Towards Mobility-Aware Predictive Radio Access: Modeling, Simulation, and Evaluation in LTE Networks

Hatem Abou-zeid
Electrical and Computer Engineering Dept.
Queen’s University
Kingston, Canada
h.abouzeid@queensu.ca

Hossam S. Hassanein
School of Computing
Queen’s University
Kingston, Canada
hossam@cs.queensu.ca

Ramy Atawia
Electrical and Computer Engineering Dept.
Queen’s University
Kingston, Canada
ramy.atawia@queensu.ca

ABSTRACT

Novel radio access techniques that leverage mobility predictions are receiving increasing interest in recent literature. The essence of these schemes is to lookahead at the future rates users will experience, and then devise long-term resource allocation strategies. For instance, a YouTube video user moving towards the cell edge can be prioritized to pre-buffer additional video content before poor coverage commences. While the potential of mobility-aware resource allocation has recently been demonstrated, several practical design aspects and evaluation approaches have not yet been addressed due to the complexity of the problem. Furthermore, since prior works have focused on specific applications there is also a strong need for a unified framework that can support different user and network requirements. For this purpose, we present a novel two-stage Predictive Radio Access Network (P-RAN) framework that can efficiently leverage both future data rate predictions in the order of tens of seconds, and instantaneous fast fading at the millisecond level. We also show how the framework can be implemented within the open source Network Simulator 3 (ns-3) LTE module, and apply it to optimize stored video delivery. A thorough set of performance tests are then conducted to assess the performance gains and investigate sensitivity to various prediction errors. Our results indicate that P-RANs can jointly improve both service quality and transmission efficiency. Additionally, we also observe that P-RAN performance can be further improved by modeling prediction uncertainty and developing robust allocation techniques. Based on our findings, we propose a simple yet representative prediction error model and discuss future directions towards robust P-RANs in conclusion of this paper.

Categories and Subject Descriptors
C.2.5 [Computer-Communication Networks]: Network Architecture and Design—Wireless communication; C.2.5 [Computer-Communication Networks]: Local and Wide-Area Networks—Access Schemes; C.4 [Performance of Systems]: Modeling Techniques

Keywords
Mobility-awareness, predictive radio access, LTE simulation, ns-3, performance under uncertainty.

1. INTRODUCTION

Advances in vehicular communications and in-vehicle telematics systems are driving an unprecedented growth of mobile traffic. Such developments will, however, introduce excessive congestion across wireless infrastructure, compelling operators to expand their networks. An alternative to expansion is to develop more efficient content delivery paradigms. In particular, alleviating Radio Access Network (RAN) congestion is paramount to operators as it postpones costly investments in radio equipment installations and new spectrum. Novel approaches in RAN design are therefore receiving increasing interest to support future vehicular connectivity at sustainable costs.

Fortunately, the predictability of human mobility [15], particularly that of vehicles [13], offers unique opportunities to devise proactive RAN transmission schemes. Knowing the routes vehicles are going to traverse enables the network to anticipate the future rates users will experience, and forecast spatio-temporal demands. To accomplish this, mobility trajectories are coupled with Radio Environment Maps (REMs) [28, 33] that provide estimates of the supported data rates at different geographical locations. This form of predictive radio access that leverages mobility information has been a recent topic of investigation. For instance, contemporary works have demonstrated its significant potential to improve network throughput and fairness [3, 25], video streaming experience [6, 29], as well as energy efficiency [5, 10]. However, due to the complexity of the problem, several practical aspects in design and evaluation have not yet been addressed, and mostly ideal assumptions have been made. The underlying complexity arises due to the long-term Resource Allocation (RA) which requires a joint optimization over a time horizon in the order of tens of seconds. Additionally, the effects of uncertainty in the predicted information such as user trajectories, and radio map accuracy remain mostly unexplored. A further requirement towards realizing the emerging Predictive Radio Access Net-
works (P-RANs) is the development of a unified framework that can accommodate different applications and network objectives. The design and practical evaluation of such a P-RAN is the focus of this paper. Our contribution towards this end is three-fold:

1. We develop a generic P-RAN that can efficiently leverage both future data rate predictions and instantaneous rate-opportunistic scheduling. This is accomplished by a two-stage design approach that i) can be easily tailored for different Quality of Service (QoS) requirements and network objectives, and ii) adapts to uncertainty in the predicted information through two levels of feedback. Details of how the framework can be implemented to extend the open source Network Simulator 3 (ns-3) LTE module [30] are also provided.

2. We define (and conduct) comprehensive sensitivity tests that investigate the robustness of P-RANs to 1) REM errors, 2) localization errors, 3) exploiting fast-fading, and 4) varying prediction window sizes. These tests can provide a guideline to assess the practical viability of future P-RAN schemes.

3. We provide insights on how REM uncertainty can be efficiently modeled. We show that simple asymmetric triangular fuzzy numbers can model REMs if the Signal to Interference plus Noise Ratio (SINR) exhibits a log-normal prediction error. Directions toward robust P-RANs that incorporate uncertainty are then highlighted to fully exploit the potential of P-RANs.

The rest of this paper is organized as follows: Section 2 provides the necessary background and reviews related work, while the system overview is covered in Section 3. The proposed two-stage P-RAN framework is then presented in Section 4 along with the corresponding ns-3 LTE implementation details. This is followed with extensive performance analysis and insights on rate uncertainty modeling in Section 5. Finally, Section 6 concludes with future directions.

2. BACKGROUND AND RELATED WORK

Before presenting the mobility-aware P-RAN framework, we first provide a brief overview on human mobility prediction and geographical bandwidth predictability. The recent advances in localization, mobility prediction, and real-time REM generation are driving forces for the development of predictive radio access mechanisms, which we then review and discuss with several use cases.

2.1 Human Mobility Predictability

LTE networks support a wide range of location positioning methods with varying levels of granularity, and a dedicated LTE Positioning Protocol (LPP) is also devised to coordinate signaling between the User Equipment (UE) and the BS [12]. Concurrently, a plethora of navigation hardware and software is also available in today’s smart phones enabling users to report their current location. Determining a mobile user’s location is therefore readily possible.

Recent analyses on human mobility traces also indicate that people tend to follow particular routes regularly, thereby enabling high predictability [15]. Several promising research results have also been presented based on real data from both cellular networks and GPS traces. While some works focus on short-term mobility prediction (next cell) [34], others predict the full trajectory [7], [13]. Approaches used include data mining [34], Markov renewal processes [7], and learning routes between common destinations via string matching [20]. User behavior profiling [9] is also a promising approach which is gaining momentum in the Location Based Services (LBS) industry as well.

2.2 Radio Environment Maps

Network operators commonly conduct road drive tests to measure radio signal strengths and other performance metrics at different locations. This information is then processed to generate maps of channel capacity (or other QoS metrics) [14]. These maps are commonly referred to as coverage, radio, or network performance maps. Openly accessible radio maps are also available online such as the OpenSignal Project [28] where signal strength is crowdsourced from users.

In addition to experimental location-rate mapping, there are several analytical studies that model the geographical correlation of channel capacity. Phillips et al. [29] present a comprehensive survey of path loss modeling approaches for different terrains/environments. Sampling strategies and interpolation techniques that incorporate practical measurements into the derived models are also presented. The authors conclude that online learning strategies and data mining approaches using measurement-based models are likely to provide the most robust radio mapping frameworks. An example of such an approach is the work in [24] which characterizes the impact of different environments and sampling positions on the wireless channel predictability.

While radio maps have typically been used to estimate the current supportable data rates, they can also enable the prediction of future rates users will experience, provided their mobility trajectories are known. Practical measurements confirming this have been recently conducted [33], [17]. Yao et al. [33] analyze bandwidth traces collected from two independent cellular providers for routes running through different radio conditions including terrestrial and underwater tunnels. Their findings confirm the correlation between user rates and location, and indicate that bandwidth uncertainty can be reduced drastically when observations from past trips are also used. Han et al. [17] also conduct a similar measurement study, and addresses other contextual factors such as user speed, time of day, and humidity to predict the available bandwidth more accurately.

Note that our focus in this paper is not to develop techniques that generate the radio maps themselves, but to propose and evaluate a predictive RAN framework that can exploit mobility predictions as illustrated next.

2.3 Predictive Radio Access Networks

The use of mobility predictions in cellular networks has been previously investigated with promising results in location management/paging [21, 32, 36], and handoff resource reservation [10, 11, 35]. However, only limited work has considered long-term resource allocation based on radio maps. The approaches proposed in handoff management are mostly concerned with optimizing resource reservation for imminent handoffs, primarily voice calls. Today, mobile application usage is changing. This is driven by both higher connectivity speeds as well as larger screen sizes of pads, tablets, and more recently phablets. For example, although watching a 20 minute sitcom on a phone screen was previously not
Rate predictions have also been proposed to develop energy efficient RAN transmission [4,5,19]. For instance, consider a video streaming user approaching the BS during low load. To save energy, only the minimum amount of content required for smooth streaming should be transmitted (i.e., without prebuffering) until the user reaches the cell center where a bulk of content can be efficiently transmitted in a short time. Thus, mobility predictions enable the network to plan spectrally efficient rate allocations without violating user streaming demands.

As previously discussed, this paper differs primarily from preceding works by 1) proposing a robust P-RAN framework that can be applied to any of the aforementioned applications, and integrated on top of traditional LTE schedulers, 2) defining necessary sensitivity tests to assess practical P-RAN performance, and 3) proposing a simple yet representative rate uncertainty model that can be used in future robust P-RAN designs.

3. SYSTEM OVERVIEW

We consider a BS with an active user set \( M \), where an arbitrary user is denoted by \( i \in M \). Users enter from the left cell edge and move in a straight line towards the other edge. The wireless link is assumed to be the bottleneck, and the requested content is always available at the BS.

The following notational conventions are used: \( X \) denotes a set and its cardinality is denoted by \( X \). Matrices are denoted with subscripts, e.g. \( x_{a,b} \in \mathbb{Z}^+_a, b \in \mathbb{Z}^+_b \).

3.1 Radio Environment Map and Mobility Information

The REM assumed to be typically available at the service provider would contain the average data rates at different network locations. In order to represent such a radio map, we use the Friis path loss propagation model as implemented in ns-3 [30]. The SINR at each \( x \) and \( y \) coordinate that users traverse is then computed and the corresponding achievable rate is determined based on the Channel Quality Indicator (CQI)-to-Modulation and Coding Scheme (MCS) mapping in 3rd Generation Partnership Project (3GPP) standards for Long Term Evolution (LTE) [1].

Fast fading is modeled according to the vehicular power delay profile defined in 3GPP standards [2].

We first assume that user mobility information is known accurately for the upcoming \( T \) seconds, which we call the prediction window. Localization errors are then introduced to investigate the impact of uncertainty. By coupling a user trajectory with the REM it is possible to determine an estimate of the upcoming rates a user can receive.

3.2 Scheduling Time Scales

Scheduling in LTE is typically performed every Time Transmission Interval (TTI) by channel and/or queue-aware schedulers based on the users’ CQIs. However, the objective of

![Figure 2: Relationship between the considered time intervals.](image-url)
P-RANs is to plan scheduling for multiple time instances jointly, optimized over a time window (of the order of tens of seconds). It is therefore not practical to design the long-term RA plan at the millisecond (TTI) level for two reasons: 1) a prediction window of only 60 seconds would generate \( \approx 60,000 \) TTIs requiring joint optimization which is not scalable, and 2) it is not practically possible to predict the average data rate at a TTI granularity. We therefore propose two time scales of operation as illustrated in Fig. 2. A prediction slot \( \tau \) is defined as the duration over which the average user rate can be assumed to be constant. A typical value of \( \tau = 1 \text{s} \), for user speeds up to 20 m/s, during which the average SINR will not vary significantly. As shown in Fig. 2, several prediction slots make up the prediction window, which is composed of \( N = T/\tau \) slots. As time progresses the prediction window slides to include more information of the future user rates. Therefore, at time slot \( n \), the set of considered prediction slots is denoted by \( N_n = \{ n, n + 1, \cdots , n + N - 1 \} \). We define the matrix of future user rates observed at time slot \( n \) by \( \hat{r} = (\hat{r}_{i,n'}: i \in M, n' \in N_n) \). This information is leveraged by the P-RAN framework to optimize scheduling as detailed next.

4. PREDICTIVE RADIO ACCESS FRAMEWORK FOR LTE

4.1 Architecture

The proposed predictive radio access framework is composed of two stages operating at different time scales but interacting dynamically as shown in Fig. 3 and discussed below.

**Long-term RA:** This upper level RA is responsible for the long-term allocation planning over a finite time horizon, based on user rate predictions. It is here that mobility information is leveraged to provide an additional level of RA diversity, and bulk transmissions are strategically planned. The output of this stage is an airtime allocation matrix \( x = (x_{i,n'} \in [0,1]: i \in M, n' \in N_n) \) which gives the fraction of time during each slot \( n' \) that the BS bandwidth is assigned to each user. The actual number of bits transmitted, at each slot, is the element-wise product \( x \odot \hat{r} \) as illustrated by the bars in the sample allocation plan in Fig. 4.

**Instantaneous Scheduling:** A lower level opportunistic scheduler operating at the TTI level with instantaneous user CQI is used to implement the long-term RA plan. The primary objective is to exploit fast fading diversity and ensure that users receive their per slot data requirements as specified by the upper allocator.

**Feedback:** Two levels of feedback are incorporated in the P-RAN. First, the instantaneous scheduler monitors the cumulative number of bits transmitted during a prediction slot \( \tau \), and only users that have not reached their target are considered for future TTIs. The second level of feedback occurs at the end of each slot \( n \) as shown in Fig. 3. Here, the actual rates \( \hat{r} \) transmitted to each user are computed. This may differ from the originally planned allocation due to rate prediction errors, i.e., if \( \hat{r} \) is not accurate. This information is then fed back to the upper level RA to update the future allocation plan based on the actual data transmitted. Since rate uncertainty will be exhibited by practical systems, incorporating feedback is important to enable a degree of robustness to such errors.

There are several advantages of this two stage architecture. First, by separating long-term RA from instantaneous scheduling, the framework can be integrated above existing traditional schedulers as detailed in the ns-3 LTE implementation in Section 4.4. And secondly, the predictive RA component can be designed to support different applications in a modular approach without affecting the underlying TTI level scheduler. Moreover, the rate of re-planning long-term RA can also be controlled to reduce overhead based on feedback. For instance a low variation between \( \hat{r} \) and \( \bar{r} \) indicates a stable operation an may require less frequent re-allocations. We now discuss each scheduling stage in detail.

4.2 Long-term Resource Allocation Planning

The objective of this stage is to make a long-term allocation plan at time slot \( n \) for the upcoming \( N \) slots. To do so, the upcoming user demand \( d_{i,n} \) at a per slot level is first defined. For example, in video streaming \( d_{i,n} \approx V \tau \) where \( V \) is the streaming bit rate and \( \tau \) is the slot duration. Other applications may be similarly defined by a minimum per slot bit rate. To optimize RA over a time horizon it is useful to define the cumulative demand requested by slot \( n \), which is denoted by \( D_{i,n} = \sum_{n'=1}^{n} d_{i,n'} \). With these defini-
tions, an application specific RA problem can be formulated to determine the optimum airtime allocations \( x_{i,n} \) based on the predicted user rate matrix \( r \), and the cumulative user demand \( D_{i,n} \). Detailed examples of such formulations have been presented to improve throughput and fairness [3], enhance video streaming QoS [6], and minimize BS airtime [4]. In this paper, we focus on applying the proposed P-RAN framework to minimize BS airtime of stored video delivery.

Minimizing Transmission Time of Stored Videos:
The essence of predictive video streaming in P-RAN is to strategically transmit content ahead of time at the UE, after which transmission can be momentarily suspended while the user consumes the buffer [4, 23]. If we consider a user requesting a stored video at slot \( n = 1 \), with a streaming rate of \( V \) [bit/s], then the cumulative content requested is \( D_{i,n} = V \tau n \). To experience smooth streaming, the cumulative allocation made to a user \( i \) by slot \( n \) should be greater than or equal to \( D_{i,n} \). BS transmission time can be minimized by leveraging future user rate knowledge \( \hat{r} \) to make bulk transmissions at times of high channel conditions, while making the minimal transmissions that ensure smooth streaming at other times. This achieves lower airtime usage, resulting in lower power consumption or more resources for other services.

The optimization problem of minimizing BS airtime at time slot \( n \) over the prediction window, without causing any streaming discontinuities can be formulated as the following Linear Program (LP):

\[
\begin{align*}
\text{minimize} & \quad \sum_{n'=n}^{n-1+M} \sum_{i=1}^{M} x_{i,n'} \\
\text{subject to:} & \quad \sum_{i=1}^{M} x_{i,n'} \leq 1, \quad \forall n' \in \mathcal{N}_n, \\
& \quad R_{i,n-1} + \sum_{n''=n}^{n'} x_{i,n''} \hat{r}_{i,n''} - D_{i,n'} \geq 0, \quad \forall i,n', \\
& \quad x_{i,n'} \geq 0 \quad \forall i \in \mathcal{M}, n' \in \mathcal{N}_n.
\end{align*}
\]

The objective in Eq. 1 minimizes the total BS airtime consumed over the prediction window. The first constraint expresses the resource limitation at each base station, while the second constraint captures the smooth video streaming requirement. The first term of this constraint denotes the cumulative rate allocated to a user in previous time slots, based on actual (i.e., not predicted) channel conditions, and is fed back to the optimizer as shown in Fig. 3, with \( R_{i,0} = 0, \forall i \). This enables the long-term allocator to re-adjust the future allocation plan at every slot, and has not been considered in previous works [4, 23] which do not incorporate feedback or a sliding window. The second and third components of this constraint denote the cumulative allocation plan and cumulative demand \( \forall n' \in \mathcal{N}_n \), respectively. Note that the advantage of defining resource allocation with the airtime fraction variable \( x_{i,n} \) (as opposed to using discrete resource blocks) is that it enables a linear RA formulation without integer variables, thereby greatly simplifying the RA problem. Another important remark is that it is not necessary to formulate the long-term RA with an optimization problem; any algorithm that achieves similar objectives by exploiting a similar formulation has been considered in [4], but without the sliding window or the cumulative rate feedback \( R \).

the predicted rates \( \hat{r} \) can be developed within the P-RAN framework and applied as the upper level allocator.

4.3 Instantaneous Opportunistic Scheduling
For each prediction slot \( n \), scheduling is performed at the TTI level in the lower level scheduler. Here the reported user CQIs are used to opportunistically schedule users at peaks of their fast fading channel responses and avoid scheduling during deep fades. The target data requirement as determined by the upper level resource allocator is provided, and users are scheduled to meet this demand. Note that the number of bits \( x_{i,n}\hat{r}_{i,n} \) required during slot \( n \), and not the airtime fraction \( x_{i,n} \) is used as this stage. The reason is that if the actual channel rate differs from the predicted rate \( \hat{r}_{i,n} \), then the airtime calculated by the long-term RA will not be correct. By using the number of bits required, the lower level scheduler provides a degree of robustness to channel prediction errors.

Let \( r^\text{TTL}_{i,t} \) denote the instantaneous rate supported by user \( i \) at TTI \( t \), within prediction slot \( n \). We can define several scheduling rules, such as

\[
i^* = \arg \max_{i \in \mathcal{M}} r^\text{TTL}_{i,t},
\]

where \( \mathcal{M}_u \) is the set of users whose allocation requirement \( x_{i,n}\hat{r}_{i,n} \) remains unsatisfied. In this rule, the scheduler will choose the user with the higher rate, which is much like a Maximum-Rate scheduler. However, once a user has met its target for slot \( n \), it will be excluded from the set \( \mathcal{M}_u \). We refer to this framework implementation as P-RAN:MaxRate. Alternatively, the following approach, dubbed P-RAN: RelativeRatio selects the user with the higher instantaneous rate \( \hat{r}_i \) to its predicted rate:

\[
i^* = \arg \max_{i \in \mathcal{M}_u} \frac{r^\text{TTL}_{i,t}}{\hat{r}_{i,n}}.
\]

The intuition of this rule is to provide a more fair and distributed scheme that waits until users achieve their own relative peaks. Note that other more complex rules/scheduling algorithms can also be applied within the P-RAN framework to define the instantaneous lower level scheduler.

At the end of the prediction slot \( n \), the actual rates transmitted \( \hat{r}_{i,n} \) are determined, and fed back to the upper level RA. The long-term RA is then re-run to account for the actual rates transmitted in the previous slot.

4.4 NS-3 LTE Implementation
We implemented the P-RAN framework within the LTE module in the open source Network Simulator (ns-3) LTE module [30]. Our extension involved modifying existing methods, and introducing a new method to implement the upper level predictive resource allocator. An optimization solver was also integrated in ns-3 to solve the BS airtime minimization problem for stored video transmission defined in Eq. 1. It is important to note that this framework may directly be used to investigate a broader category of P-RAN formulations, network objectives, and lower-level scheduling metrics, with only minor modular modifications.

Simulation Environment: The definition of system nodes (i.e., UEs and eNB), the UE-eNB assignment and physical parameters such as transmit power and bandwidth are set through the existing LteHelper class. This class contains predefined methods that enable seamless integration
between the different network entities. It is also used to select the channel model, mobility model, and scheduler, and then apply their effect on the network nodes.

**Channel Model:** The `TraceFadingLossModel` class implemented in the ns-3 LTE module is used to generate the desired channel model. This class imports MATLAB generated tracing files that model the fast fading response of the channel in different environments according to the 3GPP standards [2]. The class was modified to include a log-normal shadowing effect with a specified variance that is added to the average received power every prediction slot $n$.

**Mobility Model:** The `LteHelper` supports different types of random and fixed velocity models. The `WaypointMobilityModel` class which imports a user trajectory file was chosen in our simulation. This choice enables a direct specification of the $x$ and $y$ user coordinates at different time instances. Localization error effects are then simply implemented by adding a Gaussian random variable to the user coordinates.

**Scheduler:** The P-RAN framework illustrated in Fig. 3 is implemented by inheriting the existing `RrfFfMacScheduler` class which implements a round-robin scheduler. The three main functions in Fig. 3 are executed by modifying two existing methods and introducing a new one as follows:

- **UpdateDRIReBufferInfo:** This method already exists in the `RrfFfMacScheduler` class to decrement the allocated user’s buffer data by one Transport Block (TB) size every TTI. We modify the function to implement the cumulative data computation used for feedback in Fig. 3. To do so, the cumulative data of the current user is increased by the same amount that is decremented from the buffer.

- **DoPredictSchedule:** This method is introduced to implement the upper level long-term RA optimizer in Fig. 3. It has three main inputs: the long-term predicted rates matrix $\hat{r}$, the users’ demands matrix $D$ and the cumulative received data $R$ calculated by the aforementioned method. This method uses the C++ interface of the Gurobi Optimizer [16] to solve the predictive allocation plan over a time horizon according to the formulation in Eq. 1. This method is executed every prediction slot $n$ or when a new user arrives. Note that to integrate Gurobi in ns-3, the Gurobi model libraries are imported from their installation directory in the `wscript` file associated with the project.

- **DoSchedDITiggerReq:** This method already exists in the inherited `RrfFfMacScheduler` class and is triggered every TTI to implement the instantaneous lower-level opportunistic scheduler. It consists of two main functions: user scheduling and buffer update. In user scheduling, a `DoPredictSchedule` method is first called to obtain the data requirement for each user during each slot $n$. The scheduling metrics defined in Eq. 2 and Eq. 3 are then implemented to select the user that is scheduled during each TTI. After scheduling, the buffer update function is implemented by calling the aforementioned `UpdateDRIReBufferInfo` method to determine when users achieve their target rates as specified from the `DoPredictSchedule` method. Note that the user CQIs are available in the `DoSchedDITiggerReq` method and can be used directly to compute the scheduling metrics.

Other minor modifications related to the video application were also made in accordance with the system setup. Specifically, the eNB buffer size (MaxTxBufferSize in the `LteRl-Um` class) was increased to store the incoming video content before transmission. The BER variable in the `LteAmpc` class was also modified to reflect the application requirements.

## 5. PERFORMANCE AND DISCUSSION

The section serves two purposes. The first is to define and conduct a set of robustness tests that investigate the performance of P-RANs with: 1) REM errors, 2) fast-fading (on the different lower-level scheduler designs), 3) localization errors, and 4) varying prediction window sizes. The second objective is to provide insight on how uncertainty can be efficiently modeled in a P-RAN based on the preceding results. We then highlight the need for robust long-term optimization techniques that incorporate rate uncertainty to further improve the practical performance of P-RANs.

**Simulation Setup:** The LTE network setup follows the model parameters in Table 1. The total number of users $M = 6$ with an arrival rate 0.2/s, and a simulation time of 60 s. Average BS airtime and Video Degradation (VD) are used as network efficiency and QoS performance metrics, respectively. Video degradation is defined as the total percentage of constraint violation, i.e. the second constraint in 1. To provide a performance reference of a non-predictive RA approach, we present the results of the popular Proportional Fair (PF) scheduler [18] as implemented in [30].

### 5.1 Numerical Results

Although the potential gains of predictive RA over traditional scheduling approaches have been reported in [4, 23], more abstract models and ideal assumptions were made. We therefore focus on investigating the effects of uncertainty with-in the proposed P-RAN framework.

**Sensitivity to REM Uncertainty:** Uncertainty in the REM will lead to errors in the predicted user rate $\hat{r}$, thereby affecting the performance of the predictive RA schemes. For instance, in cases where the actual rate is less than anticipated, users may not receive their target rates. If content was not previously buffered (e.g., in cases where the channel rate is forecasted to increase), then video stalling may occur.

We model REM uncertainty as a log-normal random variable with a variance $\sigma^2$ overlaid on the predicted SINR. In Fig. 5(a) the streaming rate is $V = 5.5$ Mbit/s which we refer to as the low load scenario. Here airtime is reduced by approximately 66% with the P-RAN framework variations compared to traditional PF scheduling, while $VD \approx 0$ for all schemes (which is not shown for space limitations). The

### Table 1: Summary of Model Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BS Transmit power</td>
<td>43dBm</td>
</tr>
<tr>
<td>BW</td>
<td>5 MHz</td>
</tr>
<tr>
<td>$T$</td>
<td>60s</td>
</tr>
<tr>
<td>$\tau$</td>
<td>1s</td>
</tr>
<tr>
<td>BER</td>
<td>$10^{-5}$</td>
</tr>
<tr>
<td>Velocity</td>
<td>30km/h</td>
</tr>
<tr>
<td>Packet size</td>
<td>$8 \times 10^5$ [bits]</td>
</tr>
<tr>
<td>Packet rate (from core network to BS)</td>
<td>$10^3$s$^{-1}$</td>
</tr>
<tr>
<td>Streaming rate V at different loads</td>
<td>$0.5, 1.5, 2.25 \times 10^6$ [bit/s]</td>
</tr>
<tr>
<td>Buffer size</td>
<td>$10^9$ [bits]</td>
</tr>
</tbody>
</table>
reason is the network’s ability to delay transmissions until users approach the cell center without violating streaming constraints. Interestingly, airtime increases only slightly even at a high prediction error variance, while VD remains close to zero. This is due to the feedback present at the lower-level scheduler that adjusts the planned user airtime to meet the data requirement at each slot in the presence of errors.

Fig. 5(b) and Fig. 5(c) show the results for a higher load scenario where $V = 1.5$ Mbit/s. In this case airtime savings are not large but video degradation is reduced as shown in Fig. 5(c). Note that the airtime of the P-RAN: RelativeRatio scheme is slightly less than the P-RAN: MaxRate scheme. We can see that at a variance $\sigma^2 = 8$ the gains of P-RAN diminish due to the very high variations between the predicted and actual rates.

In Fig. 6(a)-(c) we conduct similar performance tests but this time without fast fading. The reason is that the fast fading model gives rather optimistic results at high velocities, since throughput peaks become frequent even when users are far from the cell center (we assume the reported instantaneous CQI is accurate). The results show similar trends, but the airtime and VD performance of PF is affected since it can no longer exploit high throughput spikes at the cell edge (which is generally not possible in practice). On the other hand, the P-RAN framework does not rely on such peaks, and maintains its previous performance.

**Sensitivity to Localization Uncertainty:** Localization uncertainty is modeled as a zero mean Gaussian random variable with a variance $\sigma^2_l$ in both $x$ and $y$ dimensions (in meters). Fig. 7 shows the impact of localization errors with increasing variances. We observe that even at high variance the airtime increases slightly but the VD remains unaffected. The reason is that localization errors up to 50 m do not impact the received rates significantly due to the discrete MCSs. This applies to the considered highway scenario where the REM follows a gradual increase and decrease with the distance from the BS. However, for more complex scenarios where strong shadowing or tunnels exist, the REM may vary significantly over a small geographical region. If this is the case, the localization errors will affect P-RAN performance more.

**Effect of Prediction Window Size:** In Fig. 8 we investigate the effect of varying the prediction window size for both $M = 4$ and $M = 6$ with $V = 2.25$ Mbit/s. During the defined window size perfect channel knowledge (i.e. with no errors or fast fading) is assumed to isolate the effects of a smaller windows. The trends in Fig. 8(b) are expected, where a larger window leads to a gradual decrease in the VD. However, note that with only four users, a window of 30 s is sufficient to completely eliminate VD. Increasing the window size any further brings only a slight airtime reduc-
Figure 7: Sensitivity to localization error variance (a) Average BS airtime, and (b) Video degradation.

Figure 8: Effect of prediction window size (a) Average BS airtime, (b) Video degradation, and (c) Cumulative allocated rate for a sample user with P-RAN:MaxRate, $M = 6$.

5.2 Modeling REM Uncertainty

There is still a need to model rate prediction uncertainty itself to more accurately determine where, and how, P-RANs can be effectively applied. For example, a REM map may be more accurate in a rural area and less so in an urban region. Furthermore, the accuracy of the map may change with time due to fluctuations in network dynamics at rush hours vs. other times.

In order to define a REM uncertainty model, we first investigate the effect of different error variances $\sigma^2$ on the received data rates by determining the probabilities of occurrence as shown in Fig. 8(a) for $M = 4$.

The case for $M = 6$ is more complex for increasing window sizes. While VD also decreases with increasing windows, the performance of P-RAN:RelativeRatio is considerably worse than P-RAN:MaxRate. The reason is that at this high load it is not possible to completely eliminate VD, and thus the problem in Eq. 1 is no longer feasible. In this case, a linear relaxation of constraints is made to find a solution that minimizes the error by which the constraint’s boundary is violated. As it is more spectrally efficient to schedule competing users based on their MaxRate as opposed to the RelativeRatio metric, which is equal to 1 for all the users in this case (since fast fading is not modeled as previously highlighted). The P-RAN:RelativeRatio therefore selects a user at random, leading to lower spectral efficiency. An interesting observation also appears in Fig. 8(a) where airtime decreases and then increases for P-RAN:MaxRate. The initial decrease in airtime results from a more efficient RA due to the larger window. However, increasing the window further enables the network to foresee that a higher load is anticipated in the future. This knowledge is used to trigger content prebuffering allocations to users that are in poor channel conditions. We depict this behavior in Fig. 8(c) which demonstrates how for a window size of 60 s content prebuffering is planned well in advance. While this consumes more airtime, the VD decreases as shown in Fig. 8(b). The conclusion is that varying the window size can have changing dynamics that are specific to the modeled application and metrics of interest, and should therefore be investigated thoroughly to avoid instability and undesired behaviors.

In summary, the preceding results demonstrated that significant gains are achievable even when only trends in the predicted information are available, i.e., in the presence of significant errors and limited lookahead time. However, for other road network scenarios and applications the specific P-RAN setup should be tested for robustness and compared to baseline schedulers to assess the achievable gains under practical considerations.
currence of different TB sizes. The results are illustrated in Fig. 9, where a step structure appears due to the discrete MCSs in LTE, resulting in specific TB sizes. Note that as some TBs correspond to larger SINR ranges, there are irregularities in the step structure. Fig. 9 also shows that if the TB was originally predicted to be very small (as in Fig. 9(a)) or very large (as in Fig. 9(c)), then the structure becomes asymmetric, particularly for higher error variances. Since this step structure may be difficult to express mathematically, we observe that it is possible to approximate it with a triangle as shown in Fig. 10. Here, the right \( r_u \) and left \( r_l \) most points on the x-axis define the limits of the triangle’s base, which physically represent the boundaries on the variation of the predicted rate \( \hat{r} \). In fuzzy number representation, the fuzzy predicted rate \( \hat{r} \) is a triangular membership function expressed mathematically as [26]:

\[
\mu_{\hat{r}} = \begin{cases} 
L(\hat{r}) = \frac{\hat{r} - r_l}{r_u - r_l} + 1, & \text{if } r_l \leq \hat{r} \leq \hat{r} \\
R(\hat{r}) = \frac{\hat{r} - r_u}{r_u - r_l} + 1, & \text{if } \hat{r} \leq \hat{r} \leq r_u \\
0, & \text{otherwise.}
\end{cases}
\] (4)

While other representations are surely possible, this model provides a simple way of expressing probabilities of different TB sizes for different degrees of rate uncertainty.

Robust Resource Allocation: Modeling REM uncertainty enables robust P-RAN design. For instance, to guarantee that the user streaming constraint does not get violated, the lowest rate value \( r_l \) can be used to solve Eq. 1, instead of \( \hat{r} \). Conversely, a rate value greater than \( \hat{r} \) may be assumed for users/services that have lower QoS priorities. Therefore, knowing the degree of REM uncertainty is useful to determine the slope of the triangular membership function. This may be learned in real-time by comparing the predicted and observed rates, and then applying estimation techniques such as the Kalman Filter (KF) [27]. A more detailed review on the use of fuzzy numbers in network design is available in [22, 26], and can be directly applied to develop a robust P-RAN which we leave for future work. Alternatively, stochastic models and optimization techniques can also be investigated. However, for practical implementation real-time constraints should also be considering to avoid complex theoretical solutions.

6. CONCLUSIONS

Exploiting mobility-awareness to devise predictive radio access schemes has been recently investigated with promising results [3, 5, 23, 25]. This is driven by numerous human mobility studies and analyses of bandwidth traces demonstrating that a user’s future channel states are highly reproducible [15, 17, 33]. While these studies indicate the significant potential of P-RANs, abstract implementations and ideal conditions were mostly considered for specific application scenarios.

In this paper, our goal was to introduce a more practical design and evaluation approach towards a unified P-RAN framework that can be easily adapted for different applications and network objectives, and integrated into traditional LTE schedulers. This was accomplished by a two-stage RA approach with feedback that can jointly leverage rate predictions and instantaneous opportunistic scheduling. We also implemented the framework within the Network Simulator ns-3 LTE module and demonstrated how a broad category of predictive schemes can be enabled through the framework. Extensive sensitivity tests indicated that even trends in predictive information provide considerable QoS and network efficiency gains. While these promising results validate the use of P-RANs, we foresee that robust prediction models and RA schemes are needed to exploit their potential even further.

7. REFERENCES

[1] 3GPP. LTE; evolved universal terrestrial radio access (E-UTRA); physical layer procedures. Technical