Optimal and Robust QoS-Aware Predictive Adaptive Video Streaming for Future Wireless Networks

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*—The exploitation of mobility traces and rate predictions has enabled predictive delivery of video content that can achieve optimal resource utilization and long-term Quality of Service (QoS) satisfaction. The network recognizes users moving towards poor radio conditions in order to prioritize them over other users with better future conditions. In this paper, we propose a QoS-aware predictive Dynamic Adaptive Streaming over HTTP (DASH) scheme that leverages future information to select both the resource sharing and video qualities over a time horizon. The scheme minimizes the number of quality switches while achieving a minimal average quality level with no video stops. We firstly define the maximum prediction gains under idealistic conditions by a scheme referred to as *Optimal QoS-Aware Predictive*-DASH (OQP-DASH). Then, a robust stochastic based formulation is introduced to handle the practical uncertainty in predicted information, where the scheme is denoted by *Robust QoS-Aware Predictive*-DASH (RQP-DASH). A chance constraint programming model based on Scenario Approximation (SA) is adopted to cap the risk of service degradation while using the Probability Mass Function (PMF) of predicted rates. Under idealistic conditions, OQP-DASH outperforms the non-predictive opportunistic counterpart and results in fewer quality switches. Applying estimation errors, RQP-DASH avoids QoS degradations without compromising the prediction gains which supports the application of predictive DASH in future network.**

I. Introduction

Video traffic is expected to have a compound annual growth rate of 54 % in the next five years, contributing to more than three quarters of the overall mobile traffic [1]. Content providers strive to improve the user experience by developing channel-aware streaming protocols such as DASH [2]. On the other hand, network operators are currently concerned with allocating the optimal resources for each video streaming user such that Quality of Service (QoS) level is maximized. The upsurge in video traffic prompts a shift in the current DASH strategy from receiver-centric to network-centric such that the decisions of radio resource allocator are taken into account, specially in multi-user scenarios [3].

Current network-centric DASH allocations are based either on previous or future channel conditions. *Non-predictive* DASH relies on reported measurements, from user device, to periodically calculate the amount of resources and video quality for the current time slot. On the contrary, *Predictive*-DASH (P-DASH) adopts future channel conditions to calculate the amount of resources and quality for the slots of an upcoming time horizon [4]. In essence, the P-DASH exploits the rate predictions to minimize resources allocated to users prior reaching peak radio conditions. Thus, saves resources to prebuffer the future content, with high quality, of users moving towards poor radio conditions. Preliminary work on P-DASH demonstrated that substantial improvements in resource utilization and energy-saving can be achieved under idealistic conditions [4].

The video streaming QoS is measured by different metrics that model the user experience such as average video quality (i.e. bitrate), number of quality switches, number and duration of stops and initial buffer delays, among others [2]. While mobile users experience different channel conditions, due to large-scale fading, uniform selection of segments quality with minimal stops over a time horizon is very challenging. For instance, allowing users to stream high quality videos during poor conditions will typically result in video stops since the available radio resources are below the bitrate required for such high quality. Conversely, decreasing the quality to avoid video stops during poor conditions will result in frequent quality switches and thus low QoS level. This trade-off between the QoS metrics, i.e. stops and quality switches, has to be handled by the predictive DASH scheme while allocating resources and selecting the video quality for each user at every time slot in the planing horizon.

In this paper, we propose a *QoS-Aware Predictive*-DASH (QP-DASH) that adopts future channel conditions to strategically prebuffer the video content in advance with optimal quality to achieve long-term user satisfaction. The scheme aims to minimize the number of switches while ensuring the delivery of a minimal average quality level over the time horizon. In addition, video stops due to buffer underrun or resource limitations are avoided. Nevertheless, uncertainties in future information will be handled by means of *stochastic* optimization that strikes a balance between video quality switches and number of stops.

The contributions of this paper are summarized as follows:

1) We develop an optimal QoS-aware P-DASH that calculates both the resource sharing and quality of segments for the users over an anticipated time horizon. A mixed integer linear programming model is derived, optimized by commercial solvers, to obtain benchmark allocations and quantify QoS prediction gains over opportunistic DASH schemes under idealistic conditions.
2) To account for real world uncertainty, *stochastic* optimization model is introduced to represent predicted channel rates as random variables. Thus, provides a *robust* allocation that rectifies optimistic allocations of the

978-1-5090-5019-2/17/$31.00 ©2017 IEEE

optimal scheme. A Chance Constrained Programming (CCP) model is used to ensure that QoS is satisfied at a minimal probabilistic level even when actual rates deviate from their predicted values.

3) We obtain a deterministic mixed integer linear programming model for the *robust* scheme using Scenario Approximation (SA) which adopts the rate Probability Mass Function (PMF). This generates a closed-form representation that defines the cost of robustness and evaluates the prediction gains during uncertain conditions.

The paper is organized as follows. In Section II we provide a background on DASH and robust stochastic-based optimization. Section III presents the preliminaries, system model, and the optimal formulation. Section IV introduces the robust formulation and its deterministic equivalent, simulation results are discussed in Section V and finally, we conclude the paper in Section VI.

II. Background and Related Work

A. Predictive Network-Centric DASH

Current DASH [5] splits the video file and delivers it in the form of small segments that represent a content playback time. The quality of each segment is adapted based on user selected bitrate according to the channel condition. Low-quality segments are delivered to users experiencing low channel rates (e.g. at the cell edge), in order to avoid video stalling, while higher qualities are selected during peak rates to improve the resource utilization and user experience. Research efforts are currently concerned with shifting the DASH quality selection from user-centric to be network-driven in order to bridge the gap between the objectives of individual users, on one hand, and the resource allocator decisions on the other hand [6]. Hence, future networks will jointly optimize the segments quality and the resource shares among the users [3], [6]. Conventional network-centric DASH schemes adopt frequent channel measurements, reported by the user's device, to optimize the network metrics (e.g. resource utilization) and QoS (e.g. quality and interruptions).

Advancements in data mining and machine learning supported the prediction of mobility traces and future channel conditions at the user side. This led to *Predictive*-DASH (P-DASH) [4], [7]–[12] that relies on upcoming radio conditions to derive long-term allocation of resources and video qualities. In essence, users with poor future conditions will be prioritized, i.e. allocated a higher relative proportion of the resources, in order to prebuffer their future content at a high quality. On the contrary, users with future peak rates can accept low quality segments at their current poor conditions. Compared to conventional non-predictive approaches, the P-DASH results in energy-savings, optimal resource utilization, and long-term throughput fairness.

In this paper, we exploit information on the future channel conditions in order to minimize the number of switches in the quality of segments while ensuring a minimum average quality level over the time horizon. This is in addition to avoiding video stops, due to buffer underrun and resource limitation, at each time slot. The first scheme we introduce, referred to as *Optimal QoS-Aware Predictive*-DASH (OQP-DASH), assumes an ideal scenario, with no errors in future information prediction, and can be used to quantify the prediction gains against non-predictive schemes. The scheme jointly solves for both the resource sharing and segments quality of each time slot for every user. This is unlike the user-centric P-DASH [13] where each user requests the quality based on the measured bandwidth irrespective of other users and the resource allocator decision.

B. Robust Stochastic Optimization

While network operators rely on radio measurements to determine future rates, prediction uncertainty can not be ignored [14]. Random behaviour of received signal level will result in temporal and spatial variations of network conditions, thus imperfect estimations. Such uncertainty has to be modelled by the resource allocator in order to secure optimal and robust QoS satisfaction. As channel rates fall below the predicted values, minimal allocation to the users in poor radio conditions may result in video stops. Similarly, when the channel faces outages during anticipated peak conditions, the large amount of resources devoted to users will result in suboptimal utilization. This is in addition to the increased risk of video stops when high-quality levels are selected.

Stochastic optimization is typically used to provide a probabilistic formulation of the predictive resource allocation problem in which predicted uncertain coefficients are represented as random variables [15]. For the problem at hand, we focus on Chance Constrained Programming (CCP) in which the constraints accommodating random variables are represented by a probabilistic inequality with a maximum violation degree denoted by $\epsilon \in [0, 1]$.

A deterministic equivalent form is then derived by means of Scenario Approximation (SA), Gaussian Approximation (GA) or Bernstein Approximation (BA) [8], among others. The GA assumes that all the random variables follow a normal distribution, yet can be extended to other closed-form Cumulative Density Function (CDF). The mean and variance of each single random variable, in addition to the distribution and violation degree will be used to construct a Second order Cone Programming (SoCP) model. The BA adopts the Moment Generating Function (MGF) to develop a SoCP deterministic form based on the support of random variable distribution, thus can be applied in cases with non-closed form CDF. The SA utilizes the PMF to create a scenario tree of random variables realization. The allocator has to ensure that the calculated resources satisfy $(1 - \epsilon)$ % of the scenarios.

The second scheme we introduce, referred to as *Robust QoS-Aware Predictive*-DASH (RQP-DASH), extends the OQP-DASH by handling uncertainty in predicted information based on CCP. Hence, it ensures that the probability of video stops is kept below the violation degree ϵ. SA is then

applied to obtain a deterministic equivalent form solvable by commercial solvers, that generate optimal benchmark solutions and derive the cost of robustness. In addition to the linearity of formulation, the SA is suitable for optimization problems with time horizon, since the interdependency between time slot constraints is explicitly modelled. Our earlier work in [7]–[11] focused on robust energy-efficient predictive allocation approaches, where the objective was to minimize the energy consumption at a constant video quality. Here we focus on P-DASH where the scheme calculates both the resource sharing and quality of segments such that the QoS is maximized. In essence, the QoS is said to be maximized by minimizing the number of quality switches and video stops while achieving a minimal average quality over the time horizon.

III. Optimal Predictive DASH

A. Preliminaries

We use the following notational conventions throughout the paper. $\mathcal{X}$ denotes a set and its cardinality is denoted by X. Matrices are denoted with subscripts, e.g. $\mathbf{x} = (x_{a,b} : a \in \mathbb{Z}_+, b \in \mathbb{Z}_+)$. $Pr\{x\}$ is the probability of event x. $\tilde{r}$ represents a random variable with expectation $\bar{r} = \mathbb{E}[\tilde{r}]$. The absolute value operator of variable x is denoted by $\left|x\right| = \sqrt{x^2}$

B. System Model

The system considers a group of Base Station (BS)s, each serves an active user set denoted by $\mathcal{M}$, where the user index is $i \in \mathcal{M}$. The video is transmitted from the server to the Evolved Packet Core (EPC) and then cached at the BS, thus the bottleneck is the wireless link. We assume that user's mobility trace is known for the next T time slots, where each slot is denoted by $t \in \mathcal{T} = \{1, 2, \cdots, T\}$.

1) Radio Resources and Rate Prediction: The active users share the BS resources (airtime fractions) at each time slot t. The resource allocation matrix $\mathbf{x} = (x_{i,t} \in [0,1] : i \in \mathcal{M}, t \in \mathcal{T})$ represents the fraction of resources devoted to deliver user i content during time slot 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 network. The REM contains both the user's locations and the corresponding average rate values denoted as $\bar{r}_{i,t} = \mathbb{E}[\tilde{r}_{i,t}]$. The predicted uncertain rate is modelled as a random variable and denoted by $\tilde{r}_{i,t}$. Its PMF can be calculated as in [8], where the probability of the j^{th} realization of $\tilde{r}_{i,t}$ is denoted by $p^j_{i,t}$.

2) Video Quality Selection: Each video segment can be transmitted and streamed by quality level $q \in Q$, where Q is the set of possible segment qualities. The binary decision variable $\kappa^q_{i,t}$ is equal to 1 if the video segment transmitted to user i at time slot t is encoded in quality q, and 0 otherwise. Each segment consists of v_q bytes of data, which depends on the selected quality level q.

C. Mathematical Formulation

The introduced OQP-DASH assumes perfect prediction of channel rates represented by average values and formulated using the following mixed integer linear programming model:

$$\underset{\mathbf{x},\kappa}{\text{minimize}} \quad \left\{ \sum_{\forall i \in M} \sum_{\forall t \in T} \left| \sum_{\forall q \in Q_i} \kappa^q_{i,t} v_q - \sum_{\forall q \in Q_i} \kappa^q_{i,t-1} v_q \right| \right\} \tag{1}$$

subject to:

$$\text{C1:} \quad \sum_{t'=0}^{t} \bar{r}_{i,t'} x_{i,t'} \geq \sum_{t'=0}^{t} \sum_{\forall q \in Q_i} \kappa^q_{i,t'} v_q, \qquad \forall i \in \mathcal{M}, \forall t \in \mathcal{T},$$

$$\text{C2:} \quad \sum_{t=0}^{T} \sum_{\forall q \in Q_i} \kappa^q_{i,t} v_q \geq \hat{V} \times T, \qquad \forall i \in \mathcal{M},$$

$$\text{C3:} \quad \sum_{t'=0}^{t} \sum_{\forall q \in Q_i} \kappa^q_{i,t'} \tau_{i,t'} \geq t, \qquad \forall i \in \mathcal{M}, \forall t \in \mathcal{T},$$

$$\text{C4:} \quad \sum_{q \in Q_i} \kappa^q_{i,t} \leq 1, \qquad \forall i \in \mathcal{M}, t \in \mathcal{T},$$

$$\text{C5:} \quad \kappa^q_{i,t} \in \{0,1\}, \qquad \forall i \in \mathcal{M}, t \in \mathcal{T},$$

$$\text{C6:} \quad \sum_{i=1}^{M} x_{i,t} \leq 1, \qquad \forall t \in \mathcal{T},$$

$$\text{C7:} \quad x_{i,t} \geq 0, \qquad \forall i \in \mathcal{M}, t \in \mathcal{T}.$$

The objective function aims to minimize the total quality switches over time slots of the allocation horizon. This is calculated as the difference in the bitrate selected in each two consecutive time slots. The first QoS constraint, represented in C1, ensures that the total delivered content to the user satisfies the anticipated demand (function of the selected quality). This avoids video stops due to buffer underrun, as the cumulative allocated data to user i at each time slot t is more than the total data size of the requested video segments. The second QoS constraint is related to the total video quality delivered to the user over the time horizon, where a minimum average bitrate level $\hat{V}$ has to be delivered. The constraint in C3 complements C1 to ensure that the total duration of the selected segments is greater than the elapsed playback time thus avoids video stops due to buffer underrun. Where the playback duration of each segment is denoted by $\tau_{i,t}$.

C4 and C5 ensure that, for each user, only one quality level is selected at a given time slot. The sixth constraint C6 models the limited resources at each base station by ensuring that the sum of the airtime fractions is less than 1 second which is the duration of the allocation slot. The last constraint C7 ensures the non-negativity of the decision variable.

The above formulation necessitates replacing the absolute value operator in order to obtain a closed form solution. Linear constraints and auxiliary variables are thus adopted as shown below

$$\underset{\mathbf{x},\kappa,\mathbf{y}}{\text{minimize}} \quad \left\{ \sum_{\forall i \in M} \sum_{\forall t \in T} y_{i,t} \right\} \tag{2}$$

subject to:

C1 - C7,

C8: $\sum_{\forall q \in Q_i} \kappa^q_{i,t} v_q - \sum_{\forall q \in Q_i} \kappa^q_{i,t-1} v_q \leq y_{i,t}, \quad \forall\, i \in \mathcal{M}, t \in \mathcal{T}.$

C9: $-\sum_{\forall q \in Q_i} \kappa^q_{i,t} v_q + \sum_{\forall q \in Q_i} \kappa^q_{i,t-1} v_q \leq y_{i,t}, \quad \forall\, i \in \mathcal{M}, t \in \mathcal{T}.$

Compared to opportunistic non-predictive DASH, the introduced model will guide the resource allocator to follow two main strategies:

1) **Prebuffering based uniform qualities:** User residing at the cell center and start moving to the cell edge will experience poor channel rates in the future. The non-predictive strategy may typically result in transmitting high quality segments in the beginning, at the cell center, while opting for low quality levels in poor future conditions to decrease the risk of video stops. The introduced OQP-DASH, however, will provide uniform experience over the time horizon by selecting medium qualities during peak conditions, while utilizing the remaining vacant resources in prebuffering the future content in medium quality.
2) **Bare-minimum prebuffering:** For other users starting the session at cell edge and moving to the cell center, thus anticipating peak rates, the non-predictive scheme will follow a different strategy. Overlooking the future peak rates might result in excess prebuffering during poor conditions and various quality levels according to the network load. This impacts the aforementioned prebuffering strategy to the cell center users resulting in suboptimal utilization of resources. The proposed scheme instead will balance the trade-off between the amount of prebuffered content and the quality of selected segments. In low load scenarios, high quality segments are selected with minimal prebuffering until the user reaches the peak conditions.

The user, as such, will experience smooth streaming with minimal quality switches in addition to satisfying the average quality level. This is in addition to avoiding stops in current or future poor radio condition by maximizing the resource utilization over the time horizon. However, the performance of the proposed scheme is sensitive to uncertainties in predicted rate values which increase the risk of video stops while utilizing the bare-minimum prebuffering strategy.

IV. Robust Predictive DASH

A. Stochastic Formulation

The introduced *Optimal QoS-Aware Predictive*-DASH (OQP-DASH) scheme is extended to a robust form, referred to as RQP-DASH, based on Chance Constrained Programming (CCP). The uncertainty in predicted rate increases the risk of violating the QoS constraint in C1 Eq. 1. Thus, the average rate value will be replaced by a random variable which converts the deterministic constraint C1 to a probabilistic form as depicted below:

$$\underset{\mathbf{x},\kappa,\mathbf{y}}{\text{minimize}} \quad \left\{ \sum_{\forall i \in M} \sum_{\forall t \in T} y_{i,t} \right\} \tag{3}$$

subject to:

C1: $Pr\left\{ \sum_{t'=0}^{t} \tilde{r}_{i,t'} x_{i,t'} \geq \sum_{t'=0}^{t} \sum_{\forall q \in Q_i} \kappa^q_{i,t'} v_q \right\} \geq 1 - \epsilon_{i,t}, \quad \forall\, i \in \mathcal{M}, \forall t \in \mathcal{T},$

C2 - C9 ,

$\epsilon_{i,t} \in [0,1]$ is the probability that the QoS of user i is violated at time slot t, where $\epsilon_{i,t} = 1$ results in the maximum QoS violation.

The above formulation does not have a closed form solution due to the probabilistic constraint C1 and the random variable. We adopt the Scenario Approximation (SA) to obtain a deterministic equivalent form in the next subsection.

B. Deterministic Equivalent

The SA adopts the Probability Mass Function (PMF) of the uncertain rates to replace the probabilistic constraint. The PMF of every rate $\tilde{r}_{i,t}$ contains all the realizations $r^j_{i,t}$ and their probabilities $p^j_{i,t}$ that construct the scenarios over the time horizon. The approximation ensures that both airtime allocations and quality selections satisfy the scenarios with a total probability more than the defined QoS level (i.e. $1-\epsilon$).

Each scenario corresponds to one combination of the possible realizations of the uncertain rates in C1. For example, the constraint in the second time slot includes the rates in both the first and second time slot. The scenarios will comprise all the possible combinations of the realizations of these two rates. The first scenario consists of $r^1_{1,1}$ and $r^1_{1,2}$. $r^1_{1,1}$ represents the first realization of the rate at t=1, and $r^1_{1,2}$ is the first realization of the rate at t=2, both for the first user. The probability of this scenario will be the product of the individual probabilities (i.e. $s^1_{1,2} = p^1_{1,1} \times p^1_{1,2}$). The deterministic equivalent of C1 in Eq. 3 is captured by C10-C12 below

$$\underset{\mathbf{x},\kappa,\mathbf{y}}{\text{minimize}} \quad \left\{ \sum_{\forall i \in M} \sum_{\forall t \in T} y_{i,t} \right\} \tag{4}$$

subject to:

C2 - C9

C10: $\sum_{t'=0}^{t} r^{(j)}_{i,t'} x_{i,t'} \geq \delta^j_{i,t} \sum_{t'=0}^{t} \sum_{\forall q \in Q_i} \kappa^q_{i,t'} v_q, \quad \forall\, i \in \mathcal{M}, \forall t \in \mathcal{T}, \forall j \in \mathcal{J}_{i,t},$

C11: $\sum_{j \in \mathcal{J}_{i,t}} s^j_{i,t} \delta^j_{i,t} \geq 1 - \epsilon_{i,t}, \quad \forall\, i \in \mathcal{M}, t \in \mathcal{T},$

C12: $\delta^j_{i,t} \in \{0,1\}, \quad \forall\, i \in \mathcal{M}, t \in \mathcal{T}, j \in \mathcal{J}_{i,t},$

where $r^{(j)}_{i,t}$ is the j^{th} realization of the uncertain predicted rate at time slot t for user i. $s^j_{i,t}$ is the probability of the j^{th}

TABLE I
SUMMARY OF MODEL PARAMETERS

Parameter	Value
BS transmit power	43 dBm
Bandwidth	5 MHz
Time Horizon T	60 s
Streaming rates [Mbps]	$0.5, 1, 1.5, 2, 2.5$
Minimum average quality ($\hat{V}$) [Mbps]	1
Bit Error Rate	5×10^{-5}
Shadow correlation distance (d_{cor}) [17]	$50m$
Shadow standard deviationσ [17]	4
Velocity [km/h]	25 - 40
Packet size	10^3 [bytes]
Buffer size [bits]	10^9
Maximum violation ϵ	0.05

scenario at time slot t for user i. $\mathcal{J}_{i,t}$ is the set of scenarios at time slot t, $\delta^j_{i,t}$ is a binary decision variable which equals to 1 if the j^{th} scenario at slot t must be satisfied by the resource allocation vector x and equals 0 otherwise (C12). Constraint C11 guarantees that the total probability of all the satisfied scenarios exceeds the minimal QoS level $1 - \epsilon$.

V. PERFORMANCE EVALUATION

A. Simulation Setup

The proposed QP-DASH is evaluated using Long Term Evolution (LTE) standard compliant simulations by Network Simulator 3 (ns-3). All the formulated problems are solved optimally by Gurobi commercial solver which is integrated within the ns-3 environment [16]. The 3GPP fading model in [17] is added to the received power at the user device to simulate variations in predicted rate. For users mobility, random predefined paths within the cell coverage region are defined at random selected velocity from 25 to 40 km/h which is common in urban areas. Table I summarizes the parameters and values, and 25 simulation runs are performed with different random seeds.

The non-predictive version of the proposed DASH scheme, denoted by NP-DASH, is simulated by setting the value of $T = 1$ in Eq. 1 while assuming error-free channel rates. The proposed predictive scheme will be simulated in three versions denoted by OQP-DASH, AQP-DASH and RQP-DASH. The first scheme assumed hypothetically perfect prediction of all the future channel rates to define the maximum gains, to that end Eq. 1 is used. AQP-DASH adopts the average of uncertain rates as in Eq. 1, and is simulated under erroneous rate prediction. This scheme is used for sensitivity analysis and demonstrates the importance of the robust form. The last scheme, RQP-DASH, adopts the robust formulation in Eq. 4 and is simulated under the assumption of uncertain rates.

B. Simulation Results

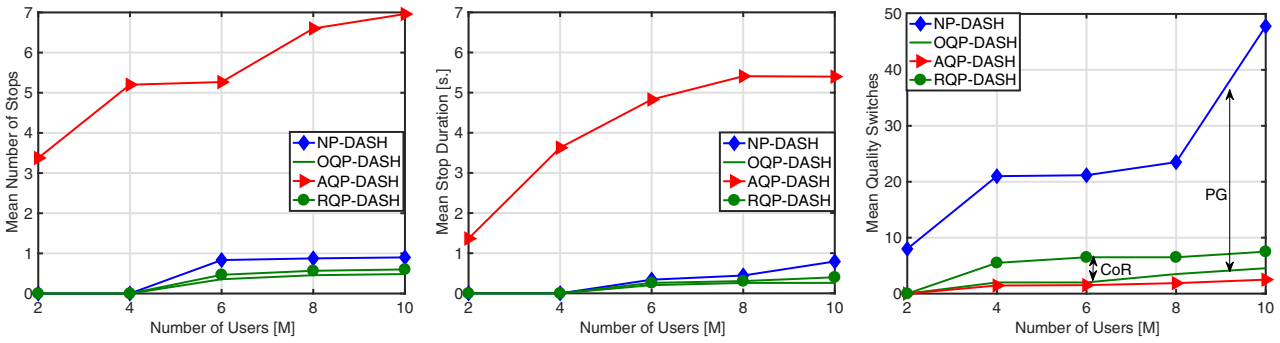
1) Prediction Gain (PG): The first metric considered in the evaluation is the PG which measures the network gains as a result of adopting future information. Quantitatively, the PG is calculated as the difference in the QoS metric (i.e. quality switches) between the NP-DASH and the OQP-DASH. In Fig. 1(c), the PG increases with the system load, i.e. number of users, due to the increased competition on the radio resources and different channel rates over the time horizon. The non-predictive scheme fairly allocates the resources to all the users for the instantaneous slot, which provides greedy quality selection of cell edge users without prioritizing users with poor future channel conditions. Thus, the latter set of users will initially receive high quality segments and then the quality will gradually decrease causing more switches. This is unlike the predictive scheme that prioritized these users and strategically buffered the future content in medium quality. An increase in the prediction gain is also observed when the average streaming rate $\hat{V}$ is increased for a fixed number of users as depicted in Fig. 2(c).

2) Cost of Robustness (CoR): The main challenge in the predictive scheme is the sensitivity to the errors in predicted rates. This is depicted by the number and duration of video stops experienced by AQP-DASH scheme in Fig. 1(a)-Fig. 1(b) for different number of users. More performance degradations are observed at higher error standard deviation and average quality as shown in Fig. 2(a)-Fig. 2(b). This is due to the bare-minimal allocation for users moving towards peak radio conditions that might violate the constraint C1 in case the channel rate falls below the adopted average value.

On the contrary, RQP-DASH adopts rate realizations with lower values than the average rate. Hence, provides more airtime for the selected quality to minimize video stops as depicted in Fig. 1(a)-Fig. 1(b) and Fig. 2(a)-Fig. 2(b). The CoR stands for the decrease in the value of objective function, compared to the optimal predictive strategy as a result of considering rate variation scenarios. In other words, CoR represents the decrease in the prediction gain. Varying the system load and prediction uncertainty level in Fig. 1(c) - Fig. 2(c) demonstrated the ability of the robust scheme to obtain stable non-conservative solutions that result in low CoR while avoiding video stops as shown in Fig. 1(a)-Fig. 1(b) and Fig. 2(a)-Fig. 2(b).

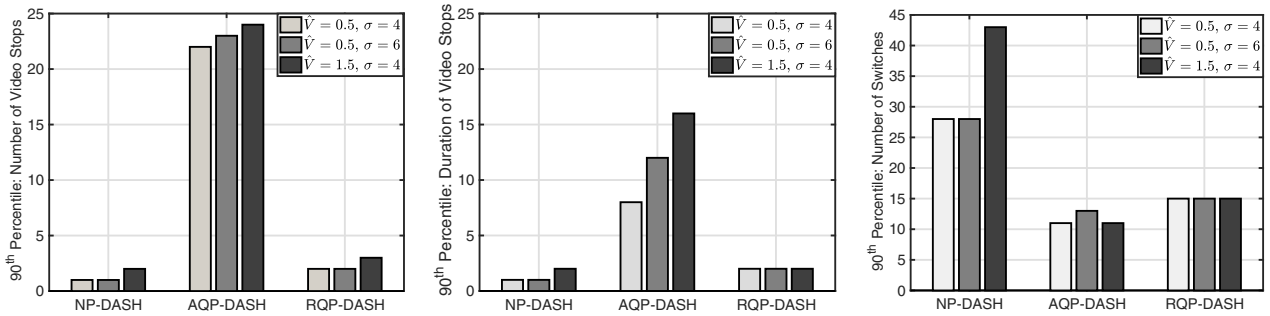
VI. CONCLUSION

We introduce a *QoS-Aware Predictive*-DASH (QP-DASH) scheme that solves for the resource sharing and video qualities jointly among the mobile users based on their future channel information. The scheme aims to improve the long-term QoS by minimizing the number of quality switches, video stops and stalling duration while satisfying a minimum predefined average quality level. Optimal predictive scheme, assuming perfect prediction, is initially introduced to define the maximum QoS gains over the non-predictive approaches. A uniform quality, with minimal switches, is thus achieved over the time horizon although users experience different channel conditions. To handle uncertainties in predicted rates, a stochastic formulation with probabilistic constraint satisfaction is introduced. Scenario Approximation (SA) is adopted to provide a deterministic equivalent form that can be solved by commercial solvers and define the cost of robustness. Performance evaluation demonstrates the ability of the robust scheme to limit the number and duration of stops as future rates



(a) Average number of video stops (b) Average video stops duration (c) Average Number Quality Switches

Fig. 1. Average performance of comparative schemes



(a) Number of video stops (b) Duration of video stops (c) Number of quality switches

Fig. 2. Performance of comparative schemes: 90^{th} percentile

fall below the predicted value. Modelling the rate PMF is of paramount significance to avoid conservative solutions which increase the cost of robustness. In addition to introducing heuristic real-time schemes, our future work will also consider extending the robust predictive DASH delivery scheme to handle uncertainties in the channel vacant capacities and user's demands.

VII. Acknowledgement

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

References

[1] CISCO, "Cisco visual networking index: Global mobile data traffic forecast update, 2016-2021." http://www.cisco.com/c/en/us/solutions/service-provider/visual-networking-index-vni/index.html, 2017. Accessed Mar. 15th, 2017.

[2] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hoßfeld, and P. Tran-Gia, "A survey on quality of experience of HTTP adaptive streaming," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 469–492, 2015.

[3] A. El Essaili, D. Schroeder, E. Steinbach, D. Staehle, and M. Shehada, "QoE-based traffic and resource management for adaptive HTTP video delivery in LTE," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 6, pp. 988–1001, 2015.

[4] H. Abou-zeid, H. S. Hassanein, and S. Valentin, "Energy-efficient adaptive video transmission: Exploiting rate predictions in wireless networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 5, pp. 2013 – 2026, 2014.

[5] T. Stockhammer, "Dynamic adaptive streaming over HTTP: standards and design principles," in *Proc. ACM on Multimedia Systems*, pp. 133–144, 2011.

[6] S. Cicalo, N. Changuel, V. Tralli, B. Sayadi, F. Faucheux, and S. Kerboeuf, "Improving QoE and fairness in HTTP adaptive streaming over LTE network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. PP, no. 99, pp. 1–1, 2015.

[7] 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.

[8] 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.

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

[10] R. Atawia, H. S. Hassanein, H. Abou-zeid, and A. Noureldin, "Robust content delivery and uncertainty tracking in predictive wireless networks," *IEEE Trans. Wireless Commun.*, pp. 1–13, 2017.

[11] R. Atawia, H. S. Hassanein, and A. Noureldin, "Energy-efficient predictive video streaming under demand uncertainties," in *Proc. IEEE ICC*, pp. 1–6, 2017.

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

[13] M. S. Mushtaq, B. Augustin, and A. Mellouk, "Regulating qoe for adaptive video streaming using bbf method," in *Proc. IEEE ICC*, pp. 6855–6860, IEEE, 2015.

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

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

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

[17] 3GPP, "LTE; evolved universal terrestrial radio access (E-UTRA); further advancements for E-UTRA physical layer aspects," Technical Report TR 36.814 V9.0.0, 3GPP, 2010.