Enabling Premium Service for Streaming Video in Cellular Networks

Xing Xu[§], Ramesh Govindan[†], Ajay Mahimkar[‡], N.K. Shankaranarayanan[‡], Jia Wang[‡], Minlan Yu[¶]
Google[§] University of Southern California[†] AT&T[‡] Harvard University[¶]

***Abstract*—Streaming video applications require high bandwidth for desired quality of experience (QoE), and they are driving rapid growth of mobile data traffic in cellular networks. Currently, cellular networks provide best-effort services to most user data applications. When there is congestion at the base station, streaming video applications will experience degraded QoE. In this paper, we take a cellular service provider's perspective and propose a premium service for improving QoE of streaming video applications. We design and implement a network adaptation scheme called SHADE, which allocates limited transmission resources at the base station among applications smartly, by (i) selecting a candidate bitrate for each streaming video application, and (ii) maintaining the downlink throughput at this targeted bitrate for better QoE, while still using the Non-Guaranteed Bit Rate traffic class which is suitable for high bit-rate streaming video. We demonstrate that SHADE can achieve this with high network utilization and improve QoE for streaming video applications, and with bounded negative performance impact to other applications. Our extensive experiments show that SHADE can significantly improve three key streaming video application QoE metrics simultaneously (up to 10 times improvement), compared to current practice. We discuss how cellular carriers and equipment vendors can adopt SHADE without major changes to current cellular network implementation.**

I. Introduction

Cellular networks provide different Quality of Service (QoS) classes for traffic flows over the radio access network. 4G LTE networks have guaranteed bitrate (GBR) classes of traffic which guarantee radio resources, and these are used for realtime applications such as Voice over LTE with low to moderate bitrate needs. The bulk of IP data flows such as video and Web content are bursty with high bit-rate requirements, and these are suitable for the non-guaranteed bitrate (NGBR) class of service which do not have any specific bit rate guarantee. With this best-effort service, application throughputs can decrease during network congestion, which typically occurs when the total user traffic demand exceeds the base station capacity. Mobile traffic is growing rapidly and is driven by streaming video applications for which users need a good and consistent Quality of Experience (QoE). Any decrease in QoE leads to lower user engagement and lower revenue [7], [40]. While it is true that streaming video content systems have been designed to adapt to changing throughput, changes in cellular network scenarios are particularly dynamic and pose a special challenge. We also expect future applications, like 360-degree video, to have even more stringent QoE requirements. Therefore, there is an urgent need for better solutions for providing good QoE to streaming video in cellular networks.

One option to address this problem is for the network operators to increase capacity by building more base stations or acquiring more radio spectrum; but this solution is expensive and also requires significant time to deploy. Other options are to provide throughput guidance feedback [1] from the network to the end users for better client adaptation. In this paper, we take the perspective of the cellular network operator, and explore the ***idea of having the cellular network adapt its resources to provide a relatively better QoE for adaptive bit rate applications such as video streaming***. Our proposal is network-centric and does not require any changes to the user equipment or applications. Since network throughput conditions can vary significantly due to radio conditions and network load, we take the approach of creating a differentiated group of premium adaptive video user flows which need to be managed for consistent QoE. With such a separate admission-controlled group, the network has greater ability to manage applications and users requiring a higher level of QoE management. Our design approach is to minimize and constrain any negative impact to regular users. We also take a pragmatic approach for easier deployment, and thus focus on building upon widely deployed non-GBR schedulers.

In today's adaptive video streaming applications, the content is broken into chunks which contain several seconds of the original content, each of which are encoded at a few different bit rates. The client-side player monitors current downlink bandwidth, and uses that information to request the next chunk of the suitable bitrate. However, when transmission resources are limited, arbitrarily increasing application throughputs is a sub-optimal solution. By empirically analyzing the relationship between downlink throughput and application QoE, we observe that a stable downlink throughput maintained at one of the appropriately chosen bitrate candidates of the application can provide better QoE.

To manage limited transmission resources, current cellular systems use schedulers in the base station to allocate resources among applications in a fair and efficient manner. For our objective of handling a set of premium adaptive video streaming user flows, our solution uses a differentiated service model. We re-design the cellular network to be *content-aware*, where the infrastructure knows which users flows are premium, and also knows the set of pre-determined suitable bitrate candidates for each adaptive video application. For the rest of the paper, we use the term ***bearer*** to capture a group of user flows or applications. For example, streaming video applications map to a premium bearer. Current LTE schedulers support scheduling on a per-bearer level. Note that, we cannot use dedicated bearers with guaranteed bit rate (e.g., Voice over LTE uses QCI-1) because it is infeasible to support this for very high bit rates. Our focus is primarily to support non-GBR data traffic with differentiated bit rates.

In this paper, we describe a system called SHADE, which Stabilizes throughput at a relatively Higher downlink bitrate to provide better ADaptive streaming video applications with good Quality of Experience. Through extensive ns-3 simulations driven by real world data collected from a large

ISBN 978-3-903176-28-7

cellular service provider, we show that SHADE can improve QoE metrics (average bitrate, rebuffering ratio, and bitrate switches simultaneously) for the premium adaptive video applications when compared to other solutions, including a strong competitor that uses the same amount of resources to promote the premium flows and also tries to maintain premium video application's downlink throughput at one of the bitrate candidates. Compared to previous approaches, SHADE achieves more than 10 times reduction on both Rebuffering Ratio and Bitrate Switches, while also improving the Average Bitrate by 18%. We make the observation that maintaining throughput at one of the bitrate candidates can provide better QoE for adaptive bitrate applications (Section IV-A).

We make following contributions in this paper. The design and implementation of SHADE (Section III) includes (i) Creation of a set of premium adaptive video users with requirement for better and consistent QoE; (ii) A bitrate selection component that selects suitable bit rate levels to meet QoE metrics, and a process for minimizing the negative impact on the set of regular users (Section IV); (iii) A throughput maintenance component that maintains each user's downlink throughput at the targeted value using non-GBR traffic class, and provides high utilization at the same time, without major changes to the widely deployed current proportional fair scheduler (Section V). We conduct an extensive evaluation of SHADE based on real world data from a large cellular service provider (Section VI).

II. Background

Cellular LTE. Our solution description is in the context of resource sharing in the LTE downlink, but the ideas can be generalized to other cellular technologies as well, and also apply to a shared uplink. In the downlink, LTE divides radio spectrum resources into orthogonal sub-carriers, each of which has a bandwidth of 15 kHz. In the time domain, LTE has frames of 10 ms which are composed of 10 Transmission Time Intervals (TTIs), each of which have two slots of 0.5 ms duration. A set of twelve consecutive sub-carriers over the duration of one slot is called a Physical Resource Block (PRB), and this is the basic scheduling unit. The scheduler in the LTE eNodeB base station can assign each PRB to any user. Each LTE user periodically measures the channel condition and provides a Channel Quality Indicator (CQI) report. On a per-user basis, the base station uses the CQI to select a Modulation and Coding Scheme (MCS) used for the radio transmission, with a higher CQI value indicating a higher MCS value which has a higher bit rate and more efficient use of the PRB. The PRB scheduling algorithm must provide good efficiency (throughput to users) and is more likely to assign a PRB to a user with a higher CQI / MCS (bit rate); but the algorithm must also be fair to ensure that users with lower MCS get an appropriate share of resources.

The most widely deployed scheduler [9], [28] is the Proportional Fair (PF) scheduler [19], which provides a balance between efficiency and fairness. For each PRB and for each user i, the PF algorithm calculates two values: first, user i's achievable rate when using this PRB (denoted by r_i); and second, user i's average data rate over a time interval in the past (denoted by M_i). This PRB is then assigned to the user with the highest value of metric m:

$$m = argmax \frac{r_i}{M_i} \quad (1)$$

The PF scheduler ensures that resources are not wasted, and is a robust approach that is equivalent to maximizing the sum of the log of each user rate. In our solution, we build on the PF scheduler by adding a weight in the numerator to assign priority. In LTE networks, each user's profile and subscriber data are stored in the EPC (Evolved Packet Core), with the usage and charging policies stored and implemented by Policy elements in the EPC. In our solution, we require the EPC to include customer data to indicate whether a bearer is part of a premium service for which the network provides a higher QoE for adaptive video application flows.

Adaptive video streaming application. Today's adaptive video streaming technologies are mainly HTTP based adaptive streaming protocols. At the server-side, video content is encoded at *a few* different bitrates (usually 5-6 of them [3]). Each bitrate version is then broken into multiple chunks that each contains several seconds of the content. Chunks of different bitrates are aligned so that the player can smoothly switch to a different bitrate at the chunk granularity.

At the client, an *adaptive bitrate algorithm (ABR)* measures the recent available bandwidth and the buffered playback to determine a suitable bitrate for the next chunk to request. A higher throughput, or a larger play out buffer, drives ABR to request chunks with higher bitrate. In this paper, we assume that user's available bandwidth is bottlenecked by the cellular downlink throughput, and thus the chunk rate request can be managed by managing the downlink throughput.

Among several QoE metrics for adaptive video streaming applications, we consider three most important ones [8], which are *average bitrate*, *rebuffering ratio* and *bitrate switches*. Average bitrate is the time average of the different bitrates that were used for the particular content over some time interval. A higher average bitrate provides better quality. Rebuffering ratio is computed as the ratio of time spent while playout is interrupted (rebuffering) to the time for which playout is smooth. A high rebuffering ratio significantly degrades QoE. Bitrate switches counts the number of quality (bitrate) changes within a time interval. A higher value of bitrate switches tends to be more distractive, and lower values are preferred.

III. SHADE: A Premium Service

Our solution consists of the following components. (a) There is a premium group or bearer of adaptive video streaming flows which is differentiated from regular traffic flows and managed separately. Admission to the bearer is controlled, and based on available capacity, (b) There is a bitrate selection mechanism to select the initial bitrate which is to be maintained, (c) There is a mechanism to maintain the throughput of a premium bearer at the target rate, and (d) There is an aggregate mechanism to ensure that resources are available for regular bearers as network scenarios change, and the target bit rates of premium bearers adjusted as needed.

We assume that some set of users are provisioned with the capability of requesting and utilizing the premium adaptive video streaming application. Consider such an application requesting to be managed as a premium bearer. We assume

that the set of suitable bitrates is known for each such bearer. The user request includes CQI information which allows the base station to estimate the PRB resources needed. An admission control module with knowledge of the current workload determines whether the new application can be admitted as a premium bearer. We do not focus on proposing the best admission control scheme in this paper. Instead, we explore admission control choices with different degrees of conservativeness. If admitted, the bearer uses the special APN, and the system strives to maintain this bearer's downlink throughput at one of the bitrate candidates of the content. While the network only manages the traffic flows in the premium set, we refer to the ***premium bearer*** in the rest of the paper when it makes sense to do so.

A. Select target bitrate

Selecting the targeted bitrate for each premium bearer is challenging because only a limited number of PRBs are available to SHADE. SHADE cannot simply select the highest bit rate for all bearers, as it may not be feasible. There are two separate challenges: the first is to bound the negative impact to regular bearers; the second is to provide good overall QoE to premium bearers. To bound the negative impact to non-premium bearers, SHADE limits the aggregate resources that can be used by premium bearers. This upper bound can effectively limit the impact on regular bearers. For the second challenge, we need a mechanism to select a bitrate for each premium bearer to provide the best QoE to premium bearers using limited resources. SHADE should perform well on all three *competing* QoE metrics: we want to achieve high average bitrate, low rebuffering ratio, and stable bitrate switches at the same time. This is challenging because improving one QoE metric normally degrades other two metrics. In addition, SHADE has to adapt to network dynamics, e.g., changes on user's mobility, requirement, and channel condition, and frequently update the bitrate selection choices. This introduces another challenge of stabilizing bitrate selection choices over time. In our experiment, we find that stabilizing selected bitrates over time is a must for good QoE performance. We describe SHADE's solution to this bitrate selection problem in Section IV.

B. Maintain downlink throughput

Traditionally, bit rates are maintained by using a reservation based approach, like the Guaranteed Bit Rate (GBR) traffic class in LTE. Using this technique to maintain the premium bearer downlink throughput has two issues. First, it is not work-conserving. Reserved resources cannot be used by others when the owner does not need them. Thus, this approach is not suitable to maintain a high throughput, because reserved but not used resources can introduce too much inefficiency to the cellular system. Second, a new GBR scheme requires a relatively big change in the base stations. To simplify deployment and maintain backward compatibility, we design SHADE to minimize the changes to current cellular system.

SHADE builds upon the existing, widely-used Proportional Fair scheduler. Proportional Fair is a sharing based (not reservation based), and thus work-conserving scheduler. By building on top of Proportional Fair, SHADE keeps the important high efficiency property of the scheduler. Potentially, there are multiple modification methods that can achieve the same goal. However, to ease the deployment, SHADE only applies a small change, i.e., a weight parameter W_i for user i to the Proportional Fair metric of Equation 1:

$$m = argmax\mathbf{W_i} \cdot \frac{r_i}{M_i} \quad (2)$$

Section V describes in detail how SHADE uses this weight parameter to maintain the downlink throughput. Intuitively, a premium bearer with weight W is equivalent to W identical non-premium bearers with weight 1. Thus, by applying W, one bearer can potentially get roughly W times of transmission resources (PRBs), within other constraints. Then, to maintain downlink throughput, SHADE dynamically adapt the number of PRBs needed to achieve the targeted throughput.

IV. Selecting Target Bitrate

In this section, we discuss how SHADE selects bitrates for premium bearers to achieve good QoE with limited impact on non-premium bearers.

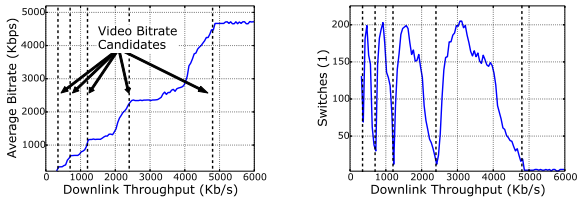
A. Video QoE in cellular networks

In cellular networks, QoE of streaming video applications are dominated by downlink throughput that user devices receive from the base station. First, when the downlink throughput is limited, video QoE can be poor even with a good ABR algorithm. Figure 1 shows our empirical observations that the average bitrate increases as the downlink throughput increases (Section VI-A). We found that the average bitrate increases in a step function pattern. It is more cost-effective to provide downlink throughput at one of the bitrate candidates shown in Figure 1a. For example, downlink throughputs of 2400Kbps and 3000Kbps yield similar average bitrates. 3000Kbps require more resources (PRBs) compared to 2400Kbps. If there is not enough resource at the base station to boost downlink throughput to 4800Kbps, providing 2400Kbps downlink throughput would be the most cost-effective. We also observe that rebuffering ratio decreases as the downlink throughput increases. When the downlink throughput is above 700Kbps, rebuffering ratio reduces to close to zero (detailed results are omitted due to space limit).

Second, unstable downlink throughput, which is common in cellular networks, can lead to fluctuation in average bitrate. Figure 1b shows that providing downlink throughput at one of the bitrate candidates yields minimal number of bitrate switches. This is because ABR algorithm determines the bitrate level of requesting video chunks based on the downlink throughput. Unstable downlink throughput between two bitrate candidates can cause the ABR algorithm to oscillate in bitrate levels of requesting chucks, and lead to poor video QoE.

In summary, to support good QoE, cellular service providers should *maintain streaming video applications downlink throughput at one of the application's bitrate candidates*. This requires cellular service providers to know bitrate candidates that are used by each streaming video applications *in advance*. Popular video content providers often use common bitrate candidates on their contents for mobile users as recommended by the HTTP Live Streaming (HLS) [3]. Alternatively, the bitrate candidates can also be obtained by parsing the application manifest[1]. In this work, we assume that

[1]When the client makes a request to a video content provider, the first response from the content provider is the manifest file, which describes available bitrates for the client to choose.



(a) Average bitrate (b) Bitrate switches

Figure 1: Video QoE vs downlink throughput (bitrate candidates: 350Kbps, 700Kbps, 1200Kbps, 2400Kbps, 4800Kbps).

the bitrate candidates for video applications is known using aforementioned methods.

B. Transmission resource allocation

Admission control. To limit the negative impact on non-premium bearers, SHADE limits the resources (i.e., PRBs) that can be used to by premium bearers to a fraction p of the total PRBs. In this paper, SHADE assumes the simplest pricing model, where all the premium bearers pay the same price. SHADE uses a simple threshold-based admission control mechanism, and the design of optimal pricing model and admission control algorithm are beyond the scope of this paper. SHADE accepts a prospective premium bearer only if some previously chosen threshold bitrate R_{AC} can be provided to all admitted premium bearers (including the new bearer), where R_{AC} is selected among bitrate candidates. The higher the value of R_{AC} is, the more conservative SHADE will be when it admits a new premium bearer.

Bitrate selection. SHADE selects a bitrate for each admitted premium bearer from its bitrate candidates. To provide better QoE to premium bearers, SHADE strives to maximize overall bitrates for premium bearers under the following three constraints. First, for a given premium bearer A_k, SHADE should provide at least bitrate $R_{k,1}$, where $R_{k,j}$ is the jth candidate bitrate for premium bearer A_k and $R_{k,1}$ is the lowest bitrate candidate for bearer A_k. Second, whenever possible, a higher bitrate should be provided to reduce the rebuffering ratio (e.g., higher than 700Kbps yielding zero rebuffering ratio). Third, and more importantly, bearer's channel conditions can change dramatically in cellular networks. SHADE needs to adapt bitrate selection accordingly. Existing bitrate optimization algorithms (e.g., Avis [12]) tend to favor bearers with good channel conditions, and thus, its bitrate selection may favor different bearers over time and lead to unstable bitrate selections. Because unstable bitrate selection can lead to poor QoE (Figure 1), it is important that SHADE stabilizes bitrate selection for premium bearers over time. Note that if there are spare resources (PRBs), SHADE will let a bearer burst at higher bit rates. In this paper, we primarily focus on the congestion scenarios at the base stations.

SHADE selects bitrate for each premium bearer independently, and allocates similar amount of PRBs to the bearer over time, regardless of other bearers' channel condition changes. There are many ways to determine how many fixed PRBs that each bearer gets initially (e.g., initial PRBs can be determined by prices model). Without loss of generality, SHADE allocates the *fair share*, which is $\frac{T}{S}p$ PRBs, for each premium bearer, where T is the total number of PRBs and S is the number of

Data: T = number of total PRBs; p = percentage of total PRBs that can be allocated to premium bearers; S = number of premium bearers; r_i = channel condition of each bearer A_i in terms of achievable rate per PRB ($i \in [1..S]$).
Result: Selected bitrate x_i for each bearer A_i .

```
1  PRB_premium = 0;
2  for i ∈ [1..S] do
       /* Assign bitrate by fair share: pT/S */
3      x_i = ClosestBitrateCandidate(r_i × pT/S);
4      PRB_premium = PRB_premium + x_i/r_i;
   /* Downgrade to use no more than pT PRBs */
5  x[] = Sort(A) /* Sort bearers in ascending
       order of channel condition r_i */
6  while PRB_premium > pT do
7      for i ∈ [1..S] do
8          if x_i ≥ R_{i,rb} then
9              PRB_premium =
                 DowngradeOneLevel(x_i, PRB_premium);
10             break;
11 while PRB_premium > pT do
12     for i ∈ [1..S] do
13         if x_i ≥ R_{i,1} then
14             PRB_premium =
                 DowngradeOneLevel(x_i, PRB_premium);
15             break;
16 while PRB_premium > pT do
17     for i ∈ [1..S] do
18         PRB_premium =
             DowngradeToNonPremium(x_i, PRB_premium);
           break;
```

Algorithm 1: SHADE's Bitrate Selection Algorithm.

premium bearers. For a given premium bearer, SHADE first uses $\frac{T}{S}p$ PRBs to estimate the achievable bitrate based on its current channel condition (denoted as "Bitrate of Fixed # of PRBs"). The fluctuation of the "Bitrate of Fixed # of PRBs" curve is due to the channel condition changes. SHADE then maps such bitrate to the closest bitrate candidate that requires minimum PRB adjustment (denoted as "Selected Bitrate").

For a given bearer A_i and its bitrate candidates $R_{i,1}$, ..., $R_{i,rb}$, ..., where bitrate $R_{i,rb}$ and above yield zero rebuffering ratio, SHADE first tries to select a bitrate that is higher than $R_{i,rb}$ to minimize rebuffering ratio. However, this may lead to use more than p portions of PRBs for all premium bearers. When SHADE uses more than p PRBs, SHADE keeps downgrading premium bearers' currently selected bitrate to a lower bitrate candidates to save PRBs, until the total used PRBs is below p.

Algorithm 1 shows SHADE bitrate selection process. In lines 1-4 SHADE maps premium bearer A_i to one of its bitrate candidates that is higher than $R_{i,rb}$, and in lines 5-19 SHADE downgrades mapped bitrate selections if needed in three steps: (i) SHADE keeps looking for premium bearer A_i whose selected bitrate is higher or equal to $R_{i,rb}$, and downgrades it to its next lower bitrate candidate (lines 6-10); (ii) When there is no such bearer, SHADE keeps looking for premium bearer A_i whose selected bitrate is higher or equal to $R_{i,1}$,

and downgrades it to its next lower bitrate candidate (lines 11-15); (iii) SHADE downgrades premium bearer A_i from $R_{i,1}$ to non-premium class (lines 16-19). The downgrading steps make sure that SHADE supports either at least $R_{i,rb}$ or at least $R_{i,1}$ to every premium bearer whenever possible. In addition, in each downgrading step, SHADE downgrades the bearer which has the worst channel condition first.

Limited impact on non-premium bearers. Assuming that there are N bearers, without SHADE, each bearer would get $\frac{T}{N}$ PRBs. With SHADE, each non-premium bearer would get $\frac{(1-p)}{(N-S)}T$ PRBs.

V. Maintain Downlink Throughput

In this section, we show how SHADE maintains downlink throughput for premium bearers by applying a weight parameter to the Proportional Fair scheduler.

A. Property of fair allocation

The Proportional Fair scheduler provides good efficiency as well as fairness. We observed in our experiments (Section VI-A) that the Proportional Fair scheduler can achieve up to 1.8x of average throughput compared to the Round Robin scheduler. The better efficiency provided by the Proportional Fair scheduler over the Round Robin scheduler is achieved by trading-off the fairness. We measure the fairness of the Proportional Fair scheduler in terms of number of PRBs allocated, and find that though the Proportional Fair yields poor fairness at short time intervals (e.g., < 1 second), it can achieve excellent fairness at longer time intervals (e.g., ≥ 1 second). For example, the Jain Fairness Index [18] is greater than 0.998 at time interval of 1 second.

Due to this fairness property, Proportional Fair scheduler's PRB allocation over a long enough time interval (e.g., ≥ 1 second) can be computed as follows. Each bearer will get the fair share, $\frac{T}{N}$ PRBs, where T is the total PRBs and N is the total number of bearers. Allocating more PRBs to a given premium bearer A than its fair share can be achieved by assigning a weight parameter W, where A will get roughly W times of its fair share. The number of PRBs allocated to bearer A will be:

$$\frac{W}{N+W-1}T \tag{3}$$

Because the total number of PRBs T is a known constant, we can tune parameter W in Equation 3 to control the amount of PRBs assigned to a given premium bearer.

B. Achieve targeted throughput for single premium bearer

SHADE achieves target throughput for a single premium bearer in two steps: First, SHADE determines the number of PRBs required to achieve the targeted throughput, at that time instant (Section V-B1). Second, SHADE determines the appropriate weight parameter W to obtain the required number of PRBs (Section V-B2).

1) Calculate required number of PRBs: Bearer achievable throughput per PRB (denoted as r) depends on its channel condition, which is measured by Modulation and Coding Scheme (MCS) Index (denoted by I_{MCS}). There is a one-to-one mapping (function f) from bearer's channel condition (I_{MCS}) to the achievable throughput per PRB: $r = f(I_{MCS})$. This mapping is given by Transport Block Size Index Table and Transport Block Size Table (Table 7.1.7.1-1 and Table 7.1.7.2.1-1 of [4]).

To achieve the targeted throughput (denoted by V) of this premium bearer, SHADE calculates the required number of PRBs (denoted by P) by: $P = \frac{V}{r} = \frac{V}{f(I_{MCS})}$. SHADE estimates premium bearer's MCS Index (I_{MCS}) based on the past MCS indices for this premium bearer.

2) Determine weight to obtain required PRBs: The next task is to calculate the weight parameter W for this premium bearer to be allocated with P PRBs. According to Equation 3, $P = \frac{W}{N+W-1}T$, we have

$$W = \frac{P(N-1)}{T-P} \tag{4}$$

Equation 4 assumes that all the bearers are *backlogged bearers*, (i.e., they require more resources than what they can get), and thus resources would be fairly distributed among all backlogged bearers. However, in reality, there can be *non-backlogged bearers* which require smaller number of PRBs than their fair shares. Let $Q(y)$ denote the required number of PRBs for a given non-backlogged bearer y, we should replace T and N in Equation 4's by T' and N', where

$$T' = T - \sum_{y \in Non-BackloggedBearers} Q(y) \tag{5}$$

$$N' = |BackloggedBearer| \tag{6}$$

SHADE needs to determine the subset of bearers that are non-backlogged. SHADE does so in iterations. At the very beginning, the fair share is $F = \frac{T}{N}$. SHADE labels bearers that require smaller than F PRBs as non-backlogged bearers. Then, F is updated accordingly using T' and N' in Equations 5 and 6. This iterative process keeps updating M', N' and F until there are no more new non-backlogged bearers (we call this condition as N' converged). SHADE's scheduler then computes $Q(y)$ based on the number of allocated PRBs to non-backlogged bearer in previous epochs. Algorithm 2 illustrates the process of determining the weight W for one premium bearer to achieve its targeted throughput.

C. Maintain throughput with network dynamics

Due to changing radio conditions and changing traffic demands, both the required PRBs $Q(\cdot)$ and the channel condition I_{MCS} keeps changing. SHADE needs to maintain throughput adapting to network and user dynamics over time. Intuitively, one could use a reactive approach, where an update happens upon detection of network/user dynamics. Instead, SHADE chooses to use a proactive approach where update happens periodically. We argue that network dynamics occurring at smaller time granularity can be handled by the streaming video applications, and hence SHADE focuses on dynamics that occur at larger time granularity. This would also help limit the update frequency at the base station and hence keeps the overhead low.

SHADE uses two different time intervals, MCS-Interval and Requirement-Interval, to update I_{MCS} and $Q(\cdot)$ ($Q(\cdot)$ also implies T and N) respectively. In particular, SHADE uses a shorter MCS-Interval to capture the channel condition variation quickly, and a longer Requirement-Interval to accurately estimate bearers' requirements. Note that, both

Function *Update-Weight()*

Data: T = total number of PRBs, N = number of bearers, V = target throughput for the premium bearer, I_{MCS} = channel condition, $Q(\cdot)$ = number of PRBs for non-backlogged bearers

Result: W = weight for the premium bearer

$T' = T,\ N' = N$;

while N' *has not converged* **do**

  $F = \frac{T'}{N'}$;

  for $x \in$ *Allbearers* **do**

    if y *is not non-backlogged and* $Q(y) < F$ **then**

      /* Non-backlogged */

      Label y as non-backlogged bearer;

      $T' = T' - Q(y)$;

      $N' = N' - 1$;

  /* Calculate weight. */

  $P = \frac{V}{f(I_{MCS})}$; $W = \frac{P(N'-1)}{(T'-P)}$;

Algorithm 2: Calculating weight parameter W for the premium bearer to achieve its targeted downlink throughput.

updates of MCS-Interval and Requirement-Interval will trigger Algorithm 2, the update of the weight parameter. The choices of using different update intervals are evaluated in Sections VI-C1 and VI-C2.

D. Support multiple premium bearers

We now extend Algorithm 2 to support multiple premium bearers, where SHADE needs to determine multiple weight parameters. To handle the dependency of the resource allocation among premium bearers, SHADE treat all premium bearers as a single virtual premium bearer. SHADE determines this virtual premium bearer's weight using Algorithm 2 first and then distributes this weight among premium bearers proportionally to their PRB requirements. Assume that there are S premium bearers, SHADE creates a virtual bearer $A_{virtual}$ by $P_{virtual} = P_1 + P_2 + P_3 + \cdots + P_S$ and $N_{virtual} = N - S + 1$. After Algorithm 2 returns the weight parameter $W_{virtual}$, SHADE calculates the i-th premium bearer's weight $W_i = W_{virtual} \cdot \frac{P_i}{P_{virtual}}$ (i = 1, 2, ...,S).

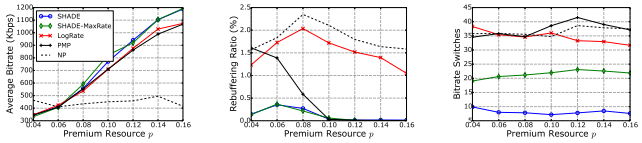
VI. Evaluation

We present our evaluation of SHADE using ns-3 simulator, real-world data from a large cellular service provider, and a streaming video simulator. We compare SHADE to other premium service baselines using three application QoE metrics (Average Bitrate, Rebuffering Ratio and Bitrate Switches) and demonstrate that SHADE outperforms all other schemes.

A. Methodology

We use the LTE module [34] of ns-3 simulator (ns-3.24.1) [2] to evaluate SHADE. We drive the simulation based on the base station location data and output TCP throughput traces for each application. We then feed the TCP throughput traces into a streaming video simulator to evaluate application QoE.

ns-3 simulation. We modify ns-3's Proportional Fair scheduler of the LTE module, by applying a weight parameter for each application and implementing bitrate selection and throughput maintenance. Our ns-3 simulation configures an evolved packet core (EPC) and a radio access network (RAN)



(a) Average Bitrate (b) Rebuffering Ratio (c) Bitrate Switches

Figure 2: QoE comparison with different premium resource percentage p and using R_1 as R_{AC} for admission control. Legends for figures (b) and (c) are the same as (a).

with multiple cell-sites. To represent a typical large cellular carrier LTE network, we use a 20 MHz channel bandwidth (to simulate the capacity of a 10MHz 2x2 MIMO system) with 70Mbits/s maximum throughput on the downlink. We also use RLC AM (Radio Link Control Acknowledged Mode) and Hybrid ARQ (HARQ). These configurations represent a typical setting of a large cellular carrier's LTE network. At the PHY layer, we model the propagation using the Friis transmission equation [33]. For each application, we configure one application server, and connect this server to the EPC using a high bandwidth link (100Gb/s) with a latency of 20ms. We configure applications' downlink traffic differently for different experiments and introduce uplink traffic (including TCP ACKs and data traffic) as interference. There are a total of 100 users per sector in the simulation and 1 bearer per user, and upto 10 among them can be premium bearers.

Cellular network data. We use the real locations of the cellular network to place cell-sites in the ns-3 simulator. We first place one cell-site consisting of three sectors at the center of the map, and then place two surrounding cell-sites as neighbors. We use sector-level traffic information from the cellular network to map the application demand in ns-3. We analyze traffic statistics over multiple time-intervals for multiple sectors in a metropolitan city and identify the peak hour traffic information. We use peak hour traffic to drive the demand simulation in ns-3. For most of our experiments, we only focus on applications served by one sector of the central cell-site, while other sectors/cell-sites are configured as interfering sectors.

Streaming video simulator. We use the TCP throughput trace as realtime bandwidth to perform a discrete event simulation of a player. The simulator chooses chunks from a set of candidate bitrates when adapting the bitrate. At the end of the simulation, it outputs three key QoE metrics: average bitrate, rebuffering ratio, and bitrate Switches. It currently allows choosing from three streaming video algorithms: a state of the art control theoretic ABR [40], a buffer based ABR [17], and a rate based ABR [21]. We use the control theoretic ABR [40] algorithm in the simulator for bitrate adaptation in our experiments. We used video content with five rates as shown in Figure 1.

B. QoE performance comparison

We first compare the overall QoE performance of SHADE with four competing baselines - non-premium (NP), Paris Metro Pricing (PMP), LogRate and SHADE-MaxRate. We refer to them below as five competitors:

1. **Non-Premium (NP)**: A baseline that all the premium bearers will be admitted as non-premium bearer, i.e., when there is no premium service.

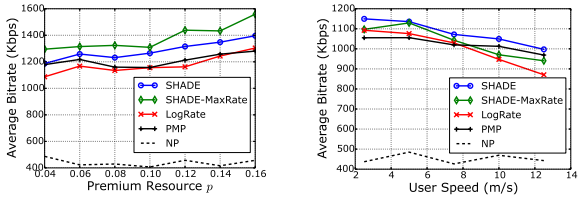
2. **Paris Metro Pricing (PMP)**: PMP [10] is a premium service that consists of two classes of identical cars with only ticket price difference. In PMP, we create the "1st class car" by reserving p of the total PRBs for premium bearers. Each admitted premium bearer will get the same amount of PRBs ($\frac{1}{S}p$), where S is the number of premium bearers.
3. **LogRate**: LogRate (a premium service) maintains downlink throughput at one of the bitrate candidates. LogRate represents Avis [12]'s bitrate selection algorithm, which maximizes the sum of the logarithm of selected bitrate.
4. **SHADE**: SHADE leverages the *stability property* and has each bearer select its bitrate independently.
5. **SHADE-MaxRate**: Similar to SHADE, SHADE-MaxRate provides a higher bitrate first to every premium bearer when possible. The difference is that for the rest of PRBs, SHADE-MaxRate is optimization based and maximizes the sum of selected bitrates among bearers.

We use the same admission control for all the competitors, and compare their overall performance using three QoE metrics. We assume that all the premium bearers share the same set of bitrate candidates $(R_1, R_2, ..., R_5)$ = (350, 700, 1200, 2400, 4800)Kbps. Each QoE metric for a competitor is calculated as the average of that metric across all the admitted premium bearers. In addition, we add a fourth metric, *Downgrade Fraction*, which is the percentage of duration that one premium bearer has been downgraded to a non-premium class. A good premium service should have the minimum value for the Downgrade Fraction. Note that, a competitor that downgrades more bearers to a non-premium class has the benefit of allocating PRBs among fewer premium bearers for better QoE performance. Downgrade Fraction helps us realize this and makes fair comparison among competitors.

1) Premium resource reservation: Figure 2 compares all the competitors with varying degrees of premium resource reservation. p captures the percentage of PRBs available for premium services. We observe that higher the value of p, all the premium services, except NP, achieve higher average bitrate (Figure 2a). This demonstrates the benefit of using premium service to improve bearer QoE. We also observe that SHADE and SHADE-MaxRate achieve the best average bitrate (up to 18% improvement compared to PMP). This performance gain comes from leveraging PRBs more efficiently.

Figure 2b shows that SHADE and SHADE-MaxRate provide the lowest rebuffering ratio. This is due to two reasons: (i) they both strive to not downgrade bearer to non-premium class, i.e., providing at least R_1 to everyone; (ii) they both strive to firstly provide R_2 when possible, which introduces 0 Rebuffering Ratio, a much better result than R_1. We observe that SHADE and SHADE-MaxRate achieves similar Downgrade Fraction. However, LogRate downgrades bearers more often. Providing minimum premium service makes sure that the Rebuffering Ratio would not be so bad, but bearers who have been downgraded to non-premium class have to compete resources with non-premium bearers and receive poor application QoE (LogRate downgrades so much and thus its Rebuffering Ratio shows some correlation with NP). The second reason is that SHADE and SHADE-MaxRate strive to firstly provide R_2 when possible, which introduces 0 Rebuffering Ratio, a much better result than R_1.

Figure 2c shows that NP and PMP have the highest number of bitrate switches. LogRate also makes many more bitrate switches because it does not satisfy the stability property that the bearers should pick their bitrates independently and although its optimization objective penalizes switches, it uses a greedy approach to round the solution of such optimization to bitrate candidates. This rounding process is very sensitive to bearers' channel conditions and introduces switches again. As a result, we see that SHADE-MaxRate can reduce bitrate switches by more than 50%. Since SHADE achieves better stability, it further reduces another 50% from SHADE-MaxRate. Thus, SHADE outperforms PMP and LogRate on all metrics.



(a) R_3 as R_{AC} (b) R_1 as R_{AC}

Figure 3: Average bit rate comparison.

2) Admission control with different conservativeness: In Figure 3a, we use R_3 as R_{AC} to admit bearers. It is a conservative admission control process as approximately it only admits $\frac{1}{3}$ of bearers compared to using R_1 (as R_1 is 350Kbps and R_3 is 1200Kbps). By admitting fewer bearers, each bearer can use more PRBs, and thus we see improved performance on average bitrate (Figure 3a) and rebuffering ratio. The competitors perform similarly for bitrate switches compared to the less conservative admission control case, except for LogRate. The reason is that when there are more PRBs, it is easier for LogRate to stabilize each bearer's bitrate selection as every bearer can get a good bitrate. However, when PRBs are scarce, LogRate keeps moving PRBs to bearers with good channel conditions, and leads to unstable bitrate selections. Interestingly, in Figure 3a, SHADE-MaxRate outperforms SHADE on Average Bitrate significantly, because it uses PRBs in a more efficient way by allocating more PRBs to good channel condition bearers, while SHADE gives every bearer similar amount of PRBs. This benefit becomes more significant when admission control is conservative. For one case, we find that SHADE selects at least R_3 to every bearer. On the other hand, SHADE-MaxRate gives R_2 to every one and then uses the rest of PRBs to promote 2 bearers with best channel conditions to R_5.

3) Channel condition changes: In Figure 3b, we present QoE comparison results by varying the degree of channel condition. We achieve this by varying the bearer's speed. Higher bearer speed introduces greater channel condition changes.[2] In Figure 3b, we observe decreasing average bitrate for SHADE-MaxRate and LogRate: when the speed is faster than 10m/s, both SHADE-MaxRate and LogRate perform worse than PMP. The reason is that when bearer's channel condition changes faster, it is more difficult to stabilize bitrate selection. Therefore, we see an increasing bitrate switches for all the competitors. We see that SHADE-MaxRate's bitrate switches is approaching PMP (LogRate is even worse), indicating poor stability of bitrate selection. On the other hand, SHADE's bitrate switches is still significantly better than PMP (56%

[2]We observe similar results when keeping bearer speed the same but shrinking the cell size to vary the degree of channel condition variation.

better for the worst case), which also benefits average bitrate performance. This result illustrates a significant benefit of stability when channel condition changes faster.

4) User mobility: Handover occurs when a mobile user changes the serving base another to another. During handover, user experiences poor performance for a few seconds. ABR usually maintains buffer of minutes of content [40], which can absorb the interruption due to handover. In our experiments, we have not observed significant QoE degradation due to introduced handovers. However, SHADE and SHADE-MaxRate react to handovers quickly. They have the benefits of quickly downgrading bearers who experience handovers to non-premium class. By doing so, PRBs will be used efficiently on other users instead of being wasted by handover users.

C. Throughput maintenance

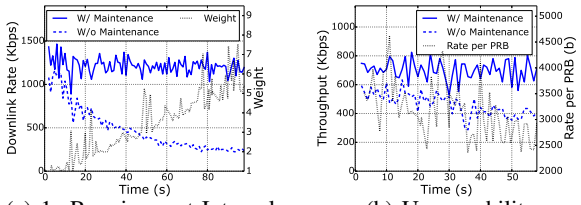
We now evaluate the throughput maintenance performance of SHADE under varying channel conditions, user dynamics and multiple premium bearers.

1) Channel condition variations: Our goal in SHADE is to react quickly to changing channel conditions and maintain the long term average downlink throughput with small variation. MCS index estimation does this by capturing the MCS indices in the past and reacting appropriately. We observe that a naive estimator that averages MCS indices on all PRBs in the previous time interval yields poor estimation compared to averaging across past *assigned PRBs*. We thus use the average of past assigned PRBs for a bearer to estimate the MCS index for the same bearer in SHADE. Choosing MCS-interval is also important to ensure faster reaction to the channel condition variations. We see that they all maintain the mean of the throughput at 1200Kbps, but using a MCS-Interval of 0.1s can reduce the standard deviation from 195Kbps to 85Kbps. We choose to use MCS-Interval of 0.1s for SHADE.

2) User dynamics: Users join and leave the network, and their requirements change all the time. These changes affect the requirement estimation, $Q(\cdot)$, of Algorithm 2. In addition, mobile users with premium bearers have varying channel condition, that in turn affects their PRB requirements for the same targeted downlink rate. SHADE updates $Q(\cdot)$ of non-premium bearers and required number of PRBs p for premium bearers for throughput maintenance, according to Algorithm 2.

Analogous to MCS indices estimation, we use the user requirement in the previous time interval to estimate for the next time interval, and we then determine the Requirement-Interval. A very high value for Requirement-Interval implies SHADE cannot capture requirement changes instantly and a very low value implies that SHADE scheduler cannot treat bearers fairly. We simulate a simple scenario with different number of bearers to vary $Q(\cdot)$, and evaluate the effect of using different Requirement-Intervals. We configure 45 bearers for the central sector. At the very beginning, there are only 5 bearers, including one premium bearer. Then, for the rest of bearers, each bearer joins the network after the previous one with a randomly chosen delay, according to a uniform distribution between 1 and 5 seconds. At the very beginning, we do not need to tune the weight as the premium bearer's throughput is higher than the targeted value, but with more bearers joining the network, we need to apply higher and higher weight for this premium bearer.

Figure 4a shows maintained downlink throughput of the best Requirement-Interval setting of 1s. Without throughput



(a) 1s Requirement-Interval (b) User mobility

Figure 4: Downlink rate, weights and rate per PRB.

maintenance, this bearer's throughput keeps going down as there are more bearers joining the network. SHADE with throughput maintenance reacts to the requirement changes and increases the weight to maintain the same targeted throughput. We observe that a low throughput point always triggers a high weight to compensate for the channel condition degradation. Finally, we model user mobility using random walk. To generate significant channel condition degradation, we show a case where a premium bearer is moving toward a neighboring base station with speed of 10m/s. In Figure 4b, we see that the throughput is maintained at 700Kbps while the channel condition is degrading continuously due to user mobility.

3) Multiple premium bearers: We explore the throughput maintenance for the case of multiple premium bearers, with all network dynamics We see that throughputs of three co-located premium bearers can be maintained at three different targets (1500Kbps, 1200Kbps, 800Kbps). Without throughput maintenance, they receive much lower throughput. In particular, we significantly increase the requirement at 30 seconds by starting a group of backlogged bearers and increasing existing bearers' requirements simultaneously. We see that without throughput maintenance, their throughput dropped by 40%. The maintained throughput case is affected by this too. However it can detect this sudden change and recover quickly.

VII. Related Work

Adaptive bitrate application. There has been a significant research literature on streaming multimedia contents with variable bandwidth conditions for a long time [35], [27], Pytheas [22], Pensieve [31]. Recent video streaming industry solutions converged on ABR based solutions, where videos are pre-encoded in different qualities, and the client makes the adaptation decision by requesting video chunks of different qualities from the video server. Researchers have proposed solutions to improve ABR based video performance from different perspectives. For example, [14], [20] focus on picking the best CDN to serve users; [17], [37], [40] propose better ABR algorithms; [5], [16], [21] studies the interaction between video player and TCP; CQIC [30] leverages user's channel condition to improve video's sending rate; QAVA [11] controls the video server and delivers appropriate video bitrate to user to not exceed her monthly data quota; AVIS [12] uses traffic shaper to achieve better fairness for video users; AGBR [39] also focuses on allocating resources among competing video users, but it is not QoE aware. Our approach is complementary to above. SHADE takes a network-centric approach and *requires no modifications to video applications or user devices*. SHADE requires minimal changes to the current cellular infrastructure, and is thus more easily deployable.

Scheduler and resource allocation. Max Rate [38], Round Robin [13] and Proportional Fair [19] are popular resource

schedulers for wireless systems. Some Proportional Fair based schedulers [24], [36] apply a weight parameter to Equation 1 to improve certain properties of scheduling, e.g., reduce latency, or reduce queue length. SHADE uses the same technique; however, for a different goal of treating the user differently and eventually maintaining users' downlink throughput. Schedulers for network virtualization [25], [26] mainly focus on allocating the right amount of resources to each slice, while SHADE controls per user resource for throughput maintenance. [23] prioritizes on the key frame of the video among users; [29] takes the deadlines of video packets into account. These content-aware, cross-layer techniques introduce significant complexity and modifications to current cellular system. On the contrary, though SHADE is also application-aware, it relies on easily accessible adaptive bitrate application information, and thus introduces minimal complexity and changes. [6] proposes weighted proportional fair scheduling in LTE networks for both non-GBR as well as GBR traffic; however SHADE primarily focuses on non-GBR.

Differentiated service. There are proposals that use different pricing schemes to enhance user performance. For example, [15] alleviates congestion by implementing a time-dependent pricing scheme to allow users defer their delay tolerant traffic to save money. [10], [32] discuss using Paris Metro Pricing scheme to support a differentiated digital service. Compared to these works, SHADE consists of a novel resource scheduling technique to maintain video users' downlink throughput.

VIII. Conclusions

In this paper, we propose a premium service for video users in cellular networks to achieve better QoE. To enable this, we describe SHADE, a new network adaptation scheme that outperforms competing schemes in improving QoE for streaming video applications. SHADE achieves better QoE by allocating more transmission resources to premium bearers to support higher downlink throughputs, and also maintaining their downlink throughputs at one of the content bitrate candidates. SHADE employs a bitrate selection component that selects bitrate for each premium bearer. It smartly leverages limited transmission resources to maximize overall QoE among premium bearers. A throughput maintenance component then maintains each premium bearer's downlink throughput at the targeted bitrate value, by dynamically adjusting the weight on the widely deployed Proportional Fair scheduler. Through extensive ns-3 simulations for a realistic LTE system, we show that SHADE's throughput maintenance component performs well under changes in radio and traffic conditions.

References

[1] Mobile throughput guidance inband signaling protocol internet-draft. https://tools.ietf.org/html/draft-flinck-mobile-throughput-guidance-04.

[2] ns-3. Available at https://www.nsnam.org/.

[3] Technical Note TN2224: Best Practices for Creating and Deploying HTTP Live Streaming Media for the iPhone and iPad. Available at https://developer.apple.com/library/ios/technotes/tn2224/_index.html.

[4] 3rd Generation Partnership Project, TS 36.213 V9.2.0, June 2010.

[5] S. Akhshabi, L. Anantakrishnan, C. Dovrolis, and A. C. Begen. Server-based traffic shaping for stabilizing oscillating adaptive streaming players. In *ACM NOSSDAV*, 2013.

[6] S. ALi, M. Zeeshan, and A. Naveed. A capacity and minimum guarantee-based service class-oriented scheduler for lte networks. In *EURASIP Journal on Wireless Communications and Networking*, 2013.

[7] A. Balachandran, V. Sekar, A. Akella, S. Seshan, I. Stoica, and H. Zhang. Developing a predictive model of quality of experience for internet video. In *ACM SIGCOMM*, 2013.

[8] A. Balachandran, V. Sekar, A. Akella, S. Seshan, I. Stoica, and H. Zhang. Developing a predictive model of quality of experience for internet video. In *ACM SIGCOMM*, 2013.

[9] T. Bu, L. Li, and R. Ramjee. Generalized proportional fair scheduling in third generation wireless data networks. In *IEEE INFOCOM*.

[10] C.-K. Chau, Q. Wang, and D.-M. Chiu. Economic viability of paris metro pricing for digital services. *ACM Trans. Internet Technol.*, 14(2-3):12:1–12:21, Oct. 2014.

[11] J. Chen, A. Ghosh, J. Magutt, and M. Chiang. Qava: Quota aware video adaptation. In *ACM CoNEXT*, 2012.

[12] J. Chen, R. Mahindra, M. A. Khojastepour, S. Rangarajan, and M. Chiang. A scheduling framework for adaptive video delivery over cellular networks. In *ACM MOBICOM* , 2013.

[13] E. Dahlman, S. Parkvall, J. Skold, and P. Beming. *3G Evolution, Second Edition: HSPA and LTE for Mobile Broadband*. Academic Press, 2 edition, 2008.

[14] A. Ganjam, F. Siddiqui, J. Zhan, X. Liu, I. Stoica, J. Jiang, V. Sekar, and H. Zhang. C3: Internet-scale control plane for video quality optimization. In *USENIX NSDI*, 2015.

[15] S. Ha, S. Sen, C. Joe-Wong, Y. Im, and M. Chiang. Tube: Time-dependent pricing for mobile data. In *ACM SIGCOMM*, 2012.

[16] T.-Y. Huang, N. Handigol, B. Heller, N. McKeown, and R. Johari. Confused, timid, and unstable: Picking a video streaming rate is hard. In *ACM IMC*, 2012.

[17] 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. *SIGCOMM CCR*, 44(4):187–198, Aug. 2014.

[18] R. Jain, A. Durresi, and G. Babic. Throughput fairness index: An explanation, 1999.

[19] A. Jalali, R. Padovani, and R. Pankaj. Data throughput of cdma-hdr a high efficiency-high data rate personal communication wireless system. In *VTC 2000-Spring Tokyo 51st*, 2000.

[20] J. Jiang, V. Sekar, H. Milner, D. Shepherd, I. Stoica, and H. Zhang. Cfa: A practical prediction system for video qoe optimization. In *USENIX NSDI*, 2016.

[21] J. Jiang, V. Sekar, and H. Zhang. Improving fairness, efficiency, and stability in http-based adaptive video streaming with festive. In *ACM CoNEXT*, 2012.

[22] J. Jiang, S. Sun, V. Sekar, and H. Zhang. Pytheas: Enabling data-driven quality of experience optimization using group-based exploration-exploitation. In *USENIX NSDI*, 2017.

[23] N. Khan, M. G. Martini, and D. Staehle. Opportunistic proportional fair downlink scheduling for scalable video transmission over lte systems. In *IEEE VTC Fall*, 2013.

[24] K. Khawam, D. Kofman, and E. Altman. The weighted proportional fair scheduler. In *QShine*, 2006.

[25] R. Kokku, R. Mahindra, H. Zhang, and S. Rangarajan. Nvs: A substrate for virtualizing wireless resources in cellular networks. *IEEE/ACM Transactions on Networking*, 20(5):1333–1346, 2012.

[26] R. Kokku, R. Mahindra, H. Zhang, and S. Rangarajan. Cellslice: Cellular wireless resource slicing for active ran sharing. In *COMSNETS*, 2013.

[27] T. V. Lakshman, A. Ortega, and A. R. Reibman. Vbr video: tradeoffs and potentials. *Proceedings of the IEEE*, 86(5):952–973, 1998.

[28] S. B. Lee, I. Pefkianakis, A. Meyerson, S. Xu, and S. Lu. Proportional fair frequency-domain packet scheduling for 3gpp lte uplink. In *IEEE INFOCOM*, pages 2611–2615, 2009.

[29] Q. Liu, Z. Zou, and C. W. Chen. Qos-driven and fair downlink scheduling for video streaming over lte networks with deadline and hard hand-off. In *IEEE Multimedia and Expo*, 2012.

[30] F. Lu, H. Du, A. Jain, G. M. Voelker, A. C. Snoeren, and A. Terzis. Cqic: Revisiting cross-layer congestion control for cellular networks. In *HotMobile*, pages 45–50, New York, NY, USA, 2015. ACM.

[31] H. Mao, R. Netravali, and M. Alizadeh. Neural adaptive video streaming with pensieve. In *ACM SIGCOMM*, 2017.

[32] A. Odlyzko. Paris metro pricing for the internet. In *ACM EC*, 1999.

[33] J. Parsons. *The mobile radio propagation channel*. Halsted Press, 1992.

[34] G. Piro, N. Baldo, and M. Miozzo. An lte module for the ns-3 network simulator. In *ICST SIMUTools*, 2011.

[35] M. v. d. Schaar and P. A. Chou. *Multimedia over IP and Wireless Networks: Compression, Networking, and Systems*. Academic Press, Inc., 2007.

[36] S. Shakkottai and A. L. Stolyar. Scheduling algorithms for a mixture of real-time and non-real-time data in hdr. In *ITC*, pages 793–804, 2017.

[37] K. Spiteri, R. Urgaonkar, and R. K. Sitaraman. BOLA: near-optimal bitrate adaptation for online videos. *CoRR*, abs/1601.06748, 2016.

[38] B. S. Tsybakov. File transmission over wireless fast fading downlink. *IEEE Transactions on Information Theory*, 48(8):2323–2337, Aug 2002.

[39] D. D. Vleeschauwer, H. Viswanathan, A. Beck, S. Benno, G. Li, and R. Miller. Optimization of http adaptive streaming over mobile cellular networks. In *IEEE INFOCOM*, 2013.

[40] X. Yin, A. Jindal, V. Sekar, and B. Sinopoli. A control-theoretic approach for dynamic adaptive video streaming over http. In *ACM SIGCOMM*, 2015.