Utilization of Stochastic Modeling for Green Predictive Video Delivery Under Network Uncertainties

Ramy Atawia, *Member, IEEE*, Hossam S. Hassanein, *Fellow, IEEE*, Najah Abu Ali, *Member, IEEE* and Aboelmagd Noureldin, *Senior Member, IEEE*

***Abstract*—Predictive resource allocation (PRA) has gained momentum in the network research community as a way to cope with the exponential increase in video traffic. Existing PRA schemes have demonstrated profound energy savings and ubiquitous quality of service (QoS) satisfaction under idealistic prediction of future network states. In this paper, we relax the main assumption of existing PRA work and tackle uncertainties in predicted information which resulted from space and time variation of the network load and users demands. A robust green PRA (R-GPRA) is proposed to: model the uncertainties as random variables, ensure a probabilistic satisfaction of QoS constraints, and follow a risk-aware preallocation of future demand. A recourse programming model is used to represent the trade-off between the energy-savings and the risk of wasting resources while considering the probability of a user terminating the video session at each time slot. Thus, the scheme prevents the network from prebuffering the future video content that might be skipped by the user. Similarly, a chance constrained programming model is proposed to provide a probabilistic QoS representation to guarantee that the sum of resources, predetermined to video streaming users, do not surpass the total time-varying network capacity. We prove that a near-optimal solution is attainable by proposing a guided heuristic search with small optimality gap to numerical methods. Simulation results demonstrate the ability of R-GPRA to deliver energy-efficient video streaming with less resources than existing PRA while promising QoS satisfaction. These results provide the incentive to implement the R-GPRA in future wireless networks.**

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

I. Introduction

THE GLOBAL Internet traffic is expected to grow tremendously reaching 2.3 zettabyte by 2020 [1]. At that time, mobile devices will be contributing to more than two-thirds of the total Internet traffic, where more than three-quarters of such traffic is expected to be video content. Maximizing streaming quality while minimizing the number and durations of stops remains the ultimate goal for mobile video streaming users. This, however, will put network operators under huge pressure as they strive to meet users' QoS expectations given the available network resources and infrastructure to maximize their profit. Nevertheless, the unprecedented increase in carbon footprint and energy costs raised the need for energy-efficient video delivery over wireless networks [2], [3]. Such challenges prompt optimal design for QoS-aware resource allocation schemes that can target energy minimization during service delivery [4]. Such design augments the gains of research work done on regulating spectrum access and maximizing radio resources [5], [6].

Supported by mobility and channel predictions [7]–[9], predictive resource allocation (PRA) was recognized as a new paradigm showing great potential of remarkable energy savings and pervasive QoS satisfaction [4], [10], [11]. Today's networks adopt opportunistic schemes that perform resource allocation decisions based on current and previous measurements. On the contrary, the PRA leverages future network conditions to recognize users moving towards regions with low channel rates (typically needing more resources) and prebuffers their video content upfront. Whereas prebuffering is postponed for other users heading to regions with high channel rates. Despite the reported energy and QoS gains in the literature, the following practical challenges related to prediction uncertainty must be addressed:

- Channel Rate Variations: The first parameter used in PRA is the future channel rate of mobile users based on their trajectory. Both mobility traces and channel state prediction accommodate errors due to the noise of their raw data, adopted low-cost filters, and temporal variations of the wireless signal [12], [13]. Existing approaches in [14] and [15] considered robust heuristic decisions that exploit states of the rate predictor or adapt the allocation window size to minimize the impact of uncertainty.
- Demand uncertainty: The user demand is represented by both the streaming bitrate (i.e., video quality) and the watching duration. Users can frequently change the quality of video, skip some frames or terminate the session

Manuscript received June 15, 2017; revised September 23, 2017 and December 5, 2017; accepted January 25, 2018. Date of publication February 1, 2018; date of current version May 17, 2018. The associate editor coordinating the review of this paper and approving it for publication was E. Ayanoglu. *(Corresponding author: Ramy Atawia.)*

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

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

N. Abu Ali is with the College of Information Technology, United Arab Emirates University, Al Ain, UAE (e-mail: najah@uaeu.ac.ae).

A. Noureldin is with the Electrical and Computer Engineering Department, Royal Military College of Canada, Kingston, ON K1R 7Y6, Canada (e-mail: aboelmagd.noureldin@rmc.ca).

Digital Object Identifier 10.1109/TGCN.2018.2800708

completely before the end [16]. The Non-PRA approach in [17] and [18] controlled the buffer size during the short term decisions based on the stability level of the session. The impact of demand uncertainty is more severe in the case of PRA. Fig. 1(a) depicts an example of energy wastage as a result of terminating the session at $t = 5$. The risk of wasting resources increases as PRA maximizes prebuffering for users experiencing peak rates. Existing robust non-PRA techniques [17], [18] aim to decide when to prebuffer the video at the current slot, to save the tail energy, or postpone the delivery. The PRA, however, requires further efforts to consider the trade-off over the time horizon since postponing full video delivery requires more resources to transmit the remaining content during future poor channel conditions.

- Network resources: The stochastic arrival of users with stringent service delay requirements, such as voice calls, will decrease the total available resources for streaming users. In turn, this will increase the risk of violating QoS requirements for cell-edge video users who are allocated a small portion of the available resources. Fig. 1(b) depicts this scenario where the network follows a stingy allocation for a cell-edge user to minimize the energy consumption. The risk of violating the demand, when the user do not receive the minimum amount of data, has to be modelled by the PRA. Thus, the minimal allocation is followed during resources stability while an opportunistic strategy is adopted in uncertain conditions.

Existing PRA assumed idealistic scenarios [4], [10], [11], [19] where the average values of all three parameters are adopted. In order to maintain the prediction gains, mobile buffering capabilities have to be fully exploited while applying long-term decisions at the beginning of the time horizon. Our recent work on *robust* PRA [20]–[25] tackled the first parameter (i.e., rate uncertainty) and showed that prediction gains are still attainable when probabilistic risk-aware PRA is applied.

We focus on the second and third sources of uncertainties by considering the possibility of video termination and variations in network resources while deriving long-term energy-efficient allocations. The proposed R-GPRA should measure the risk of wasting resources and compare it to the possibility of energy savings to determine when and which content to prebuffer. This is in addition to considering the scenarios in which the total available network resources are shared with real-time users. A probabilistic metric is defined to guarantee a minimal level of QoS satisfaction and controls the impact of such scenarios.

This paper introduces, for the first time in literature, a *robust stochastic green* PRA framework that achieves energy savings and QoS satisfaction over *a time horizon* under both demand and network resources uncertainties. The framework is referred to as R-GPRA and incorporates the following contributions:

1) We capture the uncertainties in both the demand and network resources by proposing a *stochastic* optimization model. The main objective is to minimize the total allocated resources, i.e., less energy, while satisfying the time slot demand constraint. The model relies on Recourse Programming (RP) and Chance Constrained Programming (CCP) to represent the uncertainties in the objectives and constraints as random variables. This is unlike existing PRA work [11], [26] that assumed perfect prediction and used deterministic formulation based on the average value of predicted information. In essence, our RP considers the risk of wasting resources due to video streaming users terminating the session before watching the entire video [16], [27]. Similarly, the CCP controls the QoS degradations under resources fluctuations due to the random arrival of users with real-time services. The CCP allows the network operator to adjust both the maximum allowed degradation level and the energy-saving gains over the time horizon.
2) We leverage the temporal statistical data of video content and users arrival to obtain a *robust* allocation of network resources over the *time horizon*. Such data is used to develop a deterministic equivalent form for the *stochastic* RP and CCP models. In essence, the probability distribution of video watching durations is used to quantify both the possibility of energy-saving and the risk of wasting resources. Thus, the network prebuffers future demands that has high likelihood of watching, and delays the delivery of future uncertain content. Similarly, the probability distribution of users' arrival and their traffic load are used to calculate the fluctuations in the time-varying network resources. The resultant allocation ensures that the QoS degradations do not surpass predefined level in the CCP model when the remaining network resources for video users are scarce. As such, a deterministic resource allocation model is obtained and takes into account both energy-savings and the risk of QoS violation during resources uncertainty. As opposed to traditional *non-predictive robust* approaches [28], [29], our model considers allocation with RP and CCP over a *time horizon* that captures dependency between the constraints.
3) For the network to obtain an on-line solution, a guided heuristic search algorithm with polynomial complexity is developed. The algorithm exploits the trade-off between the network constraints, i.e., resources and slot demand, and prediction gains. The heuristic considers the impact of decisions at one slot on the energy savings and QoS satisfaction in future slots. This strategy generates solutions that satisfy the probabilistic QoS level defined in the CCP model without compromising the energy savings of the RP. Whilst commercial solvers do not provide solutions in real-time, they can be only used as benchmarks.

This paper is organized as follows. In Section II we provide a background on PRA and robust stochastic-based optimization. Section III presents the system model and the problem definition. Section IV introduces the robust probabilistic formulation, and *recourse* and *chance-constrained* programming based deterministic formulations. The low complexity guided heuristic is presented in Section V, simulation

results are discussed in Section VI, and finally, we conclude the paper in Section VII.

II. Background and Related Work

A. Existing PRA for Video Streaming

Extensive network measurements demonstrated the predictability of users' behaviour up to 93% [30], including human mobility and activity [9]. Meanwhile, the radio signal strength and available bandwidth are found to follow repetitive spatio-temporal patterns [12], [13]. The availability of navigation systems at current user devices (e.g., smartphones) has enabled mobile operators to correlate the radio measurements (e.g., channel rates) with geographical locations, and construct the Radio Environment Map (REM) [31]. The REM is further used to retrieve the future radio conditions (e.g., channel rate) for the predicted mobility traces enabling the PRA to derive long-term proactive decisions over the anticipated time horizon.

The video delivery approaches in [4], [10], [11], [19], and [26] have applied the concept of PRA to achieve optimal radio resource utilization. Hence, the PRA minimizes the energy consumption and maximizes the delivered video quality, in low load scenario, while maximizes the QoS satisfaction by avoiding video stops, in high load scenarios. The PRA in general tries to avoid allocating resources to users during poor radio conditions, that consume more airtime per byte, while maximizing the allocation during peak conditions by leveraging the content availability and prebuffering capabilities at the Base Station (BS) and user devices. To calculate the resource share for each user, the PRA typically follows one of two strategies: greedy prebuffering or minimal allocation. The first is applied to users experiencing their peak channel rates where the network transmits as much video content as possible to avoid transmission with future poor channel rates. This is done under the assumption that the user will be watching the whole video. Prior to reaching such peak rates, the network serves the users with the minimal amount of resources to barely satisfy the time slot demand. The base station after prebuffering the video or sending the minimal amount, can then go into the sleeping mode to save energy or serve other real-time users.

To derive performance gains over opportunistic RA, the PRA in [4], [10], [11], [19], and [26] all assumed stable user behaviour with perfect knowledge of users' demands and network resources. Thus, in low load scenarios, the BS would be able to prebuffer a long video segment for users experiencing peak channel rates. However, the user might terminate the session before watching the prebuffered content which results in suboptimal resource utilization diminishing the prediction gain compared to the opportunistic non-predictive RA. As well, the random arrival of real-time users will impact the resource availability for video streaming users. This impact increases the risk of resource scarcity for the PRA target users especially during the minimal allocation strategy, resulting in increased video stops and QoS degradations.

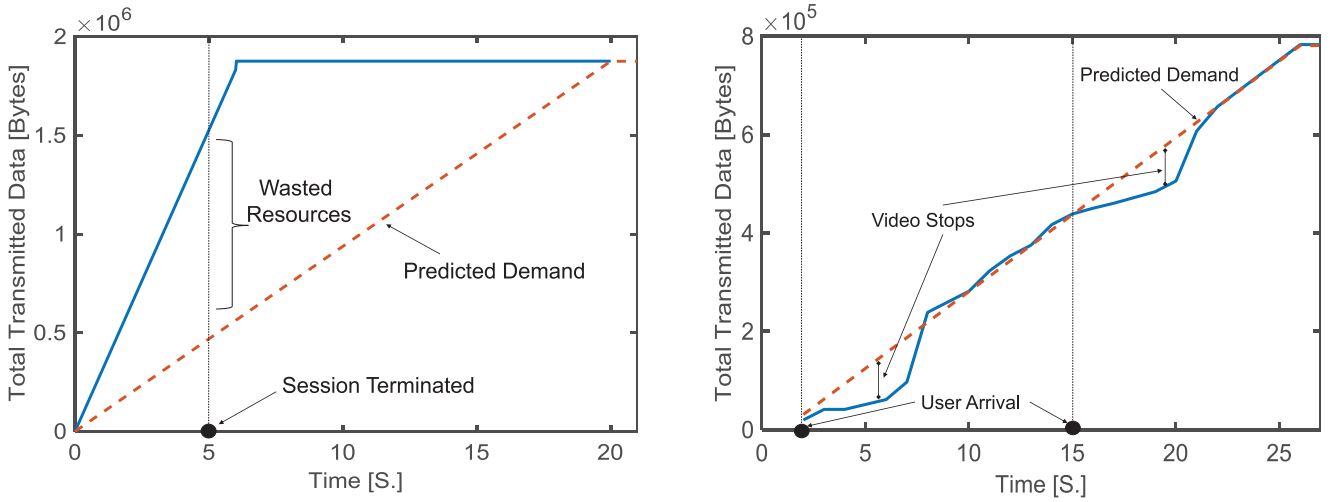
To tackle the uncertainty problem, we adopt robust stochastic optimization to measure the uncertainty in user demands and network resources. In addition to the predicted information, the stochastic optimization employs the probability of terminating the video, represented by watching time distribution, and the probability of scarce resources, represented by the user arrival distribution and load. The schematic diagrams presented in Fig. 2(a) and Fig. 2(b) illustrate the main difference between existing PRA and the proposed R-GPRA where the latter takes into consideration the uncertainties in both demand and network resources. As such, the *robust* PRA will only prebuffer the content which in all likelihood will be viewed by the user before terminating the video session and during time slots with a high probability of resource availability. We consider energy minimization where the proposed scheme will be referred to as robust green predictive resource allocation (R-GPRA) which is based on *stochastic* optimization techniques reviewed in the following subsection.

B. Robust Stochastic Optimization

Robust stochastic optimization refers to incorporating uncertain values in the mathematical programming model. In particular, these values are represented as random variables, and thus the set of constraints becomes probabilistic while the objective function appears in a non-closed scenario-based form [32]. As such, stochastic optimization provides the network designer with the flexibility of modeling the trade-off between maximizing prediction gains and minimizing the risk of violating the QoS.

The CCP is adopted to handle the probabilistic constraints by transforming them into a deterministic equivalent which is typically done using either the Probability Density Function (PDF) or Moment Generating Function (MGF) of the random variable as in Scenario Approximation (SA) and Bernstein Approximation (BA), respectively. The SA exploits the PDF of a random variable to construct all the possible realizations over the time horizon and their probabilities (i.e., probability mass function). The constraint will be represented in a form that guarantees a solution satisfying N scenarios, where the total probabilities of such N scenarios is above a minimal level denoted by β. Satisfying more scenarios will generate a conservative solution that has a lower value of the objective function and hence lower prediction gains. On the contrary, satisfying scenarios with less total probability will result in a non-robust solution that violates the QoS level.

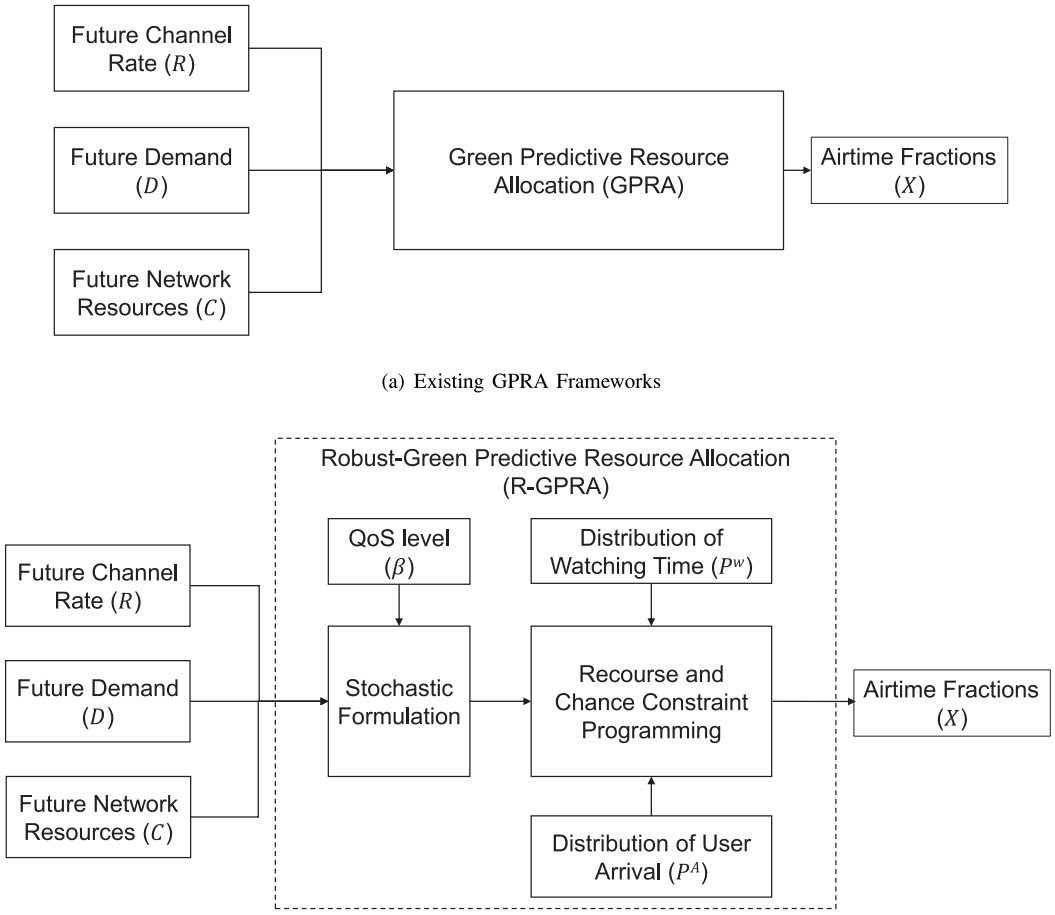
Similar to CCP, the Recourse Programming (RP) is used for handling problem uncertainties that impact the value of objective function. The RP model essentially consists of two terms in the objective function. The first term models the network gain by using the average of random variables while the second term adopts their PDF to prevent risky decisions that preserve the total gains in the network during uncertain conditions resulting in a deterministic closed form RA formulation which can be further solved by mathematical optimization techniques. As opposed to existing robust non-predictive RA [29], [33], our problem considers a time horizon that takes into account the interdependency between the resources of a single time slot and the future demands represented in a cumulative form.



(a) Example of PRA under uncertain demand

(b) Example of PRA under uncertain network resources

Fig. 1. Illustration of wasting resources and QoS degradations.



(a) Existing GPRA Frameworks

(b) Proposed R-GPRA Framework

Fig. 2. Schematic diagrams of existing and proposed frameworks.

Both CCP and RP have been discussed in the literature for non-predictive RA (without a time horizon) to handle uncertainties in the demand and channel rates [29], [33]. While the PRA typically involves decisions over a time horizon, the CCP has to account for the interdependency between the constraints. The violation of a time slot's constraint can propagate and impact the satisfaction of the next slot demand. For the RP, preventative decisions should be taken at the beginning of the time horizon to avoid suboptimal future decisions; unlike the conventional RP that reactively takes counter actions (e.g.,

releasing extra allocated resources) after the exact values of uncertain parameters are unveiled.

Motivated by the uncertainty modelling in [12] and [13], our earlier work in [20]–[22], and [34] focused on energy-efficient *robust* PRA where the uncertainty was only in future channel rates. In this paper, the proposed stochastic based R-GPRA achieves *long-term* energy savings and QoS satisfaction under demand and network uncertainty, which are demonstrated by the probability of video termination and resource unavailability at each time slot. A discrete linear formulation is obtained by RP and CCP, which can be solved by commercial solvers for benchmark solutions. Moreover, a low complexity guided heuristic for real-time allocations is introduced. This heuristic exploits the problem structure to achieve near optimal energy savings, and satisfy all the QoS and resource limitation constraints.

III. System Model and Problem Overview

A. System Model

Each BS is serving a set of video streaming users denoted by $\mathcal{M}$, where each user index is $i \in \mathcal{M}$. To achieve energy savings, a constant streaming rate is assumed which can be either manually selected by the user or decided by the network. Both the users' locations and channel rates are known for the next T time slots, where each slot index is denoted by $t \in \mathcal{T}$. The total video is available at the serving BS, at $t = 0$, and the bottleneck is assumed to be the radio link. The prediction of channel rates is performed by mapping the user's current location to the REM at the mobile operator. The REM contains both the user's locations and their corresponding channel rates $r_{i,t}$ for user i at time slot t [31].

1) Resource Allocation: The users of the same BS share the available radio resources every time slot t, where each user i is allocated a fraction of the slot's airtime denoted by $x_{i,t} \in [0, 1]$. Other real-time users are sharing the same resources, but their allocation is not handled by the R-GPRA.

2) Predicted User Demand: The average demand of user i at time slot t is denoted by $v_{i,t}$ which corresponds to the data content played back with fixed quality. We assume that the user can either terminate the session completely at any time slot t or skip part of the video and watch the subsequent frames. Unlike existing models in [35] that handle the first case only, our model allows the network to prebuffer the future content which might be watched by the user after skipping some frames of the video. Accordingly, the per slot demand is modeled as a random variable $\tilde{v}_{i,t}$ that is equal to 0 (user terminated the video) or $v_{i,t}$ (user streaming the video). The cumulative demand is denoted as a random variable $\tilde{D}_{i,t} = \sum_{t'=0}^{t} \tilde{v}_{i,t'}$.

3) Predicted Network Resources: At each time slot, the resources are shared among both the streaming users (considered by the R-GPRA) and other real-time users. The traffic of the latter is modeled using their arrival rate and demanded resources. The arrival of real-time traffic users is modeled as a Poisson distribution with mean λ, and the demand per user is denoted by C. The total airtime share allocated to real-time traffic users at time slot t is denoted by the random variable $\tilde{C}_t$. It has to be noted that we model the arrival of users' real-time traffic demand which is typically modelled as Poisson process [36], and not the mobile users admission or arrival at the BS.

4) Energy Minimization: With the current BS ON/OFF switching capabilities, the energy consumption E in the downlink is calculated using the consumed power P and the time X during which the base station was switched ON. Thus $E = P \times X = (P_W + P_D) \times X$, where P is the summation of both the power radiated over the wireless link, denoted by P_W, and the power for operating devices denoted by P_D. The value of P_W is constant since power control is not applied in LTE [37]. Similarly, the dominant part of P_D is consumed by the RF devices which is either fixed or negligibly varying and thus the power values is also constant, [38]. Thus, the energy consumption can be expressed in terms of the airtime X which corresponds to the time in which the above-mentioned devices are switched ON and the signal is radiated over the wireless link. In addition, measuring the energy consumption using the airtime will provide a common ground for energy-efficient PRA as the power term differs across BSs based on the efficiency of devices [39]. In conclusion, and similar to the existing PRA in [4] and [40], our framework will express the energy consumption as the total time fractions $\sum x_i, t$ allocated to the users.

B. Problem Description

The robust Green Predictive Resource Allocation (GPRA) scheme aims to calculate the airtime fractions $x_{i,t}$ for each user at time slot t such that the total allocated resources are minimized to achieve energy-saving or efficient bandwidth utilization. The possibility of terminating the video by the user at a certain time slot is taken into account. By doing so, this prevents the PRA from prebuffering future content to users who might terminate the video at any time slot with a certain probability. Typically, this results in more energy savings and optimal bandwidth utilization compared to existing *non-robust* PRA that assumed perfect demand prediction.

As illustrated in Fig. 3 (a), the values of predicted rates for three time slots would typically lead the PRA to prebuffer the whole content during the first slot to save energy as depicted in Fig. 3 (c). However, as shown in Fig. 3 (b), the high probability of terminating the video at the third time slot prevents the *robust* PRA from prebuffering the future content due to the high risk of wasting energy. As such, only the content of the second slot, with a low probability of video termination, is prebuffered whereas the delivery of the third slot's content will be postponed as illustrated in Fig. 3 (d). To summarize the example, delivering the rest of the video content in the third time slot costs more energy, in case of non-termination, while prebuffering all the contents causes a waste of resources in case of termination of viewing. The proposed robust PRA calculates this trade-off based on both the predicted rates and the probability of termination to perform the energy-efficient and QoS-aware allocation.

The uncertainty of future network resources, due to random user arrival, will interfere with the strategy mentioned

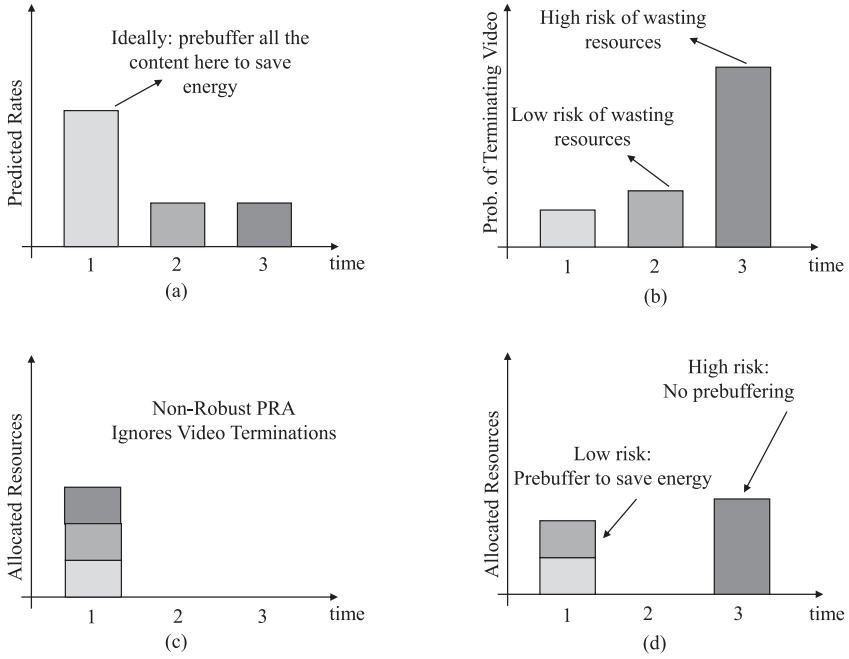


Fig. 3. Illustration of R-GPRA under Uncertain Video Streaming Demand.

earlier. Delaying the transmission in the case of high termination probability might be considered suboptimal if the future network resources are scarce. The network, in that case, will miss the chance of exploiting the current channel peaks and vacant resources, and without being able to satisfy the user demand given the future anticipated limited resources. As a result, fewer energy-savings are attained in the case of future peaks with low resources, while video stops are observed if future low channel rates are further reduced by real-time traffic user's arrival.

IV. R-GPRA Formulation Under Uncertain Demand and Resources

In this section, we mathematically formulate the problem of *robust* GPRA (R-GPRA) using stochastic optimization, and then adopt recourse and chance constraint programming to obtain deterministic equivalent forms.

A. Stochastic Formulation

The introduced *energy-efficient* robust PRA is formulated using stochastic optimization. In particular, the uncertain demand and future network resources are represented by random variables as follows:

$$\underset{\mathbf{x}}{\text{minimize}} \left\{ \sum_{\forall i \in \mathcal{M}} \sum_{\forall t \in \mathcal{T}} x_{i,t} \right\}$$

subject to:

$$\begin{aligned}
&\text{C1:} \quad \sum_{t'=0}^{t} r_{i,t'} x_{i,t'} \geq \tilde{D}_{i,t}, \qquad \forall\, i \in \mathcal{M}, \forall t \in \mathcal{T}, \\
&\text{C2:} \quad \sum_{i=1}^{M} x_{i,t} \leq 1 - \tilde{C}_t, \qquad \forall t \in \mathcal{T}, \\
&\text{C3:} \quad x_{i,t} \geq 0, \qquad \forall\, i \in \mathcal{M}, t \in \mathcal{T}. \quad (1)
\end{aligned}$$

The objective function aims to minimize the total consumed energy represented as a function of the total BS airtime [39]. The QoS constraint in C1 guarantees that the total delivered content to the user satisfies the anticipated cumulative random demand. C2 models the limited resources at each BS by ensuring that the sum of allocated airtime is less than the total available network resources (allocation slot duration) while considering the random resources allocated to the real-time traffic users. The last constraint C3 ensures the non-negativity of the decision variables. The main difference between the proposed *robust* formulation and the existing PRA work is the first and second constraints that now incorporate random demand and network resources. Such randomness has an impact on both objective function value and QoS satisfaction. In particular, when the random demand equals to $v_{i,t}$, the objective function is minimized by prebuffering the future content in peak rates. On the other hand, when the random demand becomes 0 (due to session termination) the objective function is minimized by avoiding prebuffering of future content. Similarly, the network should avoid prebuffering when available resources are low (due to the periodic arrival of real-time traffic users) as the pre-calculated resources will not be attainable.

B. Recourse and Chance Constrained Model

To represent the relation mentioned above between constraints C1, C2 and the objective function in a deterministic form, Recourse Programming (RP) and Chance Constrained Programming (CCP) models are used as depicted below:

$$\underset{\mathbf{x},\mathbf{y}}{\text{minimize}} \left\{ \sum_{\forall i \in \mathcal{M}} \sum_{\forall t \in \mathcal{T}} x_{i,t} + \mathbb{E}\big[H(\mathbf{y}, \tilde{D})\big] \right\}$$

subject to:

$$\text{C1:} \quad \sum_{t'=0}^{t} r_{i,t'} x_{i,t'} \geq \sum_{t'=0}^{t} v_{i,t'}, \qquad \forall\, i \in \mathcal{M}, \forall t \in \mathcal{T},$$

$$\begin{aligned}&\text{C2: } P_r(\sum_{i=1}^{M} x_{i,t} \leq 1-\tilde{C}_t) \geq \beta, \ \forall t \in \mathcal{T},\\ &\text{C3: } x_{i,t} \geq 0, \qquad \forall i \in \mathcal{M}, t \in \mathcal{T}.\end{aligned} \tag{2}$$

The objective function herein comprises two terms whose summation must be minimized. The first term represents the total allocated resources (similar to the non-robust approach) while the second term corresponds to the expected value of wasted resources as a result of terminating the video before watching the prebuffered content. In C2, the probability of satisfying the network resource constraint by the calculated airtime fractions is set above the QoS level β. Where $\beta \in [0, 1]$ represents the minimal probability of satisfying the QoS. Allocating more resources than the available capacity, after accounting for the real-time traffic users, will result in video stops since the users will not be able to receive the minimal data amount calculated by the R-GPRA. In the following, we show how to obtain a closed form representation for both the recourse model in the objective function, and the probabilistic constraint in C2.

1) Recourse Stage: The second term of the objective function in Eq. 2, i.e., $\mathbb{E}\big[H(\mathbf{y}, \tilde{D})\big]$, is the optimal solution of the recourse stage and formulated as follows[1]:

$$\underset{\mathbf{y},\mathbf{x}}{\text{minimize}} \ \left\{ \zeta \sum_{\forall i \in \mathcal{M}} \sum_{\forall t \in \mathcal{T}} p^{W}_{i,t} y_{i,t} \right\}$$

subject to:

$$\begin{aligned}&\text{C4: } r_{i,t-1} y_{i,t-1} + r_{i,t} x_{i,t} - v_{i,t} \leq r_{i,t} y_{i,t}, \\ &\qquad \forall i \in \mathcal{M}, \forall t \in \mathcal{T}, \\ &\text{C5: } y_{i,t} \geq 0, \quad \forall i \in \mathcal{M}, t \in \mathcal{T}.\end{aligned} \tag{3}$$

The objective function of the recourse stage in Eq. 3 minimizes the expected value of excess allocated resources (i.e., prebuffered) and calculated as a function of both the second stage decision variable $y_{i,t}$ and the probability of terminating the video denoted by $p^{W}_{i,t}$. The variable ζ is used to model the trade-off between the values of the two stages, and its value is typically less than one. The constraint in C4 is used to calculate the excess resources $r_{i,t} y_{i,t}$ after every time slot t. The first two terms on the left hand-side represent the total prebuffered and newly delivered content in this time slot, respectively. The third term represents the per slot demand in case of non-termination. The right hand-side shows the amount of excess resources after slot t which corresponds to the prebuffered future content.

2) Deterministic Equivalent: The probabilistic constraint in C2 is replaced by the following deterministic equivalent form which adopts the probability of user arrivals and their load.

$$\begin{aligned}&\text{C6: } \sum_{i=1}^{M} x_{i,t} \leq 1 - \big(C_{t,\omega} \delta_{t,\omega}\big) \quad \forall t \in \mathcal{T}, \ \forall \omega \in \Omega, \\ &\text{C7: } \sum_{\forall \omega \in \Omega} \delta_{t,\omega} {p^{A}}_{t,\omega} \geq \beta \qquad \forall \ t \in \mathcal{T}. \\ &\text{C8: } \delta_{t,\omega} \in \{0, 1\} \qquad \forall \ t \in \mathcal{T}, \ \forall \omega \in \Omega,\end{aligned} \tag{4}$$

[1]This subsection was preliminary proposed in our prior work [41].

The binary decision variable $\delta_{t,\omega}$ equals 1 if scenario ω at time slot t has to be satisfied by the airtime allocation, and equals 0 otherwise. The PDF of user arrival is used to construct the scenarios of network resources at each time slot as a result of real-time traffic user arrival. At each time slot t, the scenario ω represents the existence of ω real-time traffic users. The constraint in $C6$ demonstrates the scenarios in which the calculated airtime fractions must satisfy the vacant network resources $1 - C_{t,\omega}$. In $C7$, the total probability of satisfied scenarios must exceed the predefined QoS level β. ${p^{A}}_{t,\omega}$ is the probability of user arrival scenario ω at time slot t. When the scenario is ignored (i.e., $\delta_{t,\omega} = 0$), the right hand-side of $C6$ will be the maximum slot duration (i.e., all network resources are available), and the QoS level β will avoid ignoring the most probable scenarios.

C. Deterministic R-GPRA Linear Formulation

The complete deterministic formulation of the proposed R-GPRA can be summarized in the following closed form representation:

$$\underset{\mathbf{x},\mathbf{y},\delta}{\text{minimize}} \ \left\{ \sum_{\forall i \in \mathcal{M}} \sum_{\forall t \in \mathcal{T}} x_{i,t} + \zeta \sum_{\forall i \in \mathcal{M}} \sum_{\forall t \in \mathcal{T}} p^{W}_{i,t} y_{i,t} \right\}$$

subject to:

$$\begin{aligned}&\text{C1: } \sum_{t'=0}^{t} r_{i,t'} x_{i,t'} \geq \sum_{t'=0}^{t} v_{i,t}, \ \forall i \in \mathcal{M}, \quad \forall t \in \mathcal{T}, \\ &\text{C3: } x_{i,t} \geq 0, \qquad \forall i \in \mathcal{M}, t \in \mathcal{T}. \\ &\text{C4: } r_{i,t-1} y_{i,t-1} + r_{i,t} x_{i,t} - v_{i,t} \leq r_{i,t} y_{i,t}, \\ &\qquad \forall i \in \mathcal{M}, \ \forall t \in \mathcal{T}, \\ &\text{C5: } y_{i,t} \geq 0, \ \forall i \in \mathcal{M}, t \in \mathcal{T}. \\ &\text{C6: } \sum_{i=1}^{M} x_{i,t} \leq 1 - \big(C_{t,\omega} \delta_{t,\omega}\big) \ \forall t \in \mathcal{T}, \forall \omega \in \Omega, \\ &\text{C7: } \sum_{\forall \omega \in \Omega} \delta_{t,\omega} {p^{A}}_{t,\omega} \geq \beta \qquad \forall \ t \in \mathcal{T}. \\ &\text{C8: } \delta_{t,\omega} \in \{0, 1\} \ \forall \ t \in \mathcal{T}, \quad \forall \omega \in \Omega,\end{aligned} \tag{5}$$

The above formulation is obtained by combining Eq. 3 and Eq. 4, resulting in a mixed integer linear programming model. In the next section, we explore the possibilities and challenges of solving this NP-complete model, and propose a guided heuristic algorithm for real-time traffic allocation.

V. Real-Time Optimizer

This section reviews the numerical optimization methods that can be used to solve the formulated problem, and introduces the details of heuristic search algorithm followed by analysis of its computational complexity.

A. Optimal Solution

The robust formulation in Eq. 5 is a mixed integer linear programming model. As such, an optimal solution, which satisfies all the constraints, can be obtained using branch-and-bound or branch-and-cut, among other. Although these

techniques are capable of reaching an optimal feasible solution with a small duality gap, they suffer from low scalability and slow convergence. In particular, the complexity of such numerical optimization techniques grows exponentially with the number of decision variables [42]. These limitations are due to overlooking the problem structure and exploring a large area of the search space to avoid local optimal solutions. A guided heuristic algorithm is therefore proposed to provide a real-time feasible solution that is robust to prediction uncertainty. Commercial solvers (e.g., Gurobi [43]) that adopt these optimal techniques will be only used to evaluate the ability of the heuristic technique to maintain the prediction gains and satisfy the QoS constraints.

B. Guided Real-Time Heuristic

The proposed guided search heuristic algorithm utilizes knowledge about the problem's structure such as the interdependency and conflicts between the constraints, and their impact on the optimality of objective function. In essence, the algorithm starts by satisfying all the QoS constraints using the available radio resources while considering the distribution of user arrival and the predefined QoS level. To achieve energy minimization, resources are allocated to users that have not reached peak channel conditions. Then, the algorithm exploits the prebuffering capabilities of the mobile device for users experiencing peak channel conditions. This is done by pushing the video content in advance to avoid allocation during time slots with low channel rates or high congestion. In the next step, the value of the objective function is further minimized while examining the trade-off between possible energy savings during peak radio conditions, and the risk of wasting resources due to video termination in future time slots. The heuristic is summarized in Algorithm 1 and Algorithm 2 and detailed as follows.

In the first stage, minimal radio resources are calculated (line 2-18) in order to satisfy the QoS constraint $C1$ in Eq. 2 for each slot while considering the network resources uncertainties. The available network resources at each time slot are calculated as follows (lines 2-12):

1) The amount of resources in each scenario are initially sorted in ascending order and update the corresponding probability mass function
2) The scenarios are considered iteratively until the total probability reaches the QoS level β. Including more scenarios will result in a conservative solution that over-satisfies the QoS and deteriorates the value of the objective function.
3) The resources of the last considered scenario (i.e., the scenario that needs the maximum resources) are selected.
4) The total vacant capacity C'_t remaining for video streaming users is calculated and used in the next stage.

After satisfying constraints $C7 - C8$, the algorithm proceeds to fulfill the per slot demand constraint $C1$. This is accomplished by setting $C1$ to an equality and calculate the resource sharing $x_{i,t}$ that guarantee the satisfaction of demand. Such minimal allocation continues until the user reaches peak radio conditions (line 14). In high load scenarios, due to the large number of users or high streaming rates, the total allocated resources in a certain time slot might violate the airtime constraint $C6$ in Eq. 5. Accordingly, the preceding time slots with vacant resources will be used to prebuffer the content of the highly loaded time slots as depicted in lines 19-36 of Algorithm 1. While efficient exploitation of the radio resources is mandatory for these scenarios, the algorithm prebuffers the content of the user with the highest achievable rate. Thus, less airtime is consumed which increases the chance of satisfying the radio resource constraint $C2$. In the case of non-vacant resources, to accommodate the excess demand, the problem is said to be infeasible (lines 34-36).

To further minimize energy consumption, a calculated risk prebuffering strategy is applied by Algorithm 2. In essence, the possibility of prebuffering is checked. For each time slot following this peak, the amount of resources in the case of prebuffering and non-prebuffering is checked while considering the probability of video termination (lines 3-5) which approximates the objective function in Eq. 3. In the case of more resource saving (line 7), prebuffering is done (line 8-10). Otherwise, the risk of wasting resources is found to be high and minimal allocation is done for the demand of this slot without prebuffering in the previous slots (lines 13-16).

C. Algorithm Complexity

The first stage of the heuristic consists of sorting the scenarios (line 3) and calculating the total probability (line 6-11), each has a complexity of $O(N^2)$ in the worst case scenario. This stage is repeated for a maximum of T time slots, thus, the total complexity (lines 2-12) is $O(2T \times N^2)$. The minimal allocation in lines 13-18 has complexity of $O(MT)$, while the repairing of resources in lines 19-37 has a complexity of $O(MT^2)$ due to revisiting the preceding time slots to check the possibility of prebuffering. Similarly, the second part of the heuristic has a complexity of $O(MT^2)$ in which previous slots are also revisited for prebuffering any of the future slots with lower rates. Thus, the complexity of the whole proposed heuristic is $O(MT^2)$ which is a polynomial and significantly lower than the mathematical optimization methods whose complexity is non-polynomial and depends on the number of decision variables and constraints.

VI. Performance Evaluation

A. Simulation Environment

The proposed R-GPRA is developed in Network Simulator 3 (ns-3) Long Term Evolution (LTE) module where Gurobi (a commercial solver) is integrated to obtain benchmark solutions [44]. The probability of terminating the video at any time slot t is calculated using the model in [16]. Users follow random mobility traces within the cell coverage region at a constant velocity typical for suburban areas. The simulation parameters and numerical values are shown in Table I. The simulation is performed 25 times, and the average results of all runs are reported in the next subsections.

Algorithm 1: QoS Satisfaction Under Network Resource Uncertainty

Input : Users: $\mathcal{M}$, Time Horizon: $\mathcal{T}$, Predicted Rates: R, Demand Distribution: P, Streaming Rate: V ;
Output : X;
Initialization: $X = \emptyset$, $B = \emptyset$, $Y = \emptyset$, $Z = \emptyset$ $N_t = 0$ $\forall t \in T$;

Define: $t'_i = argmax\{r_{i,t}, \forall t \in T\}$;
for $t \in \mathcal{T}$ **do**
  $\hat{C}'_t = \text{Sort}(P^A_t \forall \omega \in \Omega)$;
  Initialize $S_t = 0$;
  Set minimum capacity $C'_t = 1$;
  while $S_t \leq \beta$ **do**
    for *omega* $\in \Omega$ **do**
      Update probability sum: $S_t = S_t + \hat{P}^A_{t,\omega}$;
      Update minimum capacity: $\hat{C}'_t = 1 - \hat{C}_{t,\omega}$;
    end
  end
end
for $i \in \mathcal{M}$ **do**
  for $t \in \mathcal{T}$ $|t \leq t'_i$ **do**
    Calculate minimal airtime $x_{i,t} = v_{i,t}/r_{i,t}$;
    Update used slot fraction $N_t = N_t + x_{i,t}$;
  end
end
for $t \in \mathcal{T}$ **do**
  if $N_t > 1$ **then**
    Set $k = t - 1$;
    while $k > 0 \& N_t > C'_t$ **do**
      if $x_{i,t} > 0 | i = argmax\{r_{i,k}, \forall i \in \mathcal{M}\}$ **then**
        Calculate the violated airtime $\Delta x_{i,t} = N_t - 1$;
        Calculate the demanded airtime $\Delta x_{i,k} = \Delta x_{i,t} \times \frac{r_{i,t}}{r_{i,k}}$;
        if $N_k + \Delta x_{i,k} \leq 1$ **then**
          Update $x_{i,k}$, $x_{i,t}$, N_t and N_k ;
          break;
        end
      end
      $k = k - 1$;
    end
  end
  if $N_t > C'_t$ **then**
    Return Infeasible Problem;
  end
end

Algorithm 2: Calculated Risk Prebuffering for Energy Minimization

Input : Users: $\mathcal{M}$, Time Horizon: $\mathcal{T}$, Predicted Rates: R, Demand Distribution: P, Streaming Rate: V ;
Output : X;
Initialization: $X = \emptyset$, $B = \emptyset$, $Y = \emptyset$, $Z = \emptyset$ $N_t = 0$ $\forall t \in T$;

Define: $t'_i = argmax\{r_{i,t}, \forall t \in T\}$;
for $t \in \mathcal{T} | t > t'_i$ **do**
  Calculate airtime without Prebuffering $x'_{i,t} = v_{i,t}/r_{i,t}$;
  for $\tau \in \mathcal{T}$ $|\tau < t, r_{i,\tau} > r_{i,t}, B_{i,t} \neq 1$ **do**
    Calculate airtime with prebuffering $z_{i,\tau} = v_{i,t}/r_{i,\tau}$;
    Calculate excess resources $y_{i,\tau} = \gamma \times p^W_{i,t} \times z_{i,t}$;
    if $x'_{i,t} > z_{i,\tau} + y_{i,\tau}$ **then**
      Update $x_{i,\tau} = x_{i,\tau} + z_{i,\tau}$;
      Update used slot fraction $N_t = N_t + z_{i,\tau}$;
      Update prebuffering status$B_{i,t} = 1$;
    end
  end
  if $B_{i,t} \neq 1$ **then**
    Update airtime without prebuffering $x_{i,t} = v_{i,t}/r_{i,t}$;
    Update used slot fraction $N_t = N_t + x_{i,t}$;
  end
end
return X

The main metric to assess the energy consumption is the total BS airtime [4], while the QoS of video streaming is quantified by the number and duration of video stops [45]. In addition, the corresponding Quality of Experience (QoE) of both number and duration of the stops is quantified by the Mean Opinion Score (MOS) [46], [47]. In essence, the QoE is a subjective metric that represents the service end-to-end performance level from the user's perspective, and can be calculated using the MOS formula in [46] and [47] depicted below:

$$MOS_{VS} = \frac{1}{|\mathcal{M}|} \sum_{i=1}^{|\mathcal{M}|} (2.99 \times e^{-0.96\eta_i} + 2.01). \quad (6)$$

$$MOS_{VD} = \frac{1}{|\mathcal{M}|} \sum_{i=1}^{|\mathcal{M}|} 4.59 \times e^{-3.44\zeta_i}. \quad (7)$$

where MOS_{VS} and MOS_{VD} are the MOS values due to number and duration of video stops, respectively. The average number and duration of video stops are denoted by η and ζ, respectively. The value of MOS varies from 1 to 5 which represents very poor to excellent service, respectively.

We adopt these metrics to evaluate the proposed R-GPRA, the existing non-robust PRA and the opportunistic RA (i.e., non-predictive). The following abbreviations are used in the next subsection:

- *PF (Non-PRA):* the traditional opportunistic proportional fair scheduler is used to represent the class of non-predictive schemes. It allocates the resources to the users based on their current channel measurements and cumulative served traffic in previous slots [48].
- *NR-GPRA:* is the existing energy-efficient predictive resource allocation that assumes perfect prediction and adopts deterministic formulations [4]. This scheme is simulated by setting the values of ζ and $C_{i,t}$ to zero in the objective function of Eq. 5, and the resultant formulation is solved using Gurobi optimizer [43].

TABLE I
SUMMARY OF MODEL PARAMETERS

Parameter	Value/Definition
BS transmit power	43 dBm
Bandwidth	5 MHz
Time Horizon T	60 s
ζ	0.99
Bit Error Rate	5×10^{-5}
Velocity	60 [kmph]
QoS level β	0.95
Packet size	10^3 [bytes]
Packet rate (from core network to BS)	$10^3 s^{-1}$
Buffer size	10^9 [bits]
Probability of watching ratio $p^W_{i,t/T}$	$2/\sigma\phi(\frac{t-\mu}{\sigma})\Phi(\alpha\frac{t-\mu}{\sigma}), \forall i \in M$
Probability of user arrival $p^A_{\omega,t}$	$\frac{\lambda^\omega e^{-\lambda}}{\omega!} \forall t \in T$
Standard deviation of watching time ratio σ	0.18
Skew parameter α	0.84
Mean of watching time ratio μ	0.27
User arrival rate λ	0.5
$\phi(x)$	PDF of normal distribution
$\Phi(x)$	CDF of normal distribution

- *PK-GPRA:* this refers to a hypothetical PRA with perfect knowledge of uncertain demand and network resources. As such it is aware of exact watching duration and amount of available resources. This is achieved by replacing the random variables in Eq. 1 by the exact values from the simulation.
- *OR-GPRA:* this represents the proposed *robust* green predictive resource allocation as formulated in Eq. 5. The probability of video termination follows the distribution in [16]. The optimal solution is obtained by the branch-and-cut methods in Gurobi optimizer [43].
- *HR-GPRA:* this refers to the heuristic version of **OR-GPRA** in which the solution is obtained by the proposed guided search in Algorithm 1 and Algorithm 2.

B. Simulation Results

1) Evaluating Demand Uncertainties: We initially evaluate the impact of uncertain demand solely on the prediction gains (i.e., energy savings). The system load, in terms of number of users and streaming rates, was configured and set below the available radio resources. Hence, no video stops were observed, and thus the QoS was satisfied by all the schemes, while the main focus remains on energy consumption. The maximum energy saving gap, referred to as prediction gain, is observed between the opportunistic non-predictive RA and hypothetical perfect knowledge PRA. As reported in the PRA literature [19], and shown in Fig. 4(a), the gain can reach up to 400 % due to the minimal allocation strategy adopted for cell edge users moving to peak radio conditions. This is in addition to maximizing the allocation for users exiting the cell.

The existing non-robust PRA (NR-GPRA), however, has diminished the gain to 150% as a result of the greedy prebuffering for cell center users exiting the cell, as yet not watching the full buffered video. On the contrary, the proposed *robust* GPRA has strategically prebuffered the video content to the users with poor future conditions, rather than transmitting their full content. Such risk-aware prebuffering strategy avoids greedy prebuffering of the future content whose delivery can be postponed until the corresponding time slots are reached, or the user arrives at time slots which have a low probability of terminating the video. This is in addition to following the minimal allocation to users experiencing poor conditions until they reach peak rate values. As such, the robust scheme was able to maintain the prediction gain at 320 %.

The same impact of uncertainty on the prediction gain was observed while increasing the streaming rate for fewer users Fig. 4(b). In this scenario, the maximum prediction gap can reach up to 150%, however, the uncertainties resulted in a 25% prediction gap as depicted by the non-robust scheme. The gain was retained to 100% by adopting the stochastic based robustness.

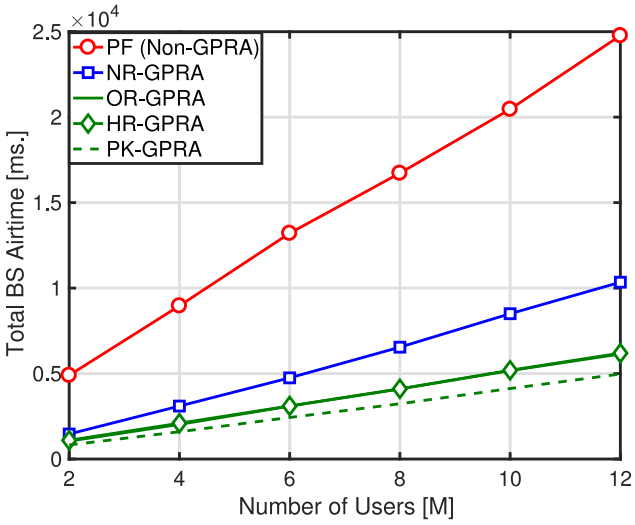
2) Evaluating Joint Demand and Resources Uncertainties: The simulations are extended to incorporate the resources uncertainties, where the QoS and QoE performance are depicted in Fig. 5(a)-Fig. 5(b) and Fig. 5(c)-Fig. 5(d), respectively.

The resources uncertainties violated the QoS level under the existing non-robust predictive scheme for a different number of users. Due to the arrival of real-time users, the network was unable to deliver the video content with the pre-calculated amount of resources. As such, the demand of cell edge users is not met by the minimal allocated resources that might be shared by the real-time users. The cell center video streaming users were not impacted due to the prebuffered content that surpasses the demand. Nevertheless, the substantial increase in the normalized number and durations of stops is attributed to the short video segments watched by the streaming users (i.e., demand uncertainty). The corresponding QoS demonstrates the exponential decay of users' experience as a result of experiencing a large number and durations of stops.

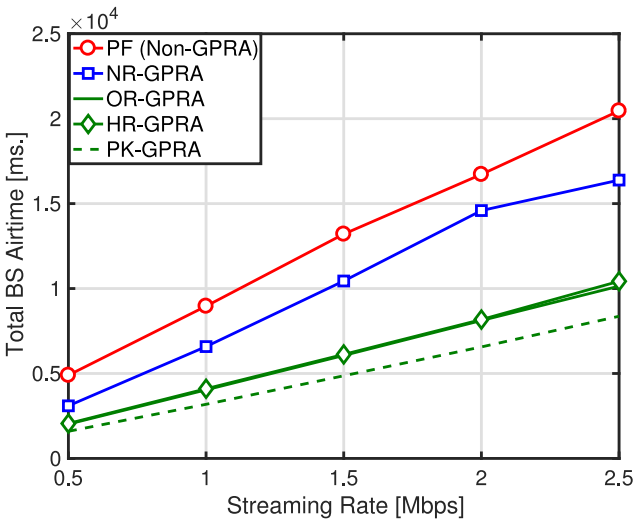
Unlike the non-robust scheme, the proposed optimal robust technique has satisfied the predefined QoS level (β) for all number of users. The robust scheme balances the amount of allocated resource to the cell edge and cell center users. Prebuffering is minimized for the cell center users and more resources can be reserved for the real-time users. As a result, the amount of allocated resources to cell edge users will be secured during the arrival of real-time users.

The performance of non-robust and robust predictive schemes is compared at different streaming rates and real-time user traffic as shown in Fig. 6(a) and Fig. 6(b). As the traffic load (streaming or real-time) increases, so does the number of unsatisfied users. With regards to energy savings and the prediction gain, the ability of robust scheme to maintain a high value was observed. Thus, the cost of robustness is said to be very low as the robust scheme avoided generating conservative solutions.

3) Performance of Heuristic: The above-mentioned observations over different system and streaming loads are also reported for the proposed heuristic. In essence, the heuristic was capable of satisfying the QoS level and maintain the prediction gap under demand and network uncertainties. The complexity of both the optimal and heuristic techniques is measured in terms of the computation time of a Quad Core i7-Processor, 3.2 GHz machine. The heuristic algorithm requires

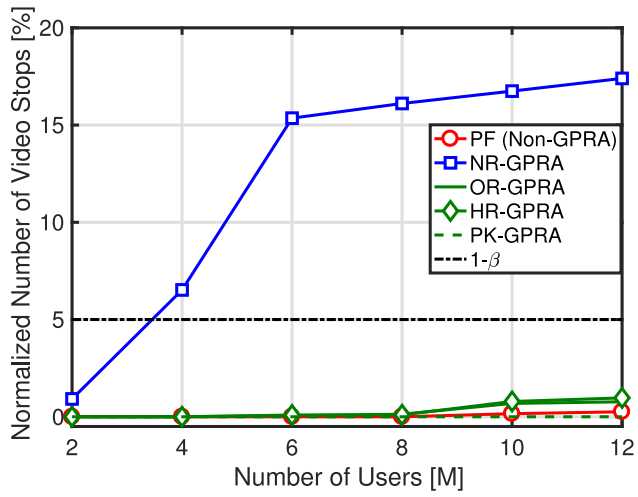


(a) Energy Consumption at V=0.5 Mbps

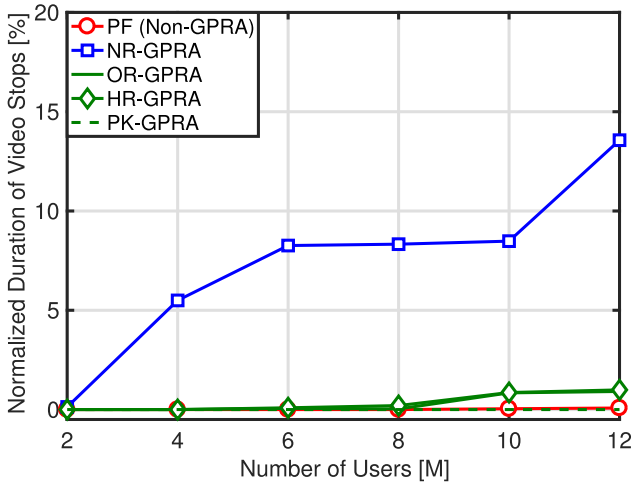


(b) Energy Consumption for 4 Users

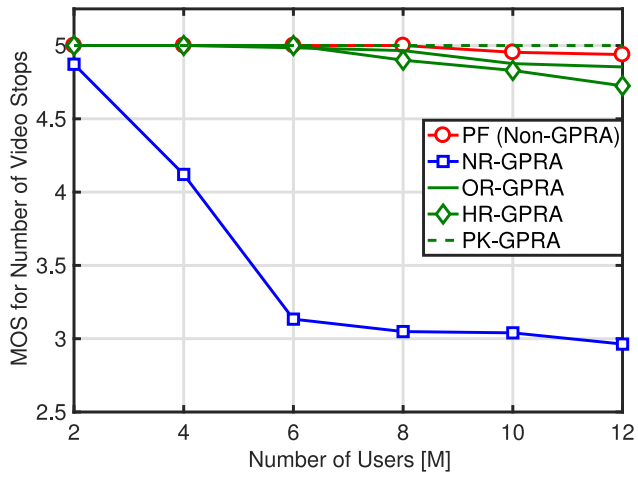
Fig. 4. Airtime-based energy consumption with uncertain demand only.



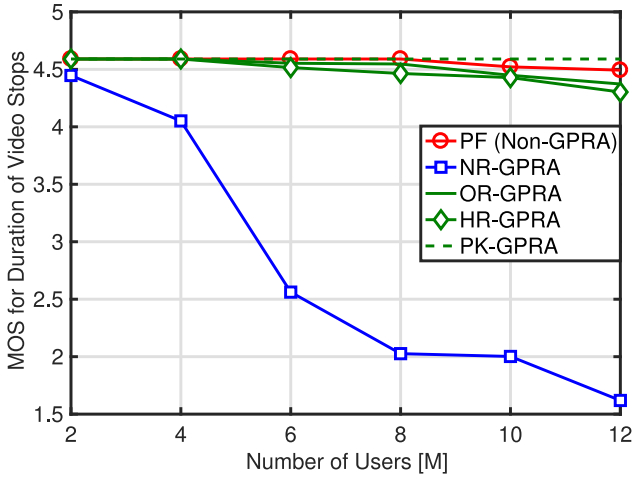
(a) Number of Stops at V=0.5 Mbps



(b) Duration of Stops at V=0.5 Mbps



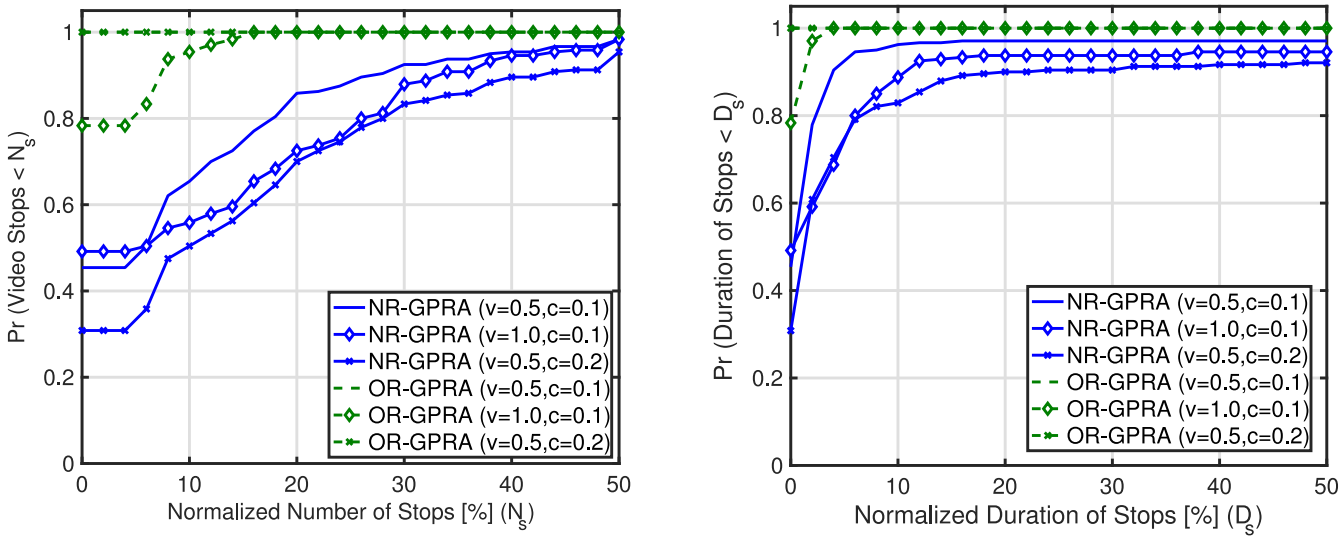
(c) QoE due to Number of Stops at V=0.5 Mbps



(d) QoE due to Duration of Stops at V=0.5 Mbps

Fig. 5. QoS and QoE for number and duration of stops with uncertain demand and network resources.

less than $0.1ms$. to solve the robust PRA formulation for all the network configurations (i.e., number of users and streaming rate values). On the other hand, the performance of Gurobi is sensitive to network load and capacity. The execution time varies from $1s$ to $15s$ depending on the number of unsatisfied users in the previous time slots, their streaming rate,



(a) Distribution of Stop Number

(b) Distribution of Stop Durations

Fig. 6. Distribution of QoS values for robust and non-robust GPRA.

and available channel capacity. Requests from users for high streaming rates while experiencing low channel capacity will result in a narrow feasibility region. Such situations are very challenging for the solver that overlooks the problem structure and generates a large number of branches and nodes to solve the integer programming model.

VII. Conclusion

We introduced a robust green predictive resource allocation (R-GPRA) scheme for video streaming that handles uncertainties in both the users' demands and network resources over a time horizon. Hence, R-GPRA avoids wasting resources and QoS violation. A stochastic formulation is proposed and a deterministic equivalent closed form representation was achieved using Recourse Programming (RP) and Chance Constrained Programming (CCP) models that adopt the probability of random video termination and arrival of real-time users. The resultant RP and CCP based formulations can be solved either by commercial solvers for benchmark solutions, or by the introduced guided heuristic search for real-time decisions. The performance evaluation, using a standard compliant simulator, demonstrated the ability of the introduced R-GPRA to maintain the energy-saving gains of PRA while satisfying the QoS levels. An increase in system load underlines the importance of having a robust scheme to avoid unnecessary excessive prebuffering for users leaving the cell center negating the high probability of terminating the video before viewing the full content. This is unlike existing PRA schemes that greedily exploit the peak radio conditions by prebuffering the whole future content without taking into consideration the users unstable demands. The proposed robust model can be extended to Dynamic Adaptive Streaming over HTTP (DASH) where the objective function can represent other QoS or QoE metrics such as average quality or quality switches. The video stops in that case will be handled by the probabilistic QoS constraints. Our future work thus considers the extension to DASH which jointly optimizes the resources and video qualities. This is in addition to considering other robust predictive forms such as long-term fairness and risk allocation for high load scenarios.

References

[1] CISCO. (2017). *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016-2021*. Accessed: Feb. 15, 2017. [Online]. Available: http://www.cisco.com/c/en/us/solutions/service-provider/visual-networking-index-vni/index.html

[2] K. Davaslioglu and E. Ayanoglu, "Quantifying potential energy efficiency gain in green cellular wireless networks," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 2065–2091, 4th Quart., 2014.

[3] Y. Bao, X. Wang, S. Zhou, and Z. Niu, "An energy-efficient client pre-caching scheme with wireless multicast for video-on-demand services," in *Proc. IEEE APCC*, 2012, pp. 566–571.

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

[5] M. B. Ghorbel, B. Hamdaoui, M. Guizani, and B. Khalfi, "Distributed learning-based cross-layer technique for energy-efficient multicarrier dynamic spectrum access with adaptive power allocation," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 1665–1674, Mar. 2016.

[6] H. Nguyen, G. Zheng, R. Zheng, and Z. Han, "Binary inference for primary user separation in cognitive radio networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 4, pp. 1532–1542, Apr. 2013.

[7] B. Hawelka, I. Sitko, P. Kazakopoulos, and E. Beinat, "Collective prediction of individual mobility traces for users with short data history," *PLoS ONE*, vol. 12, no. 1, 2017, Art. no. e0170907.

[8] 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.

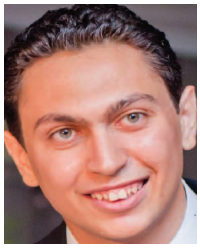
[9] A. Nadembega, A. Hafid, and T. Taleb, "An integrated predictive mobile-oriented bandwidth-reservation framework to support mobile multimedia streaming," *IEEE Trans. Wireless Commun.*, vol. 13, no. 12, pp. 6863–6875, Dec. 2014.

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

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

[12] 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*, 2015, pp. 1195–1200.

[13] N. Bui and J. Widmer, "Modelling throughput prediction errors as Gaussian random walks," in *Proc. KuVS Workshop Anticipatory Netw.*, 2014, pp. 1–3.
[14] I. Triki, R. El-Azouzi, and M. Haddad, "Anticipating resource management and QoE for mobile video streaming under imperfect prediction," in *Proc. IEEE ISM*, 2016, pp. 93–98.
[15] N. Bui, F. Michelinakis, and J. Widmer, "Mobile network resource optimization under imperfect prediction," in *Proc. IEEE WoWMoM*, 2015, pp. 1–9.
[16] Y. Chen, B. Zhang, Y. Liu, and W. Zhu, "Measurement and modeling of video watching time in a large-scale Internet video-on-demand system," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 2087–2098, Dec. 2013.
[17] W. Hu and G. Cao, "Energy-aware video streaming on smartphones," in *Proc. IEEE INFOCOM*, 2015, pp. 1185–1193.
[18] M. A. Hoque, M. Siekkinen, and J. K. Nurminen, "Using crowd-sourced viewing statistics to save energy in wireless video streaming," in *Proc. ACM MobiCom*, 2013, pp. 377–388.
[19] 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.
[20] 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.
[21] R. Atawia, H. S. Hassanein, H. Abou-Zeid, and A. Noureldin, "Robust content delivery and uncertainty tracking in predictive wireless networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 4, pp. 2327–2339, Apr. 2017.
[22] R. Atawia, H. Abou-Zeid, H. Hassanein, and A. Noureldin, "Chance-constrained QoS satisfaction for predictive video streaming," in *Proc. IEEE LCN*, 2015, pp. 253–260.
[23] R. Atawia, H. S. Hassanein, and A. Noureldin, "Robust long-term predictive adaptive video streaming under wireless network uncertainties," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 1374–1388, Feb. 2018.
[24] R. Atawia, H. Hassanein, and A. Noureldin, "Optimal and robust QoS-aware predictive adaptive video streaming for future wireless networks," in *Proc. IEEE GLOBECOM*, Dec. 2017, pp. 1–6.
[25] R. Atawia, H. Hassanein, and A. Noureldin, "Fair robust predictive resource allocation for video streaming under rate uncertainties," in *Proc. IEEE GLOBECOM*, Dec. 2016, pp. 1–6.
[26] H. Abou-Zeid, H. Hassanein, and S. Valentin, "Optimal predictive resource allocation: Exploiting mobility patterns and radio maps," in *Proc. IEEE GLOBECOM*, 2013, pp. 4877–4882.
[27] L. Chen, Y. Zhou, and D. M. Chiu, "Video browsing—A study of user behavior in online VoD services," in *Proc. IEEE ICCCN*, 2013, pp. 1–7.
[28] 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.
[29] M. Abdel-Rahman and M. Krunz, "Stochastic guard-band-aware channel assignment with bonding and aggregation for DSA networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 7, pp. 3888–3898, Jul. 2015.
[30] M. C. González, C. A. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.
[31] M. Neuland, T. Kurner, and M. Amirijoo, "Influence of positioning error on X-map estimation in LTE," in *Proc. IEEE VTC (Spring)*, 2011, pp. 1–5.
[32] P. Kall and S. W. Wallace, *Stochastic Programming*. Chichester, U.K.: Wiley, 1994.
[33] A. O. Fapojuwo, K. T. Chi, and F. C. M. Lau, "Energy consumption in wireless sensor networks under varying sensor node traffic," in *Proc. IEEE WCNC*, 2010, pp. 1–6.
[34] R. Atawia, H. Abou-Zeid, H. Hassanein, and A. Noureldin, "Robust resource allocation for predictive video streaming under channel uncertainty," in *Proc. IEEE GLOBECOM*, Dec. 2014, pp. 4683–4688.
[35] A. Finamore, M. Mellia, M. M. Munafò, R. Torres, and S. G. Rao, "YouTube everywhere: Impact of device and infrastructure synergies on user experience," in *Proc. ACM SIGCOMM*, 2011, pp. 345–360.
[36] B. Chandrasekaran, *Survey of Network Traffic Models*, Washington Univ., St. Louis, MO, USA, vol. 567, 2009.
[37] *LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures, v12.5.0*, 3GPP Standard 36.213, 2015.
[38] Y. Li, M. Reisslein, and C. Chakrabarti, "Energy-efficient video transmission over a wireless link," *IEEE Trans. Veh. Technol.*, vol. 58, no. 3, pp. 1229–1244, Mar. 2009.
[39] C. Desset *et al.*, "Flexible power modeling of LTE base stations," in *Proc. IEEE WCNC*, 2012, pp. 2858–2862.
[40] H. Abou-Zeid and H. S. Hassanein, "Toward green media delivery: Location-aware opportunities and approaches," *IEEE Wireless Commun.*, vol. 21, no. 4, pp. 38–46, Aug. 2014.
[41] R. Atawia, H. Hassanein, and A. Noureldin, "Energy-efficient predictive video streaming under demand uncertainties," in *Proc. IEEE ICC*, May 2017, pp. 1–6.
[42] G. Ausiello *et al.*, *Complexity and Approximation: Combinatorial Optimization Problems and Their Approximability Properties*. Heidelberg, Germany: Springer, 2012.
[43] Gurobi. *Gurobi Optimization*. Accessed: Sep. 29, 2016. [Online]. Available: http://www.gurobi.com/
[44] 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*, 2014, pp. 109–116.
[45] Y. Xu *et al.*, "Analysis of buffer starvation with application to objective QoE optimization of streaming services," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 813–827, Apr. 2014.
[46] L. G. M. Ballesteros *et al.*, "Energy saving approaches for video streaming on smartphone based on QoE modeling," in *Proc. IEEE CCNC*, 2016, pp. 103–106.
[47] T. Hoßfeld *et al.*, "Quantification of YouTube QoE via crowdsourcing," in *Proc. IEEE ISM*, 2011, pp. 494–499.
[48] T. Bu, L. Li, and R. Ramjee, "Generalized proportional fair scheduling in third generation wireless data networks," in *Proc. IEEE INFOCOM*, 2006, pp. 1–12.



Ramy Atawia (S'12–M'17) received the B.Sc. and M.Sc. degrees in communication engineering from the German University, Cairo, Egypt, in 2012 and 2013, respectively, and the Ph.D. degree in electrical and computer engineering from Queen's University in 2017. He is currently an Indoor Radio Solutions Developer with Ericsson Research and Development, Canada. He was a Member of Technical Staff and a Researcher with Bell Labs, Nokia. He was with Vodafone on autonomous optimization of radio networks, and was with Nokia on customer experience management and analytics. He was a teaching assistant and a guest lecturer where he delivered tutorials on optimization, wireless networks, and programming. His research work has appeared in top-tier IEEE journals and conferences, and it has led to 15 patents. His research includes stochastic optimization, predictive video streaming, machine learning, and AI in communication networks. He was a recipient of the Best Paper Award at IEEE GLOBECOM 2017. He also serves as a TPC member and a reviewer in 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. He is also the Founder and the Director of the Telecommunications Research Laboratory, Queen's University School of Computing, with extensive international academic and industrial collaborations. His record spans over 500 publications in journals, conferences, and book chapters, in addition to numerous keynotes and plenary talks in flagship venues. He was a recipient of several recognitions and best papers awards at top international conferences. 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 from 2008 to 2010).

Najah Abu Ali (M'07) received the B.Sc. and M.Sc. degrees in electrical engineering from the University of Jordan, and the Ph.D. degree from the Department of Electrical and Computer Engineering, Queen's University, Kingston, Canada, specializing in resource management in computer networks. She is currently an Associate Professor with the Faculty of Information Technology, United Arab Emirates University (UAEU). She has further co-authored a Wiley book on 4G and beyond cellular communication networks. Her general research interests include modeling wireless communications, resource management in wired and wireless networks, and reducing the energy requirements in wireless sensor networks. She has strengthened her focus on the Internet of Things, particularly at the nano-scale communications level, in addition to vehicle-to-vehicle networking. Her work has been consistently published in key publications venues for journals and conference. She has also delivered various seminar and tutorials at both esteemed institutions and flagship gatherings. She has also been awarded several research fund grants, particularly from the Emirates Foundation, ADEC, NRF/UAEU funds, and the Qatar National Research Foundation.

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 a 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 multisensor 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.