Robust Content Delivery and Uncertainty Tracking in Predictive Wireless Networks

Ramy Atawia, *Student Member, IEEE*, Hossam S. Hassanein, *Fellow, IEEE*, Hatem Abou-zeid, *Member, IEEE*, and Aboelmagd Noureldin, *Senior Member, IEEE*

***Abstract*— Predictive resource allocations (PRAs) have recently gained attention in wireless network literature due to their significant energy-savings and quality of service (QoS) gains. This enhanced performance was primarily demonstrated while assuming the *perfect* prediction of both mobility traces and anticipated channel rates. While the results are very promising, several technical challenges need to be overcome before PRAs can be practically adopted. Techniques that model the prediction *uncertainty* and provide *probabilistic* quality of service (QoS) guarantees are among such challenges. This differs from the traditional robust optimization of wireless resources, as PRAs use a time horizon with predicted demands and anticipated data rates. In this paper, we tackle this problem and present an energy-efficient *stochastic* PRAs framework that is robust to prediction uncertainty under *generic* error probability density functions. The framework is applied for video delivery, where the desired video demands are modeled as probabilistic chance constraints over the prediction time horizon, and a deterministic closed form is then derived based on the Bernstein approximation (BA). In addition to handling prediction uncertainty, mechanisms that track the variance of the channel in real-time are practically needed. Towards this end, we demonstrate how a particle filter (PF) can be adopted to effectively achieve this functionality. A low complexity guided heuristic algorithm is also integrated with the BA-based allocations, and particle filter (PF), to provide a real-time solution. Extensive numerical simulations using a standard compliant long term evolution system are then presented to examine the developed solutions under various operating conditions. Results indicate the ability of our framework to significantly reduce base station energy consumption while satisfying users' QoS under practical prediction uncertainty.**

***Index Terms*— Channel state prediction, energy efficiency, particle filter, radio access networks, resource allocation, robustness, video streaming.**

I. Introduction

MOBILE data traffic is anticipated to reach more than 24 exabytes per month by 2019, corresponding to a compound annual growth rate (CAGR) of 57% [5]. Around 70% of such growth is expected to be mobile video content that has to be delivered over the existing spectrum and infrastructure in order to maximize the operator's revenue. In addition to QoS satisfaction, mobile operators are currently concerned with energy-related operational and capital expenditures for both current and future radio networks [6]–[8]. Since Radio Access Networks (RANs) account for more than 50% of the network energy consumption, designing energy-efficient Resource Allocation (RA) strategies for video traffic are of paramount importance [7]. Opportunistic RA [9], [10] that relies on current or previous channel measurements are bound to fail in exploiting the available resources for delivering video content efficiently [1], [3]. This is because such traditional RA is not able to leverage the long-term channel characteristics, and typically minimizes only the *short-term* energy consumption by reducing the transmission power [7], [9], [10].

Recent studies on human mobility show that people tend to follow repetitive movement patterns [11], which make the channel capacities of mobile users highly predictable [12]–[14]. Consequently, Predictive Resource Allocation (PRA) that exploits future channel conditions over a time horizon, has recently been proposed to improve video streaming quality [1], and reduce transmission energy [2], [3]. PRA can prioritize users moving towards poor radio conditions, or delay transmission until a user reaches his peak radio conditions. However, the potential gains of PRAs [1]–[4] have been demonstrated under *ideal* predictions of future data rate and deterministic QoS constraints without considering uncertainties.

With the conventional approaches of PRA, evaluating system performance under real world uncertainty is challenging, and QoS guarantees are not possible. For instance, if the actual rates are less than what was predicted, user QoS may deteriorate significantly, thereby defeating the purpose of implementing PRA strategies. *Robust* PRA frameworks are therefore paramount to unleash the gains of predictions under real-life constraints. This involves 1) modeling the rate uncertainty, 2) developing models to provide probabilistic QoS guarantees, and 3) efficiently tracking the prediction uncertainty in real-time. Integrating these functionalities should enable PRAs to strike a balance between providing energy savings when predictions are accurate, and minimizing the risks associated with erroneous predictions during periods of uncertainty. In this paper, we investigate how such robustness can be practically incorporated into a PRA framework, and make the following contributions:

1) We present for the first time in literature a *stochastic* PRA framework that is *robust* to prediction uncertainty

Manuscript received May 20, 2016; revised November 5, 2016; accepted January 23, 2017. Date of publication March 13, 2017; date of current version April 7, 2017. The associate editor coordinating the review of this paper and approving it for publication was T. Taleb.

R. Atawia is with Electrical and Computer Engineering Department, Queen's University, Kingston, ON K7L 3N6, Canada (ramy.atawia@queensu.ca).

H. Hassanein is with the School of Computing, Queen's University, Kingston, ON K7L 3N6, Canada (hossam@cs.queensu.ca).

H. Abou-zeid is with the Cisco's R&D Center, Ottawa, ON K1R 7Y6, Canada (habouzei@cisco.com).

A. Noureldin is with Electrical and Computer Engineering Department, Royal Military College of Canada, Kingston, ON K7K 7B4, Canada (aboelmagd.noureldin@rmc.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2017.2662685

under *generic* error probability models. Herein, the objective is to minimize the Base Station (BS) energy for mobile video delivery while guaranteeing user QoS satisfaction.

2) We first show how the desired video QoS satisfaction levels can be modeled as probabilistic chance constraints in which the predicted rates are random variables, rather than the expected values used in PRA literature. Furthermore, in order to provide a solution that is not dependent on a particular error Probability Density Function (PDF), we adopt the Bernstein Approximation (BA) for the QoS constraint. The BA requires only the rate bounds and the average values in order to obtain a convex closed form representation for the probabilistic constraint.
3) Although the BA bounds can be calculated off-line, radio measurement studies reveal that the predictability of signal strength varies significantly with geographical location and time of day [13]. Therefore, a mechanism to *track* the variability in channel is needed for a practical solution. In this work, we also demonstrate how a particle filter (PF) can be adopted to effectively achieve this channel tracking functionality, and adapt the BA rate bounds accordingly.
4) Finally, we present a guided heuristic to provide a real-time solution for the BA formulation. This is important as, although the formulation is convex, the solution complexity is proportional to *both* the number of users and the length of the prediction window. The energy savings and QoS satisfaction levels are then compared to benchmark optimal solutions. We believe this work provides a practical direction towards the development of deployable PRAs in future generation networks.

In the following section, we review the related literature on the existing non-robust PRA and provide a background on robust resource allocation. Section III presents the system model and provides an overview for the introduced *robust* PRA framework. Section IV introduces the BA based formulation for robust energy-minimization. The guided heuristic for real-time optimization and the PF for channel tracking are then presented in Section V and Section VI, respectively. Simulation results are discussed in Section VII and finally, we conclude the paper in Section VIII.

II. Related Work

A. Existing Predictive Resource Allocation

With perfect knowledge of the future channel rates, the PRA techniques in [1]–[3] demonstrated how the total BS energy can be significantly reduced without any buffer under-runs at the user device. This was primarily achieved by leveraging knowledge of the future rate values of all the users. For instance, the BS may wait until the user reaches his peak radio conditions, and then pushes large portions of the video to avoid future allocation during the lower data rates. The BS can then go into sleep mode while the user plays back the prebuffered content. On the contrary, during poor radio conditions, no prebuffering is done and only minimal content is transmitted to sustain smooth playback. This strategy allows the PRA to transmit the video content with fewer resources compared to the traditional RA technique. The latter overlooks the future radio conditions and thus neither delays prebuffering, for cell edge user, nor prioritizes users at the cell center.

In practice, channel predictions are typically associated with uncertainties due to the low-power filters used in the mobile devices [15] and the random behaviour of the received signal level. Existing PRA techniques [1]–[4] represented such imperfections by the expected rate values, resulting in a deterministic formulation. The resultant decisions do not guarantee QoS satisfaction when predicted future rates fall below the expected values. In this case, the minimal airtime fraction allocated to the cell edge users will not be sufficient to meet their demand and buffer under-run (video stalls) occurs. In addition, when peak rates exhibit lower values than expected, energy savings will be suboptimal. The large amount of allocated airtime within these slots are not optimally utilized. We hence introduce a *robust energy-efficient* PRA that handles the deviations in predicted rates and thus avoids QoS violations and energy consumption under practical imperfect predictions.

B. Robust Optimization

Robust non-predictive RA techniques have been discussed in the literature in the context of handling uncertainties or delays in the user reported measurements [16]–[18]. Two fundamental optimization techniques namely *Fuzzy* and *Stochastic* are used to provide a robust formulation of the RA problem. In the former, the varying signal information is represented as fuzzy numbers associated by a membership function [19]. On the other hand, *Stochastic* optimization represents the uncertain values as random variables characterized by their probability density functions [20]. Although the *Fuzzy* approach does not change the order of complexity of the original non-robust formulation [19], an unsustainable conservatism is attained, resulting in suboptimal RA decisions [20]. *Stochastic* optimization, which is less conservative, was extensively adopted in non-predictive RA schemes. The main challenge remains to be the complexity.

Stochastic optimization utilizes two main techniques: Chance Constrained Programming (CCP) and Recourse Programming (RP) to handle the uncertainty in constraints and objective functions coefficients, respectively [20]. For the problem at hand, we focus on CCP to guarantee the QoS constraint satisfaction with random rates. CCP [21] represents 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. Such deterministic form should handle three main challenges: conservatism, safety and complexity. The first ensures that the constraints should not be over satisfied to avoid suboptimal network gains. The second challenge, safety, refers to the ability of capping the maximum violation probability by a certain degree denoted by ϵ. With regards to complexity, the robustness typically converts the linear RA formulation to a non-linear form. Hence, only convex and continuous formulations should be considered to obtain optimal robust solutions.

Gaussian Approximation (GA) obtains a deterministic CCP form using the inverse CDF of the random coefficients (channel rates in our case) [17], [22], [23]. However, the robustness is not guaranteed in long-term predictive RA due to the interdependence between the QoS constraints over the time horizon [24]. GA is applied only when the CDF is known and invertible, which might not be the case when multiple sources of prediction errors are present [25]. On the other hand, the Bernstein Approximation (BA), adopted in this paper, can be applied for generic CDFs of prediction errors [26], [27]. In essence, BA does not use the exact mathematical expression of the error's CDF. Instead, only the bounds (i.e. minimum and maximum values) of the error are adopted. The BA typically results in conservative solutions which have suboptimal energy savings due to the fact that the QoS constraints are over satisfied by a degree more than $1-\epsilon$ [24].

In this paper, we integrate the BA with a particle filter which tracks the channel variance based on the users' measurements. This integration balances the conservatism of the BA by allocating fewer resources to the users experiencing stable radio conditions. Using BA and PF for CCP and variation tracking, respectively, will make the framework applicable to prediction uncertainties with generic or complex error distributions. This is unlike our Robust Predictive Resource Allocation (R-PRA) works in [28] and [29] which used *fuzzy* and *GA*, respectively while assuming normally distributed and invertible prediction errors.

In general, the BA deterministic form will have a higher complexity order than the non-Robust form [24]. For instance, the BA will transform a linear CCP into a Second order Cone Programming (SoCP) which increases the computational burden [26]; due to the typically used convex optimization techniques such as Interior Point Method (IPM) [30], [31]. The robust non-predictive RA in [18] adopted the Markov inequality to approximate the CCP using a linear formulation. However, the optimal coefficients for such approximation are not easily attainable, and the degree of satisfaction ϵ will no longer model the trade-off between optimality and conservatism. Previous approaches in [16] and [17] tackled the complexity of the BA's SoCP by adopting either the first or the infinite order norms to obtain linear low-complexity deterministic forms for uplink non-predictive RA. However, both norms resulted in conservative solutions that are acceptable only for single time slot allocations (i.e. non predictive RA) to maximize the bandwidth efficiency. Moreover, the results in [24] demonstrated the high conservatism of BA when used in long-term PRA which will not allow any sort of linearization as done in [16] and [17]. The introduced framework in this paper will adopt the BA in its original SoCP form to optimize the energy savings. A real-time heuristic algorithm is thus developed to obtain near-optimal and feasible solutions by exploiting the structure of the problem at hand.

III. System Model and Framework Overview

In this section, we introduce the system model and the main blocks of the proposed robust adaptive PRA framework.

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. $\mathbf{x} = (x_{a,b} : a \in \mathbb{Z}_+, b \in \mathbb{Z}_+)$. $\tilde{r}$ represents a random variable (r.v.) and its expectation is denoted by $\mathbb{E}[\cdot]$. The $log(\cdot)$ denotes the natural logarithmic function and $\mathbb{1}_y$ is an indicator function which equals 1 if y is satisfied and 0 otherwise.

B. System Model

Each BS serves an active user set $\mathcal{M}$ where the user index is denoted by $i \in \mathcal{M}$. Each mobile user requests video with a fixed streaming rate. While current streaming standards are user driven, the network can access the files between the user and streaming server to determine the video specific information. Machine learning and data mining techniques are then exploited to predict the future bit rates that will be requested by the user. In-network caching is then adopted to request the video content ahead and store it in edge nodes near the user (e.g. at a BS) [32]. Thus, the main bottleneck is the wireless channel, and the main focus is to handle rate uncertainties, which were overlooked in existing PRA work [1], [2]. We assume that user's mobility trace is known for the next T seconds, called the prediction window $\mathcal{T}$, and at a per second granularity where $\mathcal{T} = \{1, 2, \cdots, T\}$. Future rate prediction is obtained by mapping the user's trace to the Radio Environment Map (REM) at the service provider. The REM contains the average rate for user i at time slot t and denoted as $\bar{r}_{i,t}$ [33]. While the average values are retrieved from the REM, their theoretical deviations, due to imperfect predictions or large scale fading, are calculated using Monte-Carlo simulation framework in [24]. Accordingly, the predicted uncertain rate is modelled as a random variable $\tilde{r}_{i,t} \in [r^l_{i,t}, r^u_{i,t}]$, where $r^l_{i,t}$ and $r^u_{i,t}$ are the lower and upper rate bounds, respectively, and the average value is $\bar{r}_{i,t} = \mathbb{E}[\tilde{r}_{i,t}]$. The active users can 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.

The problem addressed in this paper aims to find the optimal airtime fractions $\mathbf{x}$ based on the predicted rates such that the total BS transmission energy is minimized and video stalls are avoided.

C. System Overview

The proposed Robust Predictive Resource Allocation (R-PRA) framework aims to provide a real-time adaptive robust predictive allocation, and consists of three main blocks (see Fig. 1):

- **Robust Bernstein Formulation**: This considers satisfying the QoS constraints under uncertainties in the predicted rates. It firstly represents the constraints in a probabilistic form with a maximum violation level, denoted by ϵ, and then obtains the deterministic equivalent form. This inequality represents the relation between the unknown airtime fraction x, the average predicted

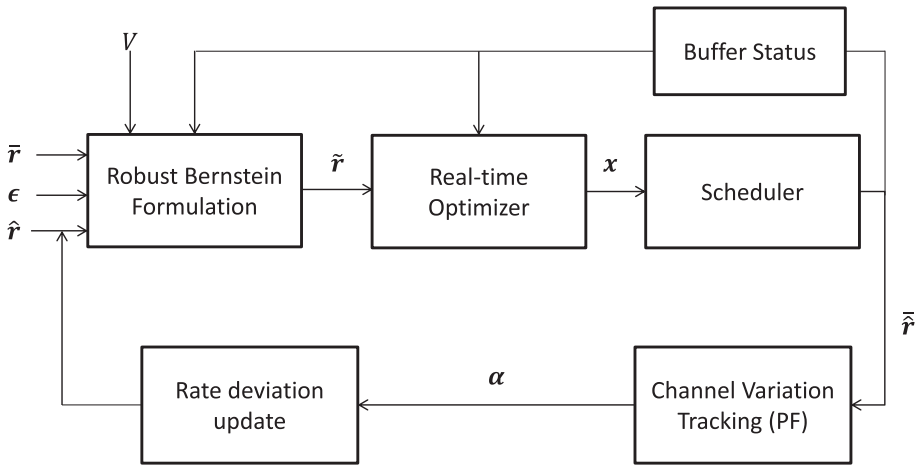


Fig. 1. Block diagram of the proposed *robust energy-efficient* PRA.

rate $\bar{r}$, maximum rate deviations $\hat{r}$, and the streaming demand V. In addition, this block continuously receives feedback from the users on their buffer status in order to either compensate for previous under-runs or avoid transmissions to full buffers. [1]

- **Real-time Optimizer**: Although the deterministic Bernstein form is convex, optimal gradient search methods cannot be adopted due to their high complexity. This module implements a low complexity local search guided algorithm that starts by satisfying the constraints and then moves on for optimizing the objective. The outcome is a real-time solution provided to schedulers and channel assignment modules.
- **Channel Variation Tracking**: The optimality of the robust form depends to a great extent on accurately modelling the rate deviations $\hat{r}$ which differ with time and location [13]. This module uses particle filter (PF) to track the degree of uncertainty α and adapts the rate deviations $\hat{r}$ based on the reported user measurements $\bar{\hat{r}}$ without prior knowledge of the channel statistics. In addition, it also allows cooperative tracking between users and thus provide real-time updates for new arriving users to the network.

In the following sections we explain the design details and challenges of each module.

IV. Bernstein Formulation for Robust Energy-Efficient PRA

In this section, we show the implementation the robust Bernstein formulation block. The *robust energy-efficient* PRA problem is firstly formulated and then Bernstein Approximation (BA) is used to obtain a deterministic equivalent form.

The uncertain predicted rate is represented as a random variable $\tilde{r}_{i,t}$ and thus the QoS constraint is probabilistic as shown in Eq. 1 below

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \ & \sum_{t=1}^{T}\sum_{i=1}^{M} x_{i,t} \\ \text{subject to: C1: } & Pr\left\{\sum_{t=0}^{t'} \tilde{r}_{i,t}x_{i,t} \geq D_{i,t'}\right\} \geq 1-\epsilon, \\ & \qquad \forall\, i \in \mathcal{M},\ t' \in \mathcal{T}, \\ & \text{C2: } \sum_{i=1}^{M} x_{i,t} \leq 1, \quad \forall t \in \mathcal{T}, \\ & \text{C3: } x_{i,t} \geq 0 \quad \forall i \in \mathcal{M},\ t \in \mathcal{T}. \end{aligned} \tag{1}$$

where $\epsilon \in [0, 1]$ is the maximum probability that video stops occur, and takes values less than 0.5 for reliable performance.

The objective function reflects the linear BS energy model in [3] and [34] which is calculated in terms of the total airtime fractions $x_{i,t}$ granted for all users. The BS is assumed to follow the Discontinuous Transmission (DTX) light sleeping mode in which some power consuming devices are turned off when no resources are allocated to the users [35]. Due to the lack of *power-control* in the current LTE 3GPP standard [36], the downlink power is constant and thus the energy consumption will depend fundamentally on the airtime.

The above formulation considers a fixed video quality, as in [37], where the media content is transmitted to each user at a constant streaming quality measured in Peak Signal-to-Noise Ratio (PSNR)[2]. Video freezing occurs when the total allocated airtime up to slot t results in delivering a total amount of video less than the corresponding cumulative streaming demand denoted as $D_{i,t} = V_i \times t$, where V_i is the fixed streaming demand of user i corresponding to the requested video quality. Accordingly, the QoS is said to be satisfied when the cumulative transmitted content is greater than the total demand $D_{i,t'}$ and thus no video stops occur. This is handled by constraint C1 which ensures smooth video playback with probability $1-\epsilon$.

[1]The robust Bernstein approximation in Section IV was first introduced in our prior work in [24].

[2]Such quality level can be guaranteed due to the availability of Channel Quality Indicator (CQI) from the users every TTI which enables the BS to select the appropriate Modulation and Coding Scheme (MCS).

The second constraint models the limited resources at each BS by ensuring that the sum of the allocated airtime is less than 1 second which is the duration of the allocation slot. Finally, C3 ensures the non-negativity of the assigned airtime fractions.

To obtain a robust deterministic form that is equivalent to Eq. 1, irrespective of the $\tilde{r}_{i,t}$ distribution, Bernstein Approximation (BA) is used. In essence, BA utilizes the marginal distribution and the moment generating function of the random variable. Generally, the chance constraint is represented as a linear summation of random variables as follows

$$Pr\left(f_0(\mathbf{x}) + \sum_{t=1}^{t'} \eta_t f_t(\mathbf{x}) \leq 0\right) \geq 1 - \epsilon, \quad \forall t' \in \mathcal{T}. \tag{2}$$

Here η_t is a random variable with marginal distribution $\mathbb{P}_t$, and $f_t(\mathbf{x})$ is a convex function containing the decision vector $\mathbf{x}$. Assuming that all the random variables η_t are independent, $\mathbb{P}_t$ has a bounded support on the interval $[-1, 1]$ $\forall t$ and the function $f_t(\mathbf{x})$ is affine in the decision vector x, a convex deterministic equivalent for Eq. 2 can be obtained as follows

$$\inf_{\lambda>0}\left[f_0(\mathbf{x}) + \sum_{t=1}^{t} \lambda \Lambda_t(\lambda^{-1} f_t(\mathbf{x})) + \lambda \log\frac{1}{\epsilon}\right] \leq 0, \quad \forall t \in \mathcal{T}. \tag{3}$$

Herein, $\Lambda_t(z)$ is the logarithm of the moment generating function $M_t(z)$ for r.v. z as depicted in Eq. 4

$$\begin{aligned} \Lambda_t(z) &= log M_t(z) \\ M_t(z) &= \mathbb{E}\left[e^{kz}\right] = \int e^{kz} d\mathbb{P}_t(k) \end{aligned} \tag{4}$$

Instead of computing the exact value of the logarithm moment generating function in Eq. 4, in addition to solving for the auxiliary variable λ, a conservative approximation using the upper bound can be adopted as in [38, eq. (5)].

$$\begin{aligned} \Lambda_t(z) &\leq max\left\{\mu_t^+ z, \mu_t^- z\right\} + \frac{\sigma_t^2}{2} z^2, \quad \forall t \in \mathcal{T} \\ -1 &\leq \mu_t^- \leq \mu_t^+ \leq 1 \end{aligned} \tag{5}$$

The variables μ_t^+, μ_t^- and σ_t are used to approximate the bounded support [38]. Therefore, a conservative deterministic equivalent for Eq. 3 is attained using Eq. 5 and the arithmetic inequality as follows

$$\begin{aligned} f_0(\mathbf{x}) + \sum_{t=1}^{t'} max\left\{\mu_t^+ f_t(\mathbf{x}), \mu_t^- f_t(\mathbf{x})\right\} \\ + \sqrt{2log(\frac{1}{\epsilon})\left(\sum_{t=1}^{t'} \sigma_t^2 f_t(\mathbf{x})^2\right)} \leq 0, \quad \forall t' \in \mathcal{T}. \end{aligned} \tag{6}$$

Finally, the robust PRA chance constraint C1 in Eq. 1 is replaced by Eq. 6 as depicted in Eq. 7

$$\begin{aligned} \sum_{t=1}^{t} \bar{r}_{i,t} x_{i,t} + \sum_{t=1}^{t'} \mu_{i,t}^- \hat{r}_{i,t} x_{i,t} \\ - \sqrt{2log(\frac{1}{\epsilon})\left(\sum_{t=1}^{t'} (\sigma_{i,t} \hat{r}_{i,t} x_{i,t})^2\right)} \geq D_{i,t'}, \quad \forall t' \in \mathcal{T}, \end{aligned} \tag{7}$$

where the random predicted rate $\tilde{r}_{i,t}$ is assumed bounded in $[r_{i,t}^l, r_{i,t}^u]$. To satisfy the assumptions for Eq. 3, this rate is normalized in $[-1, 1]$ by using the maximum deviation and the average values denoted by $\hat{r}_{i,t}$ and $\bar{r}_{i,t}$, respectively per

$$\begin{aligned} \hat{r}_{i,t} &= \frac{r_{i,t}^u - r_{i,t}^l}{2}, \quad r_{i,t}^u > r_{i,t}^l \\ \bar{r}_{i,t} &= \frac{r_{i,t}^u + r_{i,t}^l}{2} \end{aligned} \tag{8}$$

The constraint in Eq. 7 is a SoCP model which is convex for $\epsilon < 0.5$ and $x_{i,t} \in [0, 1]$ [39].

V. Real-Time Optimizer

This section discusses the implementation of the second block in Fig. 1. We firstly explore the limitations of existing optimal solvers, and then provide the details of the proposed real-time heuristic algorithm. Finally, computational complexity of both optimal and heuristic techniques is analyzed.

A. Optimal Solution

The formulation in Eq. 7 is a SoCP for $\epsilon < 0.5$ thus convex and continuous [39]. Its optimal solution can be obtained using Interior Point Method (IPM) [30] which is efficiently implemented in many commercial solvers such as Gurobi [40]. In particular, IPM searches within the set of feasible solutions for the optimal value where the latter is recognized due to its zero (or very small) duality gap. Although the IPM was proved to reach the optimality conditions in fixed number of iterations [31], the complexity per iteration hinders real-time solutions and still depends on the number of constraints. As seen from Eq. 1, the dimension of constraints increases with both the number of users M and the length of the time horizon T. In addition, the resource limitation constraint (C3) might cause the dissatisfaction of the QoS constraint (C2) especially at small values of ϵ. In this case, the QoS constraint has to be relaxed which requires extra computations. Our framework hence relies on a suboptimal heuristic algorithm to provide a real-time solution, while optimal techniques (e.g. IPM) are used for benchmarking only.

B. Guided Local Search Heuristic

The guided search algorithm proceeds by allocating the airtime that ensures exact satisfaction of QoS constraint (i.e., solves C1 in Eq. 7 as equality) to minimize the airtime. The radio capacity constraint is then checked (i.e., C2 in Eq. 1) and reallocation is done in case of violating the maximum time slot duration. Finally, the algorithm pushes all the remaining video content when the user reaches his peak radio conditions (i.e. maximum $\bar{r}$) to avoid allocation in future time slots with lower rates. The second and third steps are very challenging in multi-user scenarios where different users might experience their peak radio conditions simultaneously. The heuristic is summarized in Algorithm 1 and detailed as follows

Algorithm 1 Local-Search Guided Heuristic for Robust Allocation

Input : Users: $\mathcal{M}$, Time Horizon: $\mathcal{T}$, Average Predicted Rates: $\bar{R}$, Rate Bounds: $\hat{R}$, Maximum Violation: ϵ and Streaming Demand: D;
Output : X;
Initialization: $X = \emptyset$, $N_t = 0\ \forall t \in T$

1 **for** $i \in \mathcal{M}$ **do**
2 $\quad \hat{t}_i = \underset{t\in\mathcal{T}}{\text{argmax}}\{\bar{R}_i\}, \forall i \in \mathcal{M}$;
3 $\quad t = 0$;
4 $\quad$ **while** $t < \hat{t}_i$ **do**
5 $\quad\quad$ Transform Eq. 7 to equality and solve for $x_{i,t}$;
6 $\quad\quad N_t = N_t + x_{i,t}$;
7 $\quad$ **end**
8 **end**
9 **for** $t \in \mathcal{T}$ **do**
10 $\quad$ **if** $N_t > 1$ **then**
11 $\quad\quad j = \underset{i\in\mathcal{M}}{\text{argmax}}\left\{\frac{\bar{r}_{i,t}}{\max_{\forall t'}\{\bar{R}_i\}_{t'<t}}\right\}$;
12 $\quad\quad \Delta x_{j,t} = N_t - 1$;
13 $\quad\quad k = t - 1$;
14 $\quad\quad$ **while** $k > 0$ **do**
15 $\quad\quad\quad \Delta x_{j,k} = \Delta x_{j,t} \times \frac{\bar{r}_{j,t}}{\bar{r}_{j,k}}$ **if** $N_k + \Delta x_{j,k} \leq 1$ **then**
16 $\quad\quad\quad\quad x_{j,k} = x_{j,k} + \Delta x_{j,k}$;
17 $\quad\quad\quad\quad N_k = N_k + \Delta x_{j,k}$;
18 $\quad\quad\quad\quad N_t = 1$;
19 $\quad\quad\quad\quad k = 0$;
20 $\quad\quad\quad$ **else**
21 $\quad\quad\quad\quad k = k - 1$;
22 $\quad\quad\quad$ **end**
23 $\quad\quad$ **end**
24 $\quad$ **end**
25 **end**
26 **for** $t \in \mathcal{T}$ **do**
27 $\quad \mathcal{L} = \left\{\mathcal{M} | \hat{t}_i = t\ \forall i \in \mathcal{M}\right\}$;
28 $\quad \delta\hat{F} = 0$;
29 $\quad \hat{i} = 0$;
30 $\quad$ **for** $i \in \mathcal{L}$ **do**
31 $\quad\quad y_{i,t} = \min\left\{1 - N_t, \frac{D_{i,T} - D_{i,t}}{\max\{\bar{R}_i\}}\right\}$;
32 $\quad\quad t' = \underset{\mathcal{T}\setminus t}{\text{argmax}}\{\bar{R}_i\}, \forall i \in \mathcal{M}$;
33 $\quad\quad y_{i,t'} = \min\left\{1 - N'_t, \frac{D_{i,T} - D_{i,t}}{\max_{\mathcal{T}\setminus t}\{\bar{R}_i\}}\right\}$;
34 $\quad\quad \delta F = y_{i,t} - y_{i,t'}$;
35 $\quad\quad$ **if** $\delta F > \delta\hat{F}$ **then**
36 $\quad\quad\quad \delta\hat{F} = \delta F$;
37 $\quad\quad\quad \hat{i} = i$;
38 $\quad\quad$ **end**
39 $\quad$ **end**
40 **end**
41 return X

1) QoS Satisfaction: To minimize the energy consumption while guaranteeing QoS satisfaction, C1 in Eq. 7 is turned to equality so that the airtime exactly satisfies the demand without violating the maximum degree ϵ. This step is calculated for every time slot for each user until the peak radio conditions are reached (lines 1-8).

2) Resource Limitation Satisfaction: After calculating the airtime fractions for all users in each time slot, the resource constraint, C2 in Eq. 1, is checked. In case of violation, the excess airtime is prebuffered in a preceding time slot with vacant resources. To ensure airtime minimization, the user with the highest average predicted rate in a previous vacant time slot is chosen (lines 9-25).

3) Peak Local Search Allocation: The above allocation strategy guarantees the satisfaction of both QoS and resource constraints. Thus, minimal allocation is used until the peak data rate time slot is reached. The challenging part in this stage occurs when more than one user competes on the same time slot. Accordingly, local search is applied to select the user who will result in the highest power consumption if he is not granted this time slot. As such, the local search calculates the difference in airtime between the two scenarios: If he is allocated to this peak time slot or if the second maximum peak is selected (lines 31-34). The user with less airtime in the first scenario is selected to be served in the current slot. The algorithm terminates when all the users' cumulative demands are satisfied.

C. Optimizer Complexity Analysis

We analyze the computational complexity of the introduced local-search heuristic and compare it against the optimal. For SoCP formulations, the optimal interior point method (IPM) requires a maximum of $O(\sqrt{K})$ iterations, where K is the number of constraints. Each iteration has a complexity of $O(m^2 \sum_{i=1}^{K} n_i)$ [31], where m denotes the total number of decision variables and n_i is the dimension of the i^{th} constraint. Accordingly, the complexity measure of the IPM as a function of the number of users, M, and time slots, T, is $O(\sqrt{2TM + T}(MT)^2(MT(T + 1)/2 + M + 1)) \approx O(\sqrt{MT}(M^3T^4))$. For the heuristic in Algorithm 1, the QoS satisfaction step has a complexity of $O(MT)$. The peak allocations and solution repairing have complexities of $O(MT)$ and $O(T^2)$, respectively. Thus, the total complexity of the heuristic is $O(MT + T^2)$.

VI. Particle Filter for Tuning Rate Deviation

We extend the robustness to scenarios in which the channel variance changes over the time and location [13]. A particle filter (PF) is used to tune the rate deviations (initially obtained off-line or theoretically) in order to reflect the channel variance based on the users' measurements. This is done on two steps: Rate deviation update and PF estimation. In particular, the PF estimates the error between the measured variance and its assumed value. This error is then used to update the theoretical variance for the future allocations.

A. Rate Deviation Update

We denote the off-line calculated deviations (e.g., using Monte-Carlo [24]) as $\hat{r}_{i,t}^{(M)}$, while the final tuned deviations

using PF are denoted by $\hat{r}_{i,t}^{(P)}$ and calculated as follows

$$\hat{r}_{i,t}^{(P)} = \alpha_{i,t} \times \hat{r}_{i,t}^{(M)}, \tag{9}$$

where $\alpha_{i,t} \geq 0$ is the proportionality factor between the offline and measured rate deviations. As the channel variance changes over time and location, the value of α has to be adapted accordingly using the particle filter as shown in the next subsection.

In multi-user scenarios, cooperative tuning can also be performed where existing users in the network can propagate their estimated value of $\alpha_{i,t}$ to the recent users admitted to the same BS. Such cooperation is done using the channel correlation coefficients between the users based on their distances per Eq. 10

$$\begin{aligned} \alpha_{i,t} &= \alpha_{i,t-1} + \max_{j \in \mathcal{M}, j \neq i} \left\{ \rho_{i,j,t} \right\} \left(\alpha_{j,t-1} - \alpha_{i,t-1} \right), \\ \rho_{i,j,t} &= e^{-\frac{d_{i,j,t}}{d_{cor}}}, \end{aligned} \tag{10}$$

where $d_{i,j,t}$ and $\rho_{i,j,t}$ are the distance and distance-dependent channel correlation coefficient between user i and j at time slot t, while d_{cor} is the correlation distance. The above formula is adopted from the 3GPP channel fading model [41].

B. PF Estimation

The PF initially generates a set of values (i.e., particles) following a proposed distribution and assigns them equal weights. These weights are then tuned based on the reported user measurements according to a predefined likelihood function. A final estimate of the PF state (i.e., α) is a weighted sum of the particles' values. The measurements represent the reported deviation between the predicted and the measured channel rates.

The PF unknown posterior distribution of the state variable y given a set of previous measurements/observations Z at time t is denoted by $p(y_{t+1}|Z_t)$. This probability distribution is calculated based on a Bayesian method called Chapman-Kolmogorov defined as [42]

$$p(y_{t+1}|Z_t) = \int p(y_{t+1}|y_t) p(y_t|Z_t) dy_t \tag{11}$$

where $p(y_{t+1}|y_t)$ is used to calculate the evolution of state y over the time horizon, while $p(y_t|Z_t)$ is an initial estimate of the posteriori probability at the current time slot and calculated as follows using Bayes rule

$$p(y_t|Z_t) = \frac{p(Z_t|y_t) p(y_t|Z_{t-1})}{\int p(Z_{t+1}|y_{t+1}) p(y_{t+1}|Z_t) dy_t} \tag{12}$$

where $p(Z_t|y_t)$ represents the likelihood probability of receiving measurements as Z_t while assuming state y_t. The denominator in Eq. 12 ensures that the estimated posteriori PDF will sum up to 1 over the time horizon.

The best estimate of the state y_t in the mean square error sense is denoted by $\bar{y}_t$ and calculated as

$$\bar{y}_t = \int y_t p(y_t|Z_t) dy_t \tag{13}$$

In order to provide a tractable solution for the above equations, we apply the *Sequential Importance Sampling (SIS)* technique [43]. SIS approximates the unknown posteriori distribution by a group of generated particles where each particle is weighted by its conformity to the measurements. Such particles are drawn from a proposed distribution, based on the problem structure, that approximates the original unknown distribution using large number of particles. The particle filter methodology based on SIS is summarized as follows

1 **Initialization**

i Define the proposed distribution $p(Q)$.

ii Generate a set of N particles denoted by $Q_{t=0}$ using the distribution $p(Q_{t=0})$.

iii Initialize equal weights $(\omega_{t=0}^{i})$ for all particles.

$$\omega_{t=0}^{i} = 1/N, \quad \forall i = 1, \ldots, N, \tag{14}$$

iv Define the likelihood function $F(Q, Z)$.

2 **Measurement Phase**

i Update the weights of each particle using the measurement Z_t and the likelihood function $F(Q, Z)$:

$$\omega_t^j = \omega_{t-1}^j F(Q, Z), \quad \forall j \in 1, \ldots, N, \tag{15}$$

ii Normalize the weights:

$$\bar{\omega}_t^j = \frac{\omega_t^j}{\sum_{j=1}^{N} \omega_t^j}, \tag{16}$$

iii Calculate the best estimate:

$$\bar{y}_t = \sum_{j=1}^{N} z_t \bar{\omega}_t^j, \tag{17}$$

3 **Prediction Phase**

i Calculate the gradient:

$$\delta y_t = \frac{\partial z_t}{\partial t}, \tag{18}$$

ii Predict the future state:

$$y_{t+1} = A\bar{y}_t + B\delta y_t \delta t : \tag{19}$$

4 **Importance Sampling**

i Calculate effective samples:

$$N_{eff} = \frac{1}{\sum_{j=1}^{N} (\omega_t^j)^2}, \tag{20}$$

ii Check degeneracy then resample: If $N_{eff} < \hat{N}$ then, resample particles and set $\omega_t^j = 1/N \ \forall j \in 1, \ldots, N$,

In essence, the calculated weights ω_t^j in Eq. 15 approximate the posteriori PDF in Eq. 12, while the priori PDF in Eq. 11 is evaluated using the likelihood $F(Q, Z)$ in the initialization phase. In addition, Eq. 17 in the measurement phase implements the best estimate of the state (Eq. 13). In the prediction phase, the future state y_{t+1} in Eq. 19 is a linear weighted combination of both the best estimated state $\bar{y}_t$ and the integral of its rate of change $\delta y_t \delta t$ from the available measurements z_t. In Eq. 19, A and B are the weights of both the best estimate and integral of the rate of change, respectively.

As the PF updates the weights ω_t^j in Eq. 15 every time slot, their values may converge and few number of particles will have non-zero weights. Such situation is called degeneracy, which has to be avoided as it deviates the weight's distribution from the actual posteriori probability. Thus, the number of effective particles N_{eff} is calculated to check for the degeneracy and in case of dropping below the maximum threshold $\hat{N}$, resampling is done. Each particle contributes, based on its weight, in generating a new particle [43]. The newly generated set of particles will not contain the ones with very low wights. The new weights are equally redistributed similar to the initialization phase.

In our rate deviation tracking, the PF state y_t is the proportionality factor α_t while the measurement z_t is the reported proportionality factor $\bar{\alpha}_t$ calculated as

$$\bar{\alpha}_t = \frac{|\bar{r}_{i,t} - \mathbb{E}[r_{i,t}]|}{\hat{r}_{i,t}^{(M)}} \tag{21}$$

where $\mathbb{E}[r_{i,t}]$ is the measured channel rate by user i in the duration from slot $t-1$ to slot t.

VII. Performance Evaluation

A. Simulation Set-Up

We simulate the proposed robust PRA using the LTE module in Network Simulator 3 (ns-3) [44] integrated with Gurobi solver [45]. Gurobi uses an efficiently implemented IPM with Barrier function for solving SoCP, and its terminating condition was set to a duality gap value of 0.01%. The 3GPP correlated slow fading model and its parameters [41] are added to the received UE power in order to simulate uncertainties in predicted rates. Simulation results are averaged over 50 runs, and all the output values are 1% apart from the displayed mean. Such confidence interval is not reported in the figures for legibility and thus only the average is displayed. The simulation considered an urban area where users follow different predefined paths within the cell at varying velocities from 25 to 40 Km/h and request a video stream at a fixed quality. The user follows a path that either starts near a cell edge and moves to the center, or starts from the cell center and moves towards the edge. All the simulation parameters and their values are presented in Table I.

TABLE I
Summary of Model Parameters

Parameter	Value
BS transmit power	43 dBm
Bandwidth	5 MHz
Time Horizon T	60 s
Streaming rate V	0.25, 0.5 and 1 [Mbps]
Bit Error Rate	5×10^{-5}
Shadow correlation distance(d_{cor}) [41]	$50m$
Shadow standard deviation [41]	4
Velocity	From 25 km/h to 40 km/h
$p(Q)$	$\mathcal{U}(0,4)$
N	10000
N_{th}	$N/3$
$A = B$	0.5
μ^-	-0.5
σ_t'	$\frac{1}{\sqrt{12}}$
Feedback interval τ	$5s$.
Packet size	10^3 [bytes]
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. Comparative Schemes and Evaluation Metrics

In this evaluation study, we compare the proposed robust predictive scheme against the existing non-robust PRA and non-predictive RA schemes denoted as follows

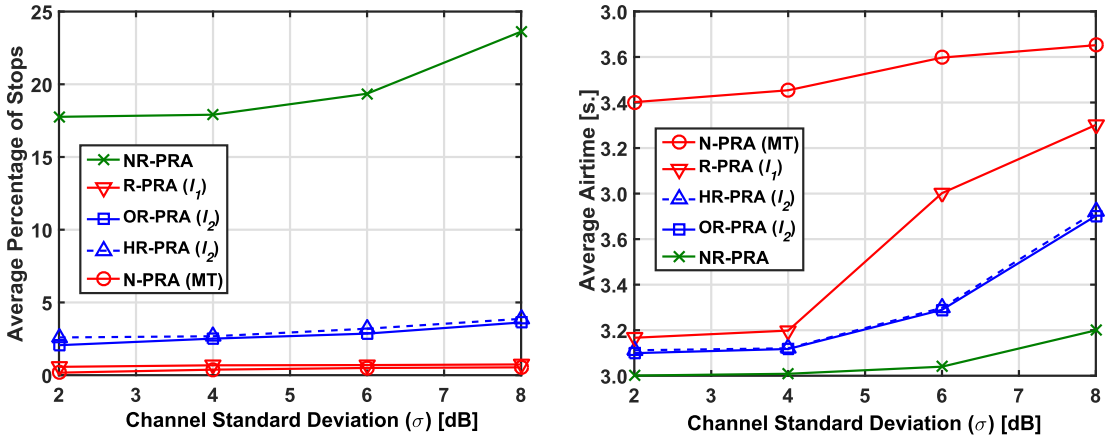
- **N-PRA (MT):** refers to a type of non-predictive RA called maximum throughput (MT) [46]. In essence, MT allocates the whole resources to the user with the current maximum channel rate regardless his future channel conditions.
- **NR-PRA:** refers to the existing non-robust PRA in [3] which only uses the average value of the predicted rate resulting in a deterministic linear formulation. The optimal solution is obtained using the simplex method implemented in Gurobi [40].
- **OR-PRA ($\mathbf{l}_2$):** refers to the introduced BA based robust PRA in this work and formulated in Eq. 1 and Eq. 7. The solution is obtained optimally using the IPM in Gurobi optimizer [40].
- **HR-PRA ($\mathbf{l}_2$):** the same as **OR-PRA ($\mathbf{l}_2$)**, but its solution is obtained using the guided local search heuristic in Section V.
- **R-PRA ($\mathbf{l}_1$):** refers to the introduced BA robust PRA in this work but linearized similar to [17] and the solution is obtained optimally using the simplex method in Gurobi optimizer [40].

All the above schemes are assessed using two main metrics: percentage of video stops and average airtime to measure the QoS satisfaction and the energy consumption, respectively. Existing predictive RA approaches revealed that playback interruptions, due to buffer under-run, are among the primary sources of user dissatisfaction with video delivery services [2], [47]. Thus video stops metric perfectly models the ability of RA to optimize the trade-off between energy-minimization and QoS satisfaction. The percentage of video stops, denoted as VD, is used to quantify the QoS degradation and calculated as the percentage of time slots in which the cumulative transmitted content ($R_{i,t}$) is less than the demand ($D_{i,t}$) per Eq. 22

$$VD = \frac{\sum_{i=1}^{M}\sum_{t=0}^{T} \mathbb{1}_{R_{i,t} < D_{i,t}}}{M \times T} \times 100, \tag{22}$$

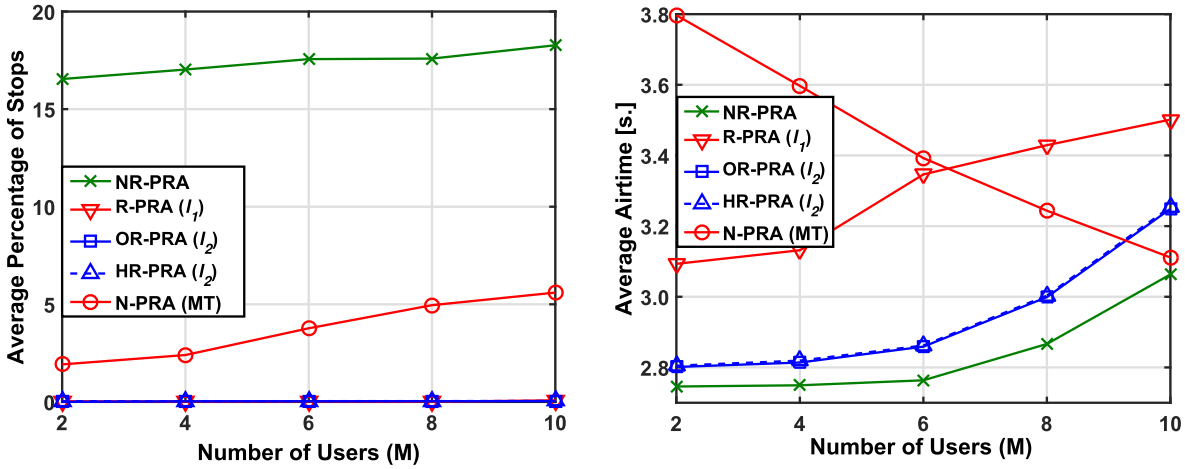
where $R_{i,t} = \sum_{t'=0}^{t} r_{i,t'} x_{i,t'}$ is the cumulative video content received by user i till time slot t while $r_{i,t}$ is the experienced channel rate by user i at timeslot t. The maximum allowed value of VD is set to the predefined constraint violation level (ϵ) $\times$ 100%.

The second metric is the average BS airtime which is used to measure the energy consumption in the network. During resource allocation, both the BS and UE consume energy in data transmission and reception. Therefore, minimizing the airtime reduces the energy consumption proportionally [34].



(a) Average Percentage of video stops. (b) Average BS airtime.

Fig. 2. Performance evaluation for different channel variances at QoS levels $(1-\epsilon)=0.9$ and 8 users requesting high quality video.



(a) Average Percentage of video stops. (b) Average BS airtime.

Fig. 3. Performance evaluation for different number of users requesting HQ at QoS levels $(1-\epsilon)=0.95$ and experiencing $\sigma=4$.

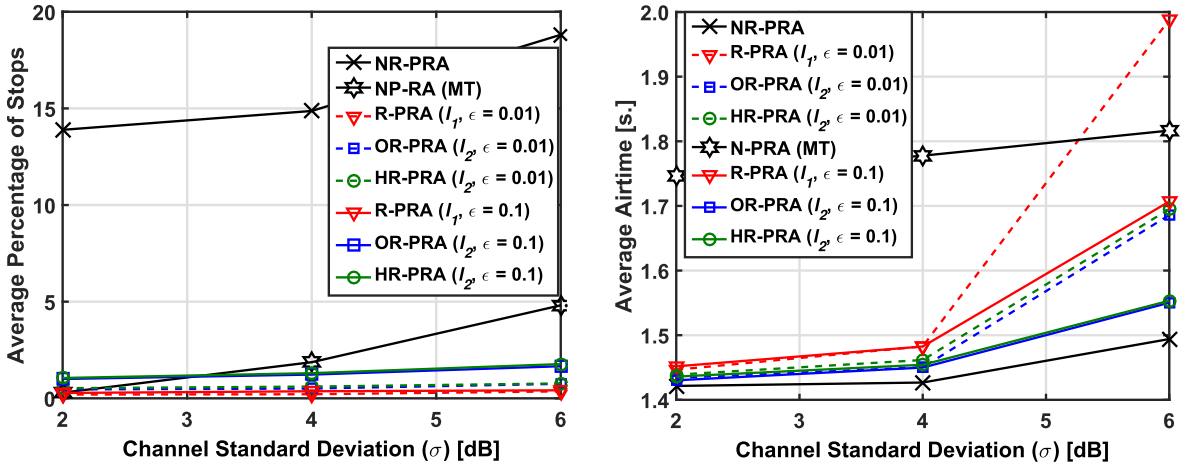
The objective function in Eq. 1 is used to quantitatively measure this metric.

C. Simulation Results

1) Comparison With Other Resource Allocators: We assume that the rate deviation $\hat{r}$ is accurately known and the focus is to show the importance of robust PRA and the heuristic solution. The first scenario considered a high quality video (i.e. $V_i = 1Mbps$) which is a high load scenario relative to the available average channel rate. The non-predictive MT continues to satisfy the QoS level independent on the channel variance as shown in Fig. 2(a). This is because the MT schedules the users based on their current reported channel rate irrespective of the variance and the future rates. The non-robust predictive technique [3] fails to satisfy the maximum VD set to 0.1 (i.e. $\epsilon = 0.1$). This QoS performance degrades with the channel variance since the measured rate deviates from the average value. The allocated minimal airtime will not be sufficient to satisfy the demand. Such deterioration is avoided by all the robust forms as the percentage of stops did not pass $\epsilon \times 100\%$ for the considered variances.

Although the non-predictive MT prioritizes users with maximum rates, its energy consumption is higher than the predictive strategies as depicted in Fig. 2(b). The MT buffers the video content for the cell peak users, which saves energy, but then turns to push the video for other users located near the cell edge rather than applying minimal allocation. On the other hand, the predictive strategy is able to minimize the energy even in the robust forms. The results also demonstrate the conservatism of the linearized BA used in [17], which decreases the energy saving gain especially at very high channel variances. The energy consumption thus increases and becomes comparable to that of the non-predictive strategy.

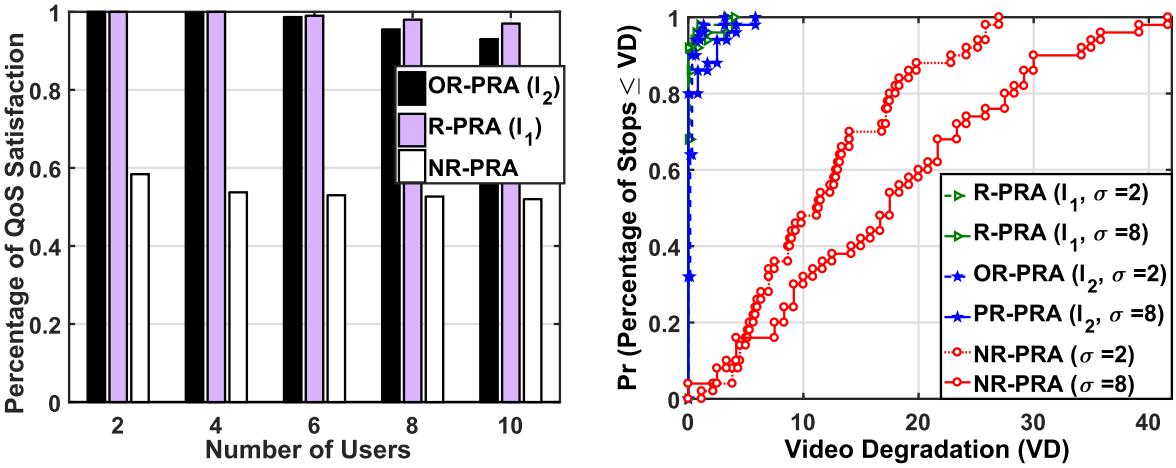
Both the load per user and the moving speed are then decreased to medium quality videos (i.e. $V_i = 0.5Mbps$) and 25 Km/h, respectively, to allow more users and higher QoS levels in the simulation scenario. The conservatism of the linearized BA becomes more significant as it consumes more energy than the non-predictive MT at high QoS level



(a) Average Percentage of video stops.

(b) Average BS airtime.

Fig. 4. Performance evaluation for different channel variances at high QoS levels and 12 users requesting MQ video.



(a) Percentage of scenarios with VD below $\epsilon = 0.05$ and $\sigma = 4$

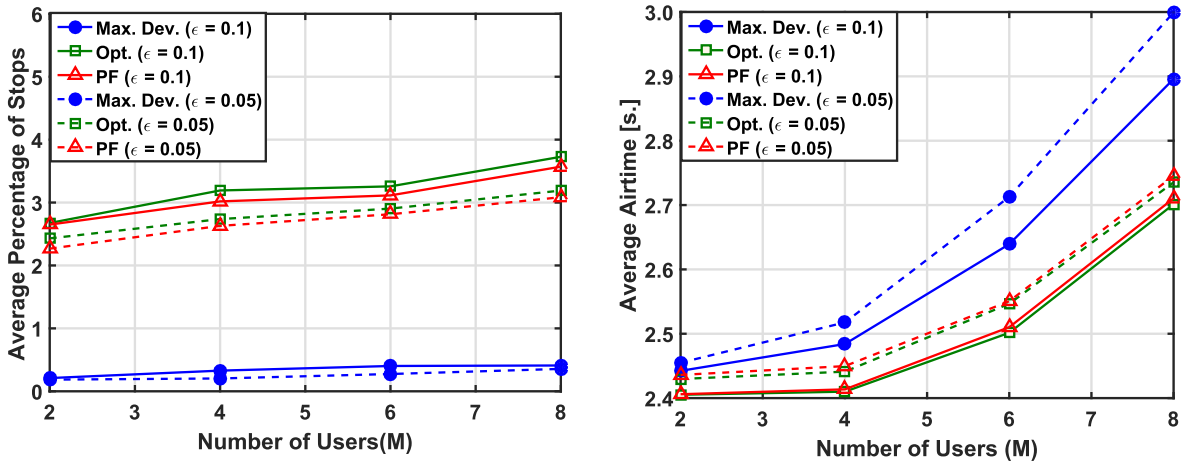
(b) CDF of video degradation for 10 users

Fig. 5. Performance evaluation for different channel variances and number of users at QoS levels $(1-\epsilon) = 0.95$ requesting high quality video.

(i.e. low ϵ) and high channel variances as in Fig. 3(b) and Fig. 4(b). The BA in its original SoCP form, however, is able to preserve the prediction gain at these high load conditions. While the energy savings gap between, the predictive and non-predictive schemes decrease for this scenario, the latter fails to meet the QoS level as shown in Fig. 3(a) and Fig. 4(a). This is because such non-predictive strategy greedily allocated the resources to the cell peak users and ignored serving the cell edge users in order to maximize the total system throughput.

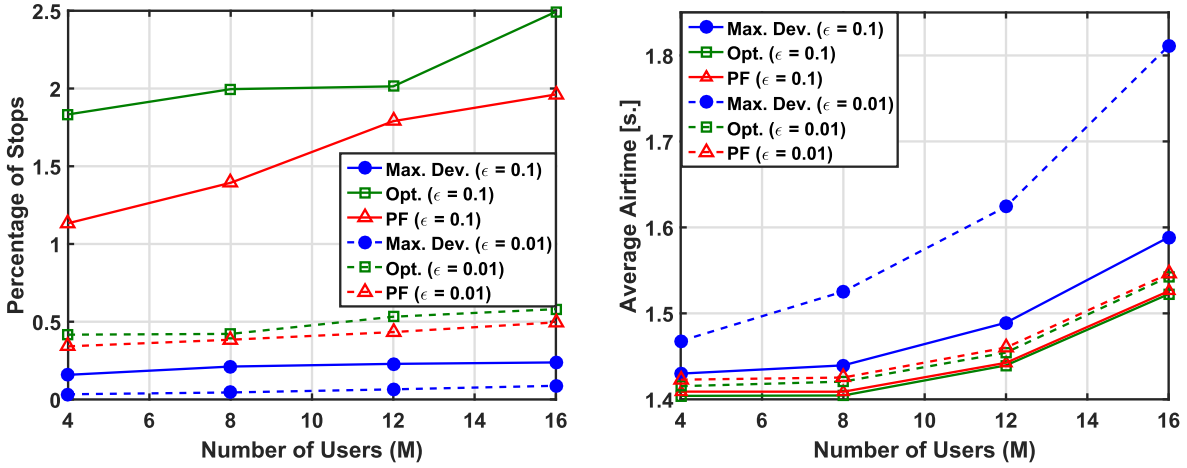
Similar observations are noted for the conservative linearized BA, NR-PRA and MT when the number of users and the QoS level are increased as shown in Fig. 4(b). The distributions of QoS satisfaction and degradation are reported in Fig. 5(a) and Fig. 5(b), respectively. The percentage of users with violated QoS levels mainly depends on their mobility traces and experienced channel rates. In Fig. 5(a), the percentage of users with violated QoS levels was around 50% in case of the non-robust PRA. This was found to be the same percentage of users who started the video streaming at the cell edge, and thus were subjected to minimal allocation strategy resulting in buffer underrun. In Fig. 5(b), the distribution of video degradation, and its maximum value, illustrate the QoS violation of non-robust PRA. Note that the robust PRA schemes experienced stable QoS performance over the system load and variance. The scenarios above demonstrate that the adopted BA SoCP based PRA formulation: 1) satisfies all QoS levels for different system loads (Fig. 3(a)) and 2) preserves the energy-saving gains of the prediction (Fig. 3(b)). In addition, the introduced heuristic shows stable performance with a very low optimality gap (<0.1 %) with respect to the optimal solution's airtime and QoS levels in all considered cases.

2) Performance of Particle Filter: In this scenario, we assess the ability of the PF to track the rate deviations while adopting the SoCP BA formulation. We compare the PF based variance is compared with both the maximum and optimal theoretical variances denoted by *Max. Dev.* and *Opt.*, respectively. The *Max. Dev.* corresponds to the maximum variance [41] that guarantees the QoS satisfaction under the highest prediction



(a) Average Percentage of video stops. (b) Average BS airtime.

Fig. 6. Performance evaluation for the robust framework with channel tracking for different number of users experiencing $\sigma = 2$ and requesting MQ video with high QoS level $(1 - \epsilon) = 0.95$.



(a) Average Percentage of video stops. (b) Average BS airtime.

Fig. 7. Performance evaluation for the robust framework with channel tracking for different number of users experiencing $\sigma = 2$ and requesting LQ video with high QoS level.

errors. The *Opt.* adopts the exact rate deviation corresponding to the current channel variance. This optimal value satisfies the QoS level without compromising the energy savings. On the other hand, the *PF* initially assumes the highest variance as the *Max. Dev.*, but continuously monitors the channel variance and adapts the rate deviation accordingly.

With regards to QoS satisfaction, the *Max. Dev.* provides a very conservative allocation that greedily satisfies the QoS at the expense of the energy saving as depicted in Fig. 6(a) and Fig. 6(b), respectively. This is not the case for *PF* which has met the constraint at nearly the exact level as the *Opt.*, resulting in high energy savings. The PF, in essence, decreases the initial maximum rate deviation to reach the lower optimal value and sometimes below. Although going below the optimal rate deviation value increases the risk of constraint violation, the conservative BA based allocation in early timeslots avoids such QoS degradation case. The energy gain of the PF-based channel tracking relative to the maximum deviation has increased in the high load scenarios (i.e. more number of users) at high QoS levels and reached up to 15 % as shown in Fig. 7(b). This adaptation mechanism results in nearly the same energy savings as the optimal deviation case and with better QoS satisfaction as less video stops have been experienced in the early slots as shown in Fig. 7(a) and Fig. 7(b).

3) Runtime Complexity: We also report the execution time of all the examined RA schemes in Table II, and measured within the simulation environment on a Quad Core i7-Processor, 3.2 GHz machine. These results highlight the efficiency of the guided heuristic solution methods for providing real-time implementation under different load scenarios. The complexity of the optimal solver increases with *both* the number of users (i.e. the problem dimensions) and the streaming rate (V) since more iterations are required to reach a feasible solution. As opposed to the solver, the guided heuristic resulted in a stable scalable performance regardless the value

TABLE II
EXECUTION TIME OF THE SIMULATED SCHEMES

Technique	Number of Users				Streaming Rate (V) [Mbps]		
	2	4	8	12	0.25	0.5	1
N-PRA (MT)	<0.1 ms	<0.1 ms	<0.1 ms	<0.1 ms	<0.1 ms	<0.1 ms	<0.1 ms
NR-PRA	1 s.	1.5 s.	2.3 s.	4 s.	4 s.	4 s.	4 s.
OR-PRA (l_2)	50 s.	80 s.	150 s.	250 s.	200 s.	250 s.	290 s.
HR-PRA (l_2)	<0.1 ms	<0.1 ms	<0.1 ms	<0.1 ms	<0.1 ms	<0.1 ms	<0.1 ms
OR-PRA (l_1)	1 s.	1.5 s.	2.3 s.	4 s.	4 s.	5 s.	5.5 s.

of the aforementioned two parameters and with a delay less than the duration of Time Transmission Interval (TTI).

VIII. CONCLUSION

In this paper, we developed a *robust* PRA framework that provides probabilistic QoS guarantees under generic error probability density functions. The solution integrates Bernstein Approximation (BA) for chance-constraint QoS modeling, a particle filter (PF) for prediction uncertainty tracking, and a guided heuristic that enables real-time implementation.

The proposed framework was applied for energy-efficient video delivery, and the results indicate it's resilience in meeting QoS constraints, while significantly reducing BS energy under practical prediction uncertainty. In particular, the BA formulation successfully satisfied the QoS level in all scenarios, unlike the existing *non-robust* PRA that rely only on *average* future rates. The results further demonstrated that *non-predictive* RA either consumes excess energy or violates the QoS level under low or high load scenarios, respectively. Handling the BA complexity through traditional linearization techniques was also investigated, but appeared to be very conservative for such long-term predictive allocations, especially under high load scenarios or tight QoS levels. However, using a guided heuristic enabled the adoption of the BA in its original less conservative SoCP form for different load and QoS levels. In addition to developing the stochastic PRA model, we demonstrated how a PF can be adopted to track the rate deviations in real-time. Such tracking enables the operator to be aggressive during periods of accurate predictions, and thereby maximize energy savings without compromising QoS.

Our future work considers the following enhancements to the robust PRA framework: 1) modeling other objectives such as maximizing the quality for adaptive video streaming during high network load, and 2) considering the uncertainties in the user's requested streaming rate (i.e. demand uncertainty).

REFERENCES

[1] 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, Jun. 2014.

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

[3] 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, Oct. 2013.

[4] R. Margolies *et al.*, "Exploiting mobility in proportional fair cellular scheduling: Measurements and algorithms," in *Proc. IEEE INFOCOM*, Apr. 2014, pp. 1339–1347.

[5] "Cisco visual networking index: Global mobile data traffic forecast update 2014–2019," Cisco, San Jose, CA, USA, White paper, 2015. [Online]. Available: http://www.cisco.com/c/en/us/solutions/service-provider/visual-networking-index-vni/index.html

[6] L. M. Correia *et al.*, "Challenges and enabling technologies for energy aware mobile radio networks," *IEEE Commun. Mag.*, vol. 48, no. 11, pp. 66–72, Nov. 2010.

[7] J. G. Andrews *et al.*, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[8] Q. Wu, M. Tao, and W. Chen, "Joint Tx/Rx energy-efficient scheduling in multi-radio wireless networks: A divide-and-conquer approach," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2727–2740, Apr. 2016.

[9] C. U. Saraydar, N. B. Mandayam, and D. J. Goodman, "Efficient power control via pricing in wireless data networks," *IEEE Trans. Commun.*, vol. 50, no. 2, pp. 291–303, Feb. 2002.

[10] L. Venturino, A. Zappone, C. Risi, and S. Buzzi, "Energy-efficient scheduling and power allocation in downlink OFDMA networks with base station coordination," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 1–14, Jan. 2015.

[11] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, Feb. 2010.

[12] A. Schulman *et al.*, "Bartendr: A practical approach to energy-aware cellular data scheduling," in *Proc. ACM Mobicom*, Apr. 2010, pp. 85–96.

[13] 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*, Mar. 2015, pp. 1195–1200.

[14] J. Yao, S. S. Kanhere, and M. Hassan, "Improving QoS in high-speed mobility using bandwidth maps," *IEEE Trans. Mobile Comput.*, vol. 11, no. 4, pp. 603–617, Apr. 2012.

[15] N. Bui and J. Widmer, "Modelling throughput prediction errors as Gaussian random walks," in *Proc. KuVS Workshop Anticipatory Netw.*, 2014, pp. 1–3.

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

[17] 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, Mar. 2013.

[18] M. J. 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. 14, pp. 3888–3898, Jul. 2015.

[19] B. Liu, *Theory and Practice of Uncertain Programming*. Heidelberg, Germany: Physica-verlag., 2002.

[20] P. Kali and S. W. Wallace, *Stochastic Programming*. Hoboken, NJ, USA: Wiley, 1994.

[21] A. Charnes and W. W. Cooper, "Chance-constrained programming," *Manage. Sci.*, vol. 6, no. 1, pp. 73–79, Oct. 1959.

[22] Y. Hu, S. Han, and C. Yang, "Context-aware energy saving with proactive power allocation," in *Proc. IEEE GlobalSIP*, Dec. 2015, pp. 53–57.

[23] R. Atawia, H. S. Hassanein, and A. Noureldin, "Fair robust predictive resource allocation for video streaming under rate uncertainties," in *Proc. IEEE GLOBECOM*, Dec. 2016, pp. 1–6.

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

[25] N. Bui, F. Michelinakis, and J. Widmer, "A model for throughput prediction for mobile users," in *Proc. Eur. Wireless*, May 2014, pp. 1–6.

[26] W. Xu, A. Tajer, X. Wang, and S. Alshomrani, "Power allocation in MISO interference channels with stochastic CSIT," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1716–1727, Mar. 2014.

[27] W. W.-L. Li, Y. J. Zhang, A. M.-C. So, and M. Z. Win, "Slow adaptive OFDMA systems through chance constrained programming," *IEEE Trans. Signal Process.*, vol. 58, no. 7, pp. 3858–3869, Jul. 2010.

[28] R. Atawia, H. Abou-Zeid, H. S. Hassanein, and A. Noureldin, "Robust resource allocation for predictive video streaming under channel uncertainty," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2014, pp. 4683–4688.

[29] R. Atawia, H. Abou-Zeid, H. S. Hassanein, and A. Noureldin, "Joint chance-constrained predictive resource allocation for energy-efficient video streaming," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1389–1404, May 2016.

[30] J. E. Mitchell, "Polynomial interior point cutting plane methods," *Optim. Methods Softw.*, vol. 18, no. 5, pp. 507–534, 2003.

[31] M. S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, "Applications of second-order cone programming," *Linear Algebra Appl.*, vol. 284, nos. 1–3, pp. 193–228, Nov. 1998.

[32] H. Abou-Zeid and H. Hassanein, "Toward green media delivery: Location-aware opportunities and approaches," *IEEE Wireless Commun.*, vol. 21, no. 4, pp. 38–46, Aug. 2014.

[33] A. Galindo-Serrano, B. Sayrac, S. B. Jemaa, J. Riihijärvi, and P. Mähönen, "Cellular coverage optimization: A radio environment map for minimization of drive tests," in *Cognitive Communication Cooperative HetNet Coexistence*, Switzerland, 2014, pp. 211–236.

[34] C. Desset *et al.*, "Flexible power modeling of LTE base stations," in *Proc. IEEE WCNC*, Apr. 2012, pp. 2858–2862.

[35] P. Frenger, P. Moberg, J. Malmodin, Y. Jading, and I. Godor, "Reducing energy consumption in LTE with cell DTX," in *Proc. IEEE VTC*, May 2011, pp. 1–5.

[36] *LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures*, document Technical Specification TS 36.213 v12.5.0, 3GPP, 2015.

[37] S. Singh, O. Oyman, A. Papathanassiou, D. Chatterjee, and J. G. Andrews, "Video capacity and QoE enhancements over LTE," in *Proc. IEEE ICC*, Jun. 2012, pp. 7071–7076.

[38] A. Ben-Tal and A. Nemirovski, "Selected topics in robust convex optimization," *Math. Programm.*, vol. 112, no. 1, pp. 125–158, Mar. 2008.

[39] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004, pp. 156–159.

[40] Gurobi. *Gurobi Optimization*, accessed on Mar. 29, 2015. [Online]. Avaiable: http://www.gurobi.com/

[41] *LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Further Advancements for E-UTRA Physical Layer Aspects*, doument TR 36.814 V9.0.0, 3GPP, 2010.

[42] K. Huber and S. Haykin, "Improved Bayesian MIMO channel tracking for wireless communications: Incorporating a dynamical model," *IEEE Trans. Wireless Commun.*, vol. 5, no. 9, pp. 2458–2466, Sep. 2006.

[43] L. Mihaylova, D. Angelova, S. Honary, D. R. Bull, C. N. Canagarajah, and B. Ristic, "Mobility tracking in cellular networks using particle filtering," *IEEE Trans. Wireless Commun.*, vol. 6, no. 10, pp. 3589–3599, Oct. 2007.

[44] G. Piro, N. Baldo, and M. Miozzo, "An LTE module for the ns-3 network simulator," in *Proc. Int. ICST Conf. Simul. Tools Techn.*, Mar. 2011, pp. 415–422.

[45] 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, Sep. 2014.

[46] F. Capozzi, G. Piro, L. A. Grieco, G. Boggia, and P. Camarda, "Downlink packet scheduling in LTE cellular networks: Key design issues and a survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 2, pp. 678–700, May 2013.

[47] A. ParandehGheibi, M. Médard, A. Ozdaglar, and S. Shakkottai, "Avoiding interruptions—A QoE reliability function for streaming media applications," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 5, pp. 1064–1074, May 2011.

Ramy Atawia (S'12) received the B.Sc. and M.Sc. degrees in communication engineering from German University, Cairo, Egypt, in 2012 and 2013, respectively. He is currently pursuing the Ph.D. degree in electrical and computer engineering with Queen's University, Canada. He previously worked with Vodafone on automating the design and optimization of radio networks, and with Nokia on customer experience management and analytics. He was a Teaching Assistant and Guest Lecturer, where he delivered tutorials on optimization, wireless networks, and programming. He is currently a member of technical staff with Bell Labs, Nokia, Belgium. His research work appeared in top-tier IEEE journals and conferences, and led to ten patents. His research includes stochastic optimization, predictive video streaming, machine learning, and AI driven management of indoor networks. He also serves as a TPC member and reviewer for the IEEE flagship conferences and journals.

Hossam S. Hassanein (S'86–M'90–SM'05–F'17) is a leading authority in the areas of broadband, wireless and mobile networks architecture, protocols, control and performance evaluation. His record spans over 500 publications in journals, conferences, and book chapters, in addition to numerous keynotes and plenary talks in flagship venues. He has received several recognition and best paper awards at top international conferences. He is also the Founder and Director of the Telecommunications Research Laboratory, School of Computing, Queen's University, with extensive international academic and industrial collaborations. He is a Former Chair of the IEEE Communication Society Technical Committee on Ad hoc and Sensor Networks. He is an IEEE Communications Society Distinguished Speaker (Distinguished Lecturer 2008–2010).

Hatem Abou-zeid (S'04–M'14) received the Ph.D. degree in electrical and computer engineering from Queen's University, Canada, in 2014, where his research was recognized with a medal nomination for innovation. He has been a Software Engineer with the Cisco R&D Center, Ottawa, Canada, since 2015. He has contributed to the development of scalable traffic engineering and IP routing protocols for service provide and data center networks. He was a Post-Doctoral Fellow prior to joining CISCO. He is also an experienced lecturer and has been granted several teaching fellowships with Queen's University to instruct freshman and senior-level engineering courses. His main research interests include SDNs, context-aware radio access networks, and adaptive video delivery for vehicular communications. He is also a technical reviewer for several prestigious conferences and journals.

Aboelmagd Noureldin (S'98–M'02–SM'08) received the B.Sc. degree in electrical engineering and the M.Sc. degree in engineering physics from Cairo University, Egypt, in 1993 and 1997, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Calgary, Alberta, Canada, in 2002. He is a Professor with the Departments of Electrical and Computer Engineering, Royal Military College of Canada (RMCC) with Cross-Appointment with the School of Computing and the Department of Electrical and Computer Engineering, Queen's University. He is also the Founder and the Director of the Navigation and Instrumentation Research Group, RMCC. His research is related to GPS, wireless location and navigation, indoor positioning, and multi-sensor fusion. He has published over 230 papers in journals and conference proceedings. His research work led to ten patents in the area of position, location and navigation systems.