Adaptive Resource Allocation for QoE-Aware Mobile Communication Networks

Mirghiasaldin Seyedebrahimi, Xiao-Hong Peng
School of Engineering and Applied Science, Aston University, Birmingham B4 7ET, UK.
{seyedebm, x-h.peng}@aston.ac.uk

Rob Harrison
Alchemy Wireless Ltd, Birmingham B14 6DT, UK.
rharrison@alchemy-wireless.com

Abstract—A real-time adaptive resource allocation algorithm considering the end user's Quality of Experience (QoE) in the context of video streaming service is presented in this work. An objective no-reference quality metric, namely Pause Intensity (PI), is used to control the priority of resource allocation to users during the scheduling process. An online adjustment has been introduced to adaptively set the scheduler's parameter and maintain a desired trade-off between fairness and efficiency. The correlation between the data rates (i.e. video code rates) demanded by users and the data rates allocated by the scheduler is taken into account as well. The final allocated rates are determined based on the channel status, the distribution of PI values among users, and the scheduling policy adopted. Furthermore, since the user's capability varies as the environment conditions change, the rate adaptation mechanism for video streaming is considered and its interaction with the scheduling process under the same PI metric is studied. The feasibility of implementing this algorithm is examined and the result is compared with the most commonly existing scheduling methods.

Keywords— Resource allocation; QoE; scheduling; adaptive video streaming; fairness; efficiency; 3GPP-LTE

I. INTRODUCTION

In spite of the extended capability of the modern communication technologies that support a wide range of communication services, ensuring a high level of quality of service (QoS) or quality of experience (QoE) for end users remains to be a big challenge for network operators and service providers. This problem is further intensified by the growing demands for video streaming over mobile smartphones and tablets due to the limitations inherited in wireless and mobile communications environments.

Maintaining a good balance between the quality of the video and the resource requirement is one of the main hindrances in video streaming services. However, it is generally possible to compromise on the quality of the video for less required resource dedication. This is especially desirable in the case of wireless communications with scarce spectrum and high demand for mobile video services [1]. The current adaptive streaming service is an example of handling this trade-off, where multiple versions of the same video content with different video code rates are made available for different user conditions and requirements.

In the 3GPP-DASH (Dynamic Adaptive Streaming over HTTP) standard [2], clients choose the code rates of the video content from the server (client-pull), without the intervention of the intermediate unit of the network, e.g. the base station in mobile networks. Collaboration between the base station and the either side of an end-to-end video streaming system (server or client) can enhance the experience of the client being served with quality. But this may entail extra information exchange among them and does not comply with the idea of the independent-client based adaptive service such as DASH. Furthermore, it may require additional processing overhead and standardization amendment which practically can be a limitation for the implementation of this idea.

To tackle this issue, a quality of experience (QoE) driven resource allocation scheme with scheduling algorithms for the last-mile scheduler is proposed in this work based on a no-reference packet based video quality metric, i.e. Pause Intensity (PI) [3]. This metric takes account of user's video code rate (required data rate) and network performance (throughput), which can realistically characterize the demand-supply relationship of video streaming services [4]. The proposed scheme provides the capability of online adjustment of system efficiency, fairness and correlation between the required and allocated data rates. The PI metric can be easily assessed by the scheduler on the network side without requiring extra information exchanged between users and the network. In addition, PI can also play a role in shaping the distribution of video code rates for adaptive video streaming and reaching the required level of QoE for clients. The proposed algorithms are examined in the context of 3GPP-LTE (Long Term Evolution [5]) for both adaptive and non-adaptive video streaming scenarios, complied with the 3GPP and related standards for streaming services[6], [7].

The rest of the paper is organized as follows: The background and related works are explained in Section II. The model description, proposed optimization system and its implementation algorithm are presented in Section III. The simulation results and analysis are discussed in Section IV and finally the conclusion is provided in Section V.

II. BACKGROUND AND RELATED WORKS

There are two quality related aspects of a video service that can be compromised for less resource allocation during the communication process: the fidelity based quality of the image and the continuity of the service. Actually these two aspects are related to each other in terms of sharing the same amount of resource. For example, the discontinuity of playback is more likely during a video service with a higher level of visual quality (given a limited amount of bandwidth) [8]. Both fidelity and continuity based quality issues are

related to QoE and can be generally assessed through subjective metrics, such as the mostly used Mean Opinion Score (MOS) [9].

Due to the subjective nature of QoE and the diversity of the related applications, a unified objective metric for QoE is still not available. Many variations of PSNR (Peak Signal to Noise Ratio) and SSIM (Structural SIMilarity) are used as video quality assessment tools to evaluate the performance of proposed solutions [10]. The occupancy of the playback buffer usually forms a base for the evaluation of continuity in video streaming. The occupancy level, probability of buffer underrun, initial delay and pause durations, pause frequency and jitter are some of the metrics which have been used to quantify the continuity of a video service [11], [12].

A communication model using the visual quality assessment metrics (e.g. PSNR) usually needs the output of the decoder and the original video reference. This type of metrics is more suitable for performance analysis rather than an online quality assessment process [13], [14]. In contrast, the continuity based quality can be evaluated without the need of the original reference and decoder output. However, most of the continuity based metrics mentioned above don not have a good relation with subjective QoE metrics such as MOS. Furthermore, a QoE-driven solution normally acquires extra information sent from the user to the network, which leads to additional control overhead or standard amendments.

Pause Intensity (PI), as described in Section I, is a reference-less metric for continuity assessment and takes both pause duration and pause frequency into account. It is highly correlated with the subjective QoE metric, MOS, which is content independent, as shown in Fig. 1. The PI value, $\mathrm{PI}(\eta,\lambda)$, can also be determined by both network performance, i.e. throughput η, and the required data rate (video code rate, λ) per user, which is expressed by [3]:

$$\mathrm{PI}(\eta,\lambda) = 1 - \frac{\eta}{\lambda} \qquad (1)$$

In a non-recorded streaming scenario η is always less than or equal to λ, i.e. $\eta \leq \lambda$ *and* $0 \leq PI \leq 1$. The description of the PI model, buffer paly-pause characterization and associated subjective tests are provided in [3] and the PI metric has been applied in the context of 3GPP-LTE in [4].

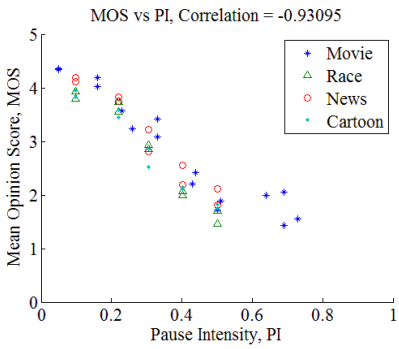

Fig. 1. Correlation between MOS and PI produced from subjective testing using different video contents [3].

PI is an objective representation of QoE which, as it will be shown in Section III, can be evaluated locally both on the user side and on the network side without additional information exchange and hence no extra overhead for this purpose.

The proposed algorithm in this paper is examined in the context of LTE/4G. The last mile resource allocation function in eNodeB (i.e. LTE's base station) plays the main role in the proposed idea alongside the link adaptation and channel status control, which will be detailed in the next section. Although the exact resource allocation policy in LTE has not been defined by 3GPP standards, the state of the art solutions are usually based on the general rationality of the resource allocation in mobile communication systems. Subsequently, the main parameters used to characterize a scheduler are the efficiency of the system as a whole and the fairness of the scheduler to each user.

The most common resource allocation algorithms used to make an efficient or fair scheduler are *best-CQI*, *proportional fair* and *MaxMin throughput* schedulers [15], [16]. The CQI (Channel Quality Indication) in LTE is a feedback from the user to the base station to indicate the capability of the user for using the allocated resources, which is related to the modulation order and the channel coding (or code) rate. The *best-CQI* scheduler (also known as *maxC/I*) is focused on the efficiency of the system by targeting the users with the highest capability in each round of the allocation. In contrast, a scheduler such as *MaxMin throughput* scheduler achieves a high degree of fairness by allocating almost the equal resource to all users regardless of their CQIs.

A balanced allocation to each user can be achieved through the consideration of the efficiency of each user alongside the history of the allocation to that user. For example, in the *proportional fair* scheduler [17] a user with higher efficiency (i.e. better channel quality) will be served more than the users with poorer channel quality. Meanwhile, the comparison of the total allocation to all users will prevent the scheduler from excessive allocation to that user and force the scheduler to serve other users as well. Later in Section IV, the performance of our proposed algorithm will be compared with these common scheduling methods.

In the next section the analytical model for the proposed QoE driven scheduler together with an implementation algorithm will be derived and explained.

III. Model Description

A. Resource allocation assumptions

LTE provides resources through a combination of Orthogonal Frequency Division Multiple Access (OFDMA) and Adaptive Modulation and Coding (AMC) techniques in a bandwidth range from 1.4MHz to 20MHz. A resource allocation unit in LTE is defined as a 'Resource Block (RB)' in a two dimensional time-frequency grid. Each RB is a 180 KHz of bandwidth allocated for one time slot of 0.5ms. Allocation will remain the same in the next time slot which

creates a 1ms Transmission Time Interval (TTI) for each transmission process.

Each client provides an evaluation of its channel status, i.e. signal-to-noise ratio (SNR) across the N_{RB} predefined resource blocks:

$$SNR \in \{SNR_{min}, \dots, SNR_{max}\}^{1\times N_{RB}} \quad (2)$$

A channel quality indicator (CQI) feedback will be generated based on this evaluation and the capability of the client's device:

$$CQI \in \{1,2, \dots, CQI_{max}\}^{1\times N_{RB}} \quad (3)$$

The value of CQI can be a result of a linear fitting of SNR value(s) or searching through a lookup table which reflects the capability of user's device with regard to different modulation and (channel) code rates (i.e. MCS) given the SNR values. CQI suggests a range of modulation and code rates for which at least 90% accuracy will be achievable at the receiver. Given the selected modulation and code rate (based on the CQI values) and the allocated resources, r_k, the total allocated data rate to user k ($k=1$ to N_{UE}) in the i^{th} round of the allocation, R_k^i, can be calculated as:

$$\begin{cases} R_k^i = C_k^T . r_k \\ C_k = C_k\big(CQI(SNR)\big) \in \mathbb{R}_{>0}{}^{1\times N_{RB}} \end{cases} \quad (4)$$

where C_k is the vector of the achievable capacities in the resource blocks for user k, given the corresponding CQI values. r_k is the vector of the allocation defined as follow:

$$\begin{cases} r_k = [r_{k,1}, r_{k,2}, .., r_{k,N_{RB}}]^T \in \{0,1\}^{N_{RB}}, \\ r_i . r_j^T = 0 \ \ \forall\, i \neq j\,, \\ \sum_{k=1}^{N_{UE}} \|r_k\|_1 \leq N_{RB} \end{cases} \quad (5)$$

$r_{k,l}=1$ indicates the allocation of the l^{th} resource block to user k and $r_{k,l}=0$ otherwise. Hence: 1) each resource block is supposed to be allocated just to one user; 2) all resource blocks can be allocated to one user; and 3) allocated resources in each round can be less than the total number of available resources (i.e. some resources may remain unused in each round).

The weighted average of the allocated data rate after the i^{th} allocation round can be assessed as:

$$\overline{R_k^i} = \left(1 - \frac{1}{t_w}\right)\overline{R_k^{i-1}} + \frac{1}{t_w} R_k^i \quad (6)$$

where t_w is the average window size and must be large enough compared to the frame duration to filter out the fluctuation of the average allocated resources and capture the average video code rate of user (e.g. $t_w=100ms$ will suffice for LTE with $10ms$ frame duration). As it is depicted in Fig. 2, the average video playback buffer incoming data rate at the receiver, η_k, can be expressed as:

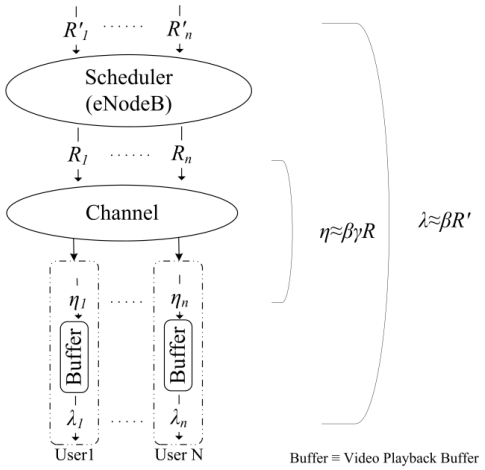

Fig. 2. Model description and data rates' assumptions

$$\overline{\eta_k^i} = \beta_k \gamma_k \overline{R_k^i} \quad (7)$$

where β_k reflects the ratio of the pure video data rate, λ_k, to the whole incoming data at the scheduler related to that user, R'_k. This usually includes extra information such as voice, metadata etc. The channel quality, the robustness of the error detection/correction techniques (i.e. HARQ and ARQ) and adequacy of the selected modulation and code rate based on the received feedback are all reflected in γ_k.

B. Proposed QoE-driven optimization method and implementation algorithm

A QoE driven allocation aims to maximize the users' satisfaction level from the service continuity's point of view, which can be interpreted as a process of lowering pause intensity during the playback. This can be expressed as:

$$\begin{cases} r^* = \arg_r \ \min\max PI \\ \begin{cases} PI = \{PI_1, PI_2, \dots, PI_{N_{UE}}\}, \\ PI_i = 1 - \frac{\eta_i}{\lambda_i}, PI \in \{x | x \in \mathbb{R},\ 0 \leq x \leq 1\}^{1\times} \\ \mathrm{H} \leq \Lambda\,, \quad \mathrm{H} = \{\eta_1, \eta_2, \dots, \eta_{N_{UE}}\}, \\ \Lambda = \{\lambda_1, \lambda_2, \dots, \lambda_{N_{UE}}\} \end{cases} \end{cases} \quad (8)$$

As it has been examined in [4] and similar to the well-known attribute of a *MaxMin throughput*, the above optimization problem tends to be extremely fair and inefficient. To restore the efficiency of the system while maintaining the effect of the user's experienced quality, the problem in (8) can be rewritten as a weighted rate scheduling algorithm as follows:

$$\begin{cases} r^* = \arg_r max \sum_{k=0}^{N_{UE}} u_k \\ u_k = PI_k{}^{\alpha} . R_k\,, \quad u_k \in \mathbb{R}_{\geq 0} \end{cases} \quad (9)$$

u_k in (9) is the utility function where its first term (i.e. PI^{α}), reflects the effect of the clients satisfaction (i.e. QoE). The first term can also be viewed as the weight for the second term, R_k, which represents the user efficiency to consume the allocated resources. The value of α defines the trade-off between the efficiency and fairness, which will be discussed in Section IV.

Fig. 3 shows the changes of the weight, ${PI_k}^{\alpha}$,of rate R_k in the proposed utility function for different value of α and versus a range of user channel status from poor to good (represented as the ratio of the achieved throughput, η, to the required data rate, λ). The depicted result justifies the trade-off between the efficiency and the fairness of the scheduler through the adjustment of α. The result shows almost the equal weight for all users in the case of smaller value of α and higher weight for users with poor channel status when α is greater. Therefore, with the value of α closer to zero, users with good channel status are expected to be more beneficiary from their achievable rate and the allocation is more efficient. In contrast, users with poor channel status are expected to be more beneficiary when the value of α increases, leading to a fairer allocation.

Usually in a wideband assessment of SNR at the receiver, a single average CQI will be generated to suggest the most suitable modulation scheme and code rate for the whole available allocation spectrum at the scheduler. Therefore all the elements of vector C_k in (4) will be equal to a certain value, $c_k(CQI)$, and in the i^{th} round of the allocation, (9) can be rewritten as a linear programing as follows:

$$\begin{cases} x^* = \arg_x \max f x^T \\ f \in \mathbb{R}^{1\times N_{UE}}, \quad f_k = {PI_k}^{\alpha}.c_k \\ x \in {\mathbb{Z}_{\geq 0}}^{1\times N_{UE}}, \quad x_k = \|r_k\|_1 \leq N_{RB} \end{cases} \tag{10}$$

where f_k combines the effect of the user's experienced quality, PI, with its achievable data rate in each resource block, c_k (i.e. user efficiency). The solution of (10), ${x^*}_k \geq 0$, represents the number of the allocated resources to user k and is an integer value while the original problem in (9) was a binary integer programming problem. The evaluation of f_k in each round, given the throughput in (7) based on its corresponding average allocated data rate in (6), is as follows:

$$f_k = {PI_k}^{\alpha}.c_k = \left(1 - \frac{\eta_k}{\lambda_k}\right)^{\alpha}.c_k \tag{11}$$

PI as a QoE metric is by definition based on the client side information, while the actual allocation process is supposed to be on the network side. PI can also be treated as a pre-decoding QoE metric, so can be evaluated merely based on the network side information to avoid additional information exchanges between users and the network. This implies that, given the network arrangement shown in Fig. 2,

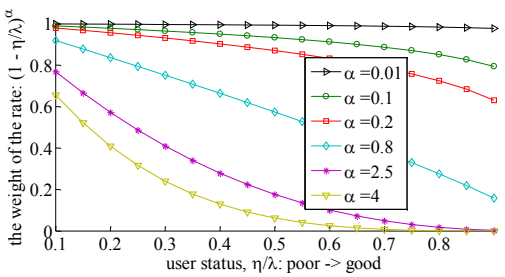

Fig. 3. The weight of the rate in utility function as a function of parameter α.

$$f_k = \left(1 - \frac{\gamma_k R_k}{R'_k}\right)^{\alpha}.c_k \tag{12}$$

Obviously, it is also possible to pull the users' evaluated PI from an element in the central network to do any specific load balancing and congestion control for those users that share the resources.

An algorithm for the implementation of the analytical models in (9) and (10) can be achieved by replacing the utility function in these models with a priority function. Based on the values of the priority function, the allocation process selects just one dominant user in each round and continues until all available resources are allocated. This algorithm can be expressed as:

$$\begin{cases} k^* = \arg_k \max u_k \\ u_k = \left(1 - \frac{\gamma_k R_k}{R'_k}\right)^{\alpha}.c_k \end{cases} \tag{13}$$

where u_k appears as a priority function and a user with the highest value of u_k will be chosen in each round of the allocation. It will be shown in Section IV that the result of the algorithmic in (13) complies with those based on the models in (9) and (10), and (13) is for the practical implementation of the proposed model.

In the next section the result of the implementation algorithm in (13) will be compared with that of the analytical model in (10). The adjustment of α for achieving a certain trade-off between efficiency and fairness will be examined. In addition, a PI based adaptive video streaming scheme will be presented in the presence of this algorithm to show the effectiveness of PI as a QoE metric for both client and network.

C. PI based rate adaptive video streaming

In a client driven rate adaptive video streaming service (e.g. 3GPP-DASH), client decides the suitable rate which has to be pulled from the server in each adaptation segment. As it is depicted in Fig. 4(a) in a shared channel with limited available resources, the user has to decide the best trade-off between the desired fidelity of the image and the minimum acceptable continuity of the service. The user will ask for each segment of the video based on the adapted rate for that time segment. The required assessment in most of the

existing technologies is based on the average incoming data rate of the playback buffer compared to a threshold (which is based on the video code rate). A simplified decision making process for the adapted rate of the $(i+1)^{th}$ segment can be expressed as:

$$\lambda_{i+1} = \begin{cases} \eta_i^*, & \eta_i < \lambda_i \\ \lambda_i + S_v, & \eta_i > \lambda_i \end{cases} \quad (14)$$

where λ represents video code rate to be adapted, η^* represents the rounded value of network throughput toward the nearest available video code rate smaller than η. S_v represents the step granularity of the video code rate. The rate adaptation condition in (14) can be reformed based on a minimum QoE threshold (a maximum acceptable discontinuity represented by $PI_{threshold}$) and expressed as:

$$PI_i > PI_{threshold} \quad (15)$$

where PI_i represents the assessed value of PI and $PI_{threshold}$ represents the maximum acceptable discontinuity of the service. Alternately (15) can be shown as:

$$\left(1 - \frac{\eta_i}{\lambda_i}\right) > PI_{threshold} \rightarrow \eta_i < (1 - PI_{threshold})\lambda_i \quad (16)$$

which resembles the initial form of the condition in (14) with the difference of the effect of the minimum desired QoE (i.e. $PI_{threshold}$). PI provides a quantitative and objective value and can be used to conduct a flexible and network oriented assessment for rate adaptation, instead of using the rigorous user oriented criteria in (14). A zero PI threshold produces the initial form in (14). As it will be shown later in Section IV, a predefined or broadcast non-zero PI threshold for users of a shared channel can produce a desired distribution of QoE from both continuity's and fidelity's points of view.

As depicted in Fig. 4 (b), a PI driven rate adaptation mechanism (on the client side) actually has an interplay with the last mile's scheduler (e.g. in eNodeB) to shape the QoE distribution among the users of the shared resources. In Section IV the results of these interactions will be discussed in more detail.

TABLE I. SIMULATION SETUP

Parameter	value
No. of Cells	1 (with the first tier interference)
Inter-site distance	2000 meters
Shadowing effect	mean=0, deviation=8 decorrelation distance=25m, inter-site correlation=0.5
Channel model	PedA, speed=3km/h
Bandwidth	5MHz, 20MHz
No. of RBs (per TimeSlot)	25, 100
Subcarrier	15KHz
Range of average SNR	-6 ~ 18 dB (CQI=1~15)
Average video code rate	156kbps ~ 1.5Mbps
No of Users	45
Each scheduling round	One TTI=1ms
Simulation time	10000*TTI (10 s)
Video stream model	Truncated Pareto for packet size and inter-arrival time

IV. SIMULATION RESULTS AND ANALYSIS

A. Simulation setup

Table I shows the settings of the simulator developed in Matlab, which is used to examine the proposed optimization method, its algorithm implementation and the QoE driven rate adaptive video streaming service. The source of the users' data is the video stream data packets generated using a truncated Pareto model (for packets' inter-arrival-time and size). No background traffics have been considered. Video code rates are in the range of standard video quality of the

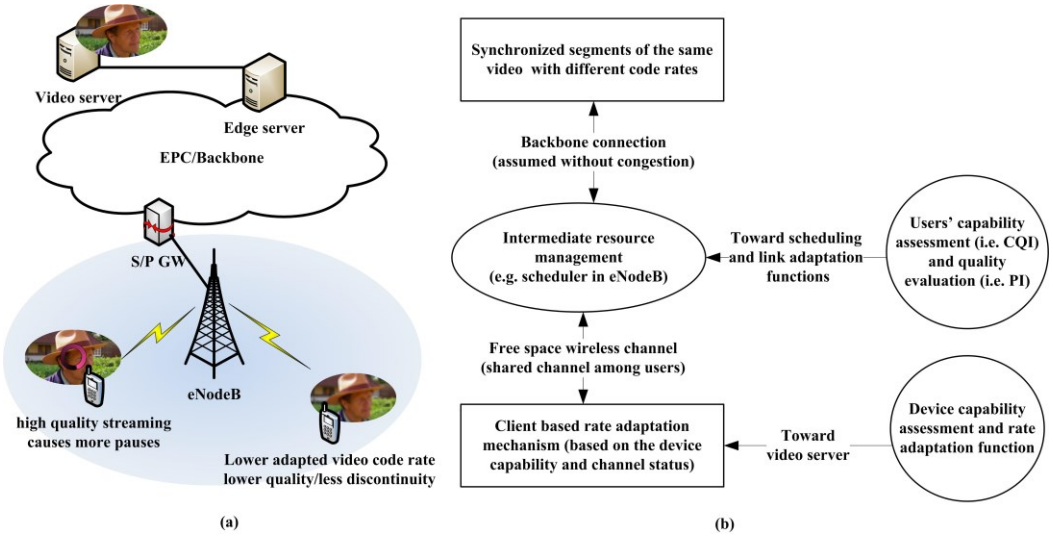

Fig. 4. Different aspects of the QoE (fidelity and continuity) in an adaptive rate video streaming service and its implementation: (a) two users with similar channel status and the trade-off between fidelity of the image and the continuity of the service (b) the implemented model of an end-to-end rate adaptive video streaning service with the conisderation of the role of last mile scheduling policy.

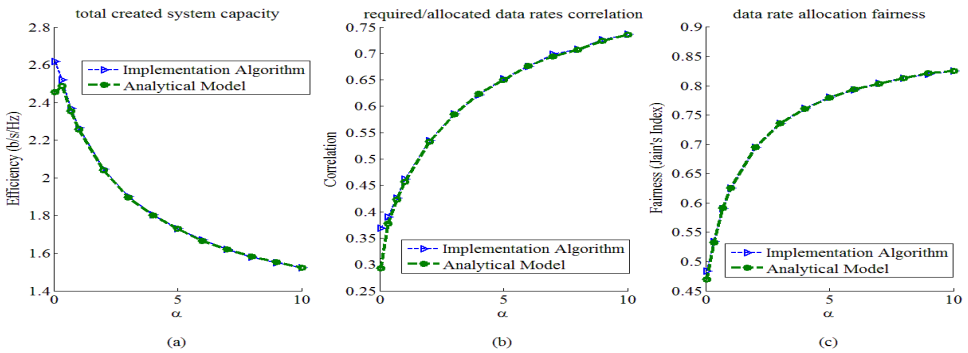

Fig. 5. The achieved trade-off between efficiency, fairness and correlation for different value α in the proposed scheduling algorithm: (a) Efficiency of the system vs. α (b) the correlation between required data rates and allocated data rates per user vs. α (c) fairness among allocated data rates to users vs. α .

state of the art technologies (e.g. for BBC-iPlayer 470 kbps - 1500 kbps is the current range of non-HD video code rates for desktop application). Users' Head-of-Line packets (HOL) are scheduled in a timely manner.

The CQI mapping table and the channel status generator presented in [16] and [18] are used in our simulator. Each video code rate is corresponding to more than one user with different SNRs in the range of the defined CQI (i.e. 1~15). This produces unbiased results with regards to the video code rate or SNR distributions. Users are distributed in one cell with the consideration of the interference from the first tier neighboring cells. Shadowing effect (inter/intra-cell spatial correlation) has been taken into account.

For the sake of comparison, MaxMin Throughput and maxCI (known as best-CQI in LTE) are taken as two extreme sides of the fairness/efficiency spectrum. The integer relaxation and rounding is employed to solve the integer linear programing problem in (10) with the constraints driven from LTE's available number of Resource Blocks for the given bandwidth in Table I. The efficiency of the system is represented in b/s/Hz and is the ratio of the total created capacity (i.e. the summation of the total allocated data rates) to the system bandwidth. Fairness is evaluated among the users' allocated data rates using the Jaine's Index. Correlation between the users' required and allocated data rates is assessed by the Pearson's Linear Correlation Coefficient.

B. *Performance of the proposed algorithm*

Fig. 5 depicts the achievable efficiency, correlation and fairness of the proposed optimization method in (10) and its implementation algorithm in (13) for different values of α. The results show the achievable trade-off between fairness and efficiency based on the value of α. Increasing α improves the fairness (Fig. 5(c)) and correlation (Fig. 5(b)), but decreasing the efficiency of the scheduling process (Fig. 5(a)). In contrast, a scheduler using smaller α will lower the levels of fairness and correlation, but increasing the efficiency. However, unlike the optimization problem in (10), the simplified algorithm in (13) allocates the resources in each round just to the dominant user (i.e. the most efficient user when α is close to zero). This leads to the over-performed efficiency of (13) compared to (10) with higher correlation between the required and the allocated data rates when α approaches zero.

Fig. 6 shows the performance of the implementation algorithm in (13) as a function of the users' channel status. The performance of the scheduler, with different values of α, lay between extremely efficient (e.g. best-CQI) and very fair (e.g. MaxMin throughput).

C. *Online adjustment of α-parameter*

The suitable value of α can be chosen based on the desired trade-off between fairness and efficiency, illustrated in Fig. 7. It can be a fixed predefined value based on the nominal characteristics of the system such as the system performance in Fig. 5. Parameter α can also be adjusted online based on the assumption about its relationship with the desired fairness or efficiency. With the assumption that the change rate of α with respect to fairness, *f*, is a constant μ (i.e. $\frac{d\alpha}{df} = \mu$), the value of α can be set online as $\alpha_i = \alpha_{i-1} + \mu(f_{target} - f_{current})$ in each iteration where f_{target} is the desired fairness, $f_{current}$ is the achieved fairness via $\alpha = \alpha_{i-1}$ and α_i is the new α to be set.

As it is depicted in Fig. 7 with two different values of μ and fairness target 0 .75, the value of αi approaches an adequate range after a transient time. It will be amended

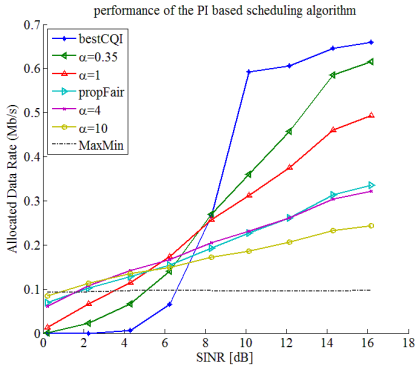

Fig. 6. The performance of the proposed scheduling algorithm based on the allocated data rates vs the status of the users' channel represented by SINR.

later, accordingly, with the changes in the situation (e.g. changes in the number of active stations, channel status, background traffics, etc.). The value of μ defines the step of the adjustment in each iteration and affects the speed of the convergence. Smaller μ produces smoother changes of α with less fluctuation in the produced fairness and efficiency (i.e. smoother change in the resource allocation) though this will extend the convergence time. The sufficiency of the achieved convergence time depends on the service demanded. Some alternative online adjustment methods, such as those suggested in [16], are available, which can be tailored for our purpose to achieve shorter transient time if necessary.

D. Client QoE-driven rate adaptation

A PI based criteria for rate adaptation for video streaming has been introduced in Subsection III-*C*. Fig. 8 provides an insight into the performance of such an adaptive scheme compared to non-adaptive video streaming from both the user's and network's points of view. The initial values of the video code rate for all of the users are a default value (780kbps in this example) and in the case of the adaptive streaming, video code rates can vary (above or below the initial value, i.e. 156kbps~1.5Mbps).

Fig. 8(a) depicts the adopted rates for two users with distinctive channel status, where the user with higher capability gradually acquires more image quality through the higher video code rate. The user with poor channel status has to reduce the requested image quality to maintain an acceptable continuity for the service. The cost of good continuity for users with bad channel status will be lower levels of fidelity for their image. However, the users with higher capability and better channel status will be served with higher video code rate. This has been shown in Fig. 8(b) where the single choice of the video code rate in non-adaptive service is expanded across a wide range of available rates higher or lower than the initial value. Fig. 8(c) shows the achieved continuity of the service in each case. Since the adaptive streaming mechanism can reduce the requested quality of the video if necessary, it maintains the continuity of the service and achieves higher probabilities of being low PI instead.

Fig. 9 depicts the interplay between the last mile scheduler (i.e. scheduling and rate adaptation functions of eNodeB in LTE) and the client side rate adaptation mechanism. The results of two distinctive efficient and fair schedulers with α=0.3 and α=3, respectively and as discussed in Section III, are provided in Fig. 9 for the purpose of comparison. On the client side, the rate adaptation mechanism chooses the desired rate of the video based on the performance of the network. As shown in Fig. 4(b), the last mile wireless channel is supposed to be the main resource bottleneck. Therefore, the scheduling policy used in eNodeB that considers the capability of the user's device is the main factor affecting the network performance.

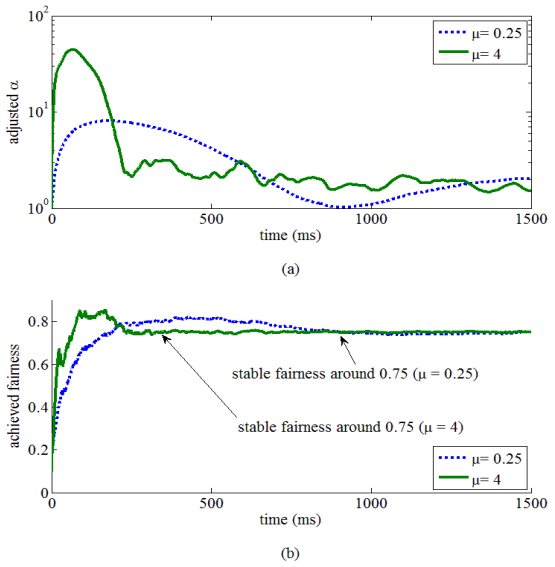

Fig. 7. Online adjustment of α for fairness target 0.75 and the step size of the amendment μ=0.25 and 4: (a) adjusted values of α vs time (b) achieved fairness using the adjusted α in the scheduler

The user is expected to choose its video code rate not only depending on its channel quality but also under the resource constraint which is related to the status of other users in the same cell (shared resources). The distribution of the QoE from the continuity's point of view (represented by Pause Intensity in Fig. 9(a)) and from fidelity point of view (represented by the spectrum of the adapted video code rates in Fig. 9(b)) are highly polarized in the case of efficient scheduler. It means that the rate of an adaptive video streaming service will be either very high or very low with

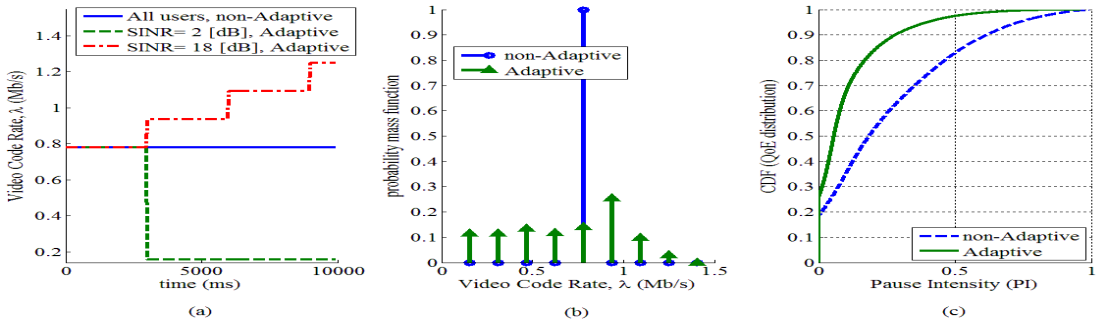

Fig. 8. Adaptive video streaming performance compared to a non-adaptive service: (a) adapted rates for two users with distinctive channel status (b) the spectrum of the adapted rates of all users compared to the fixed rate of non-adaptive service (c) overall QoE performance of adaptive service compared to the non-adaptive service from continuity point of view

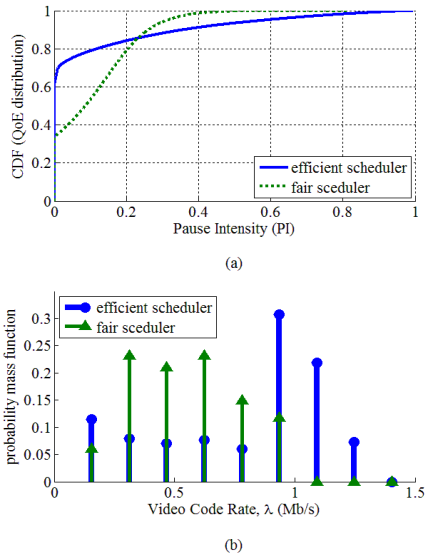

Fig. 9. The effect of the last mile sheduler over the performance of the rate adaptation process: (a) achieved QoE of an efficient scheduler compared to a fair scheduler from continuity point of view (b) a comparison between the spectrum of the adopted rates of an efficinet and a fair scheduler

low possibility for intermediate values.

An efficient scheduler provides more resources to users with better channel quality, hence higher video code rates will be fetched by them. Subsequently, users with poorer channel quality experience will limit network performance. Therefore they adopt lower video code rates to maintain the minimum desired level of continuity of the service. A wider range of video code rates will be chosen by the fair scheduler though the maximum video code rate is restricted in this case. This fact has been reflected in the result where smoother change in the distribution of QoE (for continuity) in the range of $0 \leq PI \leq 1$.

V. CONCLUSION

In this paper a QoE driven adaptive scheduler has been proposed and examined in the context of a wireless mobile communication system providing adaptive rate video streaming services. An algorithm for the implementation of the established analytical model (i.e. the LP problem) has been proposed. Pause Intensity is adopted to quantify the continuity aspect of the service with the capability of being evaluated on both client and network sides. The proposed algorithm provides a flexible tool to achieve a desired trade-off between fairness and efficiency. Furthermore, the effectiveness of the online adjustment method for the scheduler parameter to maintain the desired level of fairness or efficiency has also been shown.

PI has been used to regulate user's video code rate on the client side and to shape the distribution of the QoE related performance among users in collaboration with the scheduler on the network side. The performance trade-offs between efficient and fair schedulers in both adaptive and non-adaptive modes for streaming have been analyzed.

REFERENCES

[1] Cisco Systems, Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2013–2018., San Jose, CA, USA: Cisco Systems, Inc., Feb. 2014.

[2] 3GPP TS 26.247, "Transparent end-to-end Packet-switched Streaming Service (PSS); Progressive Download and Dynamic Adaptive Streaming over HTTP (3GP-DASH)", Rel. 12, March 2014.

[3] M. Seyedebrahimi, C. Bailey, and X.-H. Peng, "Model and Performance of a No-Reference Quality Assessment Metric for Video Streaming," IEEE Transactions on Circuits and Systems for Video Technology, vol.23, no.12, (Dec. 2013), pp.2034-2043.

[4] M. Seyedebrahimi, X.-H. Peng, and R. Harrison, "A Quality Driven Framework for Adaptive Video Streaming in Mobile Wireless Networks," Accepted for IEEE WCNC 2014.

[5] 3GPP LTE and LTE-Advanced Technology, http://www.3gpp.org/ftp/Specs/html-info/36-series.htm

[6] 3GPP TS 26.233, "Transparent end-to-end Packet-switched Streaming Service (PSS); General Description", Rel. 11, April 2014.

[7] L. De Cicco and S. Mascolo, "An adaptive video streaming control system: modeling, validation, and performance evaluation", IEEE/ACM Trans. Netw. 22, 2 (April 2014), 526-539.

[8] J. Wang, T.Z.J. Fu, D.M. Chiu; Z.B. Lei, "Perceptual quality assessment on B-D tradeoff of P2P assisted layered video streaming," Visual Communications and Image Processing (VCIP), 2011 IEEE , vol., no., pp.1,4, 6-9 Nov. 2011.

[9] Subjective Audiovisual Quality Assessment Methods for Multimedia Applications, ITU-T Rec. P.911, 1998.

[10] K. Seshadrinathan, R. Soundararajan, A.C Bovik, and L.K Cormack, "Study of Subjective and Objective Quality Assessment of Video," Image Processing, IEEE Transactions on , vol.19, no.6, pp.1427,1441, June 2010.

[11] J. Yan, W. Muhlbauer, and B. Plattner, "Analytical Framework for Improving the Quality of Streaming Over TCP," Multimedia, IEEE Transactions on , vol.14, no.6, pp.1579,1590, Dec. 2012.

[12] R. Huysegems, B. De Vleeschauwer, K. De Schepper, C. Hawinkel, W. Tingyao, K. Laevens, and W. Van Leekwijck, "Session reconstruction for HTTP adaptive streaming: Laying the foundation for network-based QoE monitoring," Quality of Service (IWQoS), 2012 IEEE 20th International Workshop on , vol., no., pp.1,9, 4-5 June 2012

[13] A. Vishwanath, P. Dutta, M. Chetlu, P. Gupta, S. Kalyanaraman, and A. Ghosh., "Perspectives on quality of experience for video streaming over WiMAX", SIGMOBILE Mob. Comput. Commun. Rev. 13, 4 (March 2010), 15-25.

[14] R. Kuschnig, I. Kofler, and H. Hellwagner, "Evaluation of HTTP-based request-response streams for internet video streaming", In Proceedings of the second annual ACM conference on Multimedia systems (MMSys '11). ACM, New York, NY, USA, 245-256.

[15] E. Dahlman, S. Parkvall, and J. Skold, "4G: LTE/LTE-Advanced for Mobile Broadband," Academic Press, 2011.

[16] S. Schwarz, C. Mehlfuhrer, and M. Rupp, "Throughput Maximizing Multiuser Scheduling with Adjustable Fairness," Communications (ICC), 2011 IEEE International Conference on , vol., no., pp.1,5, 5-9 June 2011

[17] Z. Sun, C. Yin, and G. Yue, "Reduced-Complexity Proportional Fair Scheduling for OFDMA Systems," Communications, Circuits and Systems Proceedings, 2006 International Conference on, vol.2, pp.1221-1225, 25-28

[18] Enhanced UMTS Radio Access Network Extension for ns-2, http://eurane.ti-wmc.nl/eurane/