Robust Resource Allocation for Predictive Video Streaming Under Channel Uncertainty

Ramy Atawia*, Hatem Abou-zeid*, Hossam S. Hassanein† and Aboelmagd Noureldin‡
*Electrical and Computer Eng. Dept., Queen's University, Canada, {ramy.atawia, h.abouzeid}@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*—Novel mobility-aware resource allocation schemes have recently been introduced for efficient transmission of stored videos. The essence of such mechanisms is to lookahead at the future rates users will experience, and then strategically buffer content into user devices when they are at peak radio conditions. For example, a user approaching poor coverage will be preallocated additional video segments to ensure smooth streaming. Advances in mobility prediction and real-time radio environment map updates are driving forces for such Predictive Video Streaming (PVS) mechanisms. Although previous efforts have demonstrated the large potential gains of PVS, ideal channel predictions were assumed. This paper addresses the problem of channel uncertainty in PVS, and proposes a robust resource allocation framework that 1) models channel uncertainty, 2) solves the PVS problem with a tunable level of quality of service guarantees, and 3) learns the degree of uncertainty, and adapts the channel model accordingly. Numerical results demonstrate the effectiveness of the proposed approach for PVS under channel variability.**

I. Introduction

Network operators are facing formidable resource management challenges to cope with the phenomenal growth of mobile traffic. Specifically, video accounted for over 50% of the traffic in 2012, with projections of a 14-fold increase by 2018 [1]. To address this growth, *predictive* resource allocation techniques that exploit user mobility have been recently proposed to improve throughput and fairness [2], [3], as well as video streaming delivery [4]–[7]. This is accomplished by leveraging the knowledge of the future rates users are expected to experience, and then performing *long-term* Resource Allocation (RA) plans over several seconds. By doing so, Base Stations (BSs) can schedule more resources to users during their respective peaks, and prioritize users that are headed to poor channel conditions. This is opposed to *instantaneous* RA and admission control strategies [8], [9]. Long-term RA planning is particularly useful for stored video delivery that can be strategically buffered in advance at the users' devices. For instance, if it is known a user is approaching a low coverage area, content can be prebuffered to support smooth streaming. Furthermore, as this enables efficient content prebuffering, energy is saved since transmission will not be needed during poor conditions [4], [5], [7].

The underlying assumption of Predictive Resource Allocation (PRA) approaches is that a user's future channel states are highly reproducible. This is achieved by coupling mobility information with a Radio Environment Map (REM) or Bandwidth (BW) map, typically generated with road drive tests measuring signal strength and network performance metrics at different locations. Indeed, analyses on human mobility traces reveal that people tend to follow particular routes regularly [10], [11], and several practical studies investigating the correlation between location and received data rates have also been conducted [3], [12], [13]. Yao et al. [12] analyze bandwidth traces collected from two independent cellular providers for routes running through different radio conditions including terrestrial and underwater tunnels. Their findings confirm the correlation between mobile bandwidth and location. The work in [13] conducts a similar measurement study, and addresses other contextual factors such as user speed, time of day, and humidity to predict the available bandwidth more accurately. However, while such maps provide a reasonable estimate of the wireless data rates, they do not accurately capture the dynamics of network congestion or environmental/geographical changes. For instance, a BW map may be more accurate in a rural area and less so in a more urban region. Further, the accuracy of the map may change with time due to fluctuations in network dynamics at rush hours vs. other times. Therefore, there is a need to model rate prediction uncertainty itself, and thereafter develop PRA solutions that incorporate such models. To this end, this paper presents a fuzzy-based robust RA framework for Predictive Video Streaming (PVS) under channel uncertainty. We summarize the main contributions of this paper in the following:

- We model uncertainty in the REM measurements by using triangular fuzzy numbers. We show that the triangular membership function provides a good approximation of the REM variations, if the Signal to Noise Ratio (SNR) exhibits a Gaussian prediction error.
- We develop a robust RA framework for predictive video streaming that incorporates the fuzzy REM. The framework allows the operator to control the desired degree of constraint satisfaction under rate uncertainty.
- The proposed framework also 'learns' the degree of uncertainty in the REM through feedback and prediction, via a Kalman Filter (KF), and tunes the fuzzy model to reflect the current channel variability.

As opposed to previous works on PRA [2]–[7], we also implement the system in a standard compliant Long Term Evolution (LTE) simulator [14] for more practical results.

978-1-4799-3512-3/14/$31.00 ©2014 IEEE

A. Related Work

The work in [4]–[6] are closest to this paper where rate predictions are used to minimize system utilization for stored video delivery. The authors in [4] consider the optimization problem for the multi-user single cell case, and develop optimal RA algorithms for the single user case. In [5], [6], we also discuss the potential energy savings that can be achieved by a *mobility-aware* wireless access framework. An architecture is presented with the composite functional elements and their interaction is discussed. However, in both these works, ideal channel predictions are assumed and the proposed solutions do not incorporate uncertainty, or provide robust measures to ensure streaming continuity under channel variability. This is addressed in this paper through a Robust Resource Allocation (RRA) framework for PVS. It is worth noting that robust allocation provisions have been proposed for other network functions and application as in [15], [16], where the uncertainty is in the *instantaneous* Channel Quality Indicator (CQI).

B. Paper Organization

In the following section, we present the system model. Section III presents the predictive streaming optimization problem without channel uncertainty considerations. The proposed RRA framework is then presented in Section IV, and applied to the streaming problem of Section III. We discuss the numerical results in Section V, and conclude in Section VI.

II. System Model

We use the following notational conventions: $\mathcal{X}$ denotes a set and it's cardinality is denoted by X. Matrices are denoted with bold letters as follows $\mathbf{x} = (x_{a,b} : a \in \mathbb{Z}_+, b \in \mathbb{Z}_+)$.

A. System Overview

Consider a BS with an active user set $\mathcal{M}$, where an arbitrary user is denoted by $i \in \mathcal{M}$. Users enter from the left cell edge and move in a straight line towards the other edge, requesting stored video content that is transported over HTTP (i.e. as in progressive download). We assume that the wireless link is the bottleneck, and therefore the core network bandwidth is set to 1 Gbps and the video content is always available at the BS.

B. Radio Environment Map and Mobility Information

The REM assumed to be typically available at the service provider would contain the average data rates at different network locations. In order to model such a radio map, we use the Friis Spectrum Loss propagation model in ns-3 [14]. The Signal to Interference plus Noise Ratio (SINR) at each x and y coordinate that users traverse is then computed and the corresponding achievable rate is determined based on the CQI-to-Modulation and Coding Scheme (MCS) mapping in 3rd Generation Partnership Project (3GPP) standards for LTE [17]. We assume that user mobility information is known accurately for the upcoming T seconds, which we call the prediction window, and at a per second granularity. This results in a total of T time slots within the prediction window, which we denote by the set $\mathcal{T} = \{1, 2, \cdots, T\}$. From this information, we construct a matrix of future user rates, defined by $\hat{\mathbf{r}} = (\hat{r}_{i,t} : i \in \mathcal{M}, t \in \mathcal{T})$. The values in this matrix will then be fuzzified to account for uncertainty according to the model presented in Section IV-C1.

C. Resource Sharing and Scheduling

BS airtime is shared among the active users during each slot t. We define the resource allocation matrix $\mathbf{x} = (x_{i,t} \in [0, 1] : i \in \mathcal{M}, t \in \mathcal{T})$ which gives the fraction of time during each slot t that the BS bandwidth is assigned to user i. The rate received by each user, at each slot, is the element-wise product $\mathbf{x} \odot \hat{\mathbf{r}}$. Airtime sharing is implemented as a time division rate controller over the Round-robin (RR) scheduler in ns-3 [14].

III. Predictive Video Streaming: Limitations of Crisp RA Formulations

The essence of predictive video streaming is to strategically transmit content ahead of time at the User Equipment (UE), after which transmission can be momentarily suspended while the user consumes the buffer [4], [6]. If we consider a user requesting a stored video at slot $t = 1$, with a streaming rate of V [bit/s], then the minimum cumulative video content for smooth streaming is $D_{i,t} = V \cdot t$. The cumulative allocation made to a user i by slot t is denoted by $A_{i,t} = \sum_{t_1=1}^{t} x_{i,t_1} \hat{r}_{i,t_1}$. To experience smooth streaming, $A_{i,t} \geq D_{i,t} \, \forall \, t$ for user i.

It has been illustrated in [5] how BS transmission time can be minimized by leveraging future user rate knowledge. A *predictive* scheme will wait to make bulk transmissions at times of high channel conditions, while making the minimal transmissions that avoids video stalling at other times. This achieves lower airtime usage, resulting in lower power consumption or more resources for other services. The corresponding optimization problem of minimizing BS airtime, without causing any streaming discontinuities can be formulated as the following Linear Program (LP) [6]:

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \sum_{t=1}^{T} \sum_{i=1}^{M} x_{i,t} && (1) \\ \text{subject to:} \quad & \text{C1:} \quad D_{i,t} - A_{i,t} \leq 0, && \forall \, i \in \mathcal{M}, t \in \mathcal{T}, \\ & \text{C2:} \quad \sum_{i=1}^{M} x_{i,t} \leq 1, && \forall t \in \mathcal{T}, \\ & \text{C3:} \quad x_{i,t} \geq 0 && \forall \, i \in \mathcal{M}, t \in \mathcal{T}. \end{aligned}$$

Constraint C1 ensures that the cumulative video content requirement is not violated at each time slot, while C2 expresses the resource limitation at each base station. It ensures that the sum of the airtime of all users is equal to 1 at every time slot. Finally, C3 provides the bounds for the resource allocation factor.

The solution of the LP in Eq. 1 minimizes airtime without degrading the video only if the predicted rates are accurate. For example, if the actual rate is less than the predicted rate, the airtime is minimized, but the user will suffer from video stalls. On the other hand, if the actual rate is greater than the

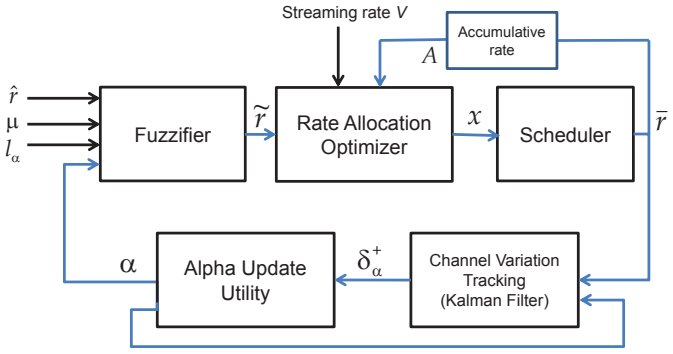


Fig. 1. Robust resource allocation framework for PVS.

predicted one, then a prebuffering opportunity is lost, resulting in relatively higher total airtime, had the user capitalized on the high rate. To capture and adapt to such variations, we present a fuzzy-based robust RA framework in the following section.

IV. Robust Resource Allocation Framework for Predictive Video Streaming

A. Overview

Fig. 1 illustrates the proposed RRA framework for predictive video streaming. The fuzzifier determines the fuzzy rate $\tilde{r}$ which is then used by the rate allocation optimizer to plan the required airtime. The scheduler implements the airtime division among users and measurements of the rates experienced $\bar{r}$ are recorded. This is fed back to a channel variation tracker that predicts the current degree of uncertainty. Based on this, the fuzzifier modifies the membership function of the fuzzy rate $\tilde{r}$ to more accurately reflect the channel variations. The values of $\bar{r}$ are also fed back to the rate allocator to re-solve the RA problem based on the received user rates. We now present the RRA framework in detail.

B. Fuzzy-Based Rate Allocation Optimization

The fuzzy rate allocation optimizer used in the RRA framework is based on the fuzzy linear programming model introduced in [18] and [19]. In this approach, the fuzzified rate $\tilde{r}$ is determined based on 1) the degree of rate uncertainty, and 2) the required level of constraint satisfaction. The formulation in Eq. 1 can be updated to account for the fuzzified rate $\tilde{r}$ by modifying constraint C1 as follows:

$$\tilde{\text{C1}}: \quad D_{i,t} - \sum_{t_1=1}^{t} \tilde{r}_{i,t_1} x_{i,t_1} \leq 0, \qquad \forall\, i \in \mathcal{M}, t \in \mathcal{T}. \tag{2}$$

Once $\tilde{r}$ is obtained, an LP solver can be used to solve the airtime minimization problem defined in Eq. 1, with the fuzzy $\tilde{\text{C1}}$ constraint. We now discuss the details of determining $\tilde{r}$.

C. Fuzzifier: Modeling Rate Uncertainty

1) Rate Membership Function: We represented the fuzzy predicted rate $\tilde{r}_{i,t}$ by a triangular membership function as shown in Fig. 2. The right r_u and left r_l most points on the x-axis define the limits of the triangle's base, which physically represent the boundaries on the variation of the predicted rate $\hat{r}$. This can be expressed mathematically as:

$$\mu_{\tilde{r}} = \begin{cases} \mathsf{L}(\tilde{r}) = \frac{\tilde{r}-\hat{r}}{\hat{r}-r_l} + 1, & \text{if } r_l \leq \tilde{r} \leq \hat{r} \\ \mathsf{R}(\tilde{r}) = \frac{\tilde{r}-\hat{r}}{\hat{r}-r_u} + 1, & \text{if } \hat{r} \leq \tilde{r} \leq r_u \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

This membership function was found to be an acceptable approximation of a Gaussian error overlaid on the predicted rate, as shown in Fig. 3. The step structure appears due to the discrete MCSs in LTE, resulting in specific Transport Block (TB) sizes. As some TBs correspond to larger SNR ranges, there are some irregularities in the step structure.

2) Defining the Degree of Rate Uncertainty: The α-cut representation of the membership function indicates the values of fuzzy numbers ($\tilde{r}_{\alpha,l} \leq \tilde{r}_\alpha \leq \tilde{r}_{\alpha,u}$) with a degree of membership that is equal to or greater than α [18]. The α-cut of the left side of the triangular membership in Fig. 2 can be determined by setting $\mathsf{L}(\tilde{r}_{\alpha,l}) = \alpha$, and solving for $\tilde{r}_{\alpha,l}$:

$$\tilde{r}_{\alpha,l} = \alpha(\hat{r} - r_l) + r_l. \tag{4}$$

Similarly, the α-cut of the right side is:

$$\tilde{r}_{\alpha,u} = \alpha(\hat{r} - r_u) + r_u. \tag{5}$$

Depending on the degree of rate variations, a suitable value of α can be selected to reflect the uncertainty in the predicted rate. For instance, a dynamically changing environment will suffer from wide variations from the predicted rate, and thus a small value of α should be assigned as shown in Fig. 2. The corresponding fuzzy rate would include most of the values along the triangle's base. On the contrary, a higher value of α is suitable for a stable channel, with slight rate variations.

3) Controlling Constraint Satisfaction under Uncertainty: Although each α-cut results in a range of possible rates, only one value should be selected (i.e. as the predicted one). This value is then used to solve the rate allocation problem as shown in Fig. 1. This process is performed based on the desired degree of constraint satisfaction, which is denoted by $l_\alpha \in [0, 1]$. A larger l_α corresponds to a higher requirement

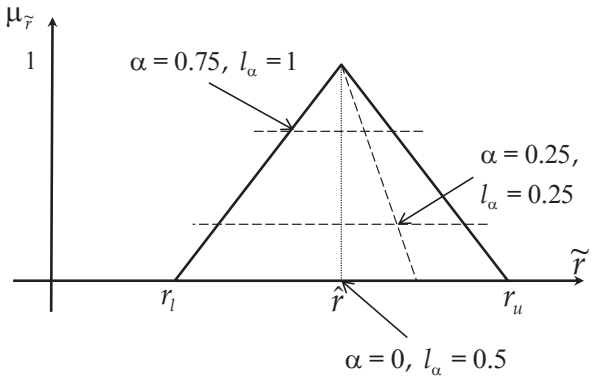


Fig. 2. Triangular membership function of the fuzzy predicted rate $\tilde{r}$ with different α-cuts, and l_α values.

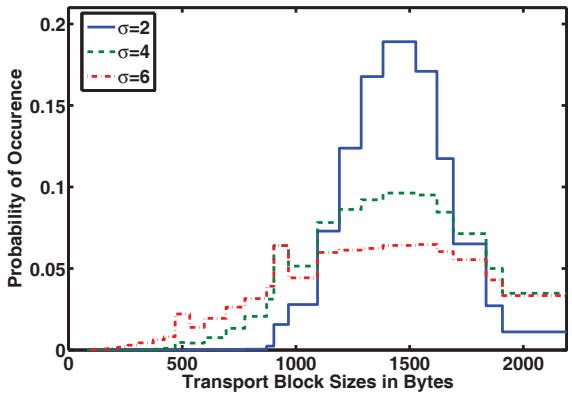


Fig. 3. Variation of TB size due to a Gaussian prediction error for different error variances σ. The originally predicted TB size is 1383 Bytes.

of constraint satisfaction [18], and therefore a very low rate should be selected from the set of the available rates in the (α)-cut. This will result in allocating more airtime to the user to ensure that the streaming constraint Eq. 2 is satisfied even with a pessimistic choice of $\tilde{r}$. On the other hand, a higher rate (i.e. an optimistic choice) of $\tilde{r}$ can be selected if the constraint satisfaction level is low. This will result in lower airtime and BS resource consumption. In other words, l_α provides an operator trade-off between guaranteeing Quality of Service (QoS) and minimizing BS airtime under rate uncertainty.

4) Determining the Fuzzified Rate ($\tilde{r}$): The degree of rate uncertainty (α), and the constraint satisfaction requirement (l_α) can be jointly coupled to determine the *fuzzified* rate ($\tilde{r}$) as illustrated in Fig. 2, and expressed mathematically as [18]:

$$\tilde{r} = l_\alpha \times \mathsf{L}(\tilde{r}_{\alpha,l}) + (1 - l_\alpha) \times \mathsf{R}(\tilde{r}_{\alpha,u}). \quad (6)$$

Here, the α-cut controls the ranges of $\mathsf{L}(\tilde{r}_{\alpha,l})$ and $\mathsf{R}(\tilde{r}_{\alpha,u})$, while the choice of l_α determines the final value selected within that range. To interpret Eq. 6 further, let us consider the follow cases:

1) **Highest Predictability** ($\alpha = 1$): If $\alpha = 1$, there is no rate uncertainty, and $\mathsf{L}(\tilde{r}_{\alpha,l}) = \mathsf{R}(\tilde{r}_{\alpha,u}) = \hat{r}$. Thus, $\tilde{r} = \hat{r}$, and the optimization problem in Eq. 1 can be solved directly.
2) **Lowest Predictability** ($\alpha = 0$): If $\alpha = 0$, the fuzzy rate will vary between the extreme values of r_l and r_u in Fig. 2. Consequently, depending on the constraint satisfaction l_α, the final value of $\tilde{r}$ is determined. For example if:
   - $l_\alpha = 0$: Such a choice is suitable if the network operator is either optimistic about the predicted rate, or has a higher preference to efficiency over QoS guarantees. In this case, Eq. 6 reduces to:

$$\tilde{r} = \mathsf{R}(\tilde{r}_{\alpha=0,u}) = r_u. \quad (7)$$

   - $l_\alpha = 1$: Now the inverse holds, and the operator wants to guarantee constraint satisfaction. According to Eq. 6, the lower bound rate is selected:

$$\tilde{r} = \mathsf{L}(\tilde{r}_{\alpha=0,l}) = r_l. \quad (8)$$

D. Adaptive α-Tuning: Tracking Rate Variability

In practice, the degree of rate variability will vary with geographical location and time. This lends a fixed value of α inefficient, and an approach is needed to predict and update α periodically based on feedback from the measured channel rates in the previous time slots. As illustrated in Fig. 1, this is accomplished in two stages discussed below.

1) Kalman Filter Rate Variability Predictor: The rate variability is determined repeatedly by the channel variation tracking function in Fig. 1. This is accomplished by comparing the difference δ_{α_t} between the previously calculated α-cut value α_{t-1}, and the measured α-cut $\bar{\alpha_t}$ corresponding to the actual rates $\bar{r_t}$ experienced by the user during the scheduling of the previous time slot as follows

$$\delta_{\alpha_t} = \bar{\alpha_t} - \alpha_{t-1}, \quad (9)$$

where $\bar{\alpha_t}$ is calculated by equating Eq. 6 to $\bar{r_t}$ and then solving for α as outlined below

$$\bar{\alpha_t} = \bar{r_t} - (l_\alpha r_l + (1-l_\alpha) r_u) l_\alpha (\hat{r} - r_l) + (1 - l_\alpha)(\hat{r} - r_u). \quad (10)$$

For a high channel variance, the measured α-cut $\bar{\alpha_t}$ will fluctuate and thus the error δ_{α_t} should be increased. On the other hand, stable channels will result in fairly equal measured α-cut $\bar{\alpha_t}$ values, and thus the error will start to decrease. Instead of calculating the error based only on the current measurements, a Kalman Filter (KF) will be used to track the error based on its previous values as well. The standard KF operations and equations are summarized below [20]:

Prediction Phase:

$$X_t^- = \Phi_t X_{t-1}^+ \quad (11)$$

$$P_t^- = \Phi_t P_{t-1}^+ \Phi_t^{'} + Q. \quad (12)$$

Measurement Phase:

$$K_t = P_t^- H_t^{'} (H_t P_t^- H_t^{'} + R)^{-1} \quad (13)$$

$$X_t^+ = X_t t^- + K_t (z_t - H_t X_t^-) \quad (14)$$

$$P_t^+ = P_t^- - K_t H_k P_t^- \quad (15)$$

where X_t^- and X_t^+ are the priori and posterior error values respectively. P_t^- and P_t^+ are the error estimation matrices respectively. H and Φ are the observation and state transition matrices respectively, while Q and R are the process and the measurement noise covariance matrices respectively, and K is the Kalman filter gain.

In our model, the priori error X_t^- represents the error in the calculated degree of uncertainty δ_α and is assumed to be the same as the corrected error of the previous time step X_{t-1}^+. Thus, the state transition matrix is set to unity. The observation z_t represents the current error in the degree of uncertainty based on the current measurements δ_{α_t} shown in Eq. 9. The observation is updated every time slot based on the average

measured rate $\bar{r_t}$. Since the observations z_t and the predicted state value X_t^- represent values for the errors in the degree of uncertainty, the state observation matrix H is set to unity. The values of Q, R and the initial value of P (P_0^+) are obtained from excessive tuning and their values are shown in Table I. In summary, the KF equations (11-15) are modified as follows:

Prediction Phase:

$$\delta_{\alpha t}^- = \delta_{\alpha t-1}^+ \tag{16}$$

$$P_t^- = P_{t-1}^+ + Q. \tag{17}$$

Measurement Phase:

$$K_t = P_t^-(P_t^- + R)^{-1} \tag{18}$$

$$\delta_{\alpha t}^+ = \delta_{\alpha_t}^- + K_t(\delta_{\alpha_t} - \delta_{\alpha_t}^-) \tag{19}$$

$$P_t^+ = P_t^- - K_t P_t^- \tag{20}$$

2) α-Tuning Utility: After estimating the rate variation from the tracked error $\delta_{\alpha t}^+$ using the KF, the value of α_t is determined. When rate variations are low, $\delta_{\alpha t}^+ \approx 0$, which should correspond to $\alpha_t \approx 1$. Contrarily, high variations result in larger values of $\delta_{\alpha t}^+$, which should lead to $\alpha_t \approx 0$. This mapping is accomplished using the following utility

$$\alpha_t = 1 - e^{-\gamma/|\delta_{\alpha t}^+|}, \tag{21}$$

where γ controls the rate of decrease with increasing $\delta_{\alpha t}^+$.

V. Performance Evaluation

A. Simulation Set-up

The simulation is performed using the LTE module in the Network Simulator (ns-3) [14], with model parameters as indicated in Table I. Gurobi [21] is used to solve the rate allocation optimization and is integrated in the simulator. Average BS airtime and video degradation (VD) are the performance metrics, where VD is the fraction of constraint violation, i.e. Eq. 2.

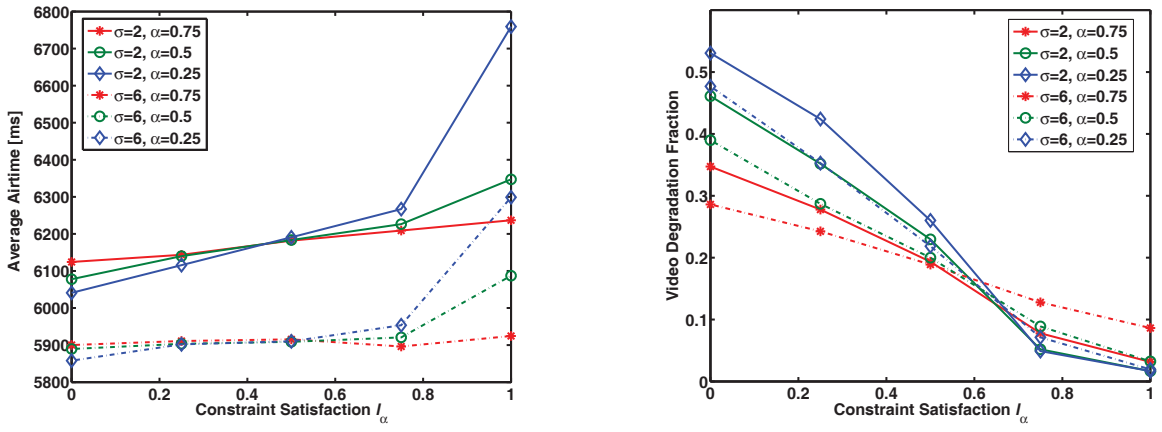
TABLE I
Summary of Model Parameters

Parameter	Value
BS transmit power	43 dBm
BW	5 MHz
T	60 s
τ	1 s
V	1 [Mbps]
BER	5×10^{-5}
Velocity	30 km/h
P_0	1
Q	0.1
R	10
α_0	0.5
δ_{α_0}	0.7
Packet size	8×10^3 [bits]
Packet rate (from core network to BS)	$10^3 s^{-1}$
Total number of packets	7.5×10^3
Buffer size	10^9 [bits]

B. Effect of Constraint Satisfaction (l_α)

In Fig. 4, we investigate the effect of the constraint satisfaction level l_α, for a single user. The results were averaged over 100 different log-normal error distributions for different variances σ, and predictability levels (α-cuts).

- Fig. 4(a) shows that as l_α increases, airtime also increases. This is because higher l_α values will result in a smaller $\tilde{r}$, thereby requiring more airtime to satisfy Eq. 2. The result is lower VD as illustrated in Fig. 4(b), since the constraints are satisfied with higher probabilities.
- To almost eliminate VD, $l_\alpha \approx 1$ and $\alpha \approx 0.25$. This cause a sharp increase in the consumed airtime.
- As the α-cuts increase, the airtime increases for $l_\alpha < 0.5$, but decreases for $l_\alpha > 0.5$. The reason is that $l_\alpha \approx 0.5$ is an inflection point, where $\tilde{r} > \hat{r}$ for $l_\alpha < 0.5$, while $\tilde{r} < \hat{r}$ for $l_\alpha > 0.5$ as illustrated in Fig. 2. As α decreases, the deviation from $\hat{r}$ increases, so the effects are more pronounced at both extremes of l_α. A similar reasoning can be applied to trends of VD in Fig. 4(b).
- Average airtime is less for higher error variances σ. Referring to Fig. 3, we can see that a higher variance has a



(a) Average BS airtime.

(b) Video degradation fraction.

Fig. 4. Video degradation fraction VD and average BS airtime for varying constraint satisfaction levels l_α, α-cuts, and error variances σ.

TABLE II
EFFECT OF ADAPTING α, WITH $l_\alpha = 0.75$ WITH MULTI-VARIANCE ERROR.

α	0	0.5	0.75	1	*Adapt.*
VD Fraction	0.08	0.01	0.17	0.19	0.09
Airtime [s]	6.2	5.9	5.89	5.86	5.7

TABLE III
EFFECT OF ADAPTING α, WITH $l_\alpha = 0.75$ WITH THE MULTI-USER, MULTI-VARIANCE ERROR SCENARIO.

α	0	0.5	0.75	1	*Adapt.*
VD Fraction	0.04	0.08	0.14	0.35	0.05
Airtime [s]	30.2	26.1	25.1	24	26.6

higher probability of large TB transmissions. Even though there is a higher probability of small TB transmissions as well, the overall effect is a reduced airtime since a few large TBs are sufficient to buffer the video. As a result, the VD is generally less for higher σ. However, this is not the case for larger degrees of constraint satisfaction l_α, where it is paramount that the perceived user rates are greater than the value of $\tilde{r}$.

C. Effect of Rate Variations and Adaptive α-Tuning

We now investigate the effect of variable degrees of channel uncertainty and the potential gains or adapting α with time. In this scenario, the error variance is initially high $\sigma = 6$, and then decreases to $\sigma = 2$ as the user approaches the cell center. The airtime and VD results are jointly presented in Table II for different values of α-cuts (at an $l_\alpha = 0.75$), and compared to the adaptive-α scheme based on the KF. We can see that lowering α reduces degradation, but at the cost of an increased airtime. However, keeping α constantly low, is not ideal in this case since the channel error variance decreases to $\sigma = 2$ during the simulation. The proposed adaptive-α approach is able to decrease its value when the channel variance is high in order to avoid VD, and then increase α when the channel variance is low in order to satisfy the constraint with lower airtime. This results in an acceptable VD with a low airtime as illustrated in Table II. This scenario was extended to the multi-user case, where 8 users enter the cell with an inter-arrival rate of 5 seconds, requesting a video of 0.5 Mbps. The results in Table III further emphasized the importance of adapting α using the KF.

VI. CONCLUSION

In this work, we developed a fuzzy-based RRA framework that incorporates channel uncertainty into the PVS problem, and provides a tunable level of service guarantees. We also find that it is important to learn the degree of uncertainty in order to meet the desired constraint satisfaction levels without unnecessary resource consumption. To this end, we incorporate feedback in the framework to learn and adapt to the degree of channel uncertainty, and re-optimize the RA in PVS. A detailed numerical analysis of the framework was conducted to investigate the effects of channel variability and provide insights to further developments. Future work includes studying the performance in more complex simulation settings, as well as investigating the use of stochastic models and other channel variability predictors within the RRA framework.

ACKNOWLEDGEMENT

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

REFERENCES

[1] CISCO, "Cisco visual networking index: Global mobile data traffic forecast update, 2013-2018," 2014. Accessed Apr. 29th, 2014.
[2] H. Abou-zeid, H. Hassanein, and S. Valentin, "Optimal predictive resource allocation: Exploiting mobility patterns and radio maps," in *Proc. IEEE GLOBECOM*, pp. 4714–4719, 2013.
[3] R. Margolies, A. Sridharan, V. Aggarwal, R. Jana, N. K. Shankaranarayanan, V. A. Vaishampayan, and G. Zussman, "Exploiting mobility in proportional fair cellular scheduling: Measurements and algorithms," in *Proc. IEEE INFOCOM*, 2014, to appear.
[4] 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.
[5] 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.
[6] H. Abou-zeid and H. S. Hassanein, "Efficient lookahead resource allocation for stored video delivery in multi-cell networks," in *Proc. IEEE Wireless Commun. and Netw. Conf. (WCNC)*, 2014, to appear.
[7] H. Abou-zeid, H. S. Hassanein, and S. Valentin, "Energy-efficient adaptive video transmission: Exploiting rate predictions in wireless networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 5, pp. 2013 – 2026, 2014.
[8] M. Katoozian, K. Navaie, and H. Yanikomeroglu, "Utility-based adaptive radio resource allocation in OFDM wireless networks with traffic prioritization," *IEEE Trans. on Wireless Commun.*, vol. 8, no. 1, pp. 66–71, 2009.
[9] J. B. Othman, L. Mokdad, and S. Ghazal, "Performance analysis of wimax networks ac," *Wireless Personal Communications*, vol. 74, no. 1, pp. 133–146, 2014.
[10] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson, "Approaching the limit of predictability in human mobility," *Scientific reports*, vol. 3, no. 2923, pp. 1–9, 2013.
[11] C. Song, Z. Qu, N. Blumm, and A. Barabasi, "Limits of predictability in human mobility," *Science*, vol. 327, pp. 1018–1021, 2010.
[12] J. Yao, S. S. Kanhere, and M. Hassan, "An empirical study of bandwidth predictability in mobile computing," in *Proc. ACM WiNTECH*, pp. 11–18, 2008.
[13] D. Han, J. Han, Y. Im, M. Kwak, T. T. Kwon, and Y. Choi, "MASERATI: Mobile adaptive streaming based on environmental and contextual information," in *Proc. ACM WiNTECH*, pp. 33–40, 2013.
[14] G. Piro, N. Baldo, and M. Miozzo, "An LTE module for the ns-3 network simulator," in *Proc. Int. ICST Conf. on Simulation Tools and Techniques*, pp. 415–422, 2011.
[15] N. Zorba and A. I. Perez-Neira, "Robust power allocation schemes for multibeam opportunistic transmission strategies under quality of service constraints," *IEEE J. Select. Areas Commun.*, vol. 26, no. 6, pp. 1025–1034, 2008.
[16] T. Q. Quek, M. Z. Win, and M. Chiani, "Robust power allocation algorithms for wireless relay networks," *IEEE Trans. Commun.*, vol. 58, no. 7, pp. 1931–1938, 2010.
[17] 3GPP, "LTE; evolved universal terrestrial radio access (E-UTRA); physical layer procedures," Technical Specification TS 36.213 v8.8.0, 2009.
[18] X. Liu, "Measuring the satisfaction of constraints in fuzzy linear programming," *Fuzzy Sets and Systems, Elsevier*, vol. 122, no. 2, pp. 263–275, 2001.
[19] B. Melián and J. L. Verdegay, "Using fuzzy numbers in network design optimization problems," *IEEE Trans. Fuzzy Systems*, vol. 19, no. 5, pp. 797–806, 2011.
[20] A. Noureldin, T. B. Karamat, and J. Georgy, *Fundamentals of Inertial Navigation, Satellite-Based Positioning and Their Integration*. Springer, 2013.
[21] Gurobi, "Gurobi Optimization." http://www.gurobi.com/. Accessed Feb. 11th, 2014.