Fair Robust Predictive Resource Allocation for Video Streaming Under Rate Uncertainties

Ramy Atawia*, Hossam S. Hassanein† and Aboelmagd Noureldin‡
*Electrical and Computer Eng. Dept., Queen's University, Canada, ramy.atawia@queensu.ca
†School of Computing, Queen's University, Canada, hossam@cs.queensu.ca
‡Electrical and Computer Eng. Dept., Royal Military College of Canada, Canada, aboelmagd.noureldin@rmc.ca

***Abstract*—Predictive Resource Allocation (PRA) has demonstrated its ability to provide smooth video delivery with minimal and fair interruptions. Recent work on PRA techniques exploited rate predictions to strategically allocate the limited radio resources for delivering video content. However, existing PRA techniques assume perfect prediction of future information in order to define the maximum attainable gains. In this paper, we introduce a probabilistic robust PRA framework that handles prediction errors. By adopting chance constraint programming we were able to define a probabilistic measure on the QoS degradation due to prediction uncertainties. A deterministic non-convex formulation is then obtained using the statistical parameters of predicted rates. Accordingly, we propose a convex approximation to the formulated fair PRA, which can be solved using optimal solvers to obtain a benchmark solution for future robust PRA schemes. We evaluate non-PRA and non-robust PRA schemes considering typical error models of the predicted rates. We found these schemes to result in suboptimal fairness and increased QoS degradations with the network load. Results further reveal the ability of the introduced robust fair PRA to reach the optimal and fair QoS satisfaction levels. Our approach provides a step towards applying PRA in future wireless networks to deliver video streaming content.**

I. Introduction

Mobile data traffic is anticipated to reach 24.3 exabytes per month by 2019, forming a compound annual growth rate (CAGR) of 57% [1]. Approximately 70% of this tremendous growth is attributed to mobile video content which produces a fourfold in the amount of traffic than other Internet applications such as browsing and e-mails. The ability of the mobile operator to exploit the purchased wireless spectrum and infrastructure is the key element of minimizing the network operational expenditures (OpEx) while delivering video traffic at high QoS levels. Recent studies emphasized the limitations of traditional resource allocation (RA) strategies in delivering video content at the users' requested quality [2]–[4].

Predictive Resource Allocation (PRA) has gained momentum as it provided promising energy savings, user fairness, and rate maximization [2], [3]. In essence, PRA utilizes the repetitive signal strength and high predictable users' mobility traces in order to forecast future channel rates [5]. Ubiquitous navigation [6], [7] further supports mobility prediction in dense urban areas with high traffic demands. In PRA, users moving towards poor radio conditions are prioritized and their content is prebuffered to provide smooth playback in the future. On the other hand, allocation is minimized to the users heading to the cell center and prebuffering is postponed till they reach the peak radio conditions. This predictive strategy guarantees both energy savings and long-term fairness over the time horizon compared to the traditional RA that allocates the users based on previous measurements while ignoring their anticipated channel conditions.

The potential gains of PRA [2]–[4] were preliminary attained under *ideal* predictions of future rates without any considerations of real world uncertainties. However, extensive channel measurements in [8] showed that the predicted average values of signal strength varies with geographical location and time of day. This practical challenge hinders the Quality of Service (QoS) satisfaction under deterministic constraints adopted in existing non-robust PRA work.

Our recent work on energy-efficient PRA [9]–[11] further studied the challenges in assuming *ideal* predictions and demonstrated the importance of *robust* QoS constraints. This robustness was able to preserve the prediction gains of the PRA while satisfying predefined probabilistic QoS levels under rate prediction errors. While PRA minimizes the Base Station (BS) energy in low load scenarios, user fairness has to be maximized in higher loads where users move to poor radio conditions and greedily request high streaming rates. This paper introduces an optimal *stochastic* PRA that is *robust* to prediction uncertainties and thus achieves *fair* QoS satisfaction among users during the time horizon. We summarize the contributions of this paper as follows:

1) We develop a stochastic based formulation that represents the predicted channel information as random variables and defines a probabilistic metric to indicate the QoS under prediction uncertainties. In essence, this metric quantifies the risk of QoS degradations given the calculated resource allocation solution. Hence, the formulation provides the operator with the flexibility to control the trade-off between fairness and QoS degradation when predictions are uncertain.
2) A deterministic equivalent form is then obtained to calculate both the resource allocation matrix and its associated QoS degradations using the rates' Cumulative Density Function (CDF) and statistical properties.
3) We then derive a convex approximation for the deterministic form using budgeted robust and fairness constraints. Thus, an optimal solution can be obtained through commercial solvers, providing a benchmark solution for future robust PRA schemes.

In the following section, we provide a background to stochastic optimization, and review the related literature. Section III presents the preliminaries and system model. Section IV introduces the *stochastic* based robust formulation, the deterministic equivalent and its convex approximation. Simulation results are discussed in Section V and finally, we conclude the paper in Section VI.

II. Related Work and Background

Existing PRA techniques [2]–[4] represented imperfect predictions by the expected values, resulting in a deterministic formulation. The corresponding decisions do not guarantee QoS satisfaction when predicted future rates fall below the expected value. This is because minimal resources allocated to less prioritized users will not be sufficient to meet the demand and thus buffer under-run occurs. On the other hand, when peak rates exhibit lower values than the average, fairness and resource utilization are suboptimal. This is due to the large amount of resources allocated to the users moving away from such varying peak radio conditions.

Robust techniques have been discussed in the literature for non-predictive RA in order to handle uncertainties or delays in the user reported measurements [12]–[14]. Our focus in this paper is on *Stochastic* Chance Constrained Programming (CCP) which models the uncertain constraints' coefficients or bounds as random variables. In essence, CCP represents these constraints in a probabilistic form with a maximum violation degree denoted as $\epsilon \in [0,1]$. A deterministic equivalent form is then derived to obtain a closed form RA formulation that can be solved by numerical optimization techniques.

The robust non-predictive RA in [14] adopted the Markov inequality to approximate the CCP and obtained a linear deterministic formulation. However, the exact optimal coefficients for this approximation are not easily attainable and do not provide a direct modelling for the trade-off between optimality and degree of constraint satisfaction. The robust uplink power allocation in [12], [13] obtained the deterministic form using the Bernstein Approximation (BA) and based on the Moment Generating Function (MGF). The resulted BA approximation was separable and convex, however the maximum violation degree was kept constant.

In this paper, our introduced *robust* PRA framework aims to optimize the violation degree ϵ for smooth video streaming and long-term fairness among users. By optimizing such a parameter, the associated risk of QoS degradation can be calculated in the beginning of the time horizon. Moreover, other performance metrics such as bandwidth efficiency and fairness can be represented using this degree of violation in the resource optimization problem. This is in addition to handling the non-convexity of the resultant formulation through robust approximations. This is unlike our robust energy-efficient PRA approaches in [9]–[11] where the objective was to minimize the energy consumption at a fixed violation degree.

III. System Model

A. Preliminaries

We use the following notational conventions throughout the paper: $\mathcal{X}$ denotes a set and its cardinality is denoted by X. Matrices are denoted with subscripts, e.g. $\mathbf{x} = (x_{a,b} : a \in \mathbb{Z}_+, b \in \mathbb{Z}_+)$. $Pr\{x\}$ is the probability of event x, and $log(.)$ denotes the natural logarithmic. $\mathbb{1}_x$ is an indicator function that is equal to one if event x is true and equals 0 otherwise. $\tilde{r}$ represents a random variable whose expectation is $\mathbb{E}[\tilde{r}]$, its CDF is dented by $\Phi_{\tilde{r}}(x) = Pr\{\tilde{r} \leq x\}$ and the Probability Density Function (PDF) is $N(\tilde{r})$. $\Phi(x)$ represents the CDF of a normally distributed random variable with zero mean and unit variance, while its inverse CDF is denoted by Φ_x^{-1} corresponding to the x^{th} percentile. The gradient and Hessian of function $\mathbf{f}(\cdot)$ are denoted by $\nabla\mathbf{f}(\cdot)$ and $\nabla^2\mathbf{f}(\cdot)$, respectively.

B. System Model

The system considers a group of BSs, each with an active user set denoted by $\mathcal{M}$ and user index $i \in \mathcal{M}$. Each mobile user requests video with a fixed streaming rate. Advanced caching techniques [15] can be utilized to store the video near the BS, which minimizes back-haul delays, and thus the main bottleneck is the limited radio resources. We assume that user's mobility trace is known for the next T timeslots, where each slot is denoted by $t \in \mathcal{T} = \{1, 2, \cdots, T\}$.

1) BS Radio Resources: The active users share the BS resources (airtime fractions) at each time slot t. The resource allocation matrix $\mathbf{x} = (x_{i,t} \in [0,1] : i \in \mathcal{M}, t \in \mathcal{T})$ gives the fraction of time slot t during which BS's bandwidth is assigned to user i.

2) Rate Prediction: For the resource allocation, prediction of rate is done by mapping the user's current location to the Radio Environment Map (REM) at the service provider. The REM contains both the user's locations and their corresponding average values of the channel rate for user i at time slot t, denoted as $\bar{r}_{i,t} = \mathbb{E}[\tilde{r}_{i,t}]$. The predicted uncertain rate is modelled as a Gaussian distributed random variable $\tilde{r}_{i,t}$ whose standard deviation is denoted by $\sigma_{i,t}$ and calculated as in [10].

3) User Satisfaction Model: The video is transmitted to user i at a constant encoding rate denoted by V_i that provides a fixed streaming quality. For smooth video playback, the total amount of buffered content should be more than the cumulative streaming demand $D_{i,t} = V_i \times t$. Otherwise video freezing occur and result in users dissatisfaction and low QoS. The number and duration of video freezes will thus be used as a user dissatisfaction metric.

The robust PRA scheme in this paper aims to calculate the airtime fractions $x_{i,t}$ such that the user QoS level is maximized. In essence, QoS is said to be satisfied when users experience very small number of stops (ideally zero) and for short durations. In case of high demand, relative to the available channel capacity, the users should experience a fair number of stops (ideally the same number of stops among all users).

IV. Optimal Robust Fair PRA Formulation

A. Stochastic Chance Constrained Formulation

The introduced robust fair PRA based on CCP is formulated in Eq. 1

$$\underset{\mathbf{x},\varepsilon}{\text{minimize}} \quad \underset{\forall i\in M,\forall t\in T}{\text{maximum}} \{\varepsilon_{i,t}\} \tag{1}$$

S.t.:

$$\begin{aligned}
&\text{C1:}Pr\left\{\sum_{t'=0}^{t}\tilde{r}_{i,t'}x_{i,t'} < D_{i,t}\right\} \leq \varepsilon_{i,t}, && \forall i\in\mathcal{M},\forall t\in\mathcal{T},\\
&\text{C2:}\sum_{i=1}^{M} x_{i,t}\leq 1, && \forall t\in\mathcal{T},\\
&\text{C3:}\varepsilon_{i,t}\leq 1, && \forall i\in\mathcal{M}, t\in\mathcal{T},\\
&\text{C4:}x_{i,t},\varepsilon_{i,t}\geq 0, && \forall i\in\mathcal{M}, t\in\mathcal{T}.
\end{aligned}$$

Where $\varepsilon_{i,t}\in[0,1]$ is the probability that the QoS of user i is unsatisfied at time slot t, e.g. $\varepsilon_{i,t}=1$ indicates a complete QoS violation.

The constraint C1 in Eq. 1 models the QoS satisfaction by ensuring that the probability of experiencing a video stop is capped by the QoS parameter $\varepsilon_{i,t}$. The objective function aims to minimize the maximum value of $\varepsilon_{i,t}$ such that the dissatisfaction is fairly distributed among the users and over the time horizon to avoid user starvation. The second constraint models the limited resources at each base station by ensuring that the sum of the total allocated airtime is less than 1 second which is the duration of the allocation slot. The third and the fourth constraints ensure the bounds for the probability level and the non-negativity of the decision variables, respectively.

The robust formulation in Eq. 1 solves for both the airtime fraction $x_{i,t}$ and the probability of video degradation $\varepsilon_{i,t}$. This results in two advantages compared to the existing PRA work in [2]–[4]. Firstly, the calculation of the airtime fraction will consider the variations in the predicted rate as will be shown in the next subsection. Particularly, more airtime can be allocated to the users with erroneous rates while granting less resources to users experiencing stable radio conditions. The second advantage is that the calculated probability of degradation will provide the network operator with the anticipated QoS by the users under the calculated allocation. This provides flexibility of prioritizing some users by setting a lower bound to their corresponding degradation probabilities. The above formulation can be further extended to consider other network metrics than the fairness such as total degradations.

B. Deterministic Equivalent

In order to determine a closed form solution for the formulated probabilistic PRA in Eq. 1, a deterministic equivalent shall be obtained. In particular, the random variable will be replaced by its mean, standard deviation and CDF. The summation of the independent normally distributed rates in C1 of Eq. 1 will result in a multivariate normal distribution with mean and standard deviations calculated as follows

$$\begin{aligned}
\mu_{i,t} &= \sum_{t'=0}^{t}\bar{r}_{i,t'}x_{i,t'},\\
\Sigma_{i,t} &= \sqrt{\sum_{t'=0}^{t}x_{i,t'}^2\sigma_{i,t'}^2},
\end{aligned} \tag{2}$$

where $\sigma_{i,t'}^2 = E[(\tilde{r}_{i,t'}-\bar{r}_{i,t'})^2]$.

The deterministic closed form of C1 in Eq. 1 can be then expressed using the cumulative distribution function of multivariate random variables as shown below.

$$\begin{aligned}
Pr\left\{\sum_{t'=0}^{t}\tilde{r}_{i,t'}x_{i,t'} < D_{i,t}\right\} &= \int_{-\infty}^{D_{i,t}} N(\mathbf{r},\mu,\Sigma)d\mathbf{r}\\
= \frac{\Phi(\frac{D_{i,t}-\mu_{i,t}}{\Sigma_{i,t}})-\Phi(\frac{-\mu_{i,t}}{\Sigma_{i,t}})}{\prod_{t'=0}^{t}\left(\Phi(\frac{r_{i,t'}^{(u)}-\bar{r}_{i,t'}}{\sigma_{i,t'}})-\Phi(\frac{r_{i,t'}^{(l)}-\bar{r}_{i,t'}}{\sigma_{i,t'}})\right)} &\leq \varepsilon_{i,t}.
\end{aligned} \tag{3}$$

Where $r_{i,t}^{(l)}$ and $r_{i,t}^{(u)}$ are the lower and upper values of the realization of the future predicted rate $\tilde{r}_{i,t}$ (i.e. the support). Typical values of the channel rates in the current and future networks are more than the corresponding variance values (i.e. $\mu_{i,t} >> \Sigma_{i,t}$) and thus $\Phi(\frac{-\mu_{i,t}}{\Sigma_{i,t}})\approx 0$. Finally, the deterministic closed form of Eq. 1 with the preceding assumptions can be summarized below

$$\underset{\mathbf{x},\varepsilon}{\text{minimize}} \quad \underset{\forall i\in M,\forall t\in T}{\text{maximum}} \{\varepsilon_{i,t}\} \tag{4}$$

S.t.:

$$\text{C1:}\Phi(\frac{D_{i,t}-\mu_{i,t}}{\Sigma_{i,t}}) \leq \varepsilon_{i,t}S_{i,t}, \quad \forall i\in\mathcal{M}, t\in\mathcal{T}.$$

C2 - C4

Where $S_{i,t}=\prod_{t'=0}^{t}\left(\Phi(\frac{r_{i,t'}^{(u)}-\bar{r}_{i,t'}}{\sigma_{i,t'}})-\Phi(\frac{r_{i,t'}^{(l)}-\bar{r}_{i,t'}}{\sigma_{i,t'}})\right)$ is used to normalize the truncated probability distribution of the random rates. Although the formulation in Eq. 4 is continuous and differentiable, it is non-convex due to the objective function and the QoS constraint C1 [11].

C. Convex Deterministic Formulation

Proposition 1: Applying the budgeted robust approximation of [12] will result in a feasible conservative solution for Eq. 4. Let $\Sigma_{i,t}^{(L)}=\sum_{t'=0}^{t}|x_{i,t'}\sigma_{i,t'}|$, thus $\Sigma_{i,t}\leq\Sigma_{i,t}^{(L)}$, and $\Phi(\frac{1}{\Sigma_{i,t}})\geq\Phi(\frac{1}{\Sigma_{i,t}^{(L)}})$. This guarantees the satisfaction of C1 by substituting $\Sigma_{i,t}$ with $\Sigma_{i,t}^{(L)}$.

Proposition 2.1: C1 is non convex due to the multiplication between functions including the two decisions variables $\varepsilon_{i,t}$ and $x_{i,t}$, thus C1 is convex in the log domain. Let $\gamma_{i,t}=\varepsilon_{i,t}S_{i,t}$, substitute $\Sigma_{i,t}$ by $\Sigma_{i,t}^{(L)}$ and by using the inverse CDF denoted by Φ^{-1}, then C1 in Eq. 4 becomes

$$\mu_{i,t}+\Phi_{\gamma_{i,t}}^{-1}\Sigma_{i,t}^{(L)}\geq D_{i,t}, \quad \forall i\in\mathcal{M}, t\in\mathcal{T}. \tag{5}$$

Let $y_{i,t}=\Phi_{\gamma_{i,t}}^{-1}$ and by applying the log operator on both sides, Eq. 5 becomes

$$\log(\mu_{i,t}-D_{i,t})\geq\log(-y_{i,t}\Sigma_{i,t}^{(L)}), \quad \forall i\in\mathcal{M}, t\in\mathcal{T}. \tag{6}$$

Proposition 2.2: Function $\mathcal{F}$ is convex if its Hessian matrix $H = \nabla^2(\mathcal{F})$ is positive definite or semidefinite.

$$\mathcal{F}_{i,t} = -log(\mu_{i,t} - D_{i,t}) + \log(\Sigma_{i,t}^{(L)}) + \log(-y_{i,t}) \quad (7)$$

$$\frac{\partial^2 \mathcal{F}_{i,t}}{\partial x_{i,t}^2} = \frac{\bar{r}_{i',t}^2}{(\mu_{i,t} - D_{i,t})^2} - \frac{\sigma_{i,t}^2}{(\Sigma_{i,t}^{(L)})^2} \quad (8)$$

$$\frac{\partial^2 \mathcal{F}_{i,t}}{\partial x_{i,t}\partial x_{j,t}} = \frac{\bar{r}_{i',t}\bar{r}_{j,t}}{(\mu_{i,t} - D_{i,t})^2} - \frac{\sigma_{i,t}\sigma_{i,t'}}{(\Sigma_{i,t}^{(L)})^2}, t \neq t' \quad (9)$$

$$\frac{\partial^2 \mathcal{F}_{i,t}}{\partial x_{i,t}\partial y_{i,t}} = \frac{\partial^2 \mathcal{F}_{i,t}}{\partial y_{i,t}\partial x_{i,t}} = 0 \quad (10)$$

$$\frac{\partial^2 \mathcal{F}_{i,t}}{\partial {y_{i,t}}^2} = \frac{1}{y_{i,t}^2} \quad (11)$$

$$H_{i,t} = \nabla^2 \mathcal{F}_{i,t} = \begin{bmatrix} Z_{T\times T} & 0_{T\times 1} \\ 0_{T\times 1} & \frac{1}{y_{i,t}^2} \end{bmatrix} \quad (12)$$

where $Z_{i,i} = \frac{\partial^2 \mathcal{F}_{i,t}}{\partial x_{i,t}^2}, Z_{i,j} = Z_{j,i} = \frac{\partial^2 \mathcal{F}_{i,t}}{\partial x_{i,t}\partial x_{j,t}}$. All the entries of the Hessian matrix in Eq. 12, calculated in Eq. 8-Eq. 11 are positive and $Z_{i,i} \geq Z_{i,j}$. The principle minors are then calculated and found to be positive or zero, hence the Hessian matrix is positive semidefinite and the log constraint is convex.

After convexifying the constraint C1, a convex form of the objective function in Eq. 4 will be obtained. Recalling $y_{i,t} = \Phi^{-1}_{\varepsilon_{i,t}S_{i,t}}$, therefore the objective function becomes: $\underset{\forall i\in M, \forall t \in T}{\text{maximum}} \{\Phi(y_{i,t})/S_{i,t}\}$

Proposition 3.1: Define and minimize an upper bound Γ for all the degradations $\Phi(y_{i,t})/S_{i,t}$. This is represented by adding constraint $\Phi(y_{i,t})/S_{i,t} \leq \Gamma,\ \forall i \in \mathcal{M}, t \in \mathcal{T}$ and replace the non-convex objective function with the upper bound Γ.

Proposition 3.2: The new fairness constraint $\Phi(y_{i,t})/S_{i,t} \leq \Gamma,\ \forall i \in \mathcal{M}, t \in \mathcal{T}$ is convex for $y_{i,t} \leq 0$.

$$\begin{aligned} \mathcal{F}_{i,t} &= \frac{\Phi(y_{i,t})}{S_{i,t}} - \Gamma, \\ \nabla^2 F_{i,t} &= \frac{-y_{i,t}e^{-0.5(y_{i,t}^2)}}{\sqrt{2\pi}S_{i,t}} \end{aligned} \quad (13)$$

As such $\nabla^2 F_{i,t} > 0,\ \forall y_{i,t} \leq 0$ and thus the added fairness constraint is convex.

The final convex equivalent for the robust predictive fair scheduler Eq. 4 is summarized as

$$\underset{\mathbf{x},\mathbf{y},\mathbf{\Gamma}}{\text{minimize}} \quad \Gamma \quad (14)$$

S.t.:

$$\begin{aligned}
&\text{C1:} \quad \log(\frac{\mu_{i,t} - D_{i,t}}{\Sigma_{i,t}^{(L)}}) \geq \log(-y_{i,t}), && \forall i \in \mathcal{M}, t \in \mathcal{T}, \\
&\text{C2:} \quad \sum_{i=1}^{M} x_{i,t} \leq 1, && \forall t \in \mathcal{T}, \\
&\text{C3:} \quad y_{i,t} \leq 0, && \forall i \in \mathcal{M}, t \in \mathcal{T}, \\
&\text{C4:} \quad x_{i,t} \geq 0, && \forall i \in \mathcal{M}, t \in \mathcal{T}. \\
&\text{C5:} \quad \frac{\Phi(y_{i,t})}{S_{i,t}} \leq \Gamma, && \forall i \in \mathcal{M}, t \in \mathcal{T}.
\end{aligned}$$

TABLE I
SUMMARY OF MODEL PARAMETERS

Parameter	Value
BS transmit power	43 dBm
Bandwidth	5 MHz
Time Horizon T	60 s
Streaming rate V	1 [Mbps]
Bit Error Rate	5×10^{-5}
Shadow correlation distance(d_{cor}) [17]	$50m$
Shadow standard deviation [17]	6
Velocity	25 - 40 [km/h]
Packet size	10^3 [bytes]
Packet rate (from core network to BS)	$10^3 s^{-1}$
Buffer size	10^9 [bits]

Where C3 represents the non-negativity constraint on the degradation $y_{i,t}$ which guarantees a reliable QoS satisfaction by probability 0.5. Another advantage of the above formulation is that the QoS constraint C1 can be also an equality and thus wider range of convex optimization technique can be adopted. Such advantage does not exist in other non-robust PRAs that require the inequality to allow video prebuffering.

V. PERFORMANCE EVALUATION

A. Simulation Set-up

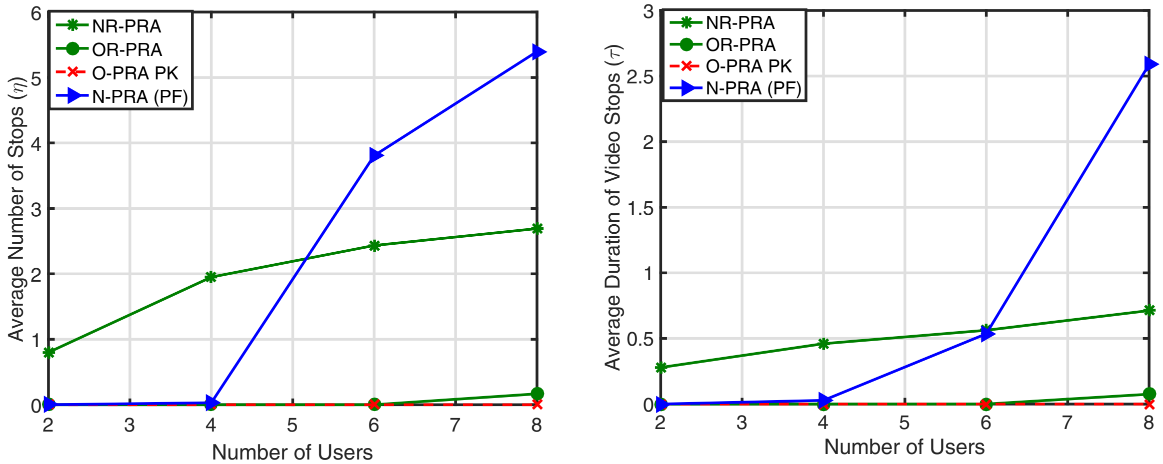
We simulate the proposed robust PRA using our modified ns-3 LTE-Gurobi module in [16]. Simulation results are averaged over 50 runs with different shadowing values. The 3GPP correlated slow fading model and its parameters [17] are added to the received UE power and thus provide predicted rate variations. Users follow different predefined paths within the cell at varying velocities from 25 to 40 Km/h, which correspond to typical values in urban areas where prediction is uncertain. All the simulation parameters and values are presented in Table I.

B. Evaluation Metrics and Comparative Schemes

In order to assess the introduced *robust-fair* PRA framework, we measure the QoS satisfaction using the number and duration of video stops denoted by η and τ, respectively, and calculated as $\eta_i = \sum_{t=0}^{T} \mathbb{1}_{R_{i,t}<D_{i,t}}$ and $\tau_i = \int_0^T \tau_{i,\kappa}\partial\kappa / \int_0^T \partial\kappa$. Where $R_{i,t} = \sum_{t'=0}^{t} r_{i,t'}x_{i,t'}$ is the cumulative video content received by user i till time slot t while $r_{i,t}$ is the experienced channel rate by user i at timeslot t. $\tau_{i,\kappa}$ equals to 1 if the user i experienced a video stop at time instant κ and $\kappa << 1$.

The overall network performance is calculated as the average of each of the above metrics over the number of users. On the other hand, the fairness is calculated using Jain's index $J = \frac{(\sum_{i=1}^{M} n_i)^2 + \epsilon}{M\sum_{i=1}^{M} {\tau_i}^2 + \epsilon}$. Where n_i is either the video stops τ_i or their durations η_i of user i. ϵ is a very small number that results in J=1 when any of the two metrics is 0 for all the users.

In this evaluation study, we compare the introduced robust scheme with the existing non-robust PRA and the traditional non-predictive RA techniques with the following abbreviations:



(a)

(b)

Fig. 1. Average values of both (a) the number of video stops and (b) video stop durations in seconds

- **N-PRA (PF):** refers to a type of non-predictive RA denoted as proportional fairness (PF) [18]. In essence, PF allocates the resources to the users based on both their current channel rates and the total amount of transmitted data in the previous slots, and regardless the future radio conditions.
- **NR-PRA:** refers to the existing non-robust fair PRA in [3] which only uses the average value of the predicted rate. The optimal solution is obtained using the simplex or interior point method implemented in Gurobi solver.
- **O-PRA (PK):** refers to a hypothetical optimal PRA with perfect channel knowledge (actual rates without errors), and thus provides an upper bound to the performance of other schemes. The solution is obtained optimally using Gurobi solver.
- **OR-PRA:** refers to the proposed optimal fair robust PRA in this work formulated in Eq. 14 while the solution is obtained optimally using the interior point method in MATLAB optimization toolbox.

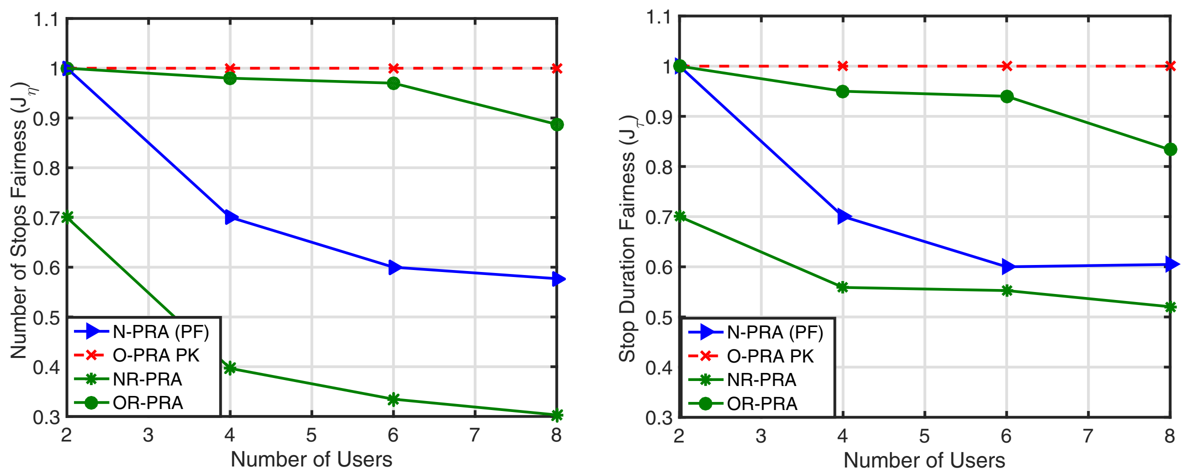
C. Simulation Results

The demands and channel rates in the scenario were typically selected such that the capacity is sufficient to satisfy the users' predefined streaming rates. Thus, the only reason that a video stop occurs is the non-robustness of the allocation and not the limited channel capacity. This is illustrated by performance of the *O-PRA (PK)* scheme in which both the number and duration of stops are zero as shown in Fig. 1(a) and Fig. 1(b). Accordingly, all stops experienced by the NR-PRA scheme are attributed to minimal allocated airtime to the cell edge users who were then moving towards the cell center. In essence, the average value of the predicted rates was adopted to allocate the exact amount of airtime fractions such that the demand is barely satisfied. Such optimistic approach with the cell edge users would ideally provide more capacity for the others with worse future channel conditions. However, under typical channel variations the cell edge users experienced low rates than the average values and thus the minimal allocated airtime was insufficient to avoid video freezes. While buffering is maximized at the users near the cell center, the probability of experiencing video stops by the cell edge users increases resulting in a low fairness index as depicted in Fig. 2(a) and Fig. 2(b).

On the other hand, the non-predictive strategy *N-PRA (PF)* greedily divides the resources among the users irrespective of their future rates. As the number of users increase, such an opportunistic approach fails to prebuffer the content of users moving towards the poor radio conditions, resulting in an increased number and durations of the stops as depicted in Fig. 1(a) and Fig. 1(b). This is in addition to the suboptimal fairness in Fig. 2(a) and Fig. 2(b) due to the lower priority, compared to the PRA, assigned to the users with poor future radio conditions.

It can be noted that both the robust PRA (*NR-PRA*) and the opportunistic *N-PRA (PF)* provide allocation that are far from the optimal scheme *O-PRA (PK)*. This is because *NR-PRA* prioritizes the users moving towards the cell edge over others heading to the cell peak, while *N-PRA (PF)* applies a converse strategy.

The introduced robust PRA (i.e. *OR-PRA*), in essence, considered both the average and variance of the future rates. This is in addition to the probabilistic bound ε which was less than 0.5 in order to satisfy constraint C3 in Eq. 14. This results in an equivalent predicted rate that is less than the average and thus more airtime fraction, compared to *NR-PRA*, will be allocated to the cell edge users to avoid buffer starvation. This increased airtime was not conservative and thus did not compromise the availability of resources to other users moving towards the cell edge, allowing the prebuffer of their future content. Hence, outperforms the non-predictive scheme and maintains the prediction gain.



(a)

(b)

Fig. 2. Fairness Index for both (a) the number of video stops and (b) average video stop duration

VI. Conclusion

We introduce an *optimal robust* PRA scheme for video delivery to handle uncertainties in predicted rates. The scheme allows the operator to calculate the probability of QoS degradations for all users and thus different metrics related to network performance and fairness can be calculated during the optimization. As this work provides a benchmark solution to other PRA and opportunistic techniques, a convex formulation was proposed. Such optimal solution can be obtained using numerical optimization techniques or commercial solvers.

Simulation results show the resilience of the proposed PRA framework to meet QoS constraints under imperfect predictions, while achieving fairness among the users. Existing non-robust PRA adopted only the average of rates and thus violated the QoS at the beginning of the time horizon. Accordingly, the non-robust PRA can only be used when all the users are predicted to experience similar average rates. On the other hand, the non-predictive RA overlooked the future poor radio conditions and thus users moving towards cell edge suffered from buffer under run. This technique can be applied for cell edge users moving to the peak radio conditions or at very low load scenarios.

VII. Acknowledgement

This research is supported by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) under grant number: STPGP 479248.

References

[1] CISCO, "Cisco visual networking index: Global mobile data traffic forecast update, 2014-2019." http://www.cisco.com/c/en/us/solutions/service-provider/visual-networking-index-vni/index.html, 2015. Accessed Mar. 15th, 2016.

[2] Z. Lu and G. de Veciana, "Optimizing stored video delivery for mobile networks: The value of knowing the future," in *Proc. IEEE INFOCOM*, pp. 2806–2814, 2013.

[3] H. Abou-zeid, H. S. Hassanein, and N. Zorba, "Enhancing mobile video streaming by lookahead rate allocation in wireless networks," in *Proc. IEEE CCNC*, pp. 768–773, 2014.

[4] H. Abou-zeid and H. S. Hassanein, "Predictive green wireless access: Exploiting mobility and application information," *IEEE Wireless Commun.*, vol. 20, no. 5, pp. 92–99, 2013.

[5] A. Schulman, V. Navda, R. Ramjee, N. Spring, P. Deshpande, C. Grunewald, K. Jain, and V. N. Padmanabhan, "Bartendr: a practical approach to energy-aware cellular data scheduling," in *Proc. ACM Mobicom*, pp. 85–96, 2010.

[6] M. Elazab, A. Noureldin, and H. Hassanein, "Integrated cooperative localization for connected vehicles in urban canyons," in *Proc. IEEE Globecom*, pp. 1–6, 2015.

[7] A. Mahmoud, A. Noureldin, and H. Hassanein, "Distributed vehicle selection for non-range based cooperative positioning in urban environments," in *Proc. IEEE ICC*, pp. 1–7, 2016.

[8] H. Abou-zeid, H. S. Hassanein, Z. Tanveer, and N. AbuAli, "Evaluating mobile signal and location predictability along public transportation routes," in *Proc. IEEE WCNC*, pp. 1195 – 1200, 2015.

[9] R. Atawia, H. Abou-zeid, H. Hassanein, and A. Noureldin, "Robust resource allocation for predictive video streaming under channel uncertainty," in *Proc. IEEE GLOBECOM*, pp. 4683–4688, Dec 2014.

[10] R. Atawia, H. Abou-zeid, H. Hassanein, and A. Noureldin, "Chance-constrained qos satisfaction for predictive video streaming," in *Proc. IEEE LCN*, pp. 253–260, 2015.

[11] R. Atawia, H. Abou-zeid, H. S. Hassanein, and A. Noureldin, "Joint chance-constrained predictive resource allocation for energy-efficient video streaming," *IEEE J. Select. Areas Commun.*, vol. 34, pp. 1389–1404, May 2016.

[12] R. Ramamonjison and V. K. Bhargava, "Sum energy-efficiency maximization for cognitive uplink networks with imperfect csi," in *Proc. IEEE WCNC*, pp. 1012–1017, 2014.

[13] N. Y. Soltani, S.-J. Kim, and G. B. Giannakis, "Chance-constrained optimization of ofdma cognitive radio uplinks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 3, pp. 1098–1107, 2013.

[14] M. Abdel-Rahman and M. Krunz, "Stochastic guard-band-aware channel assignment with bonding and aggregation for dsa networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 7, pp. 3888–3898, 2015.

[15] H. Farahat and H. Hassanein, "Optimal caching for producer mobility support in named data networks," in *Proc. IEEE ICC*, pp. 1–6, 2016.

[16] H. Abou-zeid, H. S. Hassanein, and R. Atawia, "Towards mobility-aware predictive radio access: modeling; simulation; and evaluation in lte networks," in *Proc. ACM MSWiM*, pp. 109–116, 2014.

[17] 3GPP, "LTE; evolved universal terrestrial radio access (E-UTRA); further advancements for E-UTRA physical layer aspects," 2010.

[18] H. Holma and A. Toskala, *LTE for UMTS-OFDMA and SC-FDMA based radio access*. John Wiley & Sons, 2009.