UE Blocking Probability Model for Planning 5G Guaranteed Bit Rate (GBR) RAN Slices

Oscar Adamuz-Hinojosa, Pablo Ameigeiras, Pablo Muñoz, and Juan M. Lopez-Soler

***Abstract*—Modeling the probability of blocking User Equipment (UE) sessions is key for planning in advance the amount of radio resources required by a Guaranteed Bit Rate (GBR) slice. This task is challenging since the amount of these resources depends on factors such as the channel quality of each UE, the packet scheduler discipline, and the GBR requirements. Under this context, we propose an analytical model to evaluate the UE blocking probability for a GBR slice. A key aspect of the proposed model is considering as input the distribution of the average Signal-to-Interference-plus-Noise Ratio (SINR) experienced by the UE, therefore allowing to capture the radio conditions in the cell. Our model builds a multi-dimensional Erlang-B system which meets the reversibility property. This involves our model is insensitive to the holding time distribution for the UE sessions. The reversibility property also involves the state transition probabilities have product form, so that the computational complexity of our model is low. Furthermore, the proposed model captures the SINR gain provided when the base station implements a channel-aware packet scheduler to allocate radio resources to the UE sessions. We validate the proposed model, demonstrating an estimation error for the UE blocking probability below 1.5%.**

***Index Terms*—Blocking probability, GBR service, Erlang-B, Channel-aware scheduler, RAN slicing.**

I. Introduction

The emergence of Fifth Generation (5G) mobile networks will boost a wide variety of unprecedented communication services with stringent requirements in terms of performance and functionalities [1]. Considering each communication service separately and building a Radio Access Network (RAN) accordingly would be unfeasible in terms of cost. RAN slicing is a technological solution to economically provide separate communication services over a common wireless infrastructure [2]. It consists of the provision of multiple logical networks, denominated RAN slices, each adapted to the requirements of a specific communication service.

One of the main standard developing organizations on network slicing is the Global System for Mobile Communications Association (GSMA). It has defined the Generic Network Slice Template (GST), which allows the Mobile Network Operator (MNO) to unambiguously interpret the Quality of Service (QoS) requirements of a RAN slice and negotiate the Service Level Agreement (SLA) accordingly [3]. In this paper, we focus on RAN slices which rely on data transmissions with strict Guaranteed Bit Rate (GBR) requirements. In this aspect, one of the main GST's attributes is the uplink/downlink throughput per User Equipment (UE). This parameter specifies the target GBR for each active uplink/downlink data session. Other key QoS attribute is the service availability, which represents the percentage of time the end-to-end communication service is delivered. To measure this attribute, one of the Key Performance Indicators (KPIs) is the service request success rate. This KPI represents the probability an UE data session request is successful. For a GBR service, an UE data session request can be admitted only if the MNO can ensure such session an average data rate equal to the target GBR. Otherwise, the data session is blocked.

To fulfill the SLA, the MNO must guarantee: *i*) the probability of blocking an UE session, and *ii*) the throughput per UE. To ensure these values before deploying the RAN slice, the MNO must rely on the RAN slice planning, which is a key management task executed in the RAN slice preparation phase [4]. This task mainly consists of deciding in advance the feasibility of deploying requested RAN slices and the adequate parameter configuration of the RAN to accommodate them [5]. The parameter configuration includes, among other things, the amount of guaranteed radio resources, computing resources, and networking resources for each RAN slice. In this work, we specifically concentrate on the radio interface because of the limited spectrum which the MNO has available, and in particular we focus on the downlink.

Designing a strategy for planning RAN slices with GBR requirements is challenging. For each RAN slice, the MNO must consider the spatio-temporal variations in the number of its active sessions as well as the bandwidth consumption of each session. The later depends on factors such as (a) the UE channel quality; (b) the scheduler[1] discipline considered by each serving cell in runtime; and (c) the target GBR. Furthermore, the MNO must guarantee the target UE blocking probability is below a certain threshold. Therefore, it is crucial for the MNO to rely on a mathematical model which considers all the previous aspects to accurately planning the amount of

This work was supported in part by the H2020 Research and Innovation Project Beyond 5G Multi-Tenant Private Networks Integrating Cellular, Wi-Fi, and LiFi, Powered by Artificial Intelligence and Intent Based Policy (5G-CLARITY) under Grant 871428, in part by the Spanish Ministry of Science and Innovation / State Investigation Agency under Project PID2019-108713RB-C53, and in part by the Spanish Ministry of Economic Affairs and Digital Transformation under Project TSI-063000-2021-28.

The authors are with the Department of Signal Theory, Telematics and Communications, University of Granada, Granada, Spain (e-mail: {oadamuz, pameigeiras, pabloml, juanma}@ugr.es).

[1]In our paper, the terms *schedule* and *plan* differ. The former is related to the runtime radio resource allocation in each Transmission Time Interval (TTI), i.e., in the order of one millisecond or hundreds of microseconds. The latter is related to the estimation (before deploying the RAN slice) of the number of required radio resources throughout its lifetime.

reserved radio resources for each RAN slice. In this way, the MNO would avoid the under-provisioning of radio resources, i.e., its UE blocking probability for a RAN slice would be much higher than the considered threshold, and vice-versa.

A. Related Works

The existing literature for modeling the UE blocking probability in a cell is vast. In [6], the authors provide an excellent model which is based on a multi-class processor sharing queue for a Code Division Multiple Access (CDMA)-High Data Rate cell. Since the processor sharing discipline is insensitive to the holding time distribution, arbitrary distributions can be adopted for the UE session duration. This discipline also forces an equal distribution of radio resources among the UEs, therefore this model is applicable to Round Robin and Proportional Fair (PF) scheduling disciplines. Considering these disciplines involve those UE sessions with better channel conditions could achieve a data rate equal or above the GBR, whereas those UE sessions with worse channel conditions could be rejected. This means this model is not appropriate for providing a GBR service to any UE regardless its channel conditions. In [7], this work was extended by including intra-cell UE mobility. However this improvement involves loosing the insensitivity property, thus only the exponential distribution is valid for the UE session duration.

In both works, the authors build their model based on two scenarios: (a) a single cell in isolation, and (b) multiple cells following regular topologies. In the first scenario, the authors consider the intra-cell interference is constant over the considered cell. In the second scenario, the authors consider the cells are placed either equidistantly in an infinite line or following an hexagonal distribution. These assumptions easily enable their model to capture an estimation of the average channel quality within the considered cell by using a discrete number of concentric rings. However, these assumptions limit the accuracy for modeling the distribution of the average channel quality. For instance, two UEs located at the same distance from an access node could not perceive the same channel quality. The reason is there could be different obstacles and geographical features between each UE and the serving access node, involving a different impact of the channel effects such as the shadowing.

Due to the simplicity of modeling the average channel quality by using concentric rings, other authors have also considered this approach in their models for the UE blocking probability (e.g., [8]–[13]). In [8], the authors have proposed a Markov chain-based model to compute the UE blocking probability. The main novelty of this model is the consideration of UE mobility within a cell based on Wideband CDMA. Using this model, the paper proposes an admission control mechanism for UE sessions supporting voice and data calls with a minimum data rate. In [9], the authors use a ring-based model for modeling the coverage area of a base station which implements Orthogonal Frequency-Division Multiple Access (OFDMA). In addition to the UE mobility, this model also considers the effect of handovers from adjacent cells. Based on that, the model considers the equilibrium balance equations for the UEs in each ring to compute the UE blocking probability for a service with a constant data rate. In [10], the authors propose an analytical model to capture the time-varying capacity of an OFDMA cell. Based on a Markov chain, this model also considers concentric rings to capture the data rate achieved by each UE. Furthermore, their model captures the UE mobility within the cell by assuming an exponential distribution for moving an UE from one ring to another. Using this model, the authors design an admission control mechanism which considers the UE blocking probability. In [11], the authors propose a Quality of Service-oriented resource allocation strategy for streaming flows that require a constant bit rate in a WiMaX cell. The proposed strategy considers an analytical model to derive the UE blocking probability. This model is based on a Markov chain which assumes a set of concentric rings to model the channel quality within the cell. Additionally, all the previous models present valuable contributions which have been applied by others authors which go beyond traditional mobile broadband services. For instance, the authors of [12], [13] adapt their models to IPTV services which requires GBR requirements.

Despite these works contribute significantly to the state-of-the-art solutions for modeling the UE blocking probability for GBR services, they present some drawbacks as the assumption of an exponential distribution for the UE session duration and/or the limited modeling of the UE channel quality within the cell. These restrictions reduces the accuracy of their models to estimate the UE blocking probability.

In an attempt to consider a more precise characterization of the average channel quality within the cell, the authors of [14] consider several zones distinguished by the strength of the received signal. Specifically, they use a combination of indicators such as the Reference Signal Received Power, the Reference Signal Received Quality, the Received Signal Strength Indication and the Signal-to-Interference-plus-Noise Ratio (SINR) of each UE to compute the UE blocking probability in a scenario with intra-cell UE mobility. These indicators are taken as input for a Markov chain-based model. Despite this improvement, the authors assume a reduction of the data rate for each UE when the total required bandwidth exceed the available bandwidth in the cell, thus this model considers the GBR is not met for all the UEs. Furthermore, this work assume the UE session duration follows an exponential distribution.

Finally, the state-of-the-art proposals which model the UE blocking probability for GBR services do not consider the impact on the channel gain when a packet scheduler dynamically allocates radio resources to the active UE sessions. In the literature, there exists a wide range of channel-aware strategies to schedule radio resources for GBR services. We recommend the readers to review the comprehensive survey presented in [15], where the most representative scheduling strategies for GBR services are analyzed.

B. Contributions

In this article, we assume a single 5G-New Radio (NR) OFDMA cell implements a RAN slice which provides an

arbitrary communication service with GBR requirements and without strict requirements in terms of latency. This means (*i*) the cell only admits those data sessions for which it can guarantee an specified average data rate and (*ii*) the remaining data sessions are rejected. We also consider the UEs served by this RAN slice generate data sessions following a Poisson distribution. It is well known the sum of a large number of UEs generating data sessions, each UE with an arbitrary distribution of renewal time, will tend to a Poisson process [16]. In addition, we consider the cell implements a channel-aware scheduler to allocate radio resources in runtime to the admitted data sessions. Under this context, the main contributions of this article are:

- The proposal of an analytical model to predict the UE blocking probability in a cell for a RAN slice based on the amount of available radio resources for that slice. This model is based on a Multi-dimensional Erlang-B system. It meets the reversibility property which means the proposed model allows the adoption of an arbitrary distribution for the UE session duration. Additionally, this property involves the solution for the state probabilities has product form, thus it reduces the complexity for their computation. Furthermore, the proposed model may consider as input any distribution and geometry for the average SINR which an arbitrary UE session could perceive within the cell. Unlike the state-of-the-art proposals, this approach allows a more precise characterization of the UEs' channel quality. For example, our model may use as input a probabilistic distribution of the average SINR obtained by either experimental measurements in a real cell or simulation.
- The mathematical formulation to capture the behavior of a channel-aware scheduler in the proposed model for the UE blocking probability. In this work, we take a key step forward with respect to the first version of our model [17], where we considered the scheduler was channel-agnostic. To that end, the mathematical formulation in this work captures the channel gain perceived by a UE session when the serving cell implements a channel-aware scheduler to assign UEs the radio resources which provide it a better instantaneous SINR. Specifically, this formulation relates the average data rate by an admitted UE session and the probability of allocating radio resources for such session. Additionally, using this formulation, we can establish the conditions which must be met to admit an UE data session. For comprehensibility purposes, we have considered a representative channel-aware scheduler based on the alpha-fair metric [18]. However, this does not detract the proposed model from considering any other channel-aware scheduler by following the proposed formulation.

In the provided results, we first validate the proposed multi-dimensional Erlang-B model by means of simulation, demonstrating that it exhibits an estimation error for the UE blocking probability below 1.5%. Then, based on the proposed model, we show the benefits of using a channel-aware scheduler to reduce the UE blocking probability.

The remainder of this article is organized as follow. Section II describes the system model. In Section III, we present the proposed UE blocking probability model. Then, we describe how the channel-aware scheduler meets the GBR requirements of each admitted UE session in Section IV. In Section V, we define the experimental setup, and based on that, we validate the proposed model and evaluate the impact of using a channel-aware scheduler on the UE blocking probability. Finally, Section VI draws the main conclusions.

II. System Model

In this work, we focus on the downlink operation of a single OFDMA cell. It implements a RAN slice providing a GBR service to their UEs, which dynamically request and release data sessions. This cell also implements a channel-aware scheduler which considers the channel quality perceived by the active UE sessions to dynamically allocate them radio resources. Based on this scenario, we first describe the model for the radio resources in a OFDMA cell. Then, we define the channel model. Later, we present the characteristics of the offered traffic. Finally, we define the characteristic of the channel-aware scheduler.

A. Radio Resource Model

We assume a serving OFDMA cell $i \in \mathcal{I}$ with a total bandwidth W_i. This bandwidth is divided into N OFDM sub-carriers. In turn, these sub-carriers are arranged in groups of N_{SC} sub-carriers. Each group of sub-carriers defines a Resource Block (RB), which is the smallest unit of resources that can be allocated to a single UE. The number of available RBs during a time slot is given by Eq. (1). The parameter Δf is the bandwidth between sub-carriers whereas OH denotes the overhead factor due to control plane data.

$$N_i^{RB} = \left\lfloor \frac{W_i}{N_{SC}\Delta f}(1 - OH) \right\rfloor \tag{1}$$

If the cell i employs a small sub-carrier spacing Δf and a large bandwidth W_i, the number of available RBs in a time slot could be too high. For instance, in a 5G New Radio (5G-NR) cell, the maximum number of available RBs could be 273 units [19] [20]. From the perspective of radio resource allocation, it becomes advantageous to reduce the management complexity by grouping the RBs into resource chunks, which are allocated to the UEs as indivisible units [21]. This can be done through the concept of Resource Block Group (RBG) defined in [22]. A RBG is a collection of consecutive RBs that can be allocated to a specific UE. The size of the RBG N_{size}^{RBG} (i.e., number of consecutive RBs) can be used for establishing the minimum allocation unit size. Increasing N_{size}^{RBG} may serve to reduce the signaling overhead at the expense of a loss of flexibility. Based on that, we can compute the available RBGs on a time slot in the cell i as $N_i^{RBG} = \left\lfloor N_i^{RB}/N_{size}^{RBG} \right\rfloor$.

B. Channel Model

In this work, we adopt the SINR as the metric to measure the channel quality within the cell i. Specifically, we define in Eq.

(2) the instantaneous SINR $\gamma_{u,n}(m)$ for the UE u in the RBG n and the time slot m [23]. The parameter $X_{i,u,n}(m)$ denotes the fading component which is random and varies for each time slot m and RBG n. We adopt the well-known Rayleigh-fading model for the fading component $X_{i,u,n}(m)$, leading to an exponential distribution with unit mean, i.e., $\mathrm{E}[X_{i,u,n}] = 1$. The parameters $\overline{P}_{i,u,n}$ and $\overline{P}_{j,u,n}$ denote the average received powers (i.e., without considering fast-fading) from the serving cell $i \in \mathcal{I}$ and the neighbor cell $j \in \mathcal{I}\backslash\{i\}$, respectively. Finally, P_N is the noise power measured in one RBG.

$$\gamma_{u,n}(m) = \frac{\overline{P}_{i,u,n} X_{i,u,n}(m)}{\sum_{j\in\mathcal{I}\backslash\{i\}} \overline{P}_{j,u,n} X_{j,u,n}(m) + P_N} \tag{2}$$

The average received power $\overline{P}_{i,u,n}$ from cell i depends on the path-loss $\overline{h}^{PL}_{i,u}$, the shadowing $\overline{h}^{SH}_{i,u}$ and the transmission power per RBG $P_{i,n}$, i.e., $\overline{P}_{i,u,n} = \overline{h}^{PL}_{i,u} \cdot \overline{h}^{SH}_{i,u} \cdot P_{i,n}$. We consider $\overline{h}^{PL}_{i,u}$ and $\overline{h}^{SH}_{i,u}$ remain constant throughout the duration of the UE data session. It is valid in scenarios where the UEs have reduced mobility within the cell (e.g., semi-static people in live events such as sport events or concerts, IoT sensors, or equipment for industry 4.0). Furthermore, we consider the cell transmits the same constant power for each RBG, i.e., $P_{i,n} = P_{i,m}\ \forall n \in \mathcal{N}^{RBG}, \forall m \in \mathcal{N}^{RBG}$. Hereinafter, we omit the subscript n for the transmitted power and the average received power, i.e., P_i instead of $P_{i,n}$; and $\overline{P}_{i,u}$ instead of $\overline{P}_{i,u,n}$.

To make the analysis mathematically tractable, we omit the fast-fading component in each interference term from Eq. (2), i.e., $\sum_{j\in\mathcal{I}\backslash\{i\}} \overline{P}_{j,u,n} X_{j,u,n} \approx \sum_{j\in\mathcal{I}\backslash\{i\}} \overline{P}_{j,u,n}$. This means the sum of the interference terms can be regarded as an additional source of noise [23]. Under such assumption, the instantaneous SINR $\gamma_{u,n}(m)$ follows an exponential distribution with average $\overline{\gamma}_u$. This results in the simplified Probability Density Function (PDF) $f_{\gamma_{u,n}}(\gamma)$ and Cumulative Distribution Function (CDF) $F_{\gamma_{u,n}}(\gamma)$ defined in Eqs. (3) and (4), respectively. The average SINR $\overline{\gamma}_u$ is given by Eq. (5). Since the average SINR does not depend on the RBG n, we omit the subscript n for such parameter in the remainder of this paper, i.e., $\overline{\gamma}_u$ instead of $\overline{\gamma}_{u,n}$.

$$f_{\gamma_{u,n}}[\gamma] = \frac{1}{\overline{\gamma}_u} \exp\left[\frac{-\gamma}{\overline{\gamma}_u}\right] \tag{3}$$

$$F_{\gamma_{u,n}}[\gamma] = 1 - \exp\left[\frac{-\gamma}{\overline{\gamma}_u}\right] \tag{4}$$

$$\begin{aligned}\overline{\gamma}_u &= \mathrm{E}\left[\frac{\overline{P}_{i,u} X_{i,u,n}}{\sum_{j\in\mathcal{I}\backslash\{i\}} \overline{P}_{j,u} + P_N}\right] \\ &= \frac{\overline{P}_{i,u}\mathrm{E}[X_{i,u,n}]}{\sum_{j\in\mathcal{I}\backslash\{i\}} \overline{P}_{j,u} + P_N} = \frac{\overline{P}_{i,u}}{\sum_{j\in\mathcal{I}\backslash\{i\}} \overline{P}_{j,u} + P_N}\end{aligned} \tag{5}$$

If $\overline{\gamma}_u$ is measured for a considerable amount of active UEs, it is possible to derive the PDF for the average SINR $f_{\overline{\gamma}_u}(\overline{\gamma})$ to model the average channel quality within the cell. Let us define the Geometry factor (G-factor) G_f as Eq. (6) shows [24], where the ensemble averages over short-term fading, but not shadowing. The G-factor allows to characterize the radio environment in the cell, including aspects such as the cell geometry or the propagation conditions [25]. The G-factor has been used as a tool to predict cell performance indicators such as cell capacity [26].

$$\begin{aligned}G_f &= \frac{\mathrm{E}\left[\overline{P}_{i,u} X_{i,u,n}\right]}{\mathrm{E}\left[\sum_{j\in\mathcal{I}\backslash\{i\}} \overline{P}_{j,u} X_{j,u,n} + P_N\right]} \\ &= \frac{\overline{P}_{i,u}\mathrm{E}[X_{i,u,n}]}{\sum_{j\in\mathcal{I}\backslash\{i\}} \overline{P}_{j,u}\mathrm{E}[X_{j,u,n}] + P_N} \\ &= \frac{\overline{P}_{i,u}}{\sum_{j\in\mathcal{I}\backslash\{i\}} \overline{P}_{j,u} + P_N}\end{aligned} \tag{6}$$

Under the assumption above that models other cell interference as an additional source of additive white gaussian noise, it can be seen that the average SINR $\overline{\gamma}_u$ defined in Eq. (5) equals the G-factor defined in Eq. (6) [27]. Consequently, we follow the same approach, and use the PDF of $\overline{\gamma}_u$ to characterize the geometry and radio environment in the cell.

Since the PDF $f_{\overline{\gamma}_u}(\overline{\gamma})$ is a continuous function, we split it into N_Z regions to make it tractable. Depicted in Fig. 1, each region z is defined as the set of values for the average SINR such as $\overline{\gamma} \in \left[\overline{\gamma}^{(z-1)}, \overline{\gamma}^{z}\right)$. For simplicity, we assume the session of an active UE takes place in one of these N_Z regions with probability π_z, which is provided by Eq. (7). Note that $\sum_{z=1}^{N_Z} \pi_z = 1$. Furthermore, we assume the UE sessions which take place in a region $z \in \mathcal{Z}$ have an average SINR $\overline{\gamma}_z$, which is defined in Eq. (8).

$$\pi_z = \int_{\overline{\gamma}^{(z-1)}}^{\overline{\gamma}^{z}} f_{\overline{\gamma}_u}[\overline{\gamma}]\, d\overline{\gamma} \tag{7}$$

$$\overline{\gamma}_z = \frac{1}{\pi_z} \int_{\overline{\gamma}^{(z-1)}}^{\overline{\gamma}^{z}} \overline{\gamma} f_{\overline{\gamma}_u}[\overline{\gamma}]\, d\overline{\gamma} \tag{8}$$

Note that the range of values $\overline{\gamma} \in \left[\overline{\gamma}^{(z-1)}, \overline{\gamma}^{z}\right)$ in the region z does not necessarily have to be mapped to the values of the average SINR perceived in a specific geographical region within the cell (e.g., concentric rings in [8]–[13]).

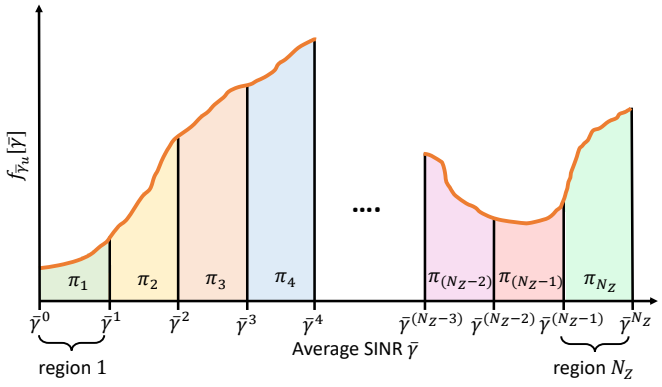


Fig. 1: Splitting the PDF for the average SINR $f_{\overline{\gamma}_u}[\overline{\gamma}]$ into N_Z regions

C. Traffic Model

To model the traffic demands within a cell, we consider the statistical distributions and the average values for the arrival rate of UE sessions and the session duration.

For the generation process of UE sessions, we assume a Poisson arrival process with an average of λ UE session requests per unit time. It is well known that in many cases, the sum of a large number of independent stationary renewal processes (i.e., in our scenario, each individual UE generating data sessions), each with an arbitrary distribution of renewal time, will tend to a Poisson process [16]. Since a Poisson process can be split into N_Z independent process [28], we can also express the average arrival rate for each region z as $\lambda_z = \lambda \pi_z$. Note that $\lambda = \sum_{z=1}^{N_Z} \lambda_z$.

With respect to the session duration $t_u^{session}$ for each UE u, we assume a random variable extracted from an arbitrary distribution. Additionally, we define $\mu = 1/E\left[t_u^{session}\right]$ as the average rate for releasing UE sessions per unit time. Note that the release session rate is independent from the region z.

Considering the tuple (λ, μ) we can define the offered traffic intensity as $\rho = \lambda/\mu$. Since the UE sessions must get a guaranteed average data rate D_{GBR}, the cell only admits those UE sessions which can achieve such average data rate in their lifetimes. This means an UE session can be rejected with a certain probability B. In Section III, we propose a model to derive the blocking probability B. Furthermore, we provide in Section IV the conditions that must be met for which a UE session can get the average data rate D_{GBR} and thus, it can be admitted.

D. Channel-Aware Scheduler

In this paper, we focus on modeling the UE blocking probability B for a RAN slice with GBR requirements. For that purpose, we assume the serving cell implements a channel-aware scheduler to allocate radio resources to the admitted UE sessions. The analysis of which channel-aware scheduler provides optimum performance for such RAN slice would deserve further investigation and therefore it is beyond the scope of this work. For this reason, we consider a representative channel-aware scheduler as the one depicted in Fig. 2. In the first step, the scheduler computes in each time slot m the metric $\hat{\gamma}_{u,n}$ defined in Eq. (9) for each UE $u \in \mathcal{U}$ in each RBG $n \in \mathcal{N}^{RBG}$. In Eq. (9), $\hat{\gamma}_{u,n}$ is based on the metric proposed in [29] and analyzed in [18]. For simplicity, we specifically consider in this metric the SINR instead of the data rate for an UE u. The fairness factor α is an adjustable parameter that controls the fairness of the RBG allocation. In this work, for simplicity reasons we assume that α is known and constant. The parameter $\tilde{\gamma}_u$ denotes the temporal average SINR for the UE u. To obtain $\tilde{\gamma}_u$, the instantaneous SINR $\gamma_{u,n}(m)$ is averaged in the last M time slots and over the cell bandwidth as Eq. (10) shows. In this work, we approximate the temporal average SINR $\tilde{\gamma}_u$ by the average SINR $\overline{\gamma}_{z_u}$ defined in Eq. (8), i.e., $\tilde{\gamma}_u \approx \overline{\gamma}_{z_u}$. The parameter z_u defines the region where the data session of the UE u was born.

$$\hat{\gamma}_{u,n} = \frac{\gamma_{u,n}}{(\tilde{\gamma}_u)^{\alpha}} \tag{9}$$

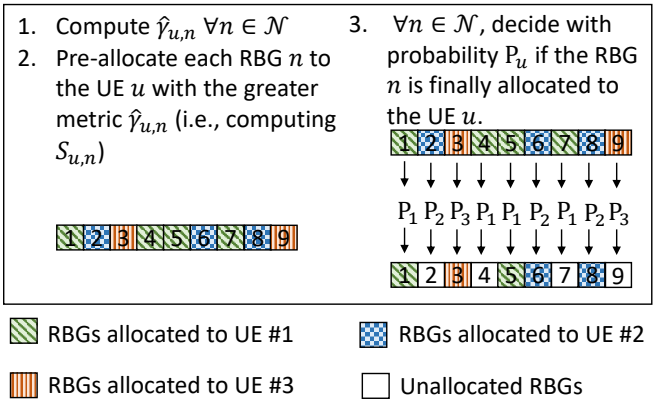


Fig. 2: Tasks performed by the channel-aware scheduler during the time slot m. For simplicity, we show an illustrative example with three UEs and nine RBGs.

$$\tilde{\gamma}_u = \frac{1}{M \cdot N_i^{RB}} \sum_{n=1}^{N_i^{RB}} \sum_{m=\tau-M}^{\tau-1} \gamma_{u,n}(m) \tag{10}$$

In this work we consider all the UE sessions require a GBR equal to D_{GBR}. However, the channel condition for each UE session is different. This means each UE session needs a specific amount of RBGs to meet its GBR requirements. Note that if the MNO sets $\alpha = 1.0$, the scheduler would implement the PF criteria. It would involve all the UE sessions have the same probability of being allocated with the same number of RBGs [30]. In this case, those UEs which perceived a low average SINR would have a lower data rate than those UEs which perceived a higher average SINR. To avoid that, the MNO could set $\alpha > 1.0$ with the goal of increasing the amount of allocated RBGs to those UEs which perceive a lower average SINR. In Section V, we show the benefits of considering this approach.

Based on $\hat{\gamma}_{u,n}$, the scheduler decides in the second step to pre-allocate each RBG n to the UE u with the greatest metric $\hat{\gamma}_{u,n}$ as Eq. (11) shows[2]. In Eq. (11), $S_{u,n}$ is a binary variable which defines with a value 1 if the RBG n is pre-allocated to the UE u and 0 otherwise. Note that $\sum_{u \in \mathcal{U}} S_{u,n} = 1$, i.e., the RBG n is pre-allocated to one UE. Furthermore, $\sum_{u \in \mathcal{U}} \sum_{n=1}^{N_i^{RBG}} S_{u,n} = N_i^{RBG}$, i.e., all the available RBGs in the cell are pre-allocated to the set of UEs $\mathcal{U}$.

$$S_{u,n} = \begin{cases} 1 & \hat{\gamma}_{u,n} \geq \max\limits_{\forall v \in \mathcal{U} \setminus \{u\}} \{\hat{\gamma}_{v,n}\} \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

Considering this pre-allocation, the admitted UE sessions achieve an average data rate equal or greater than D_{GBR}. Since we focus on a scenario where all the admitted UE sessions only get the GBR D_{GBR}, the channel-aware scheduler must limit the amount of RBGs for those UE sessions with an achievable data rate above the GBR. For this reason, this

[2]Once the RAN slice is deployed, the scheduler only needs to compute in real time the metric given in Eq. (9) and apply Eqs. (10)-(11) to derive the radio resource allocation for all its active UEs. Note that it is a slightly modification of the well known Proportional Fair discipline. Specifically, they only differ in the exponentiation operation $(\tilde{\gamma}_u)^{\alpha}$.

scheduler decides in the third step which percentage P_u of RBGs are finally allocated for each UE u. Based on that, we define the probability $\mathrm{P}[T_{u,n}=1]$ that the RBG n is finally allocated to the UE u in Eq. (12). In this equation, the parameter $T_{u,n}$ is a binary variable which takes the value 1 if the RBG n is finally allocated to the UE u and 0 otherwise. $\mathrm{P}[S_{u,n}=1]$ denotes the probability that the RBG n is pre-allocated in the second step to the UE u. Note that P_u and $\mathrm{P}[S_{u,n}=1]$ are independent. Furthermore $\sum_{u\in\mathcal{U}} T_{u,n} \leq 1$, i.e., the RBG n could be (or not) allocated to one UE. In addition, $\sum_{u\in\mathcal{U}}\sum_{n=1}^{N_i^{RBG}} T_{u,n} \leq N_i^{RBG}$, i.e., all the available RBs may not be scheduled in a time slot.

$$\mathrm{P}[T_{u,n}=1] = \mathrm{P}_u \mathrm{P}[S_{u,n}=1] \tag{12}$$

To derive $\mathrm{P}[S_{u,n}=1]$, we perform the operations described in Eq. (13). The resulting expression depends on (a) the PDF of the instantaneous SINR for the UE u; and (b) the CDFs of the instantaneous SINR for the remaining UEs [31].

$$\begin{aligned}\mathrm{P}[S_{u,n}=1] &= \mathrm{P}\left[\hat{\gamma}_{u,n} \geq \max_{\forall v\in\mathcal{U}\setminus\{u\}} \{\hat{\gamma}_{v,n}\} \,|\gamma_{u,n}=\gamma\right] \\ &= \int_0^\infty f_{\hat{\gamma}_{u,n}}[\gamma] \cdot \prod_{\forall v\in\mathcal{U}\setminus\{u\}} F_{\hat{\gamma}_{v,n}}[\gamma]\, d\gamma \\ &= \int_0^\infty \left(\overline{\gamma}_{z_u}\right)^\alpha f_{\gamma_{u,n}}\left[\left(\overline{\gamma}_{z_u}\right)^\alpha \cdot \gamma\right] \\ &\cdot \prod_{\forall v\in\mathcal{U}\setminus\{u\}} F_{\gamma_{v,n}}\left[\left(\overline{\gamma}_{z_v}\right)^\alpha \cdot \gamma\right] d\gamma\end{aligned} \tag{13}$$

Since all the UE sessions are distributed into N_z regions where each region gathers U_z UE sessions (i.e., $U_1 + U_2 + \ldots + U_{N_z} = |\mathcal{U}|$), we can rewrite $\mathrm{P}[S_{u,n}=1]$ as Eq. (14) shows.

$$\begin{aligned}\mathrm{P}[S_{u,n}=1] = \int_0^\infty &\left(\overline{\gamma}_{z_u}\right)^\alpha f_{\gamma_{u,n}}\left[\left(\overline{\gamma}_{z_u}\right)^\alpha \cdot \gamma\right] \\ &\cdot \left(F_{\gamma_{1,n}}\left[\left(\overline{\gamma}_1\right)^\alpha \cdot \gamma\right]\right)^{U_1} \\ &\cdot \ldots \cdot \left(F_{\gamma_{u,n}}\left[\left(\overline{\gamma}_{z_u}\right)^\alpha \cdot \gamma\right]\right)^{U_z-1} \\ &\cdot \ldots \cdot \left(F_{\gamma_{N_z,n}}\left[\left(\overline{\gamma}_{N_z}\right)^\alpha \cdot \gamma\right]\right)^{U_{N_z}} d\gamma\end{aligned} \tag{14}$$

Considering the PDF and the CDF of the instantaneous SINR defined in Eqs. (3) and (4), we can rewrite $\mathrm{P}[S_{u,n}=1]$ as Eq. (15) shows. The definite integral described in this equation has an analytical solution which depends on the specific value of the fairness factor α and the number of active UE sessions in each region (i.e., U_1, U_2, ..., U_{N_z}). If these values are known, such integral can be computed as described in Appendix A. Note that we have removed the subscript n since the PDF and the CDFs do not depend on the RBG n. Furthermore, we have replaced the subscript u with z as the computed probability is the same for all the UEs whose data session has been born in the region z. This also involves that P_u can be redefined as P_z and thus, the Eq. (12) can be redefined as $\mathrm{P}[T_z=1] = \mathrm{P}_z\mathrm{P}[S_z=1]$.

$$\begin{aligned}\mathrm{P}[S_z=1] = \int_0^\infty &\left(\overline{\gamma}_z\right)^{\alpha-1} \exp\left[-\left(\overline{\gamma}_z\right)^{\alpha-1}\cdot\gamma\right] \\ &\cdot \left(1-\exp\left[-\left(\overline{\gamma}_1\right)^{\alpha-1}\cdot\gamma\right]\right)^{U_1} \\ &\cdot \ldots \cdot \left(1-\exp\left[-\left(\overline{\gamma}_z\right)^{\alpha-1}\cdot\gamma\right]\right)^{U_z-1} \\ &\cdot \ldots \cdot \left(1-\exp\left[-\left(\overline{\gamma}_{N_z}\right)^{\alpha-1}\cdot\gamma\right]\right)^{U_{N_z}} d\gamma\end{aligned} \tag{15}$$

We need to compute the average number of RBGs for each UE u in a time slot. Since we assume each UE session is born in a specific region z within the cell, the average number of RBGs allocated for a single UE depends on the scheduling probability $\mathrm{P}[T_z=1]$. In Eq. (16), we see $\mathrm{P}[T_z=1]$ depends on the state $s=(U_1,U_2,\ldots,U_{N_Z}) \in \mathcal{S}'$. This state defines the total amount of admitted UE sessions as well as their distributions along the N_z regions. $\mathcal{S}'$ denotes the set of states of the system.

$$N_{z,s}^{RBG} = \left\lceil \mathrm{P}[T_z=1]\, N_i^{RBG} \right\rceil \tag{16}$$

Finally, the average number of RBGs obtained by a single UE session which is born in region z must satisfy Eq. (17). In this equation T_{slots}^u denotes the number of time slots throughout the session duration.

$$N_{z,s}^{RBG} = \frac{1}{T_{slots}^u} \sum_{t=1}^{T_{slots}^u} \sum_{n=1}^{N_i^{RBG}} T_{u,n}^{(t)} \;\; \forall u \in \mathcal{U}^z \tag{17}$$

III. UE Blocking Probability and Capacity model of an OFDMA cell

This section explains the proposed model, including the methodology to derive the UE blocking probability, the average RBG utilization, and the cell capacity for a RAN slice with GBR requirements.

A. Multi-dimensional Erlang-B model

Let us consider an OFDMA cell where the values of $f_{\overline{\gamma}_u}[\overline{\gamma}]$ for a single RAN slice are grouped into N_Z regions. To model this system, we employ a multi-dimensional Erlang-B model. In this model, we assume each UE session takes place into one region z, defined by the tuple (λ_z, μ). The offered traffic intensity in each region becomes $\rho_z = \lambda_z/\mu$, and the total offered traffic intensity is $\rho = \sum_{z=1}^{N_Z} \rho_z$.

To define the set of potential states of the system $\mathcal{S}$, we take into account (a) an active UE session in the region z consumes $N_{z,s}^{RBG}$ RBGs on average; and (b) the available RBGs in the cell are limited by N_i^{RBG}. These statements are gathered by Eq. (18), which provides the necessary condition to define a state $s \in \mathcal{S}$.

$$N_i^{RBG} - \sum_{z=1}^{N_Z} U_z N_{z,s}^{RBG} \geq 0 \;\; \forall s \in \mathcal{S} \tag{18}$$

Focusing on a single state $s \in \mathcal{S}$, it could happen that the GBR requirements could not be met for the UE sessions which fall within one or more regions, i.e., their average data rates

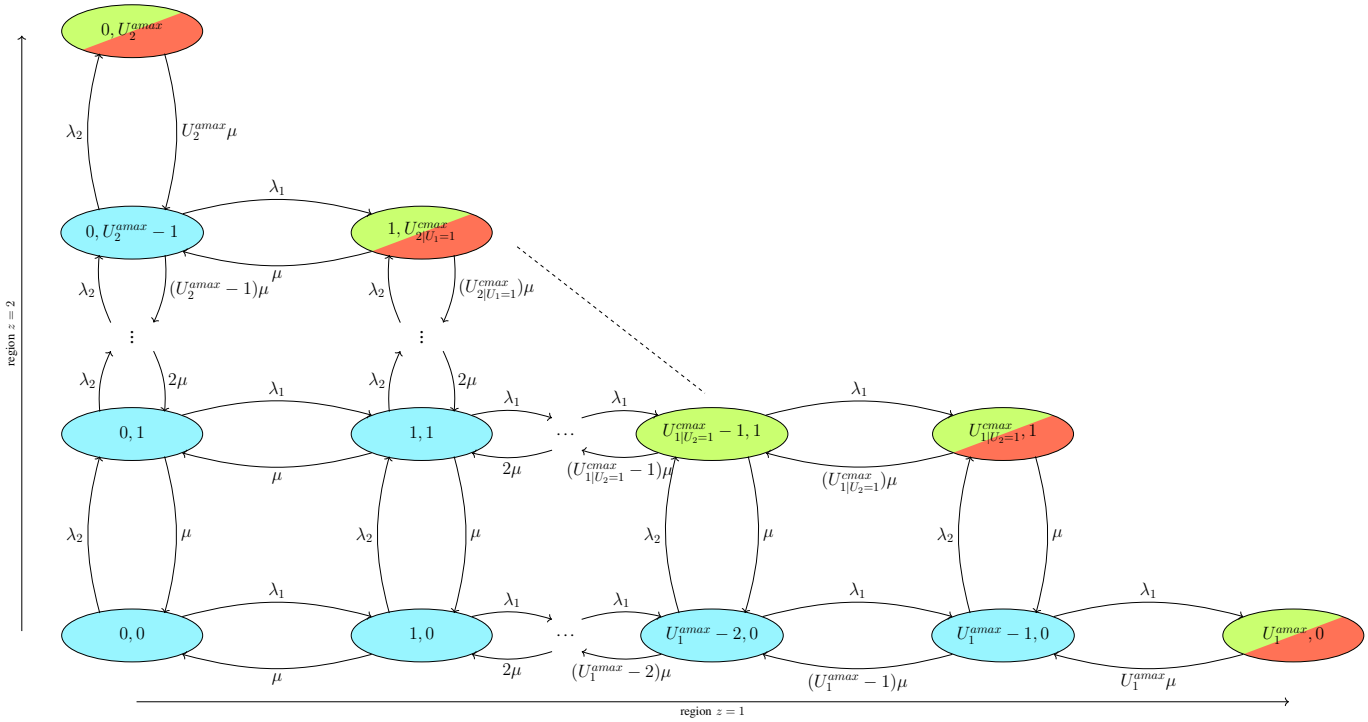


Fig. 3: State transition diagram for the proposed N_z-dimensional Erlang-B system. For comprehensibility purposes, only two regions are represented, i.e., $z = 1$ and $z = 2$. Note that red and green states correspond to $U_1 = U^{cmax}_{1|U_2}$ and $U_2 = U^{cmax}_{2|U_1}$, respectively.

could be less than D_{GBR}. This means the state s is not valid and thus, one or more UE sessions should be rejected. For this reason, we define the set $\mathcal{S}' = \mathcal{S} \setminus \mathcal{S}^{nf}$ of feasible states. $\mathcal{S}^{nf}$ denotes those states belonging to $\mathcal{S}$ where at least one UE session does not meet the GBR requirements. In Section IV, we provide details about determining a valid state $s \in \mathcal{S}'$, i.e., that one in which all the admitted UE sessions meet the GBR requirements.

Considering the set $\mathcal{S}'$ of feasible states, we can build the state transition diagram as Fig. 3 shows. Note that, for simplicity, the represented diagram only shows two dimensions, corresponding to the regions $z = 1$ and $z = 2$. In this diagram, $U^{cmax}_{z|U_y}$ denotes the maximum number of UE sessions in the region z conditioned to the number of UE sessions in the remaining regions (e.g., red states for region 1, and green states for region 2). This means that considering the state $s^{(t)} = \left(U_1, U_2, \ldots, U^{cmax}_{z|U_y}, \ldots, U_{N_z}\right) \in \mathcal{S}'\ \forall y \in \mathcal{Z}\setminus\{z\}$ implies that the state $s^{(t+1)} = \left(U_1, U_2, \ldots, U^{cmax}_{z|U_y} + 1, \ldots, U_{N_z}\right) \notin \mathcal{S}'\ \forall y \in \mathcal{Z} \setminus \{z\}$. If the remaining regions have 0 UEs, we can define the absolute maximum number of UE sessions U_z^{amax} in region z. This means that considering the state $s^{(t)} = (0, 0, \ldots, U_z^{amax}, \ldots, 0) \in \mathcal{S}'$ implies that the state $s^{(t+1)} = (0, 0, \ldots, U_z^{amax} + 1, \ldots, 0) \notin \mathcal{S}'$.

The resulting multi-dimensional Erlang-B system corresponds to a reversible Markov process (see proof in appendix B). This implies the proposed model is insensitive to the distribution of the UE session duration, which means the state probabilities depend only upon the mean service time [28]. Furthermore, the solution for the probability of being in a state $s \in \mathcal{S}'$, i.e., $p_s\left[U_1, U_2, \ldots, U_{N_Z}\right]$, has product form as Eq. (19) shows. In this equation, $p\left[U_z\right]$ is the one-dimensional truncated Poisson distribution for traffic stream in region z and K is a normalization constant.

$$\begin{aligned} p_s\left[U_1, U_2, \ldots, U_{N_Z}\right] =& K \cdot p\left[U_1\right] \cdot p\left[U_2\right] \cdot \ldots \cdot p\left[U_{N_Z}\right] \\ =& K \cdot \prod_{z=1}^{N_Z} \frac{\rho_z^{U_z}}{U_z!} \end{aligned} \tag{19}$$

To obtain the state probabilities, we need to derive K. This normalization constant can be computed by summing all the state probabilities (i.e., without being normalized) and equaling the resulting expression to 1 as Eq. (20) shows.

$$K = \frac{1}{\sum_{s\in\mathcal{S}'}\left(\prod_{z=1}^{N_Z}\frac{\rho_z^{U_z}}{U_z!}\right)} \tag{20}$$

B. UE Blocking Probability

Assuming a new UE session is born in region z, it will be blocked if there not exists a transition from the current state $s^{(t)} = (U_1, U_2, \ldots, U_z, \ldots, U_{N_Z})$ to $s^{(t+1)} = (U_1, U_2, \ldots, U_z + 1, \ldots, U_{N_Z})$. This happens when $U_z + 1 > U^{cmax}_{z|U_y}$. We define $\mathcal{S}_z^B$ as the set of states where a transition is not possible in the region z. Based on that, we can compute the UE blocking probability B_z conditioned to the region z where the new UE session is born by Eq. (21).

$$B_z = \sum_{s\in\mathcal{S}_z^B} p_s\left[U_1, U_2, \ldots, U^{cmax}_{z|U_y}, \ldots, U_{N_z}\right] \tag{21}$$

Finally, the UE blocking probability B in the cell is defined in Eq. (22). This is computed as the sum of the conditional

blocking probabilities weighted by the probability of a UE session is born in each region z.

$$B = \sum_{z=1}^{N_Z} \pi_z B_z \tag{22}$$

C. Mean number of consumed RBGs and Cell Capacity

Two key parameters derived from our model are the mean number of RBGs consumed in a cell $\overline{N}_{RBG}$, and the cell capacity D_i for a RAN slice with GBR requirements.

The mean number of RBGs $\overline{N}_{RBG}$ can be computed as Eq. (23) defines. In this equation, we multiply the probability of being in a state $s \in \mathcal{S}'$ by the amount of consumed RBGs in this state.

$$\overline{N}_{RBG} = \sum_{s \in \mathcal{S}'} p_s \left[U_1, U_2, \ldots, U_z, \ldots, U_{N_z}\right] \sum_{z=1}^{N_z} U_z N_{z,s}^{RBG} \tag{23}$$

The cell capacity D_i is provided by Eq. (24). It is derived as the product of the mean number of UEs $\overline{U}_z$ in each region z multiplied by the data rate consumed by each UE. We can compute $\overline{U}_z$ using Little's theorem [28], i.e., $\overline{U}_z = (\lambda_z/\mu)(1 - B_z) = \rho_z (1 - B_z)$.

$$D_i = \sum_{z=1}^{N_Z} \overline{U}_z D_{GBR} \tag{24}$$

IV. UE Throughput with a Channel-Aware Scheduler

Considering the state $s = (U_1, U_2, \ldots, U_z, \ldots, U_{N_z}) \in \mathcal{S}$, we compute in this section the expected throughput for a UE session in each region z and thus determining if the state s is feasible (i.e., $s \in \mathcal{S}'$) or not (i.e., $s \in \mathcal{S}^{nf}$).

Based on the channel-aware scheduler described in Section II-D, we define the expected throughput for the UE session u as Eq. (25) shows [23]. In this equation, $f_{\gamma_{u,n}|T_{u,n}=1}[\gamma]$ denotes the PDF of the instantaneous SINR $\gamma_{u,n}$ assuming the RBG n is allocated for the UE u. Furthermore, $C[\gamma]$ is a function which provides the data rate achieved by an UE when it perceives an instantaneous SINR γ.

$$R_u = \sum_{n=1}^{N_i^{RBG}} \int_0^{\infty} C[\gamma] \, f_{\gamma_{u,n}|T_{u,n}=1}[\gamma] \, \mathrm{P}[T_{u,n}=1] \, d\gamma \tag{25}$$

In Eq. (25), we can define $\mathrm{P}[T_{u,n}=1]$ in function of $\mathrm{P}[S_{u,n}=1]$ and P_u. Furthermore, since $S_{u,n} = 1$ when $T_{u,n} = 1$, we can replace $f_{\gamma_{u,n}|T_{u,n}=1}[\gamma]$ with $f_{\gamma_{u,n}|S_{u,n}=1}[\gamma]$. Considering these changes, we can rewrite the expected throughput R_u as

$$R_u = \sum_{n=1}^{N_i^{RBG}} \int_0^{\infty} C[\gamma] \, f_{\gamma_{u,n}|S_{u,n}=1}[\gamma] \, \mathrm{P}[S_{u,n}=1] \, \mathrm{P}_u d\gamma \tag{26}$$

We can consider the Bayes' theorem, i.e., $f_{\gamma_{u,n}|S_{u,n}=1}[\gamma] = \frac{f_{S_{u,n}=1|\gamma_{u,n}}[\gamma] f_{\gamma_{u,n}}[\gamma]}{\mathrm{P}[S_{u,n}=1]}$. Furthermore, with the aim of maintaining the block error rate for the UE's data below a certain threshold, the cell adopts a link adaptation technique. This technique enables the cell to adapt the UEs' Modulation and Coding Scheme (MCS) according to the experienced channel effects. In our work, we consider a total of N_c MCSs. Hence, the range of the instantaneous SINR is split into N_c intervals $[\gamma_i, \ \gamma_{i+1}]$. In each interval, the achieved spectral efficiency takes a specific value c_i. Considering these statements, we rewrite the expected throughput as Eq. (27) defines. Note that P_u is out of the integral since it does not depend on the instantaneous SINR.

$$\begin{aligned} R_u =& \mathrm{P}_u N_{SC} \Delta_f N_{size}^{RBG} \\ & \cdot \sum_{n=1}^{N_i^{RBG}} \sum_{i=1}^{N_c} c_i \int_{\gamma_i}^{\gamma_{i+1}} f_{S_{u,n}=1|\gamma_{u,n}}[\gamma] \, f_{\gamma_{u,n}}[\gamma] \, d\gamma \end{aligned} \tag{27}$$

To compute the PDF of the allocation of the RBG n for UE u under the assumption of an instantaneous SINR $\gamma_{u,n}$, i.e., $f_{S_{u,n}=1|\gamma_{u,n}}[\gamma]$, we make use of the scheduling criteria defined in Eq. (9). Specifically, we perform the steps described in Eq. (28) [31]. Note that $\mathcal{U}^s$ denotes the set of UE sessions considered in state $s = (U_1, U_2, \ldots, U_z, \ldots, U_{N_z}) \in \mathcal{S}$.

$$\begin{aligned} f_{S_{u,n}=1|\gamma_{u,n}}[\gamma] &= \mathrm{P}\left[\hat{\gamma}_{u,n} \geq \max_{\forall v \in \mathcal{U}^s \setminus \{u\}} \{\hat{\gamma}_{v,n}\} \, | \gamma_{u,n} = \gamma\right] \\ &= \mathrm{P}\left[\frac{\gamma}{\left(\overline{\gamma}_{z_u}\right)^{\alpha}} \geq \max_{\forall v \in \mathcal{U}^s \setminus \{u\}} \{\hat{\gamma}_{v,n}\}\right] \\ &= \prod_{\forall v \in \mathcal{U}^s \setminus \{u\}} F_{\hat{\gamma}_{v,n}}\left[\frac{\gamma}{\left(\overline{\gamma}_{z_u}\right)^{\alpha}}\right] \\ &= \prod_{\forall v \in \mathcal{U}^s \setminus \{u\}} F_{\gamma_{v,n}}\left[\frac{\left(\overline{\gamma}_{z_v}\right)^{\alpha} \gamma}{\left(\overline{\gamma}_{z_u}\right)^{\alpha}}\right] \end{aligned} \tag{28}$$

If we include the result of Eq. (28) in Eq. (27), we obtain the expected throughput as

$$\begin{aligned} R_u =& \mathrm{P}_u N_{SC} \Delta_f N_{size}^{RBG} \sum_{n=1}^{N_i^{RBG}} \sum_{i=1}^{N_c} c_i \int_{\gamma_i}^{\gamma_{i+1}} f_{\gamma_{u,n}}[\gamma] \\ & \cdot \prod_{\forall v \in \mathcal{U}^s \setminus \{u\}} F_{\gamma_{v,n}}\left[\frac{\left(\overline{\gamma}_{z_v}\right)^{\alpha} \gamma}{\left(\overline{\gamma}_{z_u}\right)^{\alpha}}\right] d\gamma \end{aligned} \tag{29}$$

Since the set $\mathcal{U}^s$ of UE sessions is split into N_z regions, we define in Eq. (30) the expected throughput R_z for a UE session which is born in the region z. Note that we have also replaced P_u with P_z since this probability is the same for all the UE sessions which are born in the same region z.

$$\begin{aligned} R_z =& \mathrm{P}_z N_{SC} \Delta_f N_{size}^{RBG} \sum_{n=1}^{N_i^{RBG}} \sum_{i=1}^{N_c} c_i \int_{\gamma_i}^{\gamma_{i+1}} f_{\gamma_{u,n}}[\gamma] \\ & \cdot \left(F_{\gamma_{1,n}}\left[\frac{(\overline{\gamma}_1)^{\alpha}\gamma}{(\overline{\gamma}_z)^{\alpha}}\right]\right)^{U_1} \cdot \ldots \cdot \left(F_{\gamma_{z,n}}[\gamma]\right)^{U_z - 1} \\ & \cdot \ldots \cdot \left(F_{\gamma_{N_z,n}}\left[\frac{\left(\overline{\gamma}_{N_z}\right)^{\alpha}\gamma}{(\overline{\gamma}_z)^{\alpha}}\right]\right)^{U_{N_z}} d\gamma \end{aligned} \tag{30}$$

If we consider the exponential distribution for the instantaneous SINR as shown in Eqs. (3) and (4), we can rewrite the expected throughput R_z for each region z as Eq. (31) defines.

Since the expected throughput R_z must be the same for each region z, we have $R_z = D_{GBR}$. Furthermore, we can denote as $I_z\left(\alpha, \gamma_i, \gamma_{i+1}, U_1, U_2, \ldots, U_{N_z}\right)$ the definite integral in Eq. (31). This integral has an analytical solution which depends on the value for the fairness factor α and the set of UEs $\mathcal{U}^s$ defined by the state $s = (U_1, U_2, \ldots, U_z, \ldots, U_{N_z})$. When α and $\mathcal{U}^s$ are known, it can be computed as Appendix A shows.

$$\begin{aligned} R_z = \mathrm{P}_z N_{SC} \Delta_f N_{size}^{RBG} N_i^{RBG} \sum_{i=1}^{N_c} c_i \int_{\gamma_i}^{\gamma_{i+1}} \frac{1}{\overline{\gamma}_z} \exp\left[\frac{-\gamma}{\overline{\gamma}_z}\right] \\ \cdot \left(1 - \exp\left[-\frac{(\overline{\gamma}_1)^{\alpha-1}}{(\overline{\gamma}_z)^{\alpha}}\gamma\right]\right)^{U_1} \\ \cdot \ldots \cdot \left(1 - \exp\left[-\frac{1}{\overline{\gamma}_z}\gamma\right]\right)^{U_z - 1} \\ \cdot \ldots \cdot \left(1 - \exp\left[-\frac{\left(\overline{\gamma}_{N_z}\right)^{\alpha-1}}{(\overline{\gamma}_z)^{\alpha}}\gamma\right]\right)^{U_{N_z}} d\gamma \end{aligned} \tag{31}$$

If we consider Eq. (31) for each region z, we can define a system of N_z equations with N_z unknown variables, i.e., $\mathrm{P}_1, \mathrm{P}_2, \ldots, \mathrm{P}_{N_z}$, as Eq. (32) shows. This equation system must be solved for each potential state $s \in \mathcal{S}$ (see Section III-A). If the equation system has solution, the potential state will be valid, i.e., $s \in \mathcal{S}'$, i.e., all the admitted UE sessions in the state s will get an average data rate equal to the GBR. And on the contrary, if the equation system for a state s does not have a solution, this state will not be valid, i.e., $s \notin \mathcal{S}'$. Assuming α is known and constant, the values for the probabilities P_z $\forall z \in \mathcal{Z}$ can be directly derived from Eq. (32). A potential state s is feasible only if the values of all the probabilities are equal or lower than 1.

Next, we evaluate the precision of the UE blocking probability and compare it with simulation results. Additionally, we set different values of the fairness factor α and evaluate the impact on the UE blocking probability, demonstrating the suitability of the proposed model as a planning tool for RAN slicing.

Note that the fairness factor α may be optimized if it is not considered a constant value, e.g., considering a different value for each region z and/or state s. However, such optimization problem is out of the scope of this paper.

V. Numerical Results

The state-of-the-art models for computing the UE blocking probability (see Section I-A) are not appropriate for RAN slices with GBR requirements under our assumptions. Specifically, our model may consider as input any distribution and geometry for the average SINR which an arbitrary UE session could perceive within the cell. Additionally, our model can support arbitrary distributions for the duration of the UE sessions. These differences hinder a fair comparison between the state-of-the-art models and our proposal. For this reason, in this paper we experimentally validate the proposed model by means of simulation. We also evaluate the UE blocking probability for different configurations of the channel-aware scheduler described in Section II-D.

Specifically, we first evaluate the relative error for the UE blocking probability with respect to the one obtained by simulation. Then, we evaluate UE blocking probability, the number of active UE sessions, and the radio resource utilization when the MNO sets different values for the fairness factor α considered by the channel-aware scheduler. These performance indicators are also evaluated in a baseline scenario where the cell implements a channel-agnostic scheduler as the one we assumed in our previous work [17]. Finally, we analyze the aspects that impact the execution time of our model.

A. Experimental Setup

To validate the proposed model, we use a Matlab-based simulator that resembles the arrival and departure of UE sessions for a RAN slice with GBR requirements in a single cell. This simulator generates UE sessions following a Poisson distribution. With respect to the UE session duration, we have carried out all the experiments considering an exponential distribution, an uniform distribution, and a constant duration. For all the cases, the results are equal since our model is insensitive to the holding time distribution. Focusing on a single UE session, the simulator considers (a) the region z where the session takes place and (b) the average number of allocated RBGs $N_{z,s}^{RBG}$ for such session. To determine if a new UE session can be admitted, the simulator first considers s as the state where the new UE session is admitted. Then, it checks if this state belongs to the set of feasible states, i.e., $s \in \mathcal{S}'$. If true, the new UE session is admitted. Table I summarizes the configuration parameters.

Regarding the access technology, we assume a 5G-NR cell implementing an OFDMA scheme with $\Delta f = 15$ KHz, and $N_{SC} = 12$. We also consider different scenarios where the serving cell allocates 20, 25, 30, 35, 40 and 45 RBs for the RAN slice. Additionally, we consider each UE session consumes multiples of 2 RBs, i.e., the RBG size $N_{size}^{RBG} = 2$. With respect to $f_{\overline{\gamma}_u}(\overline{\gamma})$, we have derived it by using the distribution of the G-factor experimentally measured in a macro cell [26]. Additionally, we consider different values for the number of regions for such distribution, from $N_z = 4$ to $N_z = 9$. For the GBR service provided by the RAN slice, we assume a data rate of $D_{GBR} = 0.8$ Mbps for each active UE session. We consider a low data rate since we assume the UE sessions which are born in the region z' with the worst average SINR $\overline{\gamma}_{z'}$ must require an average number of RBGs less than the number of allocated RBGs for the RAN slice, i.e., $N_{z',s}^{RBG} \leq N_i^{RBG}$ $\forall s \in \mathcal{S}'$. With respect to the channel-aware scheduler, we consider the constant values 1.0, 1.3, 1.5 and 1.7 for the fairness factor α.

Based on these configuration parameters, we have evaluated the UE blocking probability in function of the offered traffic intensity $\rho \in (0.5, 2)$.

All the experiments have been carried out on a computer with 16 GB RAM and an Intel core i7-7700HQ @ 2.80 GHz.

B. Model Validation

To validate our model, we have computed the relative error as $\epsilon_r(\%) = \frac{B_{sim} - B_{mod}}{B_{sim}} \cdot 100$, where B_{sim} and B_{mod} denote

$$\begin{cases} \mathrm{P}_1 N_{SC} \Delta_f N_{size}^{RBG} N_i^{RBG} \sum_{i=1}^{N_c} c_i I_1\left(\alpha, \gamma_i, \gamma_{i+1}, U_1, U_2, \ldots, U_{N_z}\right) & = & D_{GBR} \\ \mathrm{P}_2 N_{SC} \Delta_f N_{size}^{RBG} N_i^{RBG} \sum_{i=1}^{N_c} c_i I_2\left(\alpha, \gamma_i, \gamma_{i+1}, U_1, U_2, \ldots, U_{N_z}\right) & = & D_{GBR} \\ & \vdots & \\ \mathrm{P}_{N_z} N_{SC} \Delta_f N_{size}^{RBG} N_i^{RBG} \sum_{i=1}^{N_c} c_i I_{N_z}\left(\alpha, \gamma_i, \gamma_{i+1}, U_1, U_2, \ldots, U_{N_z}\right) & = & D_{GBR} \end{cases} \quad (32)$$

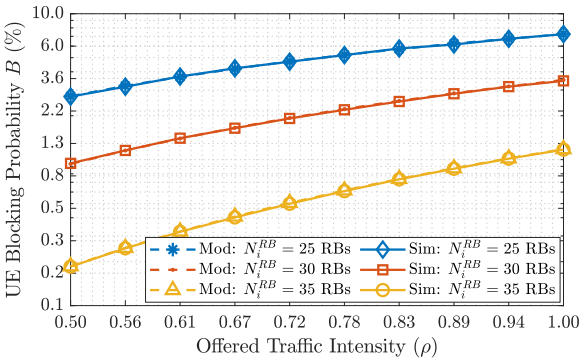
TABLE I: Configuration Parameters

Parameters	Configuration	Parameters	Configuration
Access Technology	5G-NR	Regions for the average SINR, i.e., $[\overline{\gamma}^{z-1}, \overline{\gamma}^{z})$ (in dB)	N_z = 4: [-5, 1), [1, 7), [7, 13), [13, 19) N_z = 5: [-5, -0.2), [-0.2, 4.6), [4.6, 9.4), [9.4, 14.2), [14.2 19) N_z = 6: [-5, -1), [-1, 3), [3, 7), [7, 11), [11 15), [15 19) N_z = 7: [-5, -1.574), [-1.574, 1.857), [1.857, 5.286), [5.286, 8.714), [8.714 12.143), [12.143, 15.571), [15.571, 19) N_z = 8: [-5, -2), [-2, 1), [1, 4), [4, 7), [7 10), [10, 13), [13, 16), [16, 19) N_z = 9: [-5, -2.333), [-2.333, 0.333), [0.333, 3), [3, 5.667), [5.667 8.333), [8.333, 11), [11, 13.667), [13.667, 16.333), [16.333, 19)
Subcarrier spacing Δf (OFDMA)	15 KHz	Average SINR $\overline{\gamma}_z$ (in dB) and probability π_z in each region, i.e., $(\overline{\gamma}_z, \pi_z)$	N_z = 4: [-0.907, 0.280), [3.636, 0.372), [9.496, 0.223), [15.157, 0.125) N_z = 5: [-1.601, 0.186), [1.925, 0.337), [6.776, 0.232), [11.514, 0.151), [15.944, 0.095) N_z = 6: [-2.112, 0.129), [0.944, 0.296), [4.813, 0.227), [8.907, 0.162), [12.832, 0.112), [16.3364, 0.074) N_z = 7: [-2.449, 0.094), [0.159, 0.253), [3.411, 0.217), [6.860, 0.163), [10.336 0.121), [13.757, 0.094). [16.505, 0.058) N_z = 8: [-2.981, 0.070), [-0.402, 0.210), [2.374, 0.210), [5.402, 0.162), [8.402 0.122), [11.374, 0.101). [14.598, 0.080), [16.617, 0.046) N_z = 9: [-3.037, 0.053), [-0.879, 0.179), [1.589, 0.193), [4.252, 0.159), [6.888, 0.126), [9.383, 0.104), [12.271, 0.077), [15.0187, 0.071), [16.897, 0.038)
Sub-carriers per RB N_{SC} (OFDMA)	12	Service GBR D_{GBR}	0.8 Mbps
Number of allocated RBs N_i^{RB}	20 RBs, 25 RBs, 30 RBs, 35 RBs, 40 RBs and 45 RBs	Distribution for the UE session arrival	Poisson
RBG Size N_{size}^{RBG}	2 RBs	Distribution for the UE session duration	Exponential, Uniform and Constant
Fast-fading distribution	Rayleigh with unit mean	Offered Traffic Intensity ρ	$\rho \in (0.5, 2)$
PDF average SINR in the cell: $f_{\overline{\gamma}_u}(\overline{\gamma})$	Built using the distribution of the G-factor measured in a macro cell [26]	Fairness Factor α	1.0, 1.3, 1.5 and 1.7

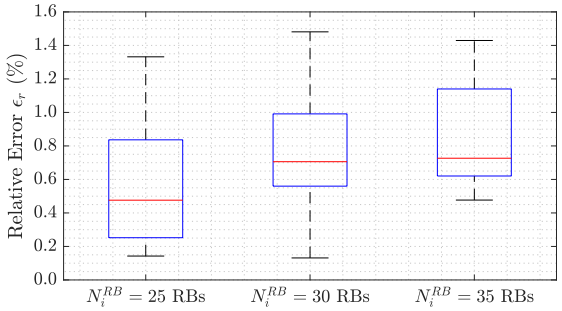
the UE blocking probability extracted from the proposed model and simulator, respectively. Aiming to get a reliable measure of B_{sim}: *i*) we have simulated the generation of approximately one million of UE data sessions[3]; and *ii*) we have repeated four times each simulation, being B_{sim} the mean of the UE blocking probabilities measured in these simulations. Additionally, we have considered three scenarios with a specific RB allocation for a RAN slice: 25 RBs, 30 RBs and 35 RBs. In all the scenarios, the number of regions considered for $f_{\overline{\gamma}_u}(\overline{\gamma})$ is $N_z = 6$.

In Fig. 4(a), we depict the curves for the UE blocking probability derived from our model and the simulator. The simulator curves are also provided with the 95 percent confidence intervals, which are gathered in Table II. Note that the interval widths is not perceptible in the simulation curves. Analyzing this figure, we observe how the UE blocking probability increases when (a) the available RBs for a RAN slice are decreased and (b) the offered traffic intensity increases. This graph is useful for MNOs to plan in advance the bandwidth of each cell (i.e., N_i^{RB}) for RAN slices with GBR requirements while a threshold for B is provided, given certain conditions for the offered traffic intensity and the channel quality, i.e., a specific $f_{\overline{\gamma}_u}(\overline{\gamma})$.

Due to the scale used for the vertical and horizontal axes in Fig. 4(a), the error between the simulation and the model cannot be observed. In Fig. 4(b), we represent the relative error by a box-and-whisker plot. A specific pair of box and whiskers



(a) UE Blocking Probability: Model vs Simulation



(b) Relative error

Fig. 4: Evaluation of the UE Blocking Probability for different cell bandwidths.

gathers the relative errors obtained for different values of the

[3]Note that one million of UE data sessions is much higher than the number of feasible states, which is 5888 at most in our experimental setup (i.e., when $N_i^{RB} = 35$ RBs are available in the cell). It allows us to accurately estimate, by means of simulation, the occurrence of each state (i.e., the probability of being in each state).

TABLE II: Reliability Analysis of Simulation Points

	$NR_i^{RB} = 25$ RBs		$NR_i^{RB} = 30$ RBs		$NR_i^{RB} = 35$ RBs	
ρ	Mean Value	95 % Confidence Interval	Mean Value	95 % Confidence Interval	Mean Value	95 % Confidence Interval
0.500	2.711	[2.683 2.738]	0.935	[0.915 0.955]	0.189	[0.183 0.195]
0.556	3.201	[3.173 3.229]	1.169	[1.151 1.182]	0.249	[0.236 0.262]
0.611	3.694	[3.671 3.718]	1.405	[1.391 1.419]	0.326	[0.311 0.340]
0.667	4.199	[4.157 4.240]	1.655	[1.620 1.691]	0.406	[0.396 0.416]
0.722	4.704	[4.649 4.760]	1.924	[1.887 1.961]	0.507	[0.496 0.518]
0.778	5.212	[5.184 5.239]	2.226	[2.201 2.251]	0.613	[0.591 0.636]
0.833	5.711	[5.679 5.743]	2.543	[2.531 2.555]	0.739	[0.729 0.749]
0.889	6.227	[6.184 6.270]	2.847	[2.822 2.871]	0.881	[0.866 0.896]
0.944	6.728	[6.687 6.764]	3.178	[3.156 3.200]	1.019	[1.006 1.032]
1.000	7.225	[7.197 7.253]	3.525	[3.479 3.570]	1.184	[1.166 1.202]

offered traffic intensity ρ when the cell has allocated a specific number of RBs. Specifically, the bottom and the top of the box represent the first and third quartiles for the measured relative errors, respectively, while the red line represents the 50th percentile. Focusing on the whiskers, the lowest and the highest lines represent the minimum and maximum measured relative errors. Observing this figure, we can conclude the relative error is below 1.2% for at least the 75% of the evaluated scenarios and it is always below 1.5% for any case.

C. Evaluation of the UE Blocking Probability with a Channel-Aware Scheduler

In this experiment, we have evaluated the UE blocking probability for different scenarios where the channel-aware scheduler sets a specific value for the fairness factor α. Specifically, we have considered α takes the values 1.0, 1.3, 1.5 and 1.7 for each scenario, respectively. In addition, we have compared these scenarios with the case of implementing a channel-agnostic scheduler as the one we considered in our previous work [17]. For all the scenarios, we have assumed there are 35 RBs allocated for a RAN slice. Furthermore, we have set $N_z = 6$ as the number of regions for $f_{\overline{\gamma}_u}(\overline{\gamma})$.

We show these results in Fig. 5. It can be seen that the UE blocking probability B is usually lower when the cell implements a channel-aware scheduler. This is because the channel-aware scheduler improves the resource utilization by allocating RBGs to the UE sessions which are less affected by the fast-fading effect, as the metric defined in Eq. (9) describes. This improvement depends on how the MNO configures the fairness factor α in the channel-aware scheduler. Below, we show how setting different values for the fairness factor α impacts the UE blocking probability.

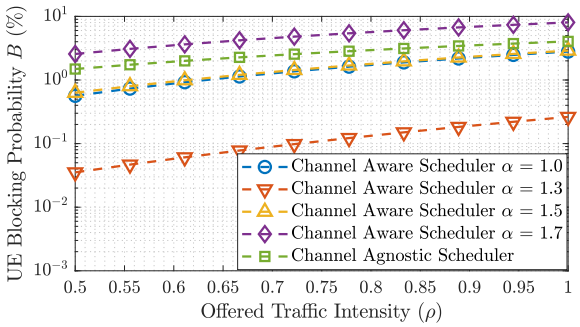


Fig. 5: UE Blocking Probability when the cell implements a scheduler with a specific configuration.

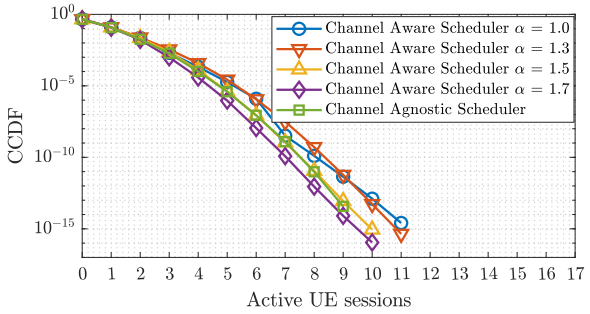
Setting $\alpha = 1.0$ (i.e., using the PF criteria) may not be an appropriate option because in the second step of the channel-aware scheduler (see Fig. 2) all the RBGs are pre-allocated with the same probability for each UE session regardless their average SINRs. Then, some RBGs may not be finally allocated after the third step (see Fig. 2) in the regions which have the best average SINRs. This means that, after the scheduler operation, there could be free RBGs for admitting more UE sessions in the regions with the lowest average SINRs. However, these RBGs cannot be used for the UE sessions which have the worst average SINRs due to the equal RBG distribution of the PF criteria. To avoid this issue, the MNO must increase the fairness factor α. In this way, the UE sessions located in the regions which have lowest average SINRs receive a greater probability for being scheduled with more RBGs. This means more UEs sessions could be admitted and thus, the UE blocking probability would decrease. For instance, we observe this fact in Fig. 5 when the MNO sets $\alpha = 1.3$. In this case, the UE blocking probability is the lowest for all the values of the offered traffic intensity.

However, if we follow increasing the fairness factor α, for instance $\alpha = 1.5$, more RBGs are scheduled for those UE sessions which were born in the regions with the lowest average SINRs. This involves more UE sessions in the regions with the greatest average SINRs are rejected. Thus the cell is using more RBGs for admitting a less amount of UE sessions and the UE blocking probability increases. Note that the UE blocking probability is very sensitive to small increments of α since this parameter is used as exponent of the average SINRs as Eq. (9) defines.

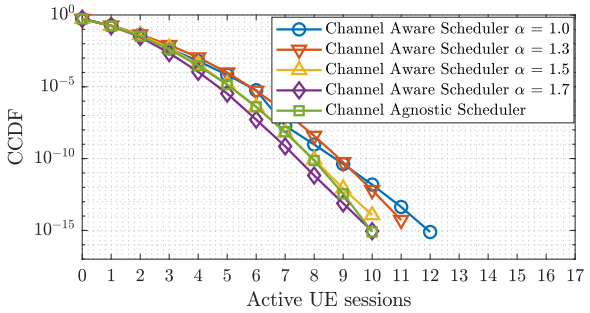
If this parameter is not properly configured, the MNO could not leverage the advantages of using a channel-aware scheduler. For instance, if we set $\alpha = 1.7$, the achieved UE blocking probability is greater than the one achieved by using a channel-agnostic scheduler.

D. Evaluation of the number of active UE sessions with a Channel-Aware Scheduler

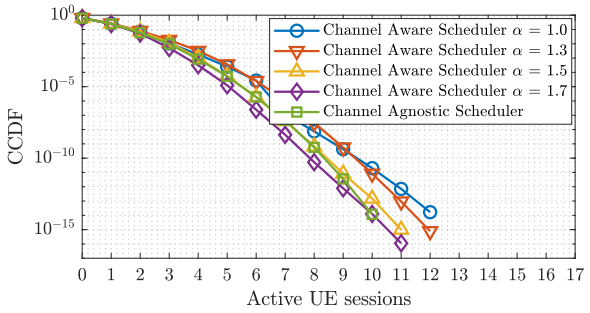
In this experiment, we have evaluated the number of active UE sessions in a RAN slice. To that end, we have considered the same scenarios presented in Section V-C. Based on them,



(a) Traffic Intensity $\rho = 1.0$



(b) Traffic Intensity $\rho = 1.5$



(c) Traffic Intensity $\rho = 2.0$

Fig. 6: CCDFs of the number of active UE sessions

Fig. 6 depicts the Complementary Cumulative Distribution Functions (CCDFs) for the number of active UE sessions when the traffic intensity ρ varies from $\rho = 1$ to $\rho = 2$. It can be observed that increasing ρ yields a greater probability that more UE sessions are active, i.e., the CCDFs are right-shifted. This means the UE blocking probability increases as Fig. 5 depicts.

If we focus on the CCDFs for a specific value of the traffic intensity ρ, we observe the probability that the number of active UE sessions is higher than a specific value (i.e., a x-axis value) is greater when the channel-aware scheduler sets the fairness factor $\alpha = 1.3$. This probability decreases (i.e., less active UE sessions) when the fairness factor α takes other values. For instance, when $\alpha = 1.3$ it is more likely to have more active UE sessions than the case of setting α with the values 1.0, 1.5 and 1.7. This means it is more difficult to reach a blocking state $\mathcal{S}^B$, thus the UE blocking probability is lower when the channel-aware scheduler sets $\alpha = 1.3$. The worst case happens when the MNO sets the fairness factor $\alpha = 1.7$.

Note that the values of the CCDF below 10^{-17} have been omitted because they are not significant.

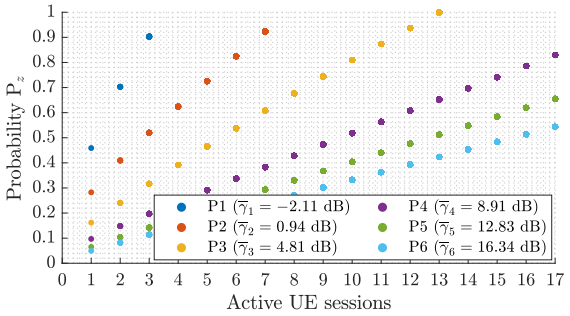
E. Analysis of the Radio Resource Utilization

In this experiment, we have evaluated the radio resource utilization for a RAN slice. To that end, we have considered the same scenarios presented in Section V-C. To measure the radio resource utilization, we consider the probabilities P_z for each region z and for each valid state $s \in \mathcal{S}'$ (see Eq. (32) and corresponding description). The reason is these probabilities define the percentage of RBGs allocated for each UE session after the scheduling criteria in Eq. (9) is applied by the channel-aware scheduler in the steps 2-3 (see Fig. 2). These probabilities are depicted in Fig. 7. Each point represents the specific value of a probability P_z when there is a specific set of active UE sessions in the system, i.e., a specific state $s \in \mathcal{S}'$. For instance, the cyan points represent these probabilities for all the UE sessions which was born in region $z = 6$, i.e., the region where $\overline{\gamma}_6 = 16.34$ dB.

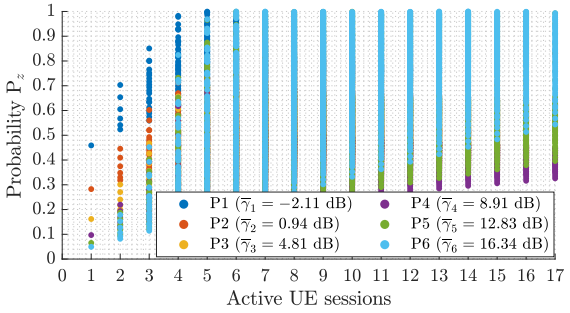
Fig. 7(a) represents these probabilities in the case the cell implements a channel-aware scheduler with $\alpha = 1.0$. This means the scheduler uses the PF criteria and thus, the RBGs are equally distributed among the UE sessions before applying these probabilities in the step 2 (see Fig. 2). If we focus on a specific amount of UE sessions (i.e., see x-axis), the probability in the region $z = 1$ (i.e., $\overline{\gamma} = -2.11$ dB) is always the greatest. Thus, the channel-aware scheduler has to allocate a higher percentage of RBGs for a UE (i.e., step 3) with respect to the amount of RBGs assigned following the PF criteria (i.e., step 2) in the region with the worst average SINR. As explained before, this means that using in a first attempt the PF criteria is not efficient because the percentage of RBGs which are not finally allocated for a UE in a region with a higher average SINR (for instance the region $z = 3$, where $\overline{\gamma} = 4.81$ dB) could be used to admit more UE sessions in regions which have a worst average SINR (for instance the region $z = 1$, where $\overline{\gamma} = -2.11$ dB).

Considering the MNO sets $\alpha = 1.3$ for all the valid states $\mathcal{S}'$, Fig. 7(b) shows how in some states the probabilities for region $z = 6$ (i.e., $\overline{\gamma} = 16.34$ dB) are closer to the value 1 in comparison with the probabilities obtained when the MNO sets $\alpha = 1.0$, i.e., the probabilities depicted in Fig. 7(a). However, almost all the probabilities in region $z = 1$ (i.e., $\overline{\gamma} = 5$ dB) are greater than the probabilities in the remaining regions in each single state. This phenomena can be observed in the left side of Fig. 7(b). This involves some RBGs are not used in the regions with a higher average SINR, and thus they could be used for admitting UE sessions in other states.

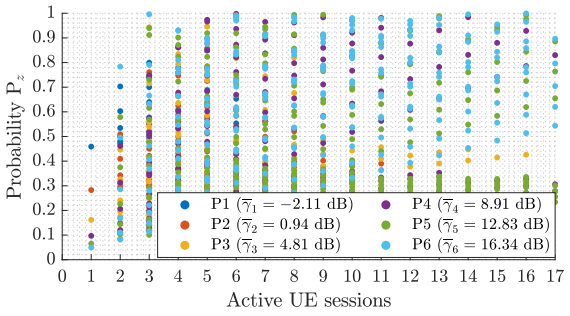
When the MNO sets $\alpha = 1.7$ for all the valid states $\mathcal{S}'$, Fig. 7(c) shows how the values of these probabilities have decreased in comparison with the values depicted in Fig. 7(b). This means that increasing the value of the fairness factor α involves that an excessive amount of RBGs are allocated for the lowest regions (i.e., those which have the lowest average SINRs). This involves not all the RBGs allocated for a UE



(a) Channel-Aware Scheduler with $\alpha = 1.0$



(b) Channel-Aware Scheduler with $\alpha = 1.3$



(c) Channel-Aware Scheduler with $\alpha = 1.7$

Fig. 7: Values of P_z per each region z and for each valid state $s \in \mathcal{S}'$

session which falls in the worst region are used, thus these RBGs are wasted and they are not used for admitting other UE sessions in regions with a better average SINR.

F. Execution Time Evaluation

We have assessed the time complexity of our analytical model in two scenarios. In the former, we have covered several amount of radio resources allocated for a RAN slice from 20 RBs to 45 RBs, considering $N_Z = 6$. In the latter, we have considered different number of regions from $N_Z = 4$ to $N_Z = 9$, with 35 RBs allocated for such RAN slice. In both scenarios, the fairness factor α is set to 1.3 and the offered traffic intensity $\rho = 1$. The results for both scenarios are shown in Table III.

We observe the execution time grows exponentially with the number of regions and the number of RBs. The reason is using higher values for both parameters involves an increment in the number of potential states $\mathcal{S}$ in the Markov chain as Eq. (18) shows. This means, the Eq. (32) must be solved more times. Specifically, one per state $s \in \mathcal{S}$ to check if this state is feasible (i.e., $s \in \mathcal{S}'$) or not (i.e., $s \in \mathcal{S}^{nf}$). Note that the number of integrals in Eq. (32) is also equal to the number of regions. Furthermore, since the set $\mathcal{S}'$ of feasible states will increase, the computation time for Eqs. (19), (20), and (21) will also increase. Additionally, although it is not shown in Table III, we have experimentally verified that setting a different value for the fairness factor could slightly change the number of states in the Markov chain, thus the execution time could accordingly increase or decrease. Note that the execution time does not depend on the offered traffic intensity since it does not modify the number of states in the Markov chain.

VI. Conclusions

RAN slicing is a solution for providing emerging communication services over a common wireless network infrastructure. Implemented as RAN slices, some of these communication services will rely on data transmission with GBR requirements. This means all the active UE sessions require an average data rate equal to a GBR throughout their lifetimes.

When the MNO plans a GBR service, a key parameter is the probability of blocking an UE data session, i.e., a session which cannot get an average data rate equal to a GBR. To provide a certain service guarantee, the MNO must ensure the UE blocking probability for a RAN slice is below an upper bound throughout its lifetime.

Under this context, we propose an analytical model based on a multi-dimensional Erlang-B system to evaluate the UE blocking probability. The main novelty of our model is it may consider as input any distribution and geometry for the average SINR within the cell. Our model also meets the reversibility property, thus it is valid for arbitrary distributions of the UE session duration. This property also reduces the computation complexity of the model because the solution for the state probabilities has product form. Additionally, we formulate the GBR achieved by an UE session when the cell implements a channel-aware scheduler. This allows our model to consider the impact of the scheduler channel gain on the UE blocking probability.

The results show that our model exhibits an estimation error for the UE blocking probability below 1.5%. Furthermore, our model allows the MNO to determine how a channel-aware scheduler must be configured to reduce the UE blocking

TABLE III: Execution Time

$N_Z = 6$	$N_i^{RB} = 20$	$N_i^{RB} = 25$	$N_i^{RB} = 30$	$N_i^{RB} = 35$	$N_i^{RB} = 40$	$N_i^{RB} = 45$
	1.802 s	3.652 s	7.451 s	13.939 s	28.125 s	47.851 s
$N_i^{RB} = 35$	$N_Z = 4$	$N_Z = 5$	$N_Z = 6$	$N_Z = 7$	$N_Z = 8$	$N_Z = 9$
	1.299 s	3.742 s	13.939 s	57.981 s	260.926 s	912.471 s

probability when a GBR service supports a specific traffic intensity ρ. Finally, we observe the execution time of our model is suitable for a planning tool.

APPENDIX A
PRACTICAL COMPUTATION OF THE DEFINITE INTEGRAL

In this appendix, we provide the steps to solve the definite integral I shown in Eq. (33).

$$I = \int_{t_1}^{t_2} w \cdot \exp\left[xt\right] \prod_{n=1}^{N} \left(1 - \exp\left[y_n t\right]\right)^{z_n} dt \qquad (33)$$

The first step consists of expanding each nth term $\left(1 - \exp\left[y_n t\right]\right)^{z_n}$ by using the binomial theorem as Eq. (34) defines.

$$\left(1 - \exp\left[y_n t\right]\right)^{z_n} = \sum_{k=0}^{z_n} (-1)^k \binom{z_n}{k} \exp\left[k y_n t\right] \qquad (34)$$

For simplicity, we can define $v_{cf,n} = \left\{(-1)^k \binom{z_n}{k}\right\}$ as a row vector whose components are the coefficients resulting from the binomial expansion. Note that they can be easily derived by (*i*) selecting the $(z_n + 1)$th row of the Pascal's triangle and (*ii*) switching the sign of the odd coefficients of such row. Additionally, we can define $v_{arg,n} = \{k y_n\}$ as a row vector whose components are the coefficients which multiply the variable t within the exponential function.

The second step consists of performing the product of all the binomial expansions. It results in a sum of $L = \prod_{n=1}^{N}(z_n+1)$ exponential functions, each one with a specific coefficient and argument. To derive the coefficient and the argument of each exponential function, we can use iteratively (i.e., $\forall n \in [2, N]$) Eqs. (35) and (36). Note that $r_{cf,1} = v_{cf,1}$ and $r_{arg,1} = v_{arg,1}$. Furthermore, the operators $\otimes$ and $(\cdot)^T$ denote the kronecker product and the transpose of a matrix, respectively. In addition, the operator $\text{rv}\left[\cdot\right]$ transforms a matrix into a row vector.

$$r_{cf,n} = \text{rv}\left[\left(r_{cf,n-1}\right)^T \otimes v_{cf,n}\right] \qquad (35)$$

$$r_{arg,n} = \text{rv}\left[\left(r_{arg,n-1}\right)^T + v_{arg,n}\right] \qquad (36)$$

Finally, the definite integral I can be computed as Eq. (37) shows. Note that the expression within the summation represents the integral of an exponential function evaluated in the interval $[t_1, t_2]$.

$$I = \sum_{l=1}^{L} \left(\frac{w \cdot r_{cf,N}[l]}{x + r_{arg,N}[l]} \cdot \exp\left[\left(x + r_{arg,N}\left[l\right]\right) t\right] \Big|_{t=t_1}^{t=t_2} \right) \qquad (37)$$

APPENDIX B
REVERSIBILITY IN A MARKOV PROCESS

To proof the reversibility property of the proposed model, we follow the Kolmogorov cycle criteria [28]. This states that a necessary and sufficient condition for reversibility of a multi-dimensional Markov process is that for each dimension-pair, the circulation flow among four neighboring states in a square equals to zero (i.e., flow clockwise = flow counter-clockwise).

Considering four neighbor states from two arbitrary regions z and x (i.e., $z \neq x$): $s_1 = (U_1, \ldots U_z, \ldots U_x, \ldots U_{N_z})$, $s_2 = (U_1, \ldots U_z, \ldots U_x + 1, \ldots U_{N_z})$, $s_3 = (U_1, \ldots U_z + 1, \ldots U_x + 1, \ldots U_{N_z})$, and $s_4 = (U_1, \ldots U_z + 1, \ldots U_x, \ldots U_{N_z})$, we derive the clockwise and counter clockwise flows $f_{cw} = \lambda_x \cdot p_1 \cdot \lambda_z \cdot p_2 \cdot (U_x + 1)\mu \cdot p_3 \cdot (U_z + 1)\mu \cdot p_4$ and $f_{ccw} = \lambda_z \cdot p_1 \cdot \lambda_x \cdot p_4 \cdot (U_z + 1)\mu \cdot p_3 \cdot (U_x + 1)\mu \cdot p_4$, respectively. We denote p_y the probability of state s_y. If we compare both equations, we easily check that the clockwise and the counter clockwise flows are equal. Thus, the proposed Erlang-B model is reversible.

REFERENCES

[1] A. Aijaz, "Private 5G: The Future of Industrial Wireless," *IEEE Ind. Electron. Mag.*, vol. 14, no. 4, pp. 136–145, 2020.

[2] O. Adamuz-Hinojosa, P. Munoz, J. Ordonez-Lucena, J. J. Ramos-Munoz, and J. M. Lopez-Soler, "Harmonizing 3GPP and NFV Description Models: Providing Customized RAN Slices in 5G Networks," *IEEE Veh. Technol. Mag.*, vol. 14, no. 4, pp. 64–75, 2019.

[3] GSM Association, "Generic Network Slice Template (Version 7.0)," June 2022.

[4] 3GPP TS 28530 V.16.3.0, "Management and orchestration; Concepts, use cases and requirements (Release 16)," Sept. 2020.

[5] O. Adamuz-Hinojosa, P. Muñoz, P. Ameigeiras, and J. M. Lopez-Soler, "Potential-Game-Based 5G RAN Slice Planning for GBR Services," *IEEE Access*, vol. 11, pp. 4763–4780, 2023.

[6] T. Bonald and A. Proutière, "Wireless Downlink Data Channels: User Performance and Cell Dimensioning," in *MobiCom*, pp. 339–352, 2003.

[7] T. Bonald *et al.*, "Flow-level performance and capacity of wireless networks with user mobility," *Queueing Syst.*, vol. 63, no. 1-4, p. 131, 2009.

[8] S.-E. Elayoubi and T. Chahed, "Admission Control in the Downlink of WCDMA/UMTS," in *EuroNGI*, pp. 136–151, Springer, 2004.

[9] D. K. Kim *et al.*, "A novel ring-based performance analysis for call admission control in wireless networks," *IEEE Commun. Lett.*, vol. 14, no. 4, pp. 324–326, 2010.

[10] B. Sas *et al.*, "Modelling the time-varying cell capacity in LTE networks," *Telecommun. Syst.*, vol. 55, no. 2, pp. 299–313, 2014.

[11] C. Tarhini and T. Chahed, "QoS-oriented resource allocation for streaming flows in IEEE802. 16e Mobile WiMAX," *Telecommun. Syst.*, vol. 51, no. 1, pp. 65–71, 2012.

[12] A. Abdollahpouri and B. E. Wolfinger, "Measures to quantify the gain of multicast with application to IPTV transmissions via WiMAX networks," *Telecommun. Syst.*, vol. 55, no. 2, pp. 185–198, 2014.

[13] M. Li, "Queueing Analysis of Unicast IPTV With Adaptive Modulation and Coding in Wireless Cellular Networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 10, pp. 9241–9253, 2017.

[14] C. Kim *et al.*, "Mathematical Models for the Operation of a Cell With Bandwidth Sharing and Moving Users," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 744–755, 2020.

[15] F. Capozzi, G. Piro, L. Grieco, G. Boggia, and P. Camarda, "Downlink Packet Scheduling in LTE Cellular Networks: Key Design Issues and a Survey," *IEEE Commun. Surv. Tutor.*, vol. 15, no. 2, pp. 678–700, 2013.

[16] L. Kleinrock, "Queue Systems, Volume i: Theory," *Jonh Wiley & Sons*, 1975.

[17] O. Adamuz-Hinojosa, P. Ameigeiras, P. Muñoz, and J. M. Lopez-Soler, "Analytical Model for the UE Blocking Probability in an OFDMA Cell providing GBR Slices," in *IEEE WCNC*, pp. 1–7, 2021.

[18] P. Ameigeiras, Y. Wang, J. Navarro-Ortiz, P. E. Mogensen, and J. M. Lopez-Soler, "Traffic models impact on OFDMA scheduling design," *EURASIP J. Wirel. Commun. Netw.*, vol. 2012, no. 1, pp. 1–13, 2012.

[19] 3GPP TS 38.101-1 V.16.3.0, "User Equipment (UE) radio transmission and reception; Part 1: Range 1 Standalone (Release 16)," Mar. 2020.

[20] 3GPP TS 38.101-2 V.16.4.0, "User Equipment (UE) radio transmission and reception; Part 2: Range 2 Standalone (Release 16)," June 2020.

[21] P. Muñoz *et al.*, "Radio Access Network Slicing Strategies at Spectrum Planning Level in 5G and Beyond," *IEEE Access*, vol. 8, pp. 79604–79618, 2020.

[22] 3GPP TS 38.214 V.16.1.0, "NR, Physical layer procedures for data (Release 16)," Mar. 2020.

[23] D. Parruca and J. Gross, "Throughput Analysis of Proportional Fair Scheduling for Sparse and Ultra-Dense Interference-Limited OFDMA/LTE Networks," *IEEE Trans. Wirel. Commun.*, vol. 15, no. 10, pp. 6857–6870, 2016.

[24] J. Ramiro-Moreno, *System Level Performance Analysis of Advanced Antenna Concepts in WCDMA*. Aalborg Universitetsforlag, 2008.

[25] I. Z. Kovacs, K. I. Pedersen, J. Wigard, F. Frederiksen, and T. E. Kolding, "HSDPA Performance in Mixed Outdoor-Indoor Micro Cell Scenarios," in *IEEE PIMRC*, pp. 1–5, 2006.
[26] P. Mogensen *et al.*, "LTE Capacity Compared to the Shannon Bound," in *2007 IEEE VTC-Spring*, pp. 1234–1238, 2007.
[27] N. Wei, *MIMO Techniques in UTRA Long Term Evolution*. Institut for Elektroniske Systemer, Aalborg Universitet, 2007.
[28] V. B. Iversen, "Teletraffic engineering and network planning," 2015.
[29] J. Liu, A. Proutiere, Y. Yi, M. Chiang, and H. V. Poor, "Stability, Fairness, and Performance: A Flow-Level Study on Nonconvex and Time-Varying Rate Regions," *IEEE Trans. Inf. Theory*, vol. 55, no. 8, pp. 3437–3456, 2009.
[30] M. H. Ahmed, O. A. Dobre, and R. K. Almatarneh, "Analytical Evaluation of the Performance of Proportional Fair Scheduling in OFDMA-Based Wireless Systems," *JECE*, vol. 2012, Jan. 2012.
[31] D. Parruca, M. Grysla, S. Gortzen, and J. Gross, "Analytical Model of Proportional Fair Scheduling in Interference-Limited OFDMA/LTE Networks," in *IEEE VTC-Fall*, pp. 1–7, 2013.

Oscar Adamuz-Hinojosa received his B.Sc., M.Sc. and Ph.D in Telecommunications Engineering from the University of Granada (Spain) in 2015, 2017, and 2022 respectively. He was granted a Ph.D. fellowship by the Education Spanish Ministry on September 2018. Currently, he is working as Interim Assistant Professor at Department of Signal Theory, Telematics and Communication of the University of Granada. His research interests are focused on Network slicing in Beyond 5G / 6G Radio Access Network (RAN).

Pablo Ameigeiras received his M.Sc.E.E. degree in 1999 from the University of Malaga, Spain. He performed his Master's thesis at the Chair of Communication Networks, Aachen University, Germany. In 2000 he joined Aalborg University, Denmark, where he carried out his Ph.D. thesis. In 2006 he joined the University of Granada, where he has been leading several projects in the field of LTE and LTEAdvanced systems. Currently his research interests include 5G and IoT technologies.

Pablo Muñoz received the M.Sc. and Ph.D. degrees in telecommunication engineering from the University of Málaga, Málaga, Spain, in 2008 and 2013, respectively. He is currently an Assistant Professor with the Department of Signal Theory, Telematics, and Communications, University of Granada, Granada, Spain. He has published more than 50 articles in peer-reviewed journals and conferences. He is the coauthor of four international patents. His research interests include radio access network planning and management, application of artificial intelligence tools in radio resource management, and virtualization of wireless networks

Juan M. Lopez-Soler received the B.Sc. degree in physics (electronics) and the Ph.D. degree in signal processing and communications, both from the University of Granada, Granada, Spain, in 1995. He is a Full Professor with the Department of Signals, Telematics and Communications, University of Granada. During 1991—1992, he joined the Institute for Systems Research (formerly SRC), University of Maryland, College Park, MD, USA, as a Visiting Faculty Research Assistant. Since its creation in 2012, he has been the Head of the Wireless and Multimedia Networking Laboratory, University of Granada. He has participated in 11 public and 13 private funded research projects and is the coordinator in 14 of them. He has advised five Ph.D. students and has published 24 papers in indexed journals and contributed to more than 40 workshops/conferences. His research interests include real-time middleware, multimedia communications, and networking.