Energy-Efficient Predictive Video Streaming Under Demand Uncertainties

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

***Abstract*—Highly predictable users' location and traffic have enabled a new video delivery paradigm over wireless networks referred to as Predictive Resource Allocation (PRA). Existing research assumes perfect prediction of information in order to derive the performance bounds of PRA and define its gains over conventional Resource Allocation (RA). In this paper we sustain the application of energy-efficient PRA under prediction uncertainties. To that end, we propose a stochastic *robust* PRA scheme that models the uncertainty in future demands and incorporates them in the mathematical formulation. A linear Recourse Programming (RP) model is adopted in order to represent the trade-off between the energy-savings and the risk of wasting resources while considering the probability of a user terminating or skipping the video session. Thus, avoids prebuffering the video chunks that might be skipped by the user. A low complexity near optimal algorithm is then introduced to provide real-time solutions for the formulated RP model. Simulation results demonstrate the ability of the introduced robust PRA to deliver energy-efficient video streaming with lower resources than the existing PRA while promising QoS satisfaction. These results provide the impetus to implement the robust PRA in future wireless networks.**

I. Introduction

The prodigious increase in mobile video traffic has posed new challenges to network operators. Users are constantly demanding high Quality of Service (QoS) content which comes in the form of high definition videos with minimal interruptions and stops in future wireless networks [1]. Yet, operators are seeking to satisfy such demands with minimal capital and operational expenditure cost. Efficient exploitation of the available spectrum is mandatory to minimize both the energy consumption and deployment of new access nodes. Accordingly, the need for QoS aware and energy efficient radio resource allocation schemes in the context of video delivery has grown [2].

Predictive Resource Allocation (PRA) has recently emerged in the research community [2]–[4] promising high energy-savings and ubiquitous QoS satisfaction compared to existing opportunistic RA. Essentially, the PRA adopts future channel information and intended users' demands to provide long-term optimal decisions. The predictability of user behaviour and location, and application-aware network management have supported the use of PRA for wireless video streaming [5]–[7]. For instance, PRA can recognize video users moving towards coverage holes and prebuffer their video content ahead in order to save energy. For other users, the PRA postpones their video delivery until they reach the peak radio conditions where less resources are consumed to transmit the content. In contrast, the opportunistic non-predictive strategies overlook the future conditions and thus do not prioritize users exiting the cell resulting in QoS degradation or high energy consumption. This is in addition to prebuffering the content of users before reaching peak radio conditions which does not provide energy savings.

While all the gains of PRA have been concluded under idealistic scenarios [2]–[4], uncertainties in users' behaviour are needed to be taken into account. A user's demand might change at a certain time slot by either skipping or terminating the streaming session [8] based on the duration and popularity of the video. As a consequence, the amount of excess allocated resources for prebuffering are squandered away, which increases the energy consumption and results in suboptimal resource utilization. As such, *robust* PRA must be introduced in order to consider the possibilities of video termination while deriving long-term allocations [9]. The *robust* form should measure the risk of wasting resources and compare it to the possibility of energy-savings in order to determine when and which content to prebuffer.

This paper introduces a *robust stochastic* PRA framework that ensures energy savings and QoS satisfaction under demand uncertainties. We summarize the contributions of this paper as follows:

1) A probabilistic formulation for the energy-efficient PRA is developed where the uncertainty in users' demand is modelled as a random variable. Recent research works [8] revealed that users tend to terminate the video session at each time instance with a certain probability that is a function of the video duration and its type of content.
2) A Recourse Programming (RP) model is then adopted to obtain a deterministic equivalent formulation with closed form solution over a time horizon. In essence, the RP uses the probability distribution of the uncertain demand in order to model the trade-off between energy-saving on the one hand, and amount of wasted resources coupled with QoS satisfaction on the other hand.
3) A low-complexity guided heuristic search algorithm is developed to obtain a near optimal and feasible solution for the formulated RP model. Due to the dimensionality of the problem, commercial solvers will not provide solutions in real-time and thus can be only used as benchmarks. The guided heuristic exploits the problem's structure to obtain initial feasible allocations which are

978-1-4673-8999-0/17/$31.00 ©2017 IEEE

further optimized iteratively using the main features of predictive and robust resource allocation.

II. Background and Related Work

In Predictive Resource Allocation (PRA) for video streaming, users' future locations and channel rates are known over a time horizon. Users located near the center, and experiencing peak rates, but moving towards the cell edge should be allocated most of the resources. By leveraging the peak rates, a large portion of video can be prebuffered and avoids serving the user in future poor conditions.

PRA approaches [2]–[4] assumed stable user behaviour and perfect knowledge of users' demands. Thus, in low load scenarios, the BS would be able to prebuffer a long segment of the video for any user exiting the cell. However, the user might terminate the session before watching the prebuffered content which results in suboptimal resource utilization. This consumes more energy than the opportunistic non-predictive schemes. To solve this problem, a robust optimization technique is required in order to measure the uncertainty in user demand and adopts the probability of terminating the video. The *robust* PRA, in this case, will only prebuffer the content which in all likelihood will be viewed by the user before terminating the video session.

Robust optimization techniques have been discussed in the literature for non-predictive RA (without time horizon) in order to handle uncertainties in the demands [10]. *Stochastic* optimization is typically used to provide a robust formulation of the RA problem in which the predicted uncertain values are represented as random variables [11]. For the problem at hand, we mainly focus on Recourse Programming (RP) in which the expected losses, that are due to uncertain information, must be minimized. The RP model essentially consists of two terms in the objective function. The first one models the gains of the network when using the predicted values while the second term adopts their Probability Density Function (PDF) to take counter actions that preserve the total gains in the network during uncertain conditions. This results in a deterministic closed form RA formulation which can be further solved by mathematical optimization techniques. As opposed to existing robust non-predictive RA [10], our problem considers a time-horizon that takes into account the interdependency between the resources of one time slot and the future demands.

The proposed *robust stochastic* PRA achieves *long-term* energy savings and QoS satisfaction under demand uncertainty, which is demonstrated by the probability of video termination at each time slot. A linear continuous formulation is obtained, which can be solved by commercial solvers for benchmark solutions. Moreover, a low complexity guided heuristic for real-time allocations is also introduced. This heuristic exploits the problem structure to achieve near optimal solutions, and satisfy all the QoS and resource limitation constraints. This differs from our earlier proposed energy-efficient *robust* PRA approaches [12]–[15], where the demand was assumed to be perfectly predicted and the main source of uncertainty was the future rate.

III. System Model and Problem Overview

A. System Model

The set of video streaming users is denoted by $\mathcal{M}$ where the user index is $i \in \mathcal{M}$. Each mobile user requests video with a constant streaming rate in order to ensure energy savings. The video content is available at the serving BS and the bottleneck is assumed to be the radio link. Both the users' locations and channel rates are known for the next T timeslots, where each slot index is denoted by $t \in \mathcal{T}$.

1) Radio Resources: The users of the same Base Station (BS) share the available radio resources every time slot t, where each user i is allocated a fraction of the airtime denoted by $x_{i,t} \in [0, 1]$.

2) Predicted User Demand and Channel Rate: The average demand of user i at time slot t is denoted by $v_{i,t}$ which corresponds to the amount of data content played back with fixed quality. Herein, we assume that the demand is uncertain as the user can terminate the video at any time slot. Accordingly, the per slot demand is modelled as a random variable $\tilde{v}_{i,t}$ that is equal to 0 (user terminated the video) or $v_{i,t}$ (user steaming the video). The cumulative demand is thus denoted as a random variable $\tilde{D}_{i,t} = \sum_{t'=0}^{t} \tilde{v}_{i,t'}$. For the resource allocation, prediction of rate is done by mapping the user's current location to the Radio Environment Map (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.

B. Problem Description

The robust PRA scheme in this paper aims to calculate the airtime fractions $x_{i,t}$ for each user at time slot t such that the total amount of allocated resources is 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. This prevents the PRA from prebuffering future content to unstable 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 prediction.

As illustrated in Fig. 1 (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. 1 (c). However, as shown in Fig. 1 (b), the high probability of terminating the video at the third time slot prevents the *robust* PRA from prebuffering the future content in order to avoid the high risk of wasting energy. As such, only the content of the second slot is prebuffered due to the low probability of video termination while the delivery of the third slot's content will be postponed as illustrated in Fig. 1 (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 a termination of viewing. The proposed robust PRA calculates this trade-off based on both the predicted rates and

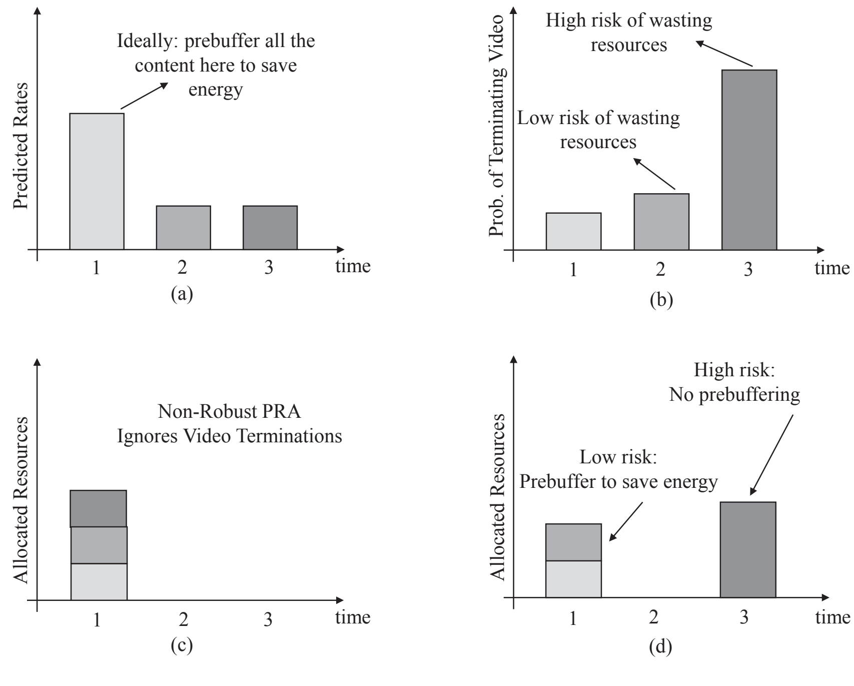


Fig. 1. Illustration of Robust PRA under Uncertain Video Streaming Demand

the probability of termination to perform energy-efficient and QoS aware allocation.

IV. PRA Formulation under Uncertain Demand

In this section we mathematically formulate the problem of *robust* PRA using stochastic optimization and recourse programming

A. Stochastic Formulation

The introduced *energy-efficient* robust PRA is formulated using stochastic optimization. In particular the uncertain demand is represented as a random variable as follows:

$$\underset{\mathbf{x}}{\text{minimize}} \quad \left\{ \sum_{\forall i \in \mathcal{M}} \sum_{\forall t \in \mathcal{T}} x_{i,t} \right\} \tag{1}$$

subject to:

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

The objective function aims to minimize the total consumed energy represented as a function of airtime. The QoS constraint in C1 guarantees that the total delivered content to the user satisfies the anticipated cumulative demand. C2 models the limited resources at each BS by ensuring that the sum of allocated airtime is less than 1 second (allocation slot duration). 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 constraint that contains the random demand. Such randomness directly affects the value of objective function. 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.

To represent the above mentioned relation between constraint C1 and the objective function, the Recourse Programming (RP) model is used.

B. Recourse Programming Model

A deterministic equivalent form for the introduced *energy-efficient* robust PRA in Eq. 1 is formulated using a Recourse Programming (RP) model as follows:

$$\underset{\mathbf{x},\mathbf{y}}{\text{minimize}} \quad \left\{ \sum_{\forall i \in \mathcal{M}} \sum_{\forall t \in \mathcal{T}} x_{i,t} + \mathbb{E}\big[H(\mathbf{y}, \tilde{D})\big] \right\} \tag{2}$$

subject to:

$$\begin{aligned}
&\text{C1:} \quad \sum_{t'=0}^{t} r_{i,t'} x_{i,t'} \geq \sum_{t'=0}^{t} v_{i,t'}, \quad \forall\, i \in \mathcal{M}, \forall t \in \mathcal{T},\\
&\text{C2:} \quad \sum_{i=1}^{M} x_{i,t} \leq 1, \qquad \forall t \in \mathcal{T},\\
&\text{C3:} \quad x_{i,t} \geq 0, \qquad \forall\, i \in \mathcal{M}, t \in \mathcal{T}.
\end{aligned}$$

The objective function herein comprises of two terms whose summation has to be minimized. The first term represents the total allocated resources (similar to the non-robust approach) while the second term corresponds to the total wasted resources as a result of terminating the video before watching the prebuffered content. In particular, the second term $\mathbb{E}\big[H(\mathbf{y},\tilde{D})\big]$ is the optimal solution of the recourse stage and formulated as follows:

$$\underset{\mathbf{y},\mathbf{x}}{\text{minimize}} \quad \Big\{\zeta \sum_{\forall i \in \mathcal{M}} \sum_{\forall t \in \mathcal{T}} p_{i,t} y_{i,t} \Big\} \tag{3}$$

subject to:

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

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 $p_{i,t}$. 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 amount of excess resources $r_{i,t}y_{i,t}$ after every time slot t. The first two terms in the left hand side represent the total prebuffered and newly allocated resources in this time slot, respectively, while the third term represents the per slot demand. The right hand side is the amount of excess resources after slot t which corresponds to the prebuffered future content.

V. Real-time Optimizer

This section firstly reviews the numerical optimization methods that can be used to solve the formulated problem, and then introduces the details of heuristic search algorithm.

A. Optimal Solution

The robust formulation in Eq. 2-Eq. 3 is linear and continuous which can be solved by numerical methods such as simplex and interior point. Although commercial solvers such as Gurobi [16] implement these optimal numerical methods, the computational complexity is non-tolerable and real-time solutions are unattainable. This is in addition to the low scalability with regards to the number of users and video duration, as the dimensions of the problem increase so does the amount of solver computation time.

B. Guided Real-time Heuristic

The proposed guided search heuristic algorithm is aware of the problem's structure and the interdependency between the constraints as well their impact on the objective function. This is in addition to exploiting the idea of robust and predictive allocation in calculating the airtime fractions. In essence, the algorithm starts by satisfying all the QoS constraints using the available radio resources while considering the minimization of objective function. Then, the algorithm exploits the prebuffering capabilities of user's device, where the video content can be pushed in advance to avoid allocation during time slots with low channel rates or high congestion. In the next step, the value of 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 detailed as follows:

In the first stage, the minimal amount of radio resources is calculated (line 4) in order to satisfy the QoS constraint C1 in Eq. 2 for each slot. Such strategy continues until the user reaches the slot with the peak channel rate and then the possibility of prebuffering is checked. For each time slot following this peak, the amount of resources in case of prebuffering and non-prebuffering is checked while considering the probability of video termination (lines 8-11) which approximates the objective function in Eq. 3. In case of more resource saving (line 12), prebuffering is done. 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 18-21).

Algorithm 1: Energy Minimization under Demand Uncertainty

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

1 **Define:** $t'_i = argmax\,\{r_{i,t}, \forall t \in T\}$;
2 **for** $i \in \mathcal{M}$ **do**
3   **for** $t \in \mathcal{T}\ |t \le t'_i$ **do**
4   Minimal airtime $x_{i,t} = v_{i,t}/r_{i,t}$;
5   Slot fraction $N_t = N_t + x_{i,t}$;
6   **end**
7   **for** $t \in \mathcal{T}|t > t'_i$ **do**
8   Without Prebuffering $x'_{i,t} = v_{i,t}/r_{i,t}$;
9   **for** $\tau \in \mathcal{T}\ |\tau < t, r_{i,\tau} > r_{i,t}, B_{i,t} \ne 1$ **do**
10   With prebuffering $z_{i,\tau} = v_{i,t}/r_{i,\tau}$;
11   Excess resources $y_{i,\tau} = \gamma \times p_{i,t} \times z_{i,t}$;
12   **if** $x'_{i,t} > z_{i,\tau} + y_{i,\tau}$ **then**
13         $x_{i,\tau} = x_{i,\tau} + z_{i,\tau}$;
14   Slot fraction $N_t = N_t + z_{i,\tau}$;
15   Prebuffering status $B_{i,t} = 1$;
16   **end**
17   **end**
18   **if** $B_{i,t} \ne 1$ **then**
19   Without prebuffering $x_{i,t} = v_{i,t}/r_{i,t}$;
20   Slot fraction $N_t = N_t + x_{i,t}$;
21   **end**
22   **end**
23 **end**
24 return X

VI. Performance Evaluation

A. Simulation Environment

We simulate the proposed robust PRA using the Long Term Evolution (LTE) module in Network Simulator 3 (ns-3) integrated with Gurobi commercial solver [17] to obtain optimal solutions for the formulated problem. The probability of terminating the video at any time slot t is calculated using the model in [8]. Users follow random predefined paths within the cell coverage region at a constant velocity of $60kmph$, which corresponds to typical values in urban and suburban areas. All the simulation parameters and values are presented in Table I. The average of all output results, over 50 simulation runs, is reported in the following subsections. In this evaluation, we compare the variants of the introduced *robust* PRA scheme with each other and with an existing non-robust PRA technique. The main metric is the energy consumption measured by the total airtime [2]. The **NR-PRA** refers to an existing non-robust PRA scheme in [2] that assumes complete streaming of the video without any probability of termination. **PRA-PP** refers to a hypothetical optimal PRA with perfect demand knowledge (i.e. exact duration of video that will be streamed). The proposed robust PRA in this work formulated in Eq. 2-Eq. 3 is referred to as **R-PRA (Optimal)** where the probability of video termination follows the distribution in [8]. The solution is obtained optimally using Gurobi optimizer [16]. **R-PRA (Heuristic)** is the same as **R-PRA (Optimal)** but the solution is obtained using the introduced guided heuristic in Algorithm 1.

TABLE I
Summary of Model Parameters

Parameter	Value
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]
Packet size	10^3 [bytes]
Packet rate (from core network to BS)	$10^3 s^{-1}$
Buffer size	10^9 [bits]

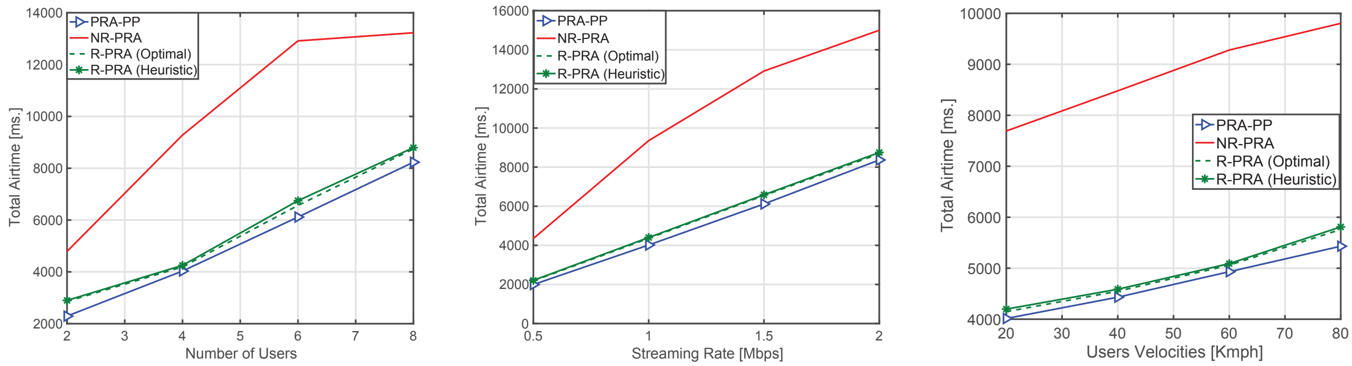
B. Simulation Results

1) Comparison to Non-Robust PRA: We set the system load (i.e. number of users and steaming rates) to a value comparable to or below the available radio resources. Accordingly, no video stops were observed and thus the QoS is said to be satisfied by all the schemes. With regards to the amount of energy, the existing non-robust PRA (i.e. NR-PRA) results in high consumption compared to the lower bound *PRA-PP* as shown in Fig. 2(a)-Fig. 2(c). This is because the video was predicted to be fully played back by the user and thus a large amount of resources was used in order to prebuffer the content at the cell center. However, such prebuffered portion was not streamed by the users who tend to terminate the video according to the predefined probability. The amount of wasted resources increases with both the system load (number of users and streaming rates) as depicted in Fig. 2(a)-Fig. 2(b) and the users velocity in Fig. 2(c). Particularly, increasing the number of users will result in prebuffering more content in order to provide vacant capacities for other users at the cell edge. Similarly, the content of fast moving users leaving the cell center has to be prebuffered before reaching the cell edge. Thus, the gap between the *NR-PRA* and the optimal *PRA-PP* increases. In contrast, the introduced *R-PRA (optimal)* has prebuffered the content for cell center users while considering the probability of video termination. This strategy avoids risky prebuffering of the future content whose delivery can be postponed until their corresponding time slots are reached or the user arrives at time slots with low probability of terminating the video. Such termination aware prebuffering resulted in lower energy consumption which is close to the perfectly predicted scheme as depicted in Fig. 2(a)-Fig. 2(c).

2) Evaluation of the Heuristic: The performance of the introduced heuristic is reported for different number of users, streaming rates and velocities in Fig. 2(a)-Fig. 2(c). Similar to its optimal form, *R-PRA (heuristic)* was able to satisfy the QoS in addition to achieving energy-savings with a very small optimality gap less than 0.1%. The complexity of both the optimal and heuristic techniques is measured in terms of the computation time. The heuristic algorithm only requires less than $0.1ms$. to solve the robust PRA formulation irrespective of the network load (i.e. number of users and streaming rates). This is as opposed to the commercial solver which required approximately $10s$. in low load scenarios. The solver's execution time increases with the number of users due to the larger problem dimension, and also with the streaming rate due to the tight feasibility region.

VII. Conclusion

We introduced a *robust* PRA scheme for video streaming that handles uncertainties in the users' demands and avoids wasting resources. A linear continuous RP model is adopted to provide closed form representation while adopting the probability of random termination. The RP model 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 robust PRA to maintain the energy-saving gains of PRA while satisfying the QoS levels. An increase in system load underlines the importance of the robust scheme to avoid unnecessary excessive prebuffering for users leaving the cell center but with 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 unstable users' demands. Our future work considers the extension to Dynamic Adaptive Streaming over HTTP (DASH) which jointly optimizes the resources and video qualities under demand uncertainty. This is in addition to considering the uncertainties in the channel vacant capacities resulted from unexpected users arrival.



(a) $V = 1.5$ Mbps (b) 6 Users (c) 6 Users with $V = 1$ Mbps

Fig. 2. Total Energy Consumption of Simulated Schemes

VIII. Acknowledgment

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

References

[1] J. C. Fernandez, T. Taleb, M. Guizani, and N. Kato, "Bandwidth aggregation-aware dynamic qos negotiation for real-time video streaming in next-generation wireless networks," *IEEE Trans. Multimedia*, vol. 11, no. 6, pp. 1082–1093, 2009.

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

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

[4] R. Margolies, A. Sridharan, V. Aggarwal, R. Jana, N. Shankaranarayanan, V. A. Vaishampayan, and G. Zussman, "Exploiting mobility in proportional fair cellular scheduling: Measurements and algorithms," in *Proc. IEEE INFOCOM*, pp. 1339–1347, 2014.

[5] C. Song, Z. Qu, N. Blumm, and A. Barabasi, "Limits of predictability in human mobility," *Science*, vol. 327, pp. 1018–1021, 2010.

[6] M. Elazab, A. Noureldin, and H. S. Hassanein, "Integrated cooperative localization for vehicular networks with partial gps access in urban canyons," *Vehicular Communications*, 2016.

[7] T. Taleb, K. Samdanis, and A. Ksentini, "Towards elastic application-oriented bearer management for enhancing qoe in lte networks," in *Proc. IEEE WCNC*, pp. 1–6, 2016.

[8] L. Chen, Y. Zhou, and D. M. Chiu, "Video browsing-a study of user behavior in online vod services," in *Proc. IEEE ICCCN*, pp. 1–7, 2013.

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

[10] 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, pp. 3888–3898, July 2015.

[11] P. Kali and S. W. Wallace, *Stochastic programming*. Springer, 1994.

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

[13] 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.*, pp. 1–14, 2017.

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

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

[16] Gurobi, "Gurobi Optimization." http://www.gurobi.com/. Accessed Sept. 29th, 2016.

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