliers associated with the re-buffering constraint, and hence, determines the risk of re-buffering for each user.

In NOVA, rate allocation is performed by the base station at the beginning of each time slot, and quality selection is performed by individual users whenever they request a new segment. At every time slot, the rate vector that maximizes $\sum_{i\in\mathcal{N}} b_i r_i$ is determined, subject to the link capacity constraint, where b_i denotes the value of the virtual buffer for user i. Therefore, users with better channel states and higher virtual buffer (higher risk of re-buffering) are prioritized for rate allocation. After allocating rates, the virtual buffers for all users are incremented by an amount proportional to τ_{slot}.

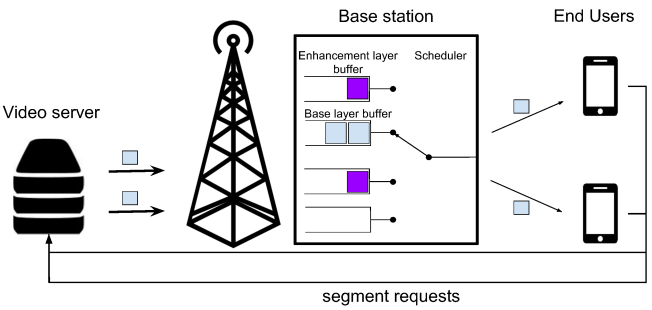


Fig. 3: The delivery procedure for streamloading with RAQUEL for two users. Each user is assigned two buffers at the base station for storing base and enhancement layer segments, respectively. Users request segments which the video server delivers to the base station and the base station schedules the segments for delivery.

A user who finishes downloading a segment, requests the quality level of the next segment s to be delivered by maximizing $q_{i,s} - \eta(q_{i,s} - m_{i,s})^2 - \frac{b_i}{1+\beta_{NOVA}} f_i(q_{i,s})$, where $m_{i,s}$ keeps track of the average video quality of the segments delivered to user i up to segment s. This objective function implies that since higher quality downloads require longer delivery times, high quality segments should be requested only if the risk of re-buffering is low, otherwise the requested quality should be decreased. The user who has finished downloading a segment, updates its virtual buffer by decreasing it proportional to τ_{seg}. By following this procedure and constantly updating the virtual buffer, the obtained video quality is maximized without violating a constraint called β_{NOVA}, on the fraction of time spent re-buffering.

We use NOVA to simulate the streaming and downloading service models in Section VII. In RAQUEL, we devise a similar strategy as will be explained in Section V-B. Before going through the details of the algorithm itself, we first describe the procedure for delivering base and enhancement layer segments.

A. Delivery Procedure

In streamloading, base and enhancement layers are requested and delivered separately. As described in Section IV-B, segment requests in SVC-based video delivery are flexible in the sense that at any decision epoch, multiple layers of multiple segments can be requested by the user. Such flexibility allows for adaptive streaming schemes, like prefetching lower layers ahead and backfilling higher layers later [24], [25]. However, in designing RAQUEL, we assume that all requested enhancement layers of a particular segment are delivered together, and downloading additional enhancement layers for segments for which some enhancement layers have already been downloaded in the past, is not possible.

The segment delivery procedure that we propose in RAQUEL is illustrated in Figure 3. We divide the joint rate allocation and quality selection into two separate tasks. The

quality selection is done at the user end, where each user makes sequential requests on what segment to receive next. Upon receiving a request from a user, the video server sends the requested segment to the base station, which in turn schedules the users and delivers the segments. Here, we make the simplifying assumption that the end to end TCP connection between the server and the user can keep up with the segment delivery, hence, no congestion occurs at the link between server and base station.

In our proposed quality selection scheme, the users prioritize base layer over enhancement layer segments to reduce the likelihood of re-buffering. Users keep requesting base layer segments until they reach the buffer limit. Once the limit is reached, they request enhancement layers for the segments that have not been played back. The number of enhancement layers that each user requests for a particular segment is determined by the procedure explained in Section V-B1. New enhancement layers are requested whenever the previously requested enhancement layers are fully delivered. New base layer segments are periodically requested as the playback continues and the buffered segments are consumed.

The rate allocation is done at the base station. As can be seen in Figure 3, the base station assigns two buffers per user for the requested base and enhancement layer segments, respectively. The base station scheduler gives absolute priority to the base layer buffers and first tries to deliver all outstanding base layer segments to the users. Once no base layer segment is left at the base station, the scheduler starts allocating resources to deliver the buffered enhancement layer segments. The way the users are scheduled in each of these cases is explained in Section V-B2.

By following the above procedure, segment quality selection and network resource allocation are independently and asynchronously performed by end users and the base station, respectively. Next, we explain each of these two steps in detail.

B. RAQUEL

In this section, we explain the two tasks of RAQUEL, namely rate allocation at the base station (RA) and the quality selection at the user end (QUEL). Since the quality of each segment is a function of the number of layers requested for that segment, layer selection and quality selection are used interchangeably throughout the rest of the paper.

Similar to NOVA, we use two variables as the virtual buffer representation for the base and enhancement layers, which are dynamically updated on a per slot basis and determine the allocated rate and selected quality for each user throughout the streamloading process. The virtual buffer for the base layer, b^b, is an indicator of the risk of violating the re-buffering constraint (11). A higher value for b^b occurs whenever the occupancy of the base layer buffer is low and hence the danger of re-buffering is high. Similarly, the virtual buffer for the enhancement layer b^e indicates the risk of violating the segment loss constraint (13). Due to the dynamic nature of the wireless channel, and also the varying buffer level at the user end, the virtual buffer values for both base and enhancement layers should be constantly updated. As the video plays back at the user end, the downloaded data in the buffer is consumed. Hence, as long as no new segment arrives at the user, the risk of draining the buffer constantly increases. Thus, at every time slot, the virtual buffer should increase proportional to the slot duration $b_i^b = b_i^b + \epsilon\tau_{slot}$ (same for b_i^e), where ϵ is a positive constant determining the rate of update. However, if new segments are delivered to the user, the risk of re-buffering (and segment loss) decreases proportional to the segment duration and the corresponding virtual buffer is updated as $b_i^b = \max\{b_i^b - \epsilon\tau_{seg}, 0\}$ (same for b_i^e). The updated values are then used to perform RA and QUEL.

1) Quality Selection QUEL: As explained in Section V-A, whenever user i fully receives the enhancement layers it had requested for segment $s-1$, a decision has to be made about the quality level for the next segment s. The request indicates how many enhancement layers user i should download for segment s, according the following maximization:

$$\textbf{QUEL}(b_i^e):$$
$$l^* = \max_{l\in\{0,\cdots,L\}} \Bigg\{ q_{i,s}^l - \eta(q_{i,s}^l - m_{i,s})^2 - \frac{b_i^e}{1+\beta_{sl}}(f_i(q_{i,s}^l) - f_i(q_{i,s}^0)), i \in \mathcal{N} \Bigg\}, \quad (19)$$

where $q_{i,s}^l$ is the video quality user i would see if it had l enhancement layers in addition to the base layer, and $q_{i,s}^0$ is the minimum segment quality provided by the base layer for segment s. The average quality up to segment s is denoted by $m_{i,s}$. It can be easily verified that the objective function is concave and solved by simply trying all possible levels l.

The right hand side of QUEL consists of three terms, where the second one accounts for the user sensitivity to variability in video quality, and thereby ensures that the requested quality is close to the average of the previously requested segments. The third term acts as a penalty on the requested quality to avoid enhancement layer loss when the system is congested, i.e., b_i^e is high. It can be observed that the larger the size of the segments become and the higher the risk of starving the enhancement layer buffer gets, the more penalty is enforced on the requested segment quality. In addition to that, a key parameter in QUEL is β_{sl} that aims at adjusting the sensitivity of the layer selection process to segment loss. For larger β_{sl} values, the system is less sensitive and as a result, tends to request more enhancement layers. This can result in a larger number of enhancement layer segment losses due to late arrival which leads to them being discarded. On the other hand, setting β_{sl} to a low value increases the sensitivity to segment loss and results in a more conservative layer selection policy. Hence, changing the value of β_{sl} has a two sided effect on the performance of QUEL. In Section VII we analyze the impact of varying β_{sl} on the overall video quality and enhancement layer segment loss, by showing that there exists an optimum value for β_{sl} which trades off between aggressive layer selection and segment loss. A similar parameter, β_{NOVA}, is used in the NOVA algorithm. However, because of the fundamental differences between streamloading and the other DASH based streaming schemes discussed in Section IV-B, β_{NOVA} determines the sensitivity of quality selection on re-

buffering. In fact, it can be easily verified that the two sided effect we observe with respect to β_{sl} for streamloading does not hold with β_{NOVA} for NOVA. Instead, increasing β_{NOVA} will constantly increase the quality of requested segments while increasing the overall re-buffering time.

QUEL also takes into account the possibility of requesting no enhancement layers for a segment. In such a scenario, user i requests the minimum quality for segment s, and only the base layer will be delivered. Since no enhancement layer is requested for this segment, the enhancement layer download frontier shifts one segment ahead, followed by a request for segment $s+1$, according to (19). This procedure repeats until the user requests one or more enhancement layers for a segment. To account for this, we update $b_i^e = \max\{b_i^b - \epsilon\tau_{seg}, 0\}$ as suggested before. In other words, we treat it as a *complete download of zero enhancement layers*.

2) Resource Allocation RA: At the beginning of each time slot, the base station decides how to allocate the available resources to each of the end users. As mentioned in Section V-A, the scheduler keeps two buffers per user for base and enhancement layer segments, respectively. Resources are allocated with absolute prioritization of base layer segments. This means that if there are base layer segments left to transmit to the users at the base station, the base station schedules those first. This is done by solving RA^b as shown below:

$$\mathbf{RA^b}(b^b): \max_{\mathbf{r}} \quad \sum_{i\in\mathcal{N}'} b_i^b r_i \qquad s.t. \quad c(\mathbf{r}) \leq 0, i \in \mathcal{N}' \tag{20}$$

where $\mathcal{N}' \subset \mathcal{N}$ is the set of all the users who have base layer data left at the base station. If there is no base layer segment left to transmit to the users, the scheduler allocates resources to deliver the queued enhancement layer segments by solving RA^e as follows:

$$\mathbf{RA^e}(b^e): \max_{\mathbf{r}} \quad \sum_{i\in\mathcal{N}} b_i^e r_i \qquad s.t. \quad c(\mathbf{r}) \leq 0, i \in \mathcal{N} \tag{21}$$

where r_i is the allocated rate to user i. According to (20) and (21), between users with the same achievable data rate, the one with larger virtual buffer has higher scheduling priority, and between users with equal virtual buffer, the one with higher achievable rate gets scheduled first.

Equations (19)-(21), capture how RAQUEL operates and Algorithm 1 shows a detailed description including all the involved steps. RAQUEL is sub-optimal, but very simple to apply and does not require information about the future states of the channel. Furthermore, the allocation and layer selection steps can be independently and asynchronously performed at the base station and mobile stations, respectively.

Algorithm 1 Streamloading Online Algorithm RAQUEL

Initialization: Let $\epsilon > 0$, and for each $i \in \mathcal{N}$, let $b_{i,0}^b \geq 0$, $b_{i,0}^e \geq 0$, and $0 \leq m_{i,0} \leq q_{max}$.

for all time slots k **do**

ALLOCATE (RA):

if Base layer segments present at base station **then**

$\mathbf{RA^b}$ determines the optimal rate allocation vector $\mathbf{r}_k^b$.

else

$\mathbf{RA^e}$ determines optimal rate allocation vector $\mathbf{r}_k^e$.

end if

Update virtual buffer as follows:

$$b_{i,k+1}^b = b_{i,k}^b + \epsilon\tau_{slot} \tag{22}$$

$$b_{i,k+1}^e = b_{i,k}^e + \epsilon\tau_{slot} \tag{23}$$

SELECT (QUEL):

if Base layer segments buffered up to the limit **then**

for $\forall i \in \mathcal{N}$ **do**

if user i finished downloading enhancement layers for s_i **then**

Solve (19) for user i to obtain $l_{s_i+1}^*$.

while $l_{s_i+1}^* = 0$ **do**

$$m_{i,s_i+1} = m_{i,s_i} + \epsilon(q_{i,s}^0 - m_{i,s_i})^2, \tag{24}$$

$$b_{i,k+1}^e = \max\{b_{i,k}^e - \epsilon\tau_{seg}, 0\}, \tag{25}$$

$$s_i = s_i + 1 \tag{26}$$

Solve (19) for user i to obtain $l_{s_i+1}^*$.

end while

$$m_{i,s_i+1} = m_{i,s_i} + (q_{i,s}^{l_{s_i+1}^*} - m_{i,s_i})^2, \tag{27}$$

$$b_{i,k+1}^b = \max\{b_{i,k}^b - \epsilon\tau seg, 0\}, \tag{28}$$

$$s_i = s_i + 1 \tag{29}$$

end if

end for

else

for $\forall i \in \mathcal{N}$ **do**

if user i finished downloading current base layer **then**

Request next base layer

$$b_{i,k+1}^b = \max\{b_{i,k}^b - \epsilon\tau_{seg}, 0\} \tag{30}$$

end if

end for

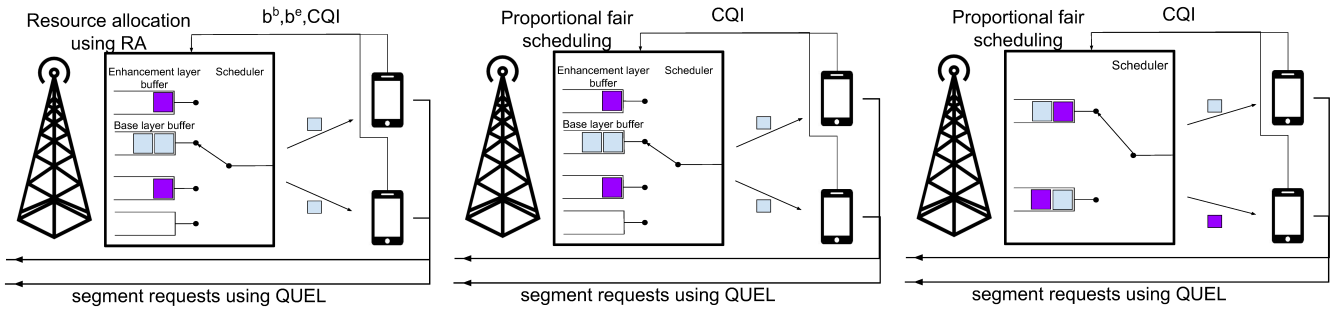
end if

end for

VI. Practical Implications

A goal for our streamloading algorithm is that its implementation be feasible on practical networks. Any algorithm for rate allocation needs to take into account the feasibility and practical limitations that exist on real base stations. Current base stations perform scheduling of video streaming data at the MAC layer without considering the playback buffer state of each user. Such a cross-layer capability would increase the complexity of the base station design but offers substantial performance gains.

In this section, we evaluate practical implications emerging from implementing streamloading using RAQUEL in terms of signaling overhead, cross-layer requirements, and complexity.



(a) RA-based scheduling with two buffers per user at base station.

(b) Proportional fair scheduling with two buffers per user at base station (PF1).

(c) Proportional fair scheduling with single buffer per user at base station (PF2).

Fig. 4: Three different schemes for scheduling in terms of base station functionality. In all three cases, the quality request procedure is based on QUEL. From (a) to (c) complexity and signaling overhead decrease with the scheme presented in (c) not requiring any packet inspection at the base station.

In order to compare RAQUEL with state-of-the-art scheduling mechanisms, we also present two schemes with decreased complexity and cross-layer functionality. All schemes follow the general procedure shown in Figure 3 and only the functionality of the scheduler changes as illustrated in Figure 4.

In Figure 4a, the base station performs rate allocation using RA as explained in Section V-B2, where two buffers are assigned to each user to store base and enhancement layer data, respectively. Hence, the base station needs to detect what layer the incoming packet belongs to, in order to properly implement RA. This can be done simply by either reading one bit from the application header of incoming packets or by using the standardized Differentiated Services (DiffServ) [26] field from the IP header. In addition to that, the user needs to send the updated values for the virtual buffer states b^b and b^e along with Channel Quality Indicator (CQI) reports back to the base station at the end of every time slot. Note that updates for b^b and b^e, i.e., Equations (22) and (23) can be done independently at the base station, while Equations (25) and (28) require knowledge that a segment download has completed, which is easier to detect at the end users. The user can in turn send the updated values to the base station accordingly. Note that the latter only occurs at segment completion and is thus relatively infrequent. As a result, the overhead would be relatively low.

Figure 4b illustrates a similar system, with the difference that here the rate allocation is replaced by a simple proportional fair scheduler without the need of the virtual buffer values being fed back to the base station. Proportional Fairness (PF) is a simple standard scheduling mechanism for wireless networks in which in each time slot, the available rate is allocated to users in proportion to the average rate they have been allocated to date [27]. Similar to the previous scenario, base layer data is prioritized over enhancement layer data during scheduling. We denote this scheduling scheme as PF1 for the remainder of the paper. The computational complexity of PF1 is similar to RA but the signaling overhead is reduced to only sending back CQI messages.

A third possible scheme is shown in Figure 4c, where not only proportional fairness is used for scheduling, but also the base and enhancement layer buffers are replaced by a common buffer that is filled with data from different layers in the order in which they are requested. In this case, the benefits that result from prioritizing base layer over enhancement layer segments such as lower re-buffering time will be lost. In this scenario, no packet inspection is required by the base station. We denote this scheme as PF2 for the rest of the paper. The computational complexity and signaling overhead of PF1 is the same as PF2, but with reduced base station cross-layer functionality.

In Section VII, we evaluate the QoE achieved under deploying these three schemes and discuss the trade-offs that exist between introducing additional complexity and the enhancement in user's QoE.

VII. Simulation Results

In this section, we evaluate the performance of RAQUEL and compare it with the streaming and downloading service models via simulation. As discussed in Section V, we implement NOVA to evaluate the performance of state-of-the-art streaming and downloading schemes. More precisely, since NOVA in its most general form does not have any buffer limit, it is best suited for the download service model. For the streaming scenario, when a user reaches the buffer limit, it stops requesting more segments until buffer space becomes available and NOVA resumes.

A. Simulation Setting

The channel model under consideration follows capacity constraints in the form of $c_k(\mathbf{r}_k) = \sum_{i \in \mathcal{N}} \frac{r_{i,k}}{\rho_{i,k}} - 1$ in each time slot k, where $\rho_{i,k}$ is the maximum achievable rate for user i in slot k. These peak rates are drawn from a rate distributions based on real HSDPA rate traces with correlation [11].

The video to be delivered is a 20 minute long sequence from the open source Valkaama video which is divided into one second long segments. Each segment is encoded into six quality levels ranging from 100kbps to 1.5Mbps. For the DASH-based streaming and downloading scenarios, these different levels correspond to different quality representations, whereas for the SVC-based streamloading scenario, each of these levels corresponds to one additional enhancement layer. For example, consider a segment that is encoded into two representations of size 100kbps and 200kbps. We assume that

the equivalent SVC representation of this segment consists of a base layer and one enhancement layer of equal size. Hence, the quality obtained from downloading the 200kbps representation for DASH, is equal to the quality resulting from the base and one enhancement layer for SVC. The same holds true for additional enhancement layers. This example is a simplification of a real SVC video. Because of the overhead imposed by SVC, we add an extra 10% to the size of each layer [28].

For each segment, we assume one base layer compressed at 100kbps and up to 5 enhancement layers, the rates of which are shown in Table II. In order for the quality-rate trade-off to capture the video quality that users perceive, we use a model for the Differential Mean Opinion Score (DMOS), see [29]. In the absence of actual DMOS values, a proxy DMOS can be used to map each segment representation to the corresponding quality. Our proxy is based on the MSSSIM-Y metric for each segment according to a model presented in [30]. Our simulations are performed using the parameters in Table II unless stated otherwise.

Parameter	Value
ϵ	0.05
η	0.1
τ_{slot}	10ms
τ_{seg}	1s
τ_{lim}	50s
β_{sl}	0
β_{NOVA}	-0.2
video length	20min
SVC overhead per enhancement layer	10%
number of enhancement layers	5
base layer bit-rate	100kbps
enhancement layer bit-rates	100,100,300,300,600kbps

TABLE II: Simulation parameters

B. Improvements in Video Quality

Our primary goal is to evaluate the performance of RAQUEL by comparing the resulting quality metrics of streamloading using RAQUEL with streaming and downloading. For this comparison, we evaluate our objective video quality metric, as well as the average re-buffering time. The video quality depends on the quality of each delivered segment minus a constant times the standard deviation of the video segment quality to account for users' negative response to variability [31]. Average re-buffering time is calculated as the cumulative amount of time that the video playback stops due to buffer starvation.

Figure 5 shows the video quality obtained under each of the service models versus the number of users in the network. The buffer limit imposed on streaming and streamloading is set to 50 seconds. The figure shows that streamloading performs as well as unconstrained downloading for lightly loaded networks and at least as well as streaming for heavily loaded networks. It outperforms conventional streaming by a large margin, e.g., for a video quality of 25, the number of users that can be supported is doubled. This shows that streamloading provides video quality close to downloading, while still being legally classified as a streaming scheme.

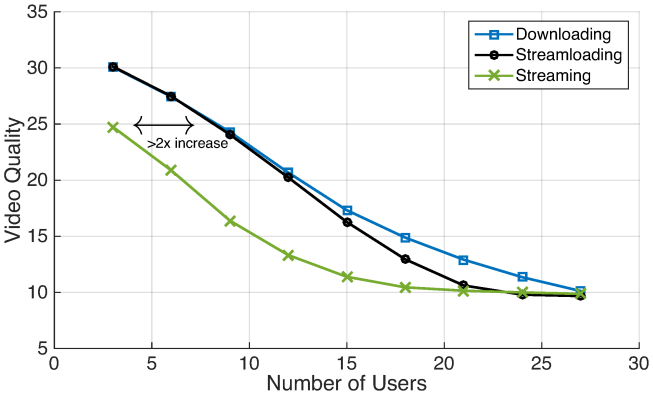


Fig. 5: Video quality comparison between streaming, downloading, and streamloading with a buffer limit equal to 50 seconds for streaming and streamloading.

Figure 6 shows the average re-buffering time for each of the three service models. It can be seen that despite larger segment sizes due to encoding overhead, streamloading has shorter re-buffering times on average, as compared with streaming and downloading. Filling the base layer buffer prior to downloading enhancement layers is the reason for the lower re-buffering time.

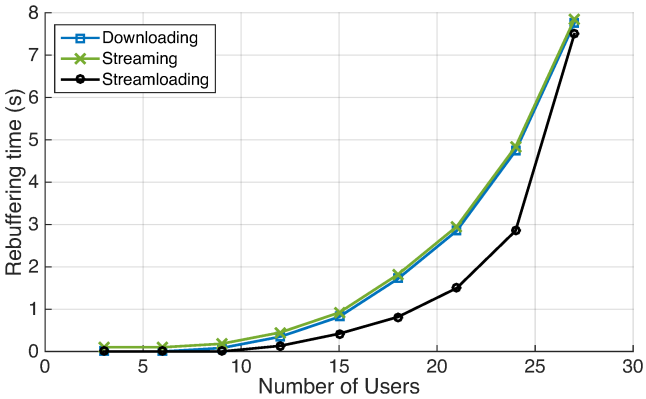


Fig. 6: Average re-buffering time for streaming, downloading, and streamloading with a buffer limit equal to 50 seconds for streaming and streamloading.

A negative side effect of downloading too many segments ahead of playback is that if users stop watching the video before it ends (abandonment), the resources that are used to deliver the abandoned segments are wasted. The downloading and streamloading service models are prone to this wastage because of their pre-fetching functionality. However, streamloading has the advantage that the pre-fetched segments do not have the base layer, therefore, it causes less wastage of resources compared to downloading. This negative effect can be further mitigated if, similar to the base layer, a limit is set for pre-fetching enhancement layer segments. This limit should obviously be set to a larger value than the base layer buffer limit to gain the benefits of streamloading. Figure 7 shows the video quality obtained from streamloading if the number of enhancement layers that can be pre-fetched is limited. It can be seen that even for an enhancement layer

buffer limit of only 100s, the streamloading quality is higher than regular streaming. Furthermore, by setting this limit above 150s, streamloading performs almost as well as the case with unlimited pre-fetching. It should be noted that limiting pre-fetching to the values depicted in Figure 7 does not change the average re-buffering time. Hence, this limit can be set according to the trade-off between avoiding resource wastage and increasing video quality.

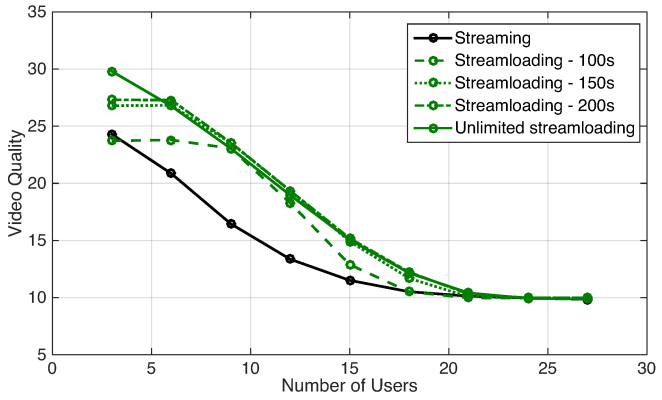


Fig. 7: Video quality comparison between streaming and streamloading with different limits for pre-fetching enhancement layer segments. The buffer limit for streaming as well as for base layers in streamloading is set to 50s.

The two sided effect of β_{sl} is shown in Figure 8 under different network loads. As discussed in Section V-B1, increasing β_{sl} results in more aggressive layer selection which in turn increases the number of lost segments. According to this figure, up to a certain value for β_{sl}, the added aggressiveness results in higher video quality. However, increasing it further causes too many segment losses which decrease the video quality and increase bandwidth wastage. The optimum value for β_{sl} is roughly constant under various network loads.

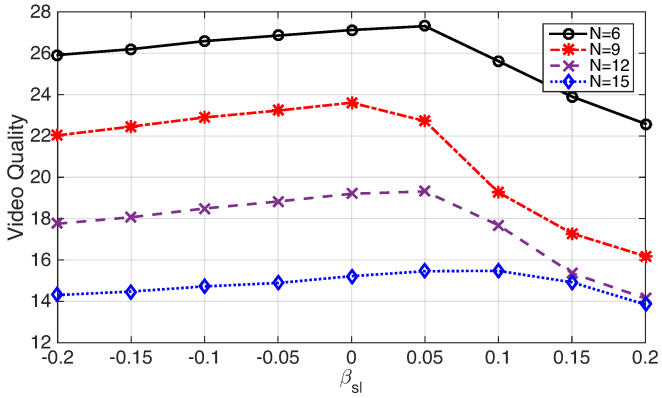


Fig. 8: Variations in video quality with respect to the segment loss sensitivity parameter β_{sl} for different number of users. The trade-off between aggressive layer selection and segment loss, suggests an optimal value for β_{sl} for different network loads.

C. RAQUEL vs. Baseline Algorithms

Let us now evaluate the performance of RAQUEL by comparing it with two widely deployed algorithms, namely proportional fairness scheduling for resource allocation, and rate matching for quality selection. Rate Matching (RM), is a quality adaptation scheme in which the next selected segment is one whose bit rate is the closest matching to the average rate the user estimates it has seen to date [32]. For the rate allocation part, we investigate PF1 and PF2 as described in Section VI.

Figures 9 and 10 demonstrate the performance of RAQUEL when compared to the cases where streamloading is done using the alternative schemes. For comparison, we adopted five different combinations for rate allocation and quality selection. In the first scheme, RAQUEL is applied to both tasks (RAQUEL). The second method uses PF1 for rate allocation while QUEL is used for quality selection (PF1-QUEL). In the third approach, PF1 and rate matching replace RAQUEL in both tasks (PF1-RM). The fourth and fifth scheme are similar to the second and third where PF1 is replaced by PF2.

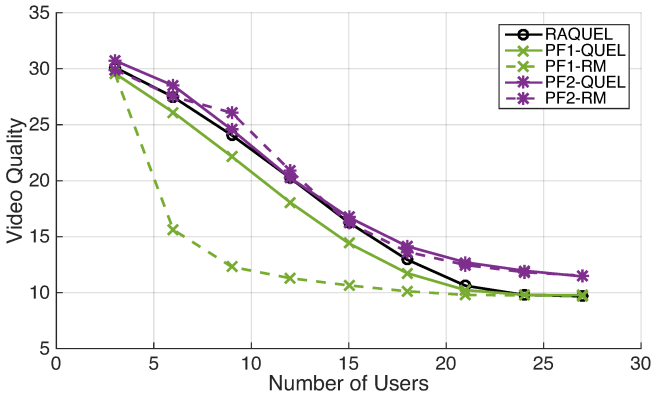


Fig. 9: Video quality comparison between RAQUEL, and the conventional proportional fairness and rate matching methods for streamloading. The base layer limit is fixed at 50 seconds.

As it can be seen from Figure 9, RAQUEL results in higher video quality than the two schemes that are based on PF1. However, the PF2-based algorithms perform slightly better than RAQUEL in terms of video quality. The reason for this is that since each user has a single buffer at the base station, scheduling is not done based on base layer prioritization. Hence, enhancement layer segments can be opportunistically downloaded and aggressively pre-fetched causing higher segment quality. However, not giving priority to base layer segments causes delay in their delivery and consequently, results in re-buffering. Figure 10 shows that RAQUEL greatly outperforms all other schemes in terms of average re-buffering time. In fact, in PF2-based schemes, users are re-buffering almost 25% of the entire streaming time. This shows that in PF1 and PF2, because scheduling is solely based on channel quality and the state of the buffer is not taken into account, re-buffering avoidance is not incorporated in the resource allocation, whereas in RAQUEL, the inclusion of the buffer level in the resource allocation mechanism through the virtual buffers reduces the re-buffering time. However, PF1-QUEL may be an acceptable compromise on performance if lower complexity at the base station is an important consideration.

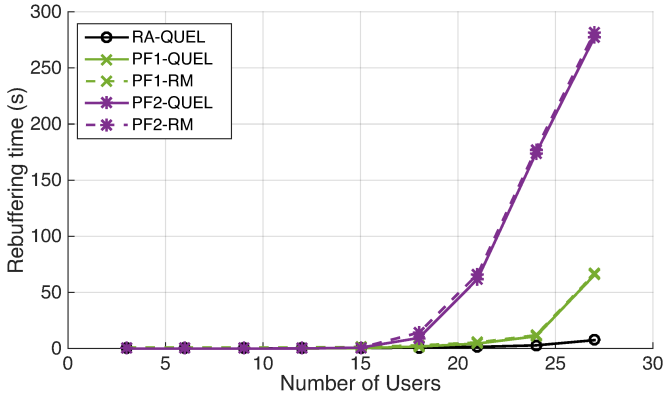


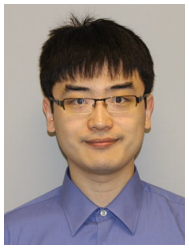
Fig. 10: Re-buffering comparison between RAQUEL and the conventional proportional fairness and rate matching methods for streamloading. The base layer limit is fixed at 50 seconds.

VIII. Conclusions and Future Work

In this work, we have proposed an online procedure for asynchronous rate allocation and quality selection for streamloading. Streamloading is shown to provide low priced, high quality video to users watching copyright restricted content. This is done by pre-fetching enhancement layers ahead of real-time, and streaming base layer segments in real-time. Our simulation results show that streamloading improves video quality over state-of-the-art streaming methods, while still satisfying the legal classification of a streaming service. Also, adding simple cross-layer functionality at the base station in order to distinguish between base and enhancement layer packets can enhance the QoE of the streamloading experience significantly.

The scope of this work can be further extended to different network types such as heterogeneous networks consisting of high capacity femtocells or Wi-Fi hotspots, which can be leveraged for more efficient pre-fetching of enhancement layers. Furthermore, the benefit of streamloading can be explored in a network with user dynamics, which includes users joining and leaving the network. The variability resulting from such traffic dynamics can be exploited by speeding up the download of enhancement layers when the network is lightly loaded, in order to increase video quality when the network load is high. These issues are subjects for future research.

S. Amir Hosseini received the B.S. degree in Electrical Engineering from Sharif University of Technology, Iran in 2011. Since 2012, he is a PhD candidate at NYU Tandon School of Engineering, USA. His research interests include wireless communications, wireless networking and video delivery optimization.



Zheng Lu received his B.E. degree in Electronics Engineering from Tsinghua University in China in 2009. He received his M.S.E. and Ph.D. in Electrical and Computer Engineering from The University of Texas at Austin in 2011 and 2015 respectively. His research focuses on algorithms and architectures to enhance perceived video quality for video streaming, and resource allocation in Device-to-Device networks to optimize system and user perceived performance. He interned at Intel Labs, Hillsboro during summer 2013. And since 2015, he has been working at Cisco Systems in San Jose, CA.

Gustavo de Veciana (S'88-M'94-SM'01-F'09) received his B.S., M.S, and Ph.D. in electrical engineering from the University of California at Berkeley in 1987, 1990, and 1993 respectively, and joined the Department of Electrical and Computer Engineering where he is currently a Cullen Trust Professor of Engineering. He served as the Director and Associate Director of the Wireless Networking and Communications Group (WNCG) at the University of Texas at Austin, from 2003-2007. His research focuses on the analysis and design of communication and computing networks; data-driven decision-making in man-machine systems, and applied probability and queueing theory. Dr. de Veciana served as editor and is currently serving as editor-at-large for the IEEE/ACM Transactions on Networking. He was the recipient of a National Science Foundation CAREER Award 1996 and a co-recipient of five best paper awards including: IEEE William McCalla Best ICCAD Paper Award for 2000, Best Paper in ACM TODAES Jan 2002-2004, Best Paper in ITC 2010, Best Paper in ACM MSWIM 2010, and Best Paper IEEE INFOCOM 2014. In 2009 he was designated IEEE Fellow for his contributions to the analysis and design of communication networks. He currently serves on the board of trustees of IMDEA Networks Madrid.

Shivendra S. Panwar is a Professor in the Electrical and Computer Engineering Department at NYU Tandon School of Engineering. He received the B.Tech. degree in electrical engineering from the Indian Institute of Technology Kanpur, in 1981, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Massachusetts, Amherst, in 1983 and 1986, respectively. He is currently the Director of the New York State Center for Advanced Technology in Telecommunications (CATT), the Faculty Director of the NY City Media Lab, and member of NYU WIRELESS. He spent the summer of 1987 as a Visiting Scientist at the IBM T.J. Watson Research Center, Yorktown Heights, NY, and has been a Consultant to AT&T Bell Laboratories, Holmdel, NJ. His research interests include the performance analysis and design of networks. Current work includes wireless networks, switch performance and multimedia transport over networks. He is an IEEE Fellow and has served as the Secretary of the Technical Affairs Council of the IEEE Communications Society. He is a co-editor of two books, Network Management and Control, Vol. II, and Multimedia Communications and Video Coding, both published by Plenum. He has also co-authored TCP/IP Essentials: A Lab based Approach, published by the Cambridge University Press. He was awarded, along with Shiwen Mao, Shunan Lin and Yao Wang, the IEEE Communication Society's Leonard G. Abraham Prize in the Field of Communication Systems for 2004. He was also awarded, along with Zhengye Liu, Yanming Shen, Keith Ross and Yao Wang, the Best Paper in 2011 Multimedia Communications Award.

REFERENCES

[1] "Cisco visual networking index: Global mobile data traffic forecast update, 2014-2019," Tech. Rep., February 2015.

[2] T. Stockhammer, "Dynamic adaptive streaming over HTTP: standards and design principles," in *Proceedings of the second annual ACM Conference on Multimedia Systems (MMSys 2011)*. ACM, 2011, pp. 133–144.

[3] "Netflix Terms of Use," https://www.netflix.com/TermsofUse, September 15, 2014.

[4] "Amazon Instant Video Terms of Use," http://www.amazon.com/gp/help/customer/display.html?nodeId=201422760, April 12, 2013.

[5] A. Rath, S. Goyal, and S. Panwar, "Streamloading: low cost high quality video streaming for mobile users," in *Proceedings of the 5th ACM Internationsl Workshop on Mobile Video (MoVid 2013)*. ACM, pp. 1–6.

[6] Jeanne Fromer, Professor, NYU School of Law, Personal communication, May 16, 2014.

[7] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H. 264/AVC standard," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 17, no. 9, pp. 1103–1120, 2007.

[8] C. Zhou, C.-W. Lin, X. Zhang, and Z. Guo, "A control-theoretic approach to rate adaption for DASH over multiple content distribution servers," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 24, no. 4, pp. 681–694, 2014.

[9] K. J. Ma and R. Bartos, "HTTP live streaming bandwidth management using intelligent segment selection," in *Global Telecommunications Conference (GLOBECOM 2011)*. IEEE, 2011, pp. 1–5.

[10] J. Jiang, V. Sekar, and H. Zhang, "Improving fairness, efficiency, and stability in HTTP-based adaptive video streaming with FESTIVE," in *Proceedings of the 8th International Conference on Emerging Networking Experiments and Technologies (CoNEXT 2012)*. ACM, 2012, pp. 97–108.

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

[12] "Vidyo contributes scalability to HEVC (H.265)," Press Release, October 2012, http://www.vidyo.com/company/news-and-events/press-releases/vidyo-contributes-scalability-to-hevc- h-265/.

[13] "Vidyo and Google collaborate to enhance video quality within WebRTC," Press Release, August 2013, http://www.vidyo.com/company/news-and-events/press-releases/vidyo-and-google-collaborate-to- enhance-video -quality-within-webrtc/.

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

[15] Y. Sánchez de la Fuente, T. Schierl, C. Hellge, T. Wiegand, D. Hong, D. De Vleeschauwer, W. Van Leekwijck, and J. Y. Le Bouédec, "iDASH: improved dynamic adaptive streaming over HTTP using scalable video coding," in *Proceedings of the second annual ACM Conference on Multimedia Systems (MMSys 2011)*. ACM, 2011, pp. 257–264.

[16] C. Sieber, T. Hosfeld, T. Zinner, P. Tran-Gia, and C. Timmerer, "Implementation and user-centric comparison of a novel adaptation logic for DASH with SVC," in *Integrated Network Management (IM 2013), 2013 IFIP/IEEE International Symposium on*. IEEE, 2013, pp. 1318–1323.

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

[18] M. S. Talebi, A. Khonsari, and M. H. Hajiesmaili, "Utility-proportional bandwidth sharing for multimedia transmission supporting scalable video coding," *Computer Communications*, vol. 33, no. 13, pp. 1543–1556, 2010.

[19] H. Zhang, Y. Zheng, M. A. Khojastepour, and S. Rangarajan, "Cross-layer optimization for streaming scalable video over fading wireless networks," *Selected Areas in Communications, IEEE Journal on*, vol. 28, no. 3, pp. 344–353, 2010.

[20] T. Kim and M. H. Ammar, "Optimal quality adaptation for scalable encoded video," *Selected Areas in Communications, IEEE Journal on*, vol. 23, no. 2, pp. 344–356, 2005.

[21] T. Schierl, Y. S. De La Fuente, R. Globisch, C. Hellge, and T. Wiegand, "Priority-based media delivery using SVC with RTP and HTTP streaming," *Multimedia Tools and Applications*, vol. 55, no. 2, pp. 227–246, 2011.

[22] V. Joseph and G. de Veciana, "Jointly optimizing multi-user rate adaptation for video transport over wireless systems: Mean-fairness-variability tradeoffs," in *INFOCOM, 2012 Proceedings IEEE*, 2012, pp. 567–575.

[23] V. Joseph, "Mean-variability-fairness tradeoffs in resource allocation with applications to video delivery," Ph.D. dissertation, University of Texas at Austin, 2013.

[24] S. A. Hosseini, F. Fund, and S. S. Panwar, "(Not) yet another policy for scalable video delivery to mobile users," in *Proceedings of the 7th ACM International Workshop on Mobile Video (MoVid 2015)*. ACM, 2015, pp. 17–22.

[25] T. Andelin, V. Chetty, D. Harbaugh, S. Warnick, and D. Zappala, "Quality selection for dynamic adaptive streaming over http with scalable video coding," in *Proceedings of the 3rd ACM annual Conference on Multimedia Systems (MMSys 2012)*. ACM, 2012, pp. 149–154.

[26] S. Blake, F. Baker, and D. Black, "Definition of the differentiated services field (DS field) in the IPv4 and IPv6 headers," *RFC2474, IETF Proposed Standard*, 1998.

[27] 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*, pp. 237–252, 1998.

[28] T. Wiegand, "Further results for an RD-optimized multi-loop SVC encoder (jvt-w071)," in *JVT Meeting (Joint Video Team of ISO/IEC MPEG & ITU-T VCEG, San Jose*, 2007.

[29] A. K. Moorthy, K. Seshadrinathan, R. Soundararajan, and A. C. Bovik, "Wireless video quality assessment: A study of subjective scores and objective algorithms," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 20, no. 4, pp. 587–599, 2010.

[30] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 23, no. 4, pp. 684–694, 2013.

[31] C. Yim and A. C. Bovik, "Evaluation of temporal variation of video quality in packet loss networks," *Signal Processing: Image Communication*, vol. 26, no. 1, pp. 24–38, 2011.

[32] S. Akhshabi, S. Narayanaswamy, A. C. Begen, and C. Dovrolis, "An experimental evaluation of rate-adaptive video players over HTTP," *Signal Processing: Image Communication*, vol. 27, no. 4, pp. 271–287, 2012.