Disclaimer:
This work has been accepted for publication in the IEEE Transactions on Wireless Communications.
Citation information: 10.1109/TWC.2022.3203937

A Stochastic Network Calculus (SNC)-based model for planning B5G uRLLC RAN slices

Oscar Adamuz-Hinojosa, Vincenzo Sciancalepore, *Senior Member, IEEE,* Pablo Ameigeiras, Juan M. Lopez-Soler, Xavier Costa-Pérez, *Senior Member, IEEE,*

Abstract

Radio Access Network (RAN) slicing involves several challenges. In particular, the Mobile Network Operator (MNO) must ensure —before deploying each slice—that corresponding requirements can be met throughout its lifetime. For ultra-Reliable Low Latency Communication (uRLLC) slices, the MNO must guarantee the packet transmission delay within a delay budget with a certain probability. Most existing solutions focus on allocating dynamically radio resources to maximize the number of packets, whose transmission delay is within the delay budget. However, these solutions do not ensure the violation probability is below a target value in the long term. In this paper, we focus on slicing from a planning perspective. Specifically, we propose a Stochastic Network Calculus (SNC)-based model, which given the amount of radio resources allocated for a uRLLC slice, the target violation probability and the traffic demand distribution, provides the delay bound for such conditions. Additionally, we propose heuristics for planning uRLLC slices. Interestingly, such heuristics benefit from the proposed SNC-based model to compute the amount of radio resources to be assigned to each slice while its delay bound, given a target

O. Adamuz-Hinojosa, P. Ameigeiras and J. M.Lopez-Soler are with Department of Signal Theory, Telematics and Communications, University of Granada, Calle Periodista Daniel Saucedo Aranda s/n, 18071 Granada, Spain. E-mails: {oadamuz, pameigeiras, juanma}@ugr.es.

V. Sciancalepore is with NEC Laboratories Europe GmbH, Kurfürsten-Anlage 36, 69115 Heidelberg, Germany. E-mail: vincenzo.sciancalepore@neclab.eu.

X. Costa-Pérez is with NEC Laboratories Europe GmbH., Kurfürsten-Anlage 36, 69115 Heidelberg, Germany, and i2CAT Foundation and ICREA, Carrer del Gran Capità 2, 08034 Barcelona, Spain, Email: xavier.costa@ieee.org.

This work is partially supported by the H2020 research and innovation project 5G-CLARITY (Grant No. 871428); the Spanish Ministry of Economy and Competitiveness, The European Regional Development Fund (Project PID2019-108713RB-C53); and the Spanish Ministry of Education, Culture and Sport (FPU Grant 17/01844).

Corresponding author: oadamuz@ugr.es. Manuscript received XXX XX, XXXX; revised XXX XX, XXXX.

violation probability, is within the delay budget. We validate the SNC-based model and demonstrate the effectiveness of the heuristics.

Index Terms

B5G, RAN slicing, uRLLC services, stochastic network calculus, delay bound modeling

I. Introduction

The 5th Generation (5G) of mobile networks brings digitalization of the industrial vertical segment. It is supposed to boost a wide variety of unprecedented communication services with stringent requirements in terms of performance and functionalities. There is a general consensus identifying three main categories: enhanced Mobile Broadband (eMBB), Machine Type Communication (mMTC) and ultra-Reliable Low Latency Communication (uRLLC) [1]. The latter is of particular interest in 4.0 industry scenarios where multiple communication services, each one with particular requirements in terms of latency and reliability, may coexist [2]. Example of these services are: monitoring and control of cyber-physical systems, or industrial Augmented Reality (AR)/Virtual Reality (VR) applications.

However, considering each communication service as an independent monolithic network instance and building a dedicated Radio Access Network (RAN) infrastructure to accommodate such stringent requirements would be costly and unaffordable. Therefore, RAN slicing has emerged as a potential solution [3] to economically provide the means to manage and successfully orchestrate such services onto a shared network infrastructure. It consists of providing logically-separated self-contained RANs, denominated RAN slices, tailored onto the requirements of a specific communication service over a common physical infrastructure.

Differently from core network slicing, where virtual machines or containers might be instantiated to accommodate chains of virtual network functions, the RAN slicing concept comes at no negligible costs: *how to assign radio resources in advance to multiple RAN slices (i.e., much before deploying them)* in such a way the Mobile Network Operator (MNO) can ensure their performance requirements will be completely satisfied throughout their lifetimes? In particular, RAN slices providing uRLLC services can further exacerbate the resource assignment problem as the MNO must guarantee that the packet transmission delay is within the overall delay budget with a certain probability.

Most of the state-of-the-art solutions on uRLLC focus on dynamic radio resource allocation, i.e., distributing on the fly the available radio resources among multiple uRLLC User Equipments (UEs) to satisfy their current traffic demands. However, the problem of computing in advance the amount of radio resources which ensure the performance requirements of multiple RAN slices in the long term has received less attention. In this regard, the 3rd Generation Partnership Project (3GPP) has defined the preparation phase within the RAN slice management and orchestration [4]. In this phase, among other things, the MNO must plan in advance the deployment of the new RAN slices and prepare (if it was necessary) the RAN infrastructure to accommodate these RAN slices along with the already deployed RAN slices. In addition, the 3GPP has also recently standardized a set of policy ratios [5] to guarantee an amount of radio resources for each RAN slice throughout their lifetimes. Although these policy ratios must be computed in the preparation phase, they must also be considered later in the operation of each RAN slice [6]. Specifically the packet scheduler must use them as radio resource quotas when it dynamically allocates radio resources for the UEs attached to these RAN slices. In this paper, we focus on the computation of the Radio Resource Management (RRM) Policy Dedicated Ratio. It defines the dedicated radio resource quota for an instantiated RAN slice in a single cell. For more information, we recommend the reader to see [3].

When the MNO executes a procedure for planning multiple uRLLC RAN slices in advance, it computes a dedicated radio resource quota per RAN slice and cell. To this end, the MNO must rely on *i*) a mathematical model, which automatically checks whether the performance requirements of a RAN slice within a cell are met in the long term by allocating it dedicated radio resources, and *ii*) an algorithm, based on the suggested model, which plans the dedicated radio resource quotas for each RAN slice.

A. Related Works

In the literature, there exists queueing theory-based models to derive the transmission delay of an uRLLC packet in the radio interface. Despite their valuable contributions, the proposed models present some drawbacks, which do not make them suitable for planning uRLLC RAN slices. One drawback is that some models assume well-known but not enough-accurate statistical distributions for modeling the packet arrival rate and the packet transmission rate in the radio interface. For instance, works such as [7]–[9] consider an exponential distribution for the packet transmission rate of a uRLLC service. This assumption along with the consideration of a Poisson distribution for the packet arrival rate has the advantage that some models provide close-form expressions for the Cumulative Distribution Function (CDF) of the packet transmission delay. However, if more complex distributions are considered [10], computing the CDF using queueing theory is mathematically intractable. In such cases, queueing theory-based models can only provide average values for the packet transmission delay.

Another mathematical tool to derive the packet transmission delay is Stochastic Network Calculus (SNC). This tool allows us to compute non-asymptotic statistical performance bounds of the type $\text{Probability}\left[\text{delay} > \text{budget}\right] \leq$ violation probability while considering complex stochastic processes [11]. This generality comes at the expense of exacts solutions for such bounds. Instead, SNC provides a conservative estimation for such bounds. In the literature, there exists a wide variety of solutions based on SNC to compute the transmission delay of an uRLLC packet. Since this tool is ideal for scenarios where there exists multiple network nodes in tandem, most of the solutions focus on computing the upper bound of the packet delay in multi-hop cellular networks. Examples can be found in [12], [13]. However, they either do not consider the radio interface or consider simple models for the radio interface without deepening into the impact of channel effects such as path loss, fast fading and/or shadowing. Conversely, other works focus on the radio interface considering specific access schemes. For instance [14] considers an Orthogonal Frequency-Division Multiple Access (OFDMA) scheme and [15], [16] assume non-orthogonal multiple access schemes. However, these works do not consider the RAN slicing technology. This means the mechanisms to *i*) guarantee performance isolation among RAN slices in the long term (e.g., by dedicating radio resources for each RAN slice) and *ii*) how they impact on the cell capacity are completely omitted. There exist SNC-based solutions, which consider network slicing such as [17], however they focus on the transport and core networks, simplifying the behavior of the radio interface. To the best of our knowledge, there are no works that analyze the packet delay bound in the radio interface for an uRLLC RAN slice.

Focusing on solutions for provisioning uRLLC services, we can find works such as [7]–[9]. In these works the authors consider latency and reliability as the main service performance requirements, however these solutions omit the RAN slicing technology. Others works such as [18]–[22] consider RAN slicing to offer uRLLC services. However, they are mainly focused on the operation of each RAN slice instead of its planning. In these works, the authors provide an algorithm to dynamically allocate radio resources for each RAN slice in every transmission time interval or within a short time windows. However, how the latency requirements could be guaranteed in the long term for each RAN slice is out of their scope. Additionally, some of these works do not consider any model for estimating the packet transmission delay. Instead, these works propose online algorithms, which directly observe the packet buffers to decide the amount of radio resources allocated for each uRLLC slice. Therefore, the solutions proposed in these works omit the radio resource quotas, which must previously be derived during a planning procedure.

B. Contributions

In this article, we assume a multi-cellular environment where the MNO must plan in advance the deployment of multiple uRLLC RAN slices. Furthermore, each RAN slice has specific requirements in terms of latency and reliability. Specifically, each RAN slice requires its packet transmission delay is within a delay budget with a certain probability. To summarize:

- We have proposed a SNC-based model which provides a bound for the packet transmission delay of an uRLLC RAN slice in a single cell. This bound is derived by considering the following inputs: *i*) the amount of dedicated radio resources for this RAN slice, *ii*) the probability the packet delay is above the delay bound, i.e., the violation probability, *iii*) the CDF for the Signal-to-Interference-plus-Noise Ratio (SINR) experienced by the users served by this RAN slice, and *iv*) the traffic demand of this RAN slice, i.e., the distribution of the packet arrival rate, and the distribution of the packet size. In our model, the packet size distribution could be arbitrary.
- To compute the CDF for the SINR perceived by an UE, which is served by a specific uRLLC RAN slice, we use a model based on stochastic geometry. This model considers the impact of the interference incurred by multiple RAN slices deployed in neighbor cells on the capacity the serving cell offers to the serving RAN slice.
- We have proposed heuristics that—using the proposed SNC-based model—plan in advance the deployment of multiple RAN slices with different requirements in terms of traffic demand, latency and reliability. To that end, the proposed heuristics compute the dedicated radio resource quotas which minimizes the difference between the delay bounds achieved with such quotas and the target delay budgets.

In the provided results, we first validate the proposed SNC-based model by means of an exhaustive simulation campaign, demonstrating that it always provides an upper estimation of the real delay bound of an uRLLC RAN slice. This makes our model suitable for planning uRLLC RAN slices. Then, based on this model, we show the benefits of using the proposed heuristics to plan the deployment of multiple uRLLC RAN slices over a multi-cellular scenario with radio resource scarcity.

The remainder of this article is organized as follows. Section II provides a background on SNC. Section III describes the system model. In Section IV, we propose a SNC-based model for modeling the packet delay bound of an uRLLC RAN slice in a single cell. In Section V, we formulate the radio resource planning for several RAN slices in a multi-cell environment. To solve this problem, we propose novel heuristics. In Section VI, we validate the proposed SNC-based model and provide the performance results for the proposed heuristics. Finally, Section VII summarizes the conclusions.

For comprehensibility purposes, we provide in Table I the key notations used in the paper.

II. Background on Stochastic Network Calculus (SNC)

A. Fundamentals

Network calculus has been developed along two tracks: Deterministic Network Calculus (DNC) and SNC. Focusing on DNC, its main principles are [23], [24]:

- For each individual network node, DNC mainly considers: (a) the accumulative arrival process $A(\tau,t)$ of a specific service; and (b) the accumulative service process $S(\tau,t)$. These stochastic processes are considered in the time interval $(\tau,t]$.
- Characterizing $A(\tau,t)$ and $S(\tau,t)$ by upper and lower bounds. These bounds are known as arrival curve $\alpha(\tau,t)$ and service curve $\beta(\tau,t)$, respectively.
- Using $\alpha(\tau,t)$ and $\beta(\tau,t)$, performance parameters such as the backlog bound B and the delay bound W of each network node can be analyzed. The definition of backlog bound B comprises the maximum number of bits which may be stored in the network node's buffer, whereas the delay bound W is the maximum waiting time which a bit in the buffer may experience. The backlog bound B in a network node is defined in Eq. (1).

$$B = \max_{\tau\in[0,t]} \{\alpha(\tau,t) - \beta(\tau,t)\} \tag{1}$$

Assuming First-come First-served (FCFS) order, the delay bound W in a network node is given by Eq. (2).

$$W = \min\{\omega \geq 0 : \max_{\tau\in[0,t]} \{\alpha(\tau,t) - \beta(\tau,t+\omega)\}\} \tag{2}$$

TABLE I: Key Notations

Notation	**Meaning**	**Notation**	**Meaning**
$\mathcal{I}$, $\mathcal{M}$, $\mathcal{U}$, $\mathcal{R}$	Set of cells, RAN slices, UEs, and RBs.	$A_{i,m}(\tau,t)$, $S_{i,m}(\tau,t)$	Accumulative arrival and service processes for RAN slice m in cell i, respectively.
$\alpha_{i,m}(\tau,t)$, $\beta_{i,m}(\tau,t)$	Affine arrival and service envelopes for RAN slice m in cell i, respectively.	$B_{i,m}$, $W_{i,m}$	Backlog and delay bounds for RAN slice m in cell i, respectively.
$\rho_{A_{i,m}}$, $\rho_{S_{i,m}}$	Rate parameters of $\alpha_{i,m}(\tau,t)$ and $\beta_{i,m}(\tau,t)$, respectively for RAN slice m in cell i.	$M_{A_{i,m}}(\theta)$, $M_{S_{i,m}}(\theta)$	MGF for $A_{i,m}(\tau,t)$ and $S_{i,m}(\tau,t)$, respectively.
$b_{A_{i,m}}$, $b_{S_{i,m}}$	Burst parameters of $\alpha_{i,m}(\tau,t)$ and $\beta_{i,m}(\tau,t)$, respectively for RAN slice m in cell i (EBB and EBF models).	$\sigma_{A_{i,m}}$, $\sigma_{S_{i,m}}$	Burst parameters of $\alpha_{i,m}(\tau,t)$ and $\beta_{i,m}(\tau,t)$, respectively for RAN slice m in cell i (MGF models).
θ, δ	Free parameters MGF model.	$\varepsilon'_{A_{i,m}}$, $\varepsilon'_{S_{i,m}}$	Overflow and deficit profiles, respectively for RAN slice m in cell i.
$\lambda_{i,m}$	Average batch arrivals for RAN slice m in cell i.	$\omega_{i,m}$	Probability an UE is served by RAN slice m in cell i.
$N_{i,m}^{pkt}$, $p_{i,m,u}$	Number of packets simultaneously generated for RAN slice m in cell i and its PMF, respectively.	L, $p_{m,l}$	Packet size for RAN slice m and its PMF, respectively.
W_m^{th}, ε'_m	Delay budget and violation probability, respectively for RAN slice m.	BW_i	Bandwidth in cell i.
N_i	OFDM subcarriers in cell i.	N_{SC}	Subcarriers per RB.
t^{slot}	Timeslot.	$\mathcal{R}_i$	Set of RBs available in cell i.
μ_{5G}	5G numerology.	Δf	Subcarrier spacing.
OH	Overhead factor	$\gamma_{u,r}^{(t)}$	Instantaneous SINR for UE u in RB r.
P_{TX}	Transmitted power.	h_r, χ	Fast and shadow fading gains, respectively.
μ_χ, σ_χ	Mean and standard deviation shadow fading distribution, respectively.	$d_{u,i}$	Distance from UE u to gNB i.
α	Pathloss exponent.	P_N	Noise power.
$\xi_{j,r}$	Binary variable which indicates if the RB r is used by the cell i in a timeslot.	κ_{gNB}, κ_{UEs}, $\kappa_{RBs,i,m}$	gNB density, UE density and density of the required number of RBs for transmitting a packet.
f_{CDF}	CDF instantaneous SINR.	γ_{th}	Target SINR.
p_{suc}	Probability $\gamma_{u,r} > \gamma_{th}$	p_{sel}	Probability RB r is scheduled to UE u.
p_{off}	Probability an arbitrary cell does not transmit data in a RB.	$\overline{R}_{i,m}^{pkt}$	Average number of required RBs to transmit a packet.
$SE_{z,i,m}$, $\pi_{z,i,m}$	Spectral efficiency and its PMF, respectively.	$f_{SE\to\gamma}$	Function to translate spectral efficiency into SINR.
$f_{SE\to\gamma}^{-1}$	Inverse function of $f_{SE\to\gamma}$.	ε_{dec}	Decoding error probability for a URLLC packet.
n_{block}	Length of codeword block.	$R_{r',i,m}^{pkt}$, $p_{r'}$	Number of RBs to transmit a packet for the RAN slice m in cell i and its PMF, respectively.
$N_{i,m}^{batch}$	Number of batch arrivals.	y_b	Number of bits generated in a batch.
N^{slots}	Number of accumulated time slots.	c_q, p_q	Number of bits which a gNB could transmit in a timeslot and its PMF, respectively.

DNC considers the worst-case scenario to compute $\alpha(\tau,t)$ and $\beta(\tau,t)$, i.e., $\alpha(\tau,t) > A(\tau,t)$ and $\beta(\tau,t) < S(\tau,t)\ \forall\tau \in (0,t]$. This means DNC ignores the effects of statistical multiplexing and therefore, it leads to an overestimation of the resource requirements for the service to be deployed in the network node. This fact has motivated the development of SNC, which extends the DNC to a probabilistic setting by considering $P[A(\tau,t) > \alpha(\tau,t)] > 0$ and $P[S(\tau,t) < \beta(\tau,t)] > 0$. In SNC $\alpha(\tau,t)$ $\beta(\tau,t)$ are commonly named arrival envelope and service envelope, respectively. For more details, we recommend the reader to see [25].

A widely practice in SNC to compute the backlog bound B and delay bound W in a network node consists of assuming an affine function (i.e., linear function plus a constant) to define $\alpha(\tau,t)$ and $\beta(\tau,t)$ as Eqs. (3) and (4) show, respectively. The parameters $\rho_A > 0$ and $b_A \geq 0$ are the rate and burst parameters for $\alpha(\tau,t)$, whereas $\rho_S > 0$ and $b_S \geq 0$ are the rate and burst parameters for $\beta(\tau,t)$. Additionally, $[x]_+$ denotes $\max\{0,x\}$. Note that, $\beta(\tau,t) = 0$ when $t-\tau \leq b_S/\rho_S$.

$$\alpha(\tau,t) = \rho_A(t-\tau) + b_A \tag{3}$$

$$\beta(\tau,t) = \rho_S\left[t-\tau-\frac{b_S}{\rho_S}\right]_+ \tag{4}$$

Considering $\alpha(\tau,t)$, $\beta(\tau,t)$, and Eqs. (1) and (2), we can compute the backlog and delay bounds in Eqs. (5) and (6), respectively. Fig. 1 depicts a graphical representation of the derived backlog and delay bounds. The backlog bound B is the vertical deviation between $\alpha(\tau,t)$ and $\beta(\tau,t)$ whereas the delay bound W is the horizontal deviation between these envelopes.

$$B = \frac{\rho_A}{\rho_S}b_S + b_A \tag{5}$$

$$W = \frac{b_A + b_S}{\rho_S} \tag{6}$$

In sections II-B, and II-C, we summarize the methods and equations proposed in [11] to obtain $\alpha(\tau,t)$ and $\beta(\tau,t)$. For more details about them, we recommend the readers to see [11, Sections II.A and II.B].

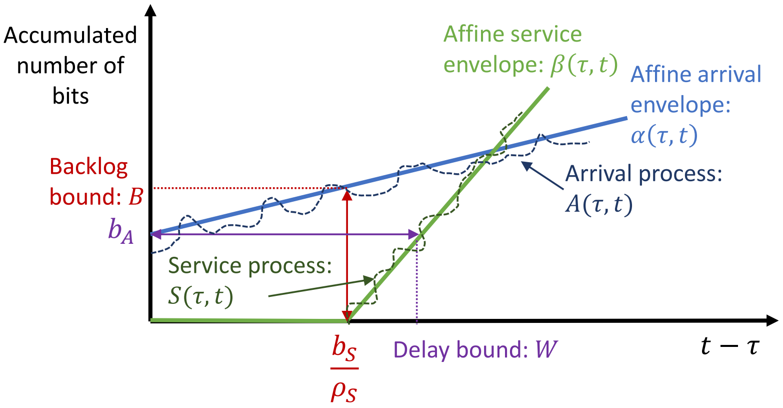

Accumulated number of bits
Affine service envelope: $\beta(\tau,t)$
Affine arrival envelope: $\alpha(\tau,t)$
Backlog bound: B
b_A
Arrival process: $A(\tau,t)$
Service process: $S(\tau,t)$
$\frac{b_S}{\rho_S}$
Delay bound: W
$t-\tau$

Fig. 1: Graphical representation of backlog bound B and delay bound W.

B. Affine Arrival Envelope

Some stochastic traffic models use Moment Generating Functions (MGFs) to uniquely determine the distribution of a random process. The MGF of $A(\tau,t)$ is defined as $M_A(\theta) = E[e^{\theta A(\tau,t)}]$ with free parameter $\theta \geq 0$ [26]. Considering an affine function, Eq. (7) defines an upper bound for $M_A(\theta)$. The variables $\rho_A > 0$ and $\sigma_A \geq 0$ are the rate and burst parameters.

$$M_A(\theta) \leq e^{\theta[\rho_A(t-\tau)+\sigma_A]} \tag{7}$$

Other approach is to use the Exponentially Bounded Burstiness (EBB) model, which provides the guarantee expressed in Eq. (8). This approach relaxes the deterministic arrival curve described in Section II-A by defining an overflow profile $\varepsilon_A \geq 0$. The overflow profile ε_A is the probability the accumulative arrival process $A(\tau,t)$ is greater than the affine arrival envelope $\alpha(\tau,t)$ throughout the time interval $(\tau,t]$.

$$P[A(\tau,t) > \alpha(\tau,t)] \leq \varepsilon_A \tag{8}$$

Comparing the arrival envelope defined in Eq. (8) with the analogous deterministic arrival curve described in Section II-A, we notice a difference arises with respect to the computation of the backlog bound. The deterministic arrival curve can be immediately applied in Eq. (1), however the EBB envelope cannot. The reason is Eq. (1) evaluates all $\tau \in [0,t]$, where the $\tau = \tau^*$ that attains the maximum is a random variable. In contrast, the EBB envelope only provides a guarantee for an arbitrary, yet, fixed $\tau \in [0,t]$. To overcome this problem, a sample path argument $\delta > 0$ is required as Eq. (9) shows [11]. This means including δ in the affine arrival envelope $\alpha(\tau,t)$. To that end, we can replace ρ_A with $\rho'_A = \rho_A + \delta$ in Eqs. (3) and (5).

$$P\left[\exists\tau \in [0,t] : A(\tau,t) > (\rho_A + \delta)(t-\tau) + b_A\right] \leq \varepsilon'_A \tag{9}$$

The EBB and MGF models are directly connected by the Chernoff bound [26] as Eq. (10) shows [11]. In this expression, we obtain the overflow profile ε'_A.

$$\varepsilon'_A = \frac{e^{\theta\sigma_A}e^{-\theta b_A}}{1-e^{-\theta\delta}} \tag{10}$$

Finally, based on Eq. (10), we can obtain a mathematical expression for b_A as Eq. (11) shows.

$$b_A = \sigma_A - \frac{1}{\theta}\left[\ln(\varepsilon'_A) + \ln\left(1-e^{-\theta\delta}\right)\right] \tag{11}$$

C. Affine Service Envelope

Considering an affine function, Eq. (12) defines a upper bound for the negative MGF of $S(\tau,t)$.

$$M_S(-\theta) \leq e^{-\theta(\rho_S(t-\tau)-\sigma_S)} \tag{12}$$

Using the analogous EBB model for the service envelope, also known as the Exponentially Bounded Fluctuation (EBF), we provide the guarantee defined in Eq. (13). In this equation, $\beta(\tau,t)$ is the affine service envelope defined in Eq. (4). Furthermore, ε_S' is the deficit profile which represents the probability the accumulative service process $S(\tau,t)$ is lower than $\beta(\tau,t)$ throughout the time interval $(\tau,t]$. This model also includes the concept of sample path δ similarly as Eq. (9). To include δ in $\beta(\tau,t)$, we need to replace ρ_S with $\rho_S' = \rho_S - \delta$ in Eqs. (4), (5) and (6).

$$P\left[\exists\tau \in [0,t] : S(\tau,t) < \beta(\tau,t)\right] \leq \varepsilon_S' \tag{13}$$

In a similar way as the affine arrival envelope, the Chernoff bound is used in Eq. (13) to derive the deficit profile ε_S' as Eq. (14) describes. Finally, using this expression the burst parameter b_S is derived as Eq. (15) shows.

$$\varepsilon_S' = \frac{e^{\theta\sigma_S} e^{-\theta b_S}}{1-e^{-\theta\delta}} \tag{14}$$

$$b_S = \sigma_S - \frac{1}{\theta}\left[\ln(\varepsilon_S') + \ln\left(1-e^{-\theta\delta}\right)\right] \tag{15}$$

III. System Model

In this work, we focus on the downlink (DL) operation of a 5G-New Radio (NR) multi-cell environment with several RAN slices. Each RAN slice provides an uRLLC service with specific requirements in terms of delay budget and violation probability. Furthermore, we consider each UE is only served by one RAN slice. Additionally, each Next generation NodeB (gNB) supports Link Adaptation (LA), thus these gNBs consider the channel quality perceived by each UE to allocate them radio resources. Under this scenario, we first describe the network model. Then, we define the characteristics of the offered DL traffic for an uRLLC service. Next, we present the radio resource model. Finally, we describe the channel model for a single cell.

A. Network Model

We consider a MNO owns a RAN infrastructure consisting of a set $\mathcal{I}$ of cells. Defining $\mathcal{M}$ as the set of RAN slices, the RAN slice orchestrator, e.g., implemented in the 3GPP standardized RAN Network Slice Subnet Management Function (NSSMF) [27], will execute a radio resource planning procedure to compute in advance the dedicated radio resource quotas which will guarantee the performance requirements of each RAN slice in the long term. These RAN slices will serve a set $\mathcal{U}$ of UEs, being *i)* $\mathcal{U}^m \subseteq \mathcal{U}$ the subset of UEs served by the RAN slice m, *ii)* $\mathcal{U}_i \subseteq \mathcal{U}$ the subset of UEs served by the cell $i \in \mathcal{I}$, and *iii)* $\mathcal{U}_i^m = \mathcal{U}^m \cap \mathcal{U}_i$ the intersection of both subsets. Finally, we consider each UE is attached to the nearest cell.

B. uRLLC Traffic Model

We assume the traffic demand of each RAN slice is non-uniformly distributed over the considered RAN infrastructure. Focusing on the traffic demand of a single RAN slice $m \in \mathcal{M}$, we consider statistical distributions for its arrival packet rate and the size of their packets.

Regarding the packet arrival, we assume the RAN slice m transmits packets to their UEs in batches (i.e., simultaneous transmission of packets for multiple UEs). Specifically, we consider the RAN slice m has an average of λ_m batch arrivals per unit time following a Poisson distribution. Since a Poisson process can be split into independent processes [28], we can also express the average batch arrival for each gNB as $\lambda_{i,m} = \omega_{i,m}\lambda_m$. The variable $\omega_{i,m}$ denotes the probability an UE $u \in \mathcal{U}^m$ is served in the cell $i \in \mathcal{I}$.

Focusing on a individual batch arrival, a set of $N_{i,m}^{pkt}$ packets are simultaneously generated. $N_{i,m}^{pkt}$ is a discrete random variable which ranges from 1 to $|\mathcal{U}_i^m|$, and follows an arbitrary distribution with Probability Mass Function (PMF) $p_{i,m,u}$. With respect to the transmitted packets in a RAN slice m, each packet presents a specific size of $L \in \mathcal{L}^m$ bits. L is also a discrete random variable which follows an arbitrary distribution with PMF $p_{m,l}$.

We assume each cell uses a packet scheduler per RAN slice. The scheduling policy used by each scheduler is the Earliest Deadline First (EDF). This discipline is the optimal policy for scheduling delay-sensitive traffic [29]. We also assume all the packets for the RAN slice m have the same deadline W_m^{th}. Since W_m^{th} is the same for all the packets, the EDF discipline is equivalent to the First In First Out (FIFO) policy.

Finally, we consider the MNO also imposes a value ε_m' for the violation probability before signing the Service Level Agreement (SLA) with the tenant which requests the RAN slice m. The violation probability is the probability that an individual uRLLC packet suffers a transmission delay above the packet delay budget W_m^{th}.

C. Radio Resource Model

We assume OFDMA as the accessing scheme for the cells. Focusing on the cell i, it supports a total bandwidth BW_i. In turn, this bandwidth is divided into N_i OFDM subcarriers, which are grouped in groups of $N_{SC} = 12$ subcarriers. Each group defines a Resource Block (RB), which is the smallest unit of resources that can be allocated to a UE. The set of RBs available in the cell i during each timeslot t^{slot} is denoted by $\mathcal{R}_i$, and the amount of these RBs is given by Eq. (16). Since a cell supports scalable numerologies ($\mu_{5G} = 0, 1, \ldots, 4$), *i*) the duration of a timeslot is computed as $t^{slot} = 10^{-3}/2^{\mu_{5G}}$ seconds, and *ii*) the subcarrier spacing is derived as $\Delta_f = 2^{\mu_{5G}} \cdot 15$ KHz. The parameter OH denotes the overhead factor due to control plane data [30].

$$|\mathcal{R}_i| = \left\lfloor \frac{BW_i}{N_{SC}\Delta f} (1 - OH) \right\rfloor \tag{16}$$

Finally, we denote *i*) $\mathcal{R}_i^m \subseteq \mathcal{R}_i$ as the subset of RBs allocated to the RAN slice m in cell i, and *ii*) $\mathcal{R}_u \subseteq \mathcal{R}_i^m$ as the subset of RBs allocated for the UE u.

D. Channel Model

To measure the channel quality perceived by an arbitrary UE $u \in \mathcal{U}_i^m$ in the RB r, we consider the instantaneous SINR, i.e., $\gamma_{u,r}^{(t)}$. This parameter is described in Eq. (17), where P_{TX} is the transmitted power in a single RB. We assume all the gNBs transmit the same power for each RB. The parameter h_r denotes the gain due to the fast fading. We consider Rayleigh fading, thus h_r is exponentially distributed with mean one. The parameter χ is the gain due to the shadowing and follows a log-normal distribution characterized by the mean μ_χ and standard deviation σ_χ, both in dB. The parameter $d_{u,i}$ denotes the distance from the gNB i to the UE u, with α the pathloss exponent. The summation terms in the denominator gather the interference suffered by the UE u in the RB r, and P_N is the noise power measured in a single RB. Focusing on a specific interference term j, the parameter $\xi_{j,r}$ denotes a binary variable that takes the value 1 when the neighbor cell j transmits data in the RB r and the value 0 otherwise. The value for $\xi_{j,r}$ will depend on the radio resource allocation performed by the corresponding gNBs in the neighbor cells. We consider that only the RBs allocated to a specific RAN slice can be scheduled to its UEs.

$$\gamma_{u,r}^{(t)} = \frac{P_{TX} h_r \chi d_{u,i}^{-\alpha}}{\sum_{j \in \mathcal{I} \setminus \{i\}} \xi_{j,r} P_{TX} h_r \chi d_{u,j}^{-\alpha} + P_N} \tag{17}$$

Considering $\gamma_{u,r}^{(t)}$ is measured for a considerable amount of time for the UEs attached in the cell i and served by the RAN slice m, we can obtain the CDF of the SINR to model the channel quality for this RAN slice in this cell. In the literature, stochastic geometry has been used as a powerful analytical tool for modeling the CDF of the SINR [31]. In this work, we rely on several state-of-the-art works on stochastic geometry (i.e., [32]–[36]) to provide a closed-form expression for the CDF of the SINR. In [32], the authors have defined this CDF as Eq. (18) shows. The parameter γ_{th} is the target SINR. The parameter p_{suc} is the probability that an arbitrary UE perceives a SINR $\gamma_{u,r}$ higher than γ_{th}, whereas p_{sel} is the probability the gNB schedules a RB to transmit data for this UE.

$$f_{CDF}(\gamma_{u,r}, \gamma_{th}) = P[\gamma_{u,r} \leq \gamma_{th}] = 1 - p_{suc} \cdot p_{sel} \tag{18}$$

From the results of [33], and considering the pathloss exponent is $\alpha = 4$, we get p_{suc} by Eq. (19). In this equation $\kappa_{gNB,\chi} = \kappa_{gNB} E\left[\sqrt{\chi}\right]$, where κ_{gNB} is the gNB density and $E\left[\sqrt{\chi}\right]$ is the fractional moment of the log-normal distribution. $E\left[\sqrt{\chi}\right]$ models the shadowing channel power as Eq. (20) shows. The authors of [35] use this fractional moment to incorporate the shadowing effect in Eq. (18). The parameter $\kappa_{j,\chi} = \kappa_j E\left[\sqrt{\chi}\right]$, where κ_j denotes the density of the neighbor cells interfering in an arbitrary RB. It is defined as $\kappa_j = \kappa_{gNB}\left(1 - p_{off}\right)$, where the parameter p_{off} is the probability that a generic cell does not transmit in a RB. The parameter $\upsilon\left(\gamma_{th}\right)$ is given by Eq. (21). Finally, the function $Q(\cdot)$ denotes the Gaussian Q-function.

$$p_{suc} = \sqrt{\frac{\pi P_{TX}}{\gamma_{th} P_N}} \exp\left(\frac{\left(\pi\left[\kappa_{gNB,\chi} + \kappa_{j,\chi}\upsilon\left(\gamma_{th}\right)\right]\right)^2 P_{TX}}{4\gamma_{th} P_N}\right) Q\left(\frac{\pi\left[\kappa_{gNB,\chi} + \kappa_{j,\chi}\upsilon\left(\gamma_{th}\right)\right]}{\sqrt{\frac{2\gamma_{th} P_N}{P_{TX}}}}\right) \tag{19}$$

$$E\left[\sqrt{\chi}\right] = \exp\left(\frac{\ln(10)\,\mu_\chi}{20} + \frac{1}{2}\left(\frac{\ln(10)\,\sigma_\chi}{20}\right)^2\right) \tag{20}$$

$$\upsilon\left(\gamma_{th}\right) = \sqrt{\gamma_{th}}\left[\frac{\pi}{2} - \arctan\left(\frac{1}{\sqrt{\gamma_{th}}}\right)\right] \tag{21}$$

To compute p_{sel} and p_{off}, we need to assume a specific model for the cell load (i.e., the fraction of RBs that are being scheduled on average for the attached UEs). In [32], the authors assume each cell has available only one RB whereas the authors of [34] extend the definition of p_{sel} and p_{off} given in [32] by considering an arbitrary number of RBs in each cell. Based on the cell load model described in [34, Section II.E], and assuming the density of UEs κ_{UEs} is much higher than the density of cells (i.e., $\kappa_{UEs} \gg \kappa_{gNBs}$), we have in Eqs. (22) and (23) the mathematical expressions for p_{sel} and p_{off}. Note that, $\Gamma(\cdot)$ denotes the gamma function.

$$p_{sel} = |\mathcal{R}_i| \left(\frac{\kappa_{UEs}}{\kappa_{gNB,\chi}} \right)^{-1} \tag{22}$$

$$p_{off} = \frac{4}{63} \frac{3.5^{3.5}}{\Gamma(3.5)} \frac{\Gamma(4.5 + |\mathcal{R}_i|)}{\Gamma(1 + |\mathcal{R}_i|)} \left(\frac{\kappa_{UEs}}{\kappa_{gNB,\chi}} \right)^{-3.5} \tag{23}$$

These parameters were formulated under the assumption the UE density is the same in the entire RAN. However, a RAN slice could present a different amount of UEs in each cell. Furthermore, the UE density for each RAN slice could be different. For this reason, we define the UE density for the RAN slice m in the cell i as $\kappa_{UE,i,m}$. In addition, the authors of [34] consider the UEs attached to a cell are randomly chosen for transmission in an arbitrary RB. This means a cell can schedule one RB to transmit data for a single UE. In our work, we consider each UE could receive data from multiple RBs in a timeslot. Depending the RAN slice which serves this UE, the length of each packet could take a different value (see Sec. III-B). Furthermore, each transmitted packet could require a specific amount of RBs according to the channel quality. All this involves computing the density of the required amount of RBs for transmitting a packet in a specific RAN slice and cell, i.e., $\kappa_{RBs,i,m}$ instead of $\kappa_{UEs,i,m}$. For simplicity, we consider the average number of required RBs, i.e., $\overline{R}_{i,m}^{pkt}$, to compute $\kappa_{RBs,i,m}$ as Eq. (24) shows.

$$\kappa_{RBs,i,m} = \kappa_{UEs,i,m} \overline{R}_{i,m}^{pkt} \tag{24}$$

Under the above assumptions, we reformulate p_{sel} and p_{off} as Eqs. (25) and (26) show, respectively. In Eq. (26), the second summation gathers the probability that a neighbor gNB j does not transmit in a RB for each RAN slice. Since the result of the second summation is different for each neighbor gNB j, we sum the weighted result of each neighbor gNB. To consider the impact of each neighbor gNB, we define the weight ι_j according to the pathloss from the neighbor gNB j to the considered gNB i, i.e., $\iota_j = d_{i,j}^{-\alpha} / \sum_{j \in \mathcal{I} \setminus \{i\}} d_{i,j}^{-\alpha}$.

$$p_{sel} = |\mathcal{R}_i^m| \left(\frac{\kappa_{RBs,i,m}}{\kappa_{gNB,\chi}} \right)^{-1} \tag{25}$$

$$p_{off} = \sum_{j \in \mathcal{I} \setminus \{i\}} \iota_j \sum_{m \in \mathcal{M}} \frac{4}{63} \frac{3.5^{3.5}}{\Gamma(3.5)} \frac{\Gamma\left(4.5 + |\mathcal{R}_j^m|\right)}{\Gamma\left(1 + |\mathcal{R}_j^m|\right)} \left(\frac{\kappa_{RBs,j,m}}{\kappa_{gNB,\chi}} \right)^{-3.5} \tag{26}$$

Once the CDF for the SINR $f_{CDF}(\gamma_{u,r}, \gamma_{th})$ is known, we can compute the PMF for the spectral efficiency $\pi_{z,i,m}$ achieved by an UE served by the RAN slice m in the cell i. Note that, an arbitrary UE could achieve N_z values for the spectral efficiency in a time slot. In this work, we denote each possible value as $SE_{z,i,m}$ and it depends on the Modulation and Coding Scheme (MCS) pair selected by the gNB after this UE reports the corresponding Channel Quality Indicator (CQI), i.e., the maximum spectral efficiency for which the Block Error Rate (BLER) is equal or below 10 %. The N_z values for $SE_{z,i,m}$ $\forall z \in \mathcal{Z}$ can be found in [37, Table 5.2.2.1-3].

To compute the probabilities $\pi_{z,i,m}$ of reporting certain CQIs, i.e., the PMF of $SE_{z,i,m}$, we can use the CDF of the SINR $f_{CDF}(\gamma_{u,r}, \gamma_{th})$ as Eq. (27) shows. In this equation, $f_{SE\rightarrow\gamma}(SE)$ is a function which translates the spectral efficiency SE to the instantaneous SINR γ. In a first attempt, we could consider the Shannon's capacity formula to perform this translation. However, since uRLLC packets are very short, the achievable spectral efficiency cannot be accurately capture by this formula [20]. Instead, the spectral efficiency in uRLLC falls in the finite blocklength channel coding regime, which is derived as Eq. (28) shows [38]. In this equation, $f^{-1}_{SE\rightarrow\gamma}(\gamma)$ represents the inverse of $f_{SE\rightarrow\gamma}(SE)$, i.e., it provides the achievable spectral efficiency SE given an instantaneous SINR γ. Furthermore, $Q^{-1}(\cdot)$ is the inverse of the Gaussian Q-function. The parameter ε_{dec} is the decoding error probability, which could range from 10^{-3} to 10^{-7} [39]. Finally, n_{block} denotes the blocklength. In practice, we can tabulate the function $f^{-1}_{SE\rightarrow\gamma}(\gamma)$ to map the achievable spectral efficiency SE to the instantaneous SINR γ and vice versa.

$$\pi_{z,i,m} = \begin{cases} f_{CDF}(\gamma_{u,r}, f_{SE\rightarrow\gamma}(SE_{1,i,m})) & \text{if } z = 1 \\ f_{CDF}(\gamma_{u,r}, f_{SE\rightarrow\gamma}(SE_{z+1,i,m})) - f_{CDF}(\gamma_{u,r}, f_{SE\rightarrow\gamma}(SE_{z,i,m})) & \text{if } 1 < z < N_z \\ 1 - f_{CDF}(\gamma_{u,r}, f_{SE\rightarrow\gamma}(SE_{N_z,i,m})) & \text{if } z = N_z \end{cases} \tag{27}$$

$$f^{-1}_{SE\rightarrow\gamma}(\gamma) = \log_2(1+\gamma) - \sqrt{\frac{1-\frac{1}{(1+\gamma)^2}}{n_{block}}} Q^{-1}(\varepsilon_{dec}) \log_2(e) \tag{28}$$

Since $\pi_{z,i,m}$ is independent from the PMFs for the packet length, i.e., $p_{m,l}$ $\forall m \in \mathcal{M}$, we can compute the amount of required RBs for transmitting a packet as Eq. (29) defines. The variable r' denotes a specific combination of a packet size $L \in \mathcal{L}^m$ and a spectral efficiency value $SE_{z,i,m}$. The PMF of the random variable $R^{pkt}_{r',i,m}$, i.e., $p_{r'}$ is the joint PMF of the random variables L and $SE_{z,i,m}$, whose PMFs are $p_{m,l}$ and $\pi_{z,i,m}$, respectively.

$$R^{pkt}_{r',i,m} = \left\lceil \frac{L_{r'}}{t^{slot} N_{SC} \Delta f SE_{r'}} \right\rceil \tag{29}$$

Finally, we can compute the average number of required RBs for transmitting a packet as Eq. (30) shows.

$$\overline{R}_{i,m}^{pkt} = \sum_{r' \in \mathcal{R}_m^{req}} p_{r'} R_{r',i,m}^{pkt} \tag{30}$$

Note that, there is a recursive relationship between $\overline{R}_{i,m}^{pkt}$ and $\pi_{z,i,m}$. On the one hand, $\overline{R}_{i,m}^{pkt}$ depends on $\pi_{z,i,m}$. Specifically, $p_{r'}$ in Eq. (30) depends on $p_{m,l}$ and $\pi_{z,i,m}$. On the other hand, $\pi_{z,i,m}$ is computed by using f_{CDF} as Eq. (27) shows. Among other things, this CDF depends on the parameter $\kappa_{RBs,i,m}$, which in turn it depends on $\overline{R}_{i,m}^{pkt}$ as Eq. (24) defines. To solve this recursion, we need first to assume initial values for $\pi_{z,i,m}$ and then, we need to iteratively recompute $\overline{R}_{i,m}^{pkt}$ and $\pi_{z,i,m}$ until these variables converge.

IV. SNC-based Model for an uRLLC RAN slice in a cell

In this section, we propose a SNC-based model to determine the amount of RBs $|\mathcal{R}_i^m|$, which the RAN slice orchestrator defines as dedicated radio resource quota for the RAN slice m in the cell i. Considering $|\mathcal{R}_i^m|$ and the violation probability ε_m', the SNC-based model provides the delay bound $W_{i,m}$ and backlog bound $B_{i,m}$ which guarantee ε_m' for this RAN slice in this cell. To that end, the proposed model considers different stochastic processes, which characterize: *i*) the DL traffic of an uRLLC RAN slice, and *ii*) the capacity the serving cell provides for such RAN slice.

Under this context, we first describe the traffic model for an uRLLC RAN slice. Then, we present the service model for the cell capacity available for such RAN slice. Finally, we use SNC to derive the mathematical expressions for the backlog and delay bounds.

A. Traffic Model for an uRLLC RAN Slice

The arrival process $A_{i,m}(\tau,t)$ is the accumulative number of bits which arrive from the core network to the gNB i for the RAN slice m. This stochastic process is defined in Eq. (31), where $N_{i,m}^{batch}(t)$ denotes the number of batch arrivals during the interval $[0,t]$. The random variable y_b denotes the amount of bits generated in a batch arrival.

$$A_{i,m}(\tau,t) = \sum_{b=1}^{N_{i,m}^{batch}(t)} y_b \tag{31}$$

In a batch arrival, the number of bits which arrives to the corresponding gNB is given by Eq. (32). The parameter $N^{pkt}_{i,m,b}$ defines the number of packets which simultaneously arrive for different UEs. For each packet, the parameter $L_n \in \mathcal{L}^m$ represents its size in bits.

$$y_b = \sum_{n=1}^{N^{pkt}_{i,m,b}} L_n \tag{32}$$

We compute the MGF of $A_{i,m}(\tau,t)$ as Eq. (33) shows. By substitution of $\ln\left(M_{y_b}(\theta)\right) = \nu$, we have the MGF of the Poisson process $N^{batch}_{i,m}(t)$ in function of the free parameter ν.

$$\begin{aligned} M_{A_{i,m}}(\theta) &= E[e^{\theta A_{i,m}(\tau,t)}] = E[(M_{y_b}(\theta))^{N^{batch}_{i,m}(t)}] = E[e^{N^{batch}_{i,m}(t)\cdot\ln\left(M_{y_b}(\theta)\right)}] = E\left[e^{\nu N^{batch}_{i,m}(t)}\right] \\ &= M_{N^{batch}_{i,m}}(\nu) = e^{\lambda_{i,m}t(e^{\nu}-1)} \end{aligned} \tag{33}$$

If we express the resulting MGF in function of the free parameter θ as Eq. (34) shows, we observe which the MGF of $A_{i,m}(\tau,t)$ depends on the MGF of y_b, i.e., $M_{y_b}(\theta)$.

$$M_{A_{i,m}}(\theta) = e^{\lambda_{i,m}\left(M_{y_b}(\theta)-1\right)t} \tag{34}$$

We compute the MGF of y_b as Eq. (35) shows. The resulting expression is equal to the MGF of $N^{pkt}_{i,m,b}$ with free parameter ν, i.e., $M_{N^{pkt}_{i,m,b}}(\nu)$.

$$M_{y_b}(\theta) = E[e^{\theta y_b}] = E[(M_{l_n}(\theta))^{N^{pkt}_{i,m,b}}] = E[e^{N^{pkt}_{i,m,b}\cdot\ln\left(M_{l_n}(\theta)\right)}] = E\left[e^{\nu N^{pkt}_{i,m,b}}\right] = M_{N^{pkt}_{i,m,b}}(\nu) \tag{35}$$

Using the definition of MGF (see Section II-B), we define the MGF of $N^{pkt}_{i,m,b}$ as Eq. (36) shows. If we express the resulting MGF in function of the free parameter θ, we observe which the MGF of y_b depends on the MGF $M_{L_n}(\theta)$ of L_n, as Eq. (37) shows.

$$M_{N^{pkt}_{i,m,b}}(\nu) = \sum_{u=1}^{|\mathcal{U}^m_i|} e^{\nu u} p_{i,m,u} \tag{36}$$

$$M_{y_b}(\theta) = \sum_{u=1}^{|\mathcal{U}^m_i|} \left[M_{L_n}(\theta)\right]^u p_{i,m,u} \tag{37}$$

Similarly to Eq. (36), we define the MGF of L_n in Eq. (38).

$$M_{L_n}(\theta) = \sum_{l\in\mathcal{L}^m} e^{\theta l} p_{m,l} \tag{38}$$

If we include this expression in Eq. (37), and then we replace the resulting expression in Eq. (34), we obtain the MGF of $A_{i,m}(\tau,t)$ as Eq. (39) shows.

$$M_{A_{i,m}}(\theta) = e^{\lambda_{i,m}\left(\sum_{u=1}^{|\mathcal{U}^m_i|}\left[\sum_{l\in\mathcal{L}^m} e^{\theta l} p_{m,l}\right]^u p_{i,m,u}-1\right)t} \tag{39}$$

Finally, by equaling the right side of Eq. (7) with Eq. (39), we obtain Eq. (40). In this expression, we define the parameters $\rho_{A_{i,m}}$ and $\sigma_{A_{i,m}}$ of the affine arrival envelope $\alpha_{i,m}(\tau,t) = \left(\rho_{A_{i,m}} + \delta\right)[t-\tau] + \sigma_{A_{i,m}}$ which bounds the arrival process $A_{i,m}(\tau,t)$.

$$\sigma_{A_{i,m}} = 0 \tag{40a}$$

$$\rho_{A_{i,m}} = \frac{\lambda_{i,m}}{\theta}\left[\sum_{u=1}^{|\mathcal{U}_i^m|}\left[\sum_{l\in\mathcal{L}^m} e^{\theta l} p_{m,l}\right]^u p_{i,m,u} - 1\right] \tag{40b}$$

B. Service Model for a RAN Slice

The service process $S_{i,m}(\tau,t)$ is the accumulative number of bits which could be processed by the cell i for the RAN slice m. This process is described in Eq. (41).

$$S_{i,m}(t) = \sum_{n=0}^{N^{slot}(t)} C_{i,m}(n) \tag{41}$$

The deterministic variable $N^{slot}(t) = t/t^{slot}$ represents the number of accumulated timeslots in t. The random variable $C_{i,m}(n)$ denotes the amount of bits which the corresponding gNB could transmit in the timeslot n if all the allocated RBs for the RAN slice m were used. $C_{i,m}(n)$ depends on (a) the amount of allocated RBs for this RAN slice, i.e., $|\mathcal{R}_i^m|$; (b) the size $L\in\mathcal{L}^m$ of each transmitted uRLLC packet and its PMF $p_{m,l}$; and (c) the spectral efficiency $SE_{z,i,m}$ to transmit these packets and its PMF $\pi_{z,i,m}$. Considering (a)-(c), we can obtain all the possible values for $C_{i,m}(n)$, i.e., c_q $\forall q\in\mathcal{Q}_i^m$, and its PMF p_q.

For comprehensibility purposes, we provide an illustrative example in Fig. 2 to explain which $C_{i,m}(n)$ means. In this example, we assume all the uRLLC packets have a size of 1024 bits. Furthermore, we consider the achievable spectral efficiency takes the following values $SE_{i,m} =$ [2, 4.5, 6] bps/Hz. Additionally, we assume the RAN slice has allocated $|\mathcal{R}_i^m| = 3$ RBs. Based on these assumptions, we obtain by Eq. (29) the number of RBs consumed by an uRLLC packet could be $R^{pkt}_{r',i,m} =$ [1, 2, 3] RBs. This means the number of transmitted bits per RB could be 1024, 512 or 341, respectively. Under this scenario, we illustrate some cases where the gNB uses all the available RBs to transmit uRLLC packets for the considered RAN slice. For each case, we show the amount of transmitted bits. For instance, the gNB transmits $c_4 = 341+512+1024 = 1877$ bits in the scenario E. In this example, the scenarios A-J illustrate all the possible values for $C_{i,m}(n) = c_q \in [c_1,\ldots,c_{10}]$. Notwithstanding, there exists other scenarios where the gNB could

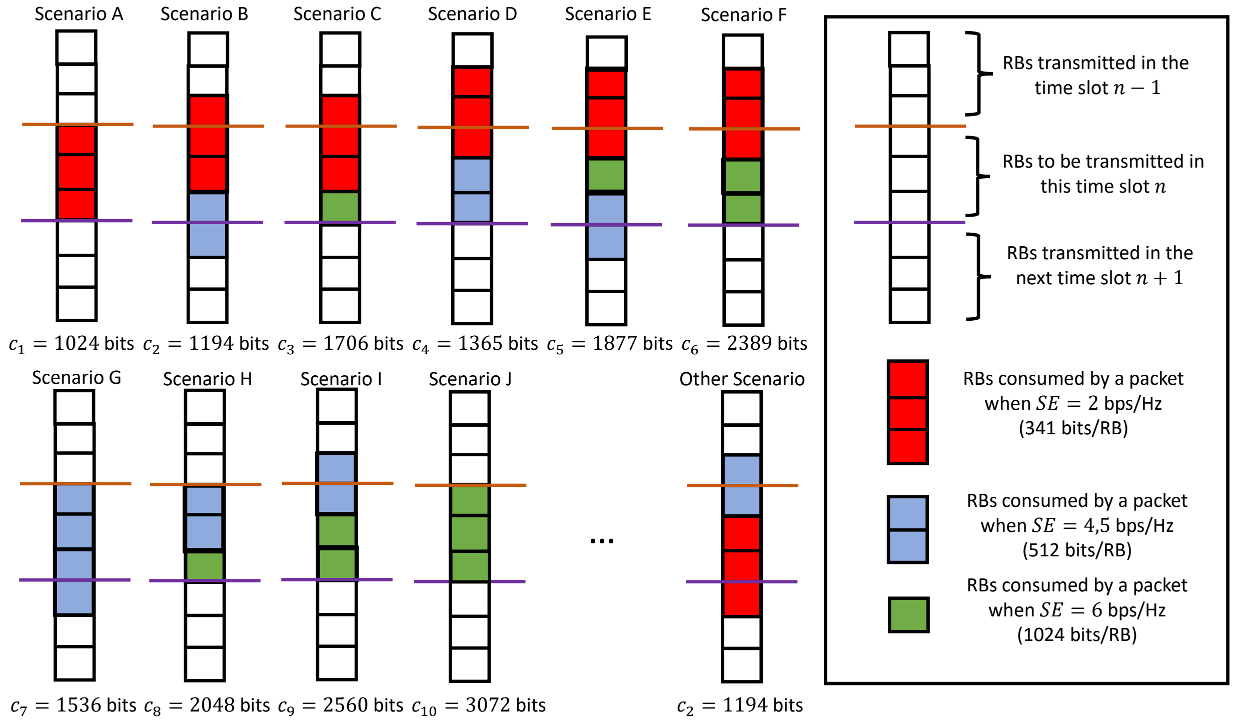


Fig. 2: Example which illustrates the capacity provided by a gNB to a RAN slice in a time slot for different scenarios.

transmit the same amount of bits, for instance scenarios B and *Other Scenario* where the gNB transmits 1194 bits. In our work, the definition of a model to obtain all the possible values for c_q and their probabilities p_q is out of the scope. Instead, we estimate them by Montecarlo simulations.

Using the definition of MGF (see Section II-B), we define the negative MGF for $C_{i,m}(n)$ as Eq. (42) shows.

$$M_{C_{i,m}}(-\theta) = E\left[e^{-\theta C_{i,m}(n)}\right] = \sum_{q\in\mathcal{Q}_i^m} e^{-\theta c_q} p_q \tag{42}$$

Considering $M_{C_{i,m}}(-\theta)$, we can compute the negative MGF of the service process $S_{i,m}(\tau,t)$ as Eq. (43) shows.

$$\begin{aligned} M_{S_{i,m}}(-\theta) &= E\left[e^{-\theta S_{i,m}(\tau,t)}\right] = E\left[\left(M_{C_{i,m}}(-\theta)\right)^{N^{slot}(t)}\right] = E\left[e^{N^{slot}(t)\ln\left(M_{C_{i,m}}(-\theta)\right)}\right] \\ &= e^{\frac{\ln\left(\sum_{q\in\mathcal{Q}_i^m} e^{-\theta c_q} p_q\right)}{t^{slot}} t} \end{aligned} \tag{43}$$

Finally, by equaling the right side of Eq. (12) with Eq. (43), we obtain Eq. (44). In this expression, we define the parameters $\rho_{S_{i,m}}$ and $\sigma_{S_{i,m}}$ of the affine service envelope $\beta_{i,m}(\tau,t) =$

$\left(\rho_{S_{i,m}} - \delta\right)[t-\tau] + \sigma_{S_{i,m}}$ which bound the service process $S_{i,m}(\tau,t)$.

$$\sigma_{S_{i,m}} = 0 \tag{44a}$$

$$\rho_{S_{i,m}} = \frac{-1}{\theta t^{slot}} \ln\left(\sum_{q\in\mathcal{Q}_i^m} e^{-\theta c_q} p_q\right) \tag{44b}$$

C. Backlog and Delay Bounds for a RAN Slice

To compute the backlog and delay bounds for the RAN slice m in the cell i, we assume the violation probability ε'_m is equally distributed into the overflow and deficit profiles, i.e., $\varepsilon'_{A_m} = \varepsilon'_{S_m} = \varepsilon'_m/2$. Considering that, we obtain in Eqs. (45) and (46) the backlog and delay bounds by (a) applying Eqs. (40a) and (44a) into Eqs. (11) and (15), respectively; and (b) using them along with Eqs. (40b) and (44b) into Eqs. (5) and (6), respectively.

$$\begin{aligned} B_{i,m} = \frac{1}{\theta}\left[\ln\left(\frac{\varepsilon'_m}{2}\right) + \ln\left(1-e^{-\theta\delta}\right)\right] \cdot \\ \left[\frac{\lambda_{i,m} t^{slot}\left(\sum_{u=1}^{|\mathcal{U}|_i^m}\left[\sum_{l\in\mathcal{L}^m} e^{\theta l} p_{m,l}\right]^u p_{i,m,u} - 1\right) + \delta\theta t^{slot}}{\ln\left(\sum_{q\in\mathcal{Q}_i^m} e^{-\theta c_q} p_q\right) + \delta\theta t^{slot}} - 1\right] \end{aligned} \tag{45}$$

$$W_{i,m} = \frac{2t^{slot}\left[\ln\left(\frac{\varepsilon'_m}{2}\right) + \ln\left(1-e^{-\theta\delta}\right)\right]}{\ln\left(\sum_{q\in\mathcal{Q}_i^m} e^{-\theta c_q} p_q\right) + \delta\theta t^{slot}} \tag{46}$$

V. Radio Resource Planning for RAN slices

Considering the proposed SNC-based model, we design a radio resource planning scheme. The RAN slice orchestrator will execute it to decide the dedicated radio resource quotas assigned for each RAN slice in each cell. First, we present the problem formulation. Then, we provide heuristics for solving this problem.

A. Problem Formulation

When the RAN slice orchestrator plans the deployment of several uRLLC RAN slices, it must guarantee that their performance requirements are met in the long term. Specifically, this means the probability that a packet k transmitted for the RAN slice m suffers a delay W_k above the packet delay budget is less than or equal to the violation probability, i.e., $\mathrm{P}\left[W_k > W_m^{th}\right] \leq \varepsilon'_m$. To meet this condition, the RAN slice orchestrator must set a dedicated radio resource quota per RAN slice and cell.

Assuming the RAN slice orchestrator establishes a specific set of RBs for each pair of RAN slice and cell, i.e., $|\mathcal{R}_i^m|$, each RAN slice m will present in each cell i an upper bound for the packet latency $W_{i,m}$ in such a way that $\mathrm{P}[W_k > W_{i,m}] = \varepsilon'_m$. With the aim of characterizing how this upper bound is close to the packet delay budget W_m^{th}, we define the parameter $\Delta W_{i,m}$ in Eq. (47). Note that $\Delta W_{i,m}$ is zero if the upper bound $W_{i,m}$ is less than the packet delay budget.

$$\Delta W_{i,m} = \max\left(W_{i,m} - W_m^{th}, 0\right) \tag{47}$$

If we consider $\Delta W_{i,m}$ in all the cells where the RAN slice m must provide the uRLLC service, we can compute the average for this parameter as Eq. (48) shows.

$$\overline{\Delta W}_m = \sum_{i \in \mathcal{I}^m} \omega_{i,m} \Delta W_{i,m} \tag{48}$$

The RAN slice orchestrator aims to compute the dedicated radio resource quotas $|\mathcal{R}_i^m|$ $\forall i \in \mathcal{I}$ $\forall m \in \mathcal{M}$ in such a way that $\overline{\Delta W}_m$ was zero for all the RAN slices. If this parameter was not zero for one or more RAN slices using all the available RBs, this would mean that the entire RAN infrastructure has not enough capacity to satisfy the latency requirements of these RAN slices. In this scenario, the MNO must prioritize which RAN slices will achieve a delay bound closer to their required packet delay bounds. To that end, we define the RAN slice priority ψ_m as a potential parameter which the MNO may tune for instance, considering an economic-based policy. With the purpose of minimizing $\psi_{m'}\overline{\Delta W}_{m'}$ for the RAN slice m' which presents the greatest value for such parameter, we formulate our problem as follows.

$$\underset{|\mathcal{R}_i^m|}{\text{minimize}} \quad \max\left(\psi_1 \overline{\Delta W}_1, \dots \psi_m \overline{\Delta W}_m, \dots \psi_{|\mathcal{M}|} \overline{\Delta W}_{|\mathcal{M}|}\right), \tag{49}$$

$$\text{subject to:} \quad \sum_{m \in \mathcal{M}} \psi_m = 1, \tag{50}$$

$$\sum_{m \in \mathcal{M}} |\mathcal{R}_i^m| \leq |\mathcal{R}_i| \tag{51}$$

B. Heuristic Algorithm Design

The objective function defined in Eq. (49) is a non-convex function and thus with at least NP-hard complexity. This fact is mainly due to the inter-cell interference. For instance, when more RBs are allocated for the RAN slice m in the cell i, the delay bound $W_{i,m}$ decreases. However, the inter-cell interference increases in each neighbor cell $j \in \mathcal{I} \setminus \{i\}$, meaning the delay bounds

$W_{j,m}$ for all the RAN slices in the neighbor cells increase, and thus the objective function may also increase. This involves the existence of multiple local minimums in the considered search space. Solving the formulated can be see as a combinatorial optimization, i.e., searching the best combination of allocated radio resources $|\mathcal{R}_i^m|$ for each RAN slice $m \in \mathcal{M}$ in each cell $i \in \mathcal{I}$ while the cost function is minimized. Performing an exhaustive search to find the optimal solution is not computationally tractable, specially when the number of cells $|\mathcal{I}|$, RBs per cell $|\mathcal{R}_i|$ and RAN slices $|\mathcal{M}|$ is considerably high. As an alternative, searching a local optimum is a better option. To that end, we propose the heuristics described in Algorithm 1. Given the performance requirements and the traffic demand of each RAN slice $m \in \mathcal{M}$ (line 1), Algorithm 1 provides the steps performed by the RAN slice orchestrator to determine the dedicated radio resource quotas for these RAN slices.

First, the algorithm equally distributes the available RBs in each cell among all the RAN slices (line 2). Furthermore the parameter $N_{not\_imp}^{ite}$ is set to zero. This parameter indicates the number of consecutive iterations which are not valid (see while loop). Based on the initial RB allocation, the algorithm derives (line 3): *i*) the PMF of $SE_{z,i,m}$ (i.e., the probabilities of reporting a certain CQI); *ii*) the possible amount of required RBs to transmit a packet; and *iii*) the PMF for this random variable. These parameters are used by the algorithm as inputs for the proposed SNC-based model (line 4). Using this model, the algorithm estimates the delay bound $W_{i,m}$ for each RAN slice and each cell. Additionally, during the SNC-based model execution, the algorithm (a) estimates the gNB capacity for a RAN slice (i.e., computing p_q and c_q $\forall q \in \mathcal{Q}_i^m$); and (b) optimizes the free parameters θ and δ to obtain $W_{i,m}$. Next, the algorithm computes the difference between the estimated delay bounds and the packet delay budget for each RAN slice (line 5). Furthermore, the algorithm averages these differences for each RAN slice. With these parameters, the algorithm evaluates the function defined in Eq. (49).

In the following steps (lines 6-20), the algorithm iteratively redistributes the RBs allocated for each RAN slice in each cell with the aim of minimizing the function defined in Eq. (49). Focusing on a single iteration, the algorithm redistributes one RB between two RAN slices in a single cell. To that end, the algorithm first determines the RAN slice m' which will receive a new RB (line 7). It is the one which maximize Eq. (49). Then, the algorithm decides the cell where one RB will be redistributed (line 8). This cell will be the one where the weighted difference between the estimated delay bound and the target packet delay bound is greater. To that end, we use the function $\text{sort}_{desc}(\cdot)$ to sort in descending direction an array where each element is given

Algorithm 1: Radio Resource Planning for $|\mathcal{M}|$ RAN slices

```
1  **Inputs:** Performance requirements (i.e., $W_m^{th}$ and $\varepsilon_m'$) and traffic distribution (i.e., $|\mathcal{U}_i^m|$,
     $p_{i,m,u}$, $\lambda_m$, $L \in \mathcal{L}^m$, and $p_{m,l}$) for each RAN slice $m \in \mathcal{M}$;
2  **Initialization:** Equal distribution of the RBs among the RAN slices. Set $N_{not\_imp}^{ite} = 0$;
3  Compute $\pi_{z,i,m}$, $R_{r',i,m}^{pkt}$, and $p_{r'}$ $\forall i \in \mathcal{I}$; $\forall m \in \mathcal{M}$. See Eqs. (27) and (29);
4  Compute $W_{i,m}$ $\forall i \in \mathcal{I}$; $\forall m \in \mathcal{M}$ by using the proposed SNC-based model. See Eq. (46);
5  $\Delta W_{i,m}$ and $\overline{\Delta W}_m$ $\forall i \in \mathcal{I}$; $\forall m \in \mathcal{M}$. Evaluate $curr\_func =$ Eq. (49);
6  **while** $curr\_func > 0$ *and* $N_{not\_imp}^{ite} < |\mathcal{I}|$ **do**
7      Select $m' = \arg\max\left(\psi_1\overline{\Delta W}_1, \dots \psi_m\overline{\Delta W}_m, \dots \psi_{|\mathcal{M}|}\overline{\Delta W}_{|\mathcal{M}|}\right)$ $\forall m \in \mathcal{M}$;
8      Set $a$ = $\mathrm{sort}_{desc}\left(\omega_{1,m'}\Delta W_{1,m'}, \dots, \omega_{i,m'}\Delta W_{i,m'}, \dots, \omega_{|\mathcal{I}|,m'}\Delta W_{|\mathcal{I}|,m'}\right)$ $\forall i \in \mathcal{I}$. Remove the
         first $N_{not\_imp}^{ite}$ elements of $a$. Select $i' = \arg(a(1))$;
9      Select $m'' = \arg\min\left(\omega_{i',1}\Delta W_{i',1}, \dots, \omega_{i',m}\Delta W_{i',m}, \dots, \omega_{i',m}\Delta W_{i',m}\right)$ $\forall m \in \mathcal{M} \setminus \{m'\}$;
10     Redistribute one RB from RAN slice $m''$ to RAN slice $m'$, i.e., $|\mathcal{R}_{i'}^{m'}| = |\mathcal{R}_{i'}^{m'}| + 1$ and
         $|\mathcal{R}_{i''}^{m''}| = |\mathcal{R}_{i''}^{m''}| - 1$;
11     Compute $\pi_{z,i,m}$, $p_{r'}$, and $R_{r',i,m}^{pkt}$ $\forall i \in \mathcal{I}$; $\forall m \in \mathcal{M}$. See Eqs. (27) and (29);
12     Set $prev\_func = curr\_func$;
13     Compute $W_{i,m}$ $\forall i \in \mathcal{I}$; $\forall m \in \mathcal{M}$ by using the proposed SNC-based model. See Eq.
         (46);
14     $\Delta W_{i,m}$ and $\overline{\Delta W}_m$ $\forall i \in \mathcal{I}$; $\forall m \in \mathcal{M}$. Evaluate $curr\_func =$ Eq. (49);
15     **if** $curr\_func < prev\_func$ **then**
16         Set $N_{not\_imp}^{ite} = 0$. Variables computed in lines 10-14 are valid;
17     **else**
18         Set $N_{not\_imp}^{ite} = N_{not\_imp}^{ite} + 1$. Variables computed in lines 10-14 are invalid;
19     **end**
20 **end**
21 **return:** $\mathcal{R}_i^m$ $\forall i \in \mathcal{I}$; $\forall m \in \mathcal{M}$
```

by $\omega_{i,m'}\Delta W_{i,m'}$. Then, we remove the first $N_{not\_imp}^{ite}$ elements of the sorted array, and select the first of the remaining elements. Note that if one or more consecutive iterations cannot improve the current value of Eq. (49), the cells selected in the previous iterations cannot be considered in the new iteration. This means the $N_{not\_imp}^{ite}$ cells which provide the greatest values for the weighted difference are not considered. Later, the algorithm decides the RAN slice m'' which will donates one RB to the RAN slice m' (line 9). This RAN slice will be the one which has the lowest weighted difference between the estimated delay bound and the packet delay budget. When the cell and the RAN slices involved in the RB redistribution are determined, the algorithm computes the new amount of RBs in such RAN slices (line 10). After that, the algorithm recomputes (lines 11-14): the PMFs for the spectral efficiency and the required amount of RBs per transmitted packet; the delay bound using the SNC-based model; and the value of the function defined in Eq. (49). Then, the algorithm checks if the value of this function has decreased with respect to its value in the previous iteration (lines 15-19). If yes, the algorithm considers the variables computed in the lines 10-14 are valid. If not, these variables are invalid, and the algorithm does not update them in this iteration. The algorithm stops when the value of Eq. (49) is either equal to zero or can no longer be minimized.

VI. Numerical Results and Discussions

In this section, we (a) validate the proposed SNC-based model and (b) evaluate the performance of the proposed heuristics, comparing it with two reference solutions. The reference solution #1 consists of the RAN slice orchestrator establishes the dedicated radio resource quotas by allocating proportionally the RBs according to the traffic demand of each RAN slice. Focusing on a specific cell, this traffic demand depends on the number of attached UEs, the average arrival of packets and the size of each packet. This means the reference solution #1 is agnostic to the latency requirements of each RAN slice. In the reference solution #2, the RAN slice orchestrator periodically recomputes the dedicated radio resource quotas for each RAN slice. Specifically, it redistributes the RBs available for each RAN slice in each cell according to its average buffer size (i.e, packets in queue waiting to be transmitted). This means the RAN slice orchestrator provides more RBs for the RAN slice which presents the greatest average buffer size. This reference solution indirectly considers the latency experienced by the packets of a RAN slice.

A. Experimental Setup

To validate the SNC-based model, we use a Matlab-based simulator that resembles the packet arrival and their transmission for an uRLLC RAN slice in a single cell. Assuming the cell has available 24 RBs, we have evaluated several scenarios where a different number of RBs are reserved to the RAN slice. We have also evaluated other scenarios where we consider a different value for the average batch arrival and the required violation probability. With respect to the

TABLE II: Simulation Parameters for SNC-based model validation

Parameter	Configuration	Parameter	Configuration
5G Numerology μ_{5G}	2 [27]	Average batch arrival rate $\lambda_{i,m}$	From 3000 to 5000 batches/s Default: 3400 batches/s
Carrier bandwidth W_i ($\lvert\mathcal{R}_i\rvert$)	20 MHz (24 RBs)	Number of UEs in the cell $\lvert\mathcal{U}_i^m\rvert$	5
RBs for the RAN slice $\lvert\mathcal{R}_i^m\rvert$	From 7 to 17 RBs Default: 10 RBs	Probability of simultaneous transmission of packets $p_{i,m,u}$	Equiprobable
Packet delay budget W_m^{th}	5 ms	Packet size $L \in \mathcal{L}^m$	[256 512 1024 2048] bits
Violation probability ε'_m	From 0.001% to 10 % Default: 0.1 %	Packet size distribution $p_{m,l}$	Equiprobable
Spectral efficiency $SE_{z,i,m}$	[0.152;0.377;0.877;1.476;1.914;2.406 2.731;3.322;3.902;4.523;5.115;5.554 6.226;6.907;7.406] bps/Hz [37, Table 5.2.2.1-3]	PMF Spectral efficiency $\pi_{z,i,m}$	[0.0007;0.0006;0.0036;0.0141;0.0844;0.1599 0.2208;0.1118;0.0816;0.0752;0.0776;0.0716 0.0414;0.0300;0.0267]

probabilities $\pi_{z,i,m}$ of reporting certain CQIs, i.e., the achieved spectral efficiency $S_{z,i,m}$, we have extracted them by using real dataset from a Long Term Evolution (LTE) network. Note that, 5G dataset is not available due to the deployment of 5G networks are already in an early stage. Table II summarizes these and other key parameters used in the model validation.

Regarding the evaluation of the proposed heuristics, we consider a RAN infrastructure composed of $|\mathcal{I}| = 7$ cells deployed over an area of 0.95 Km x 0.95 Km. We also consider the traffic demand for each RAN slice is non-uniformly distributed over this area. This means each RAN slice (a) serves a different amount of UEs $|\mathcal{U}_i^m|$ in each cell; (b) has a specific batch arrival rate λ_m for its packets; and (c) has a specific distribution for the packet size, i.e., $\mathcal{L}^m$ and p_l. In addition, each RAN slice accommodates an uRLLC service with specific performance requirements in terms of packet delay budget W_m^{th} and violation probability ε'_m. These parameters along with the channel and cell parameters are summarized in Table III.

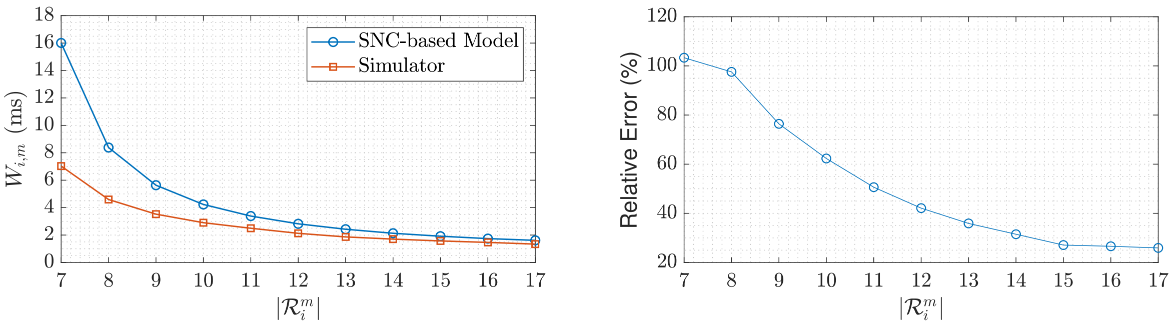
B. Validation of the Proposed SNC-based Model

In Fig. 3(a), we have evaluated the delay bound $W_{i,m}$ in function of the dedicated radio resource quota $|\mathcal{R}_i^m|$ assigned for the RAN slice m. We observe the SNC-based model always provides an upper estimation of the number of required RBs to obtain a specific delay bound, given a specific value for the violation probability ε'_m. This means that using the computed radio resource quota, the obtained delay bound in a real cell for this RAN slice would be lower than the estimated by our proposed model. This makes the proposed model suitable for ensuring the performance requirements of uRLLC RAN slices when the MNO plan them in advance. With respect to the relative error in Fig. 3(b), it seems large in a first attempt. Notwithstanding, it is acceptable for

TABLE III: Simulation Parameters for RAN slice planning

Parameter	Configuration	Parameter	Configuration
Cellular Environment	0.95 Km x 0.95 Km	Decoding Error Probability ε_{dec}	10^{-5} [39]
Number of Cells $\lvert\mathcal{C}\rvert$	7	Blocklength n_{block}	168 [38]
Cell areas $\forall i \in \mathcal{C}$	[0.1415; 0.1413; 0.1081; 0.1419 0.1255; 0.1399; 0.1060] Km^2	Number of RAN slices	3
gNB density κ_{gNBs}	$7.756 \cdot 10^{-6}$ gNBs/Km^2	Number of UEs per RAN slice and cell $\lvert\mathcal{U}_i^m\rvert$	$\lvert\mathcal{U}_i^1\rvert$ = [8; 6; 7; 10; 5; 4; 7] $\lvert\mathcal{U}_i^2\rvert$ = [10; 8; 7; 5; 8; 7; 9] $\lvert\mathcal{U}_i^3\rvert$ = [9; 10; 12; 10; 10; 13; 11]
5G Numerology μ_{5G}	2	Packet Delay Budget W_m^{th}	[15; 25; 5] ms
Carrier Bandwidth W_i ($\lvert\mathcal{R}_i\rvert$)	40 MHz (51 RBs)	Violation Probability ε_m'	[0.1; 1; 0.05] %
Cell Transmited Power $P_{TX}^{cell} = P_{TX} \cdot \lvert\mathcal{R}_i\rvert$	30 dBm [40]	RAN Slice Priority ψ_m	[0.667; 0.444; 0.889]
Shadowing Parameters	μ_χ = 0 dB σ_χ = 4 dB [34]	Packet length $l \in \mathcal{L}^m$	[256; 512; 1024; 2048] $\in \mathcal{L}^1$ bits [512] $\in \mathcal{L}^2$ bits [512; 2048] $\in \mathcal{L}^3$ bits
UE noise figure citeDiRenzo2016	10 dB PMF packet length p_l $\forall l \in \mathcal{L}^m$	[0.25 0.25 0.25 0.25] $\forall l \in \mathcal{L}^1$ [1] $\forall l \in \mathcal{L}^2$ [0.75 0.25] $\forall l \in \mathcal{L}^3$	
UE thermal noise	-174 dBm/Hz [34]	Batch arrival rate λ_m	λ_1 = 24500 batches/s λ_2 = 26950 batches/s λ_3 = 22750 batches/s
Pathloss exponent α	4 [33]	Probability of simultaneous transmission of packets p_u	Equiprobable

a model based on SNC. Specifically, modeling tools based on SNC allow us to consider more complex arrival and service processes that the ones considered by other modeling tools (e.g., Poisson arrivals and exponentially distributed service time in queue theory-based models) to get



(a) Delay bound $W_{i,m}$

(b) Relative Error (%)

Fig. 3: Evaluation of the delay bound $W_{i,m}$ in function of the number of RBs allocated to the RAN slice m, i.e., $\lvert\mathcal{R}_i^m\rvert$.

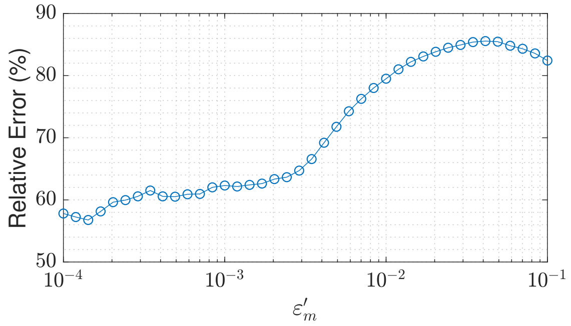


Fig. 4: Evaluation of the relative error in function of the violation probability ε'_m.

statistical performance bounds of the type $\mathrm{P}[\text{delay} > x] \leq \varepsilon$ at the expense of obtaining an exact match between the results obtained from the model and the simulator or real measurements. Instead, SNC-based models always provide an upper bound estimation for the delay bound [11]. We also notice the relative error between the SNC-based model and the simulator decreases when the number of assigned RBs increases. It is due to SNC-based models provide lower values for the relative error when the considered processing node suffers a less congestion [11]. In the case of a RAN slice, this happens when more RBs are allocated for it in a cell.

We have also evaluated the relative error in function of the violation probability as Fig. 4 shows. We observe the SNC-based model presents a better accuracy when the violation probability takes low values. This means our proposed model is ideal for uRLLC services, which have extreme requirements in terms of reliability.

C. Convergence and Computational Complexity Analysis of the Proposed Heuristics

In Fig. 5, we depict the convergence of the proposed heuristics to find a sub-optimal solution for the formulated problem in Eq. (49). Specifically, we have analyzed the convergence for different scenarios, each one characterized by a specific amount of cells. For all the scenarios, we assume each cell has available 51 RBs. If we focus on a single curve, each dot represents the value of such equation in a specific iteration (line 14 in Algorithm 1). We observe the curve is flat in the first iterations. The meaning of the flat region is the parameter $\overline{\Delta W}_m$ is infinite for one or more RAN slices. This means these RAN slices cannot accommodate the DL traffic with the amount of allocated RBs in this iteration (line 10 in Algorithm 1), i.e., with this RB allocation, the amount of packets in the buffer of one or more cells will dynamically increase up to infinity. For simplicity, we have represented this phenomena as a flat region in the curve. When all the RAN slices have enough allocated resources to accommodate their DL traffic,

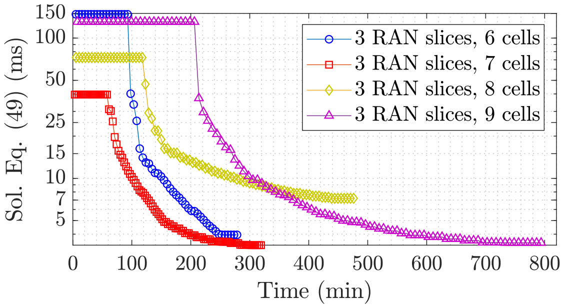
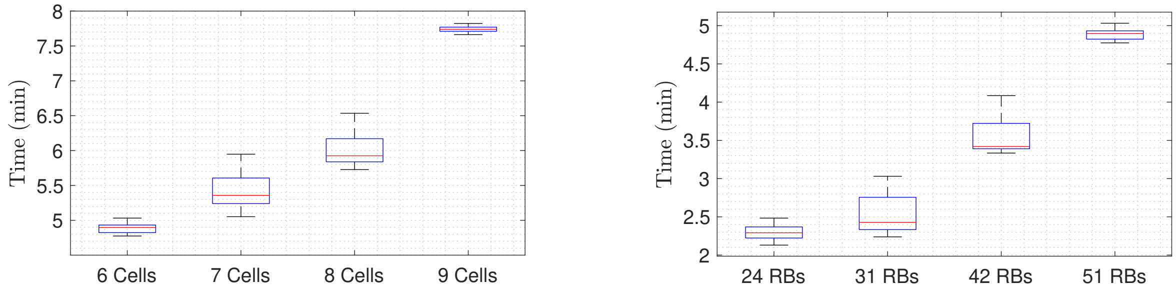


Fig. 5: Analysis of the convergence for the proposed heuristics.

the curve exponentially decreases up to the heuristics converge to a sub-optimal solution. If we compare the heuristics' convergence for the different scenarios, we observe the time between two consecutive iterations increases when the number of considered cells increases. The reason is the proposed heuristics has to call more times, one per cell, the SNC-based model to compute the delay bound for each RAN slice (line 13 in Algorithm 1).

This phenomena is better represented in Fig. 6, where a box-and-whisker plot is used for representing the statistical distribution of the execution time in an iteration. In this representation, the bottom and the top of each box represent the first and third quartiles for the measured times, 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 times. Observing Figs. 6(a) and 6(b) , we can conclude the execution time of an iteration grows exponentially with the number of cells and RBs. Although it is not depicted due to space limits, we have observed a similar behavior when the number of considered RAN slices increases.



(a) Number of cells in the RAN infrastructure. Note that, each cell has 51 RBs.

(b) Number of RBs available in each cell. Note that, the RAN infrastructure has 6 cells.

Fig. 6: Execution time of an iteration in the proposed heuristics for different scenarios.

TABLE IV: Checking if the performance requirements of each slice are met in the long term.

RAN Slices	Planning Solutions	Cell 1	Cell 2	Cell 3	Cell 4	Cell 5	Cell 6	Cell 7
Slice 1	Prop. Solution	✓	✓	✓	✓	✓	✓	✓
	Ref. Solution 1	✓	✓	✓	✓	✓	✗	✓
	Ref. Solution 2	✓	✓	✓	✗	✓	✓	✓
Slice 2	Prop. Solution	✓	✓	✓	✓	✓	✓	✗
	Ref. Solution 1	✓	✓	✓	✗	✓	✓	✓
	Ref. Solution 2	✓	✓	✓	✓	✓	✓	✓
Slice 3	Prop. Solution	✗	✓	✓	✗	✓	✓	✓
	Ref. Solution 1	✗	✓	✗	✗	✓	✓	✗
	Ref. Solution 2	✗	✗	✗	✗	✗	✗	✗

D. *Performance Analysis of the Proposed Heuristics*

With respect to the performance analysis, we have focused on an scenario where there are not enough resources to accommodate the performance requirements of three RAN slices in the long term. Under this scenario, we have compared the results from the proposed heuristics and the reference solutions 1 and 2. Specifically, we have checked if the latency and reliability requirements of each RAN slice m are met in the long term for each cell. Focusing on a cell i, this means if $\Delta W_{i,m} = 0$, i.e., $\mathrm{P}[w_k > W_m^{th}] < \varepsilon_m'$ in the long term. Note that, w_k represents the transmission delay of a individual packet k.

In Table IV, we have represented with the checkmark symbol ✓ if $\Delta W_{i,m} = 0$ and with the crossmark symbol ✗ if $\Delta W_{i,m} > 0$. On the one hand, we observe that our proposed solution obtains the highest percentage (i.e., 85.71%) of RAN slices successfully accommodated in a single cell with respect to the reference solutions 1 and 2 (i.e., 71.42% and 61.9%, respectively). On the other hand, checking if the three RAN slices can be simultaneously accommodated in a single cell, we also observe the percentage of success is higher in the proposed solution (i.e., 57.14%) with respect to the reference solutions 1 and 2 (i.e., 28.57 % and 0 %, respectively). The low performance of the reference solution 2, specially for RAN slice 3, is mainly due to the RAN slice orchestrator may allocate few radio resources in an allocation period for a RAN slice in a single cell (i.e., due to the number of enqueued packets is very low) and then, this RAN slice may suffer a traffic peak, which would involve the number of packets in the queue may grow significantly, increasing in this way the transmission delay w_k experienced by each packet k. All this demonstrates the importance of pre-establish an amount of guaranteed radio resources to ensure the latency and reliability requirements before deploying the RAN slices.

VII. Conclusions and Future Work

Deploying and running RAN slices providing uRLLC services would require an ad-hoc analysis of expected performance in terms of delay and reliability thereby driving the network resources orchestration process. Based on that, the MNO can properly take decisions on the dedicated radio resource quota to be assigned to each RAN slice within a given cell.

Under this context, we have first proposed a SNC-based model, which given *i*) the dedicated radio resource quota for a uRLLC RAN slice in a single cell, *ii*) the target violation probability, *iii*) its traffic demand and *iv*) the CDF for the SINR experienced by its attached UEs, provides the delay bound for the packet transmission delay. To derive such CDF, we relied on a model based on stochastic geometry. This model considers the impact of the interference incurred by multiple RAN slices deployed in neighbor cells on the capacity the serving cell offers to a serving RAN slice. Additionally, we have proposed heuristics to plan in advance the deployment of multiple uRLLC RAN slices in a multi-cell environment, i.e., computing the dedicated radio resource quotas which ensure their performance requirements in the long term.

We have validated the proposed SNC-based model by means of an exhaustive simulation campaign, demonstrating it provides a conservative upper estimation of the delay bound for the packet transmission delay of an uRLLC RAN slice, given its target violation probability. This makes the proposed model suitable for ensuring the performance requirements of uRLLC RAN slices when the MNO plan them in advance. Additionally, we have showed the benefits of using the proposed heuristics when the RAN slice orchestrator simultaneously plans the deployment of several uRLLC RAN slices in a multi-cellular scenario.

Regarding the future work, the proposed SNC-based model could be extended to consider the computational part of a gNB, i.e., considering the time spent by the gNB to process an uRLLC packet before its transmission via the radio interface. This extension will allow the MNO to completely characterize the delay bound of a network slice in the RAN. With respect to the proposed heuristics, it could be extended to find a suboptimal solution for the values of the weights in addition to the amount of radio resources which are dedicated for each RAN slice in each cell. In this way, the values of the weights could be also optimized in order to accommodate more uRLLC RAN slices in the MNO infrastructure.

References

[1] I. Vision, “Framework and overall objectives of the future development of IMT for 2020 and beyond,” *International Telecommunication Union (ITU), Document, Radiocommunication Study Groups*, 2015.

[2] J. Prados-Garzon, P. Ameigeiras, J. Ordonez-Lucena, P. Muñoz, O. Adamuz-Hinojosa, and D. Camps-Mur, “5G Non-Public Networks: Standardization, Architectures and Challenges,” *IEEE Access*, vol. 9, pp. 153893–153908, 2021.

[3] J. Ordonez-Lucena, P. Ameigeiras, L. M. Contreras, J. Folgueira, and D. R. López, “On the Rollout of Network Slicing in Carrier Networks: A Technology Radar,” *Sensors*, vol. 21, no. 23, 2021.

[4] 3GPP TS 28.530 V.16.1.0, “Management and orchestration; Concepts, use cases and requirements (Release 16),” Dec. 2019.

[5] 3GPP TS 28541 V.17.0.0, “Management and orchestration; 5G Network Resource Model (NRM); Stage 2 and stage 3 (Release 17),” Sept. 2020.

[6] I. Vilà, J. Pérez-Romero, O. Sallent, and A. Umbert, “A Multi-Agent Reinforcement Learning Approach for Capacity Sharing in Multi-Tenant Scenarios,” *IEEE Trans. Veh. Technol.*, vol. 70, no. 9, pp. 9450–9465, 2021.

[7] C. Guo, L. Liang, and G. Y. Li, “Resource Allocation for Vehicular Communications With Low Latency and High Reliability,” *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 3887–3902, 2019.

[8] C. Guo, L. Liang, and G. Y. Li, “Resource Allocation for High-Reliability Low-Latency Vehicular Communications With Packet Retransmission,” *IEEE Trans. Veh. Technol.*, vol. 68, no. 7, pp. 6219–6230, 2019.

[9] M. Patra, R. Thakur, and C. S. R. Murthy, “Improving Delay and Energy Efficiency of Vehicular Networks Using Mobile Femto Access Points,” *IEEE Trans. Veh. Technol.*, vol. 66, no. 2, pp. 1496–1505, 2017.

[10] L. Chinchilla-Romero, J. Prados-Garzon, P. Ameigeiras, P. Muñoz, and J. M. Lopez-Soler, “5G Infrastructure Network Slicing: E2E Mean Delay Model and Effectiveness Assessment to Reduce Downtimes in Industry 4.0,” *Sensors*, vol. 22, no. 1, 2022.

[11] M. Fidler and A. Rizk, “A Guide to the Stochastic Network Calculus,” *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 92–105, 2015.

[12] K. Katsaros, M. Dianati, R. Tafazolli, and X. Guo, “End-to-End Delay Bound Analysis for Location-Based Routing in Hybrid Vehicular Networks,” *IEEE Trans. Veh. Technol.*, vol. 65, no. 9, pp. 7462–7475, 2016.

[13] S. Ma, X. Chen, Z. Li, and Y. Chen, “Performance evaluation of URLLC in 5G based on stochastic network calculus,” *Mob. Netw. Appl.*, vol. 26, no. 3, pp. 1182–1194, 2021.

[14] Á. A. Cardoso, M. V. G. Ferreira, and F. H. T. Vieira, “Delay bound estimation for multicarrier 5G systems considering lognormal beta traffic envelope and stochastic service curve,” *Trans. Emerg. Telecommun. Technol.*, p. e4281, 2021.

[15] C. Xiao *et al.*, “Downlink MIMO-NOMA for Ultra-Reliable Low-Latency Communications,” *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 780–794, 2019.

[16] S. Schiessl, M. Skoglund, and J. Gross, “NOMA in the Uplink: Delay Analysis With Imperfect CSI and Finite-Length Coding,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 3879–3893, 2020.

[17] Q. Xu, J. Wang, and K. Wu, “Learning-Based Dynamic Resource Provisioning for Network Slicing with Ensured End-to-End Performance Bound,” *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 1, pp. 28–41, 2020.

[18] J. García-Morales, M. C. Lucas-Estañ, and J. Gozalvez, “Latency-Sensitive 5G RAN Slicing for Industry 4.0,” *IEEE Access*, vol. 7, pp. 143139–143159, 2019.

[19] T. Guo and A. Suárez, “Enabling 5G RAN Slicing With EDF Slice Scheduling,” *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2865–2877, 2019.

[20] J. Tang, B. Shim, and T. Q. S. Quek, "Service Multiplexing and Revenue Maximization in Sliced C-RAN Incorporated With URLLC and Multicast eMBB," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 881–895, 2019.

[21] D. Bega, M. Gramaglia, A. Banchs, V. Sciancalepore, and X. Costa-Pérez, "A Machine Learning Approach to 5G Infrastructure Market Optimization," *IEEE Trans. Mobile Comput.*, vol. 19, no. 3, pp. 498–512, 2020.

[22] L. Zanzi, V. Sciancalepore, A. Garcia-Saavedra, H. D. Schotten, and X. Costa-Pérez, "LACO: A Latency-Driven Network Slicing Orchestration in Beyond-5G Networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 667–682, 2021.

[23] C.-S. Chang, *Performance guarantees in communication networks*. Springer Science & Business Media, 2012.

[24] J.-Y. Le Boudec and P. Thiran, *Network calculus: a theory of deterministic queuing systems for the internet*, vol. 2050. Springer Science & Business Media, 2001.

[25] Y. Jiang, Y. Liu, *et al.*, *Stochastic network calculus*, vol. 1. Springer, 2008.

[26] S. Ross, *A First Course in Probability*. Pearson, 2014.

[27] 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 Trans. Veh. Technol.*, vol. 14, no. 4, pp. 64–75, 2019.

[28] V. B. Iversen, "Teletraffic engineering and network planning," 2015.

[29] A. Karamyshev, E. Khorov, A. Krasilov, and I. Akyildiz, "Fast and accurate analytical tools to estimate network capacity for URLLC in 5G systems," *Comput. Netw.*, vol. 178, p. 107331, 2020.

[30] 3GPP TS 38.306 V.16.2.0, "NR; User Equipment (UE) radio access capabilities (Release 16)," Oct. 2019.

[31] Y. Hmamouche, M. Benjillali, S. Saoudi, H. Yanikomeroglu, and M. D. Renzo, "New Trends in Stochastic Geometry for Wireless Networks: A Tutorial and Survey," *Proceedings of the IEEE*, pp. 1–53, 2021.

[32] S. M. Yu and S.-L. Kim, "Downlink capacity and base station density in cellular networks," in *WiOpt*, pp. 119–124, 2013.

[33] J. G. Andrews, F. Baccelli, and R. K. Ganti, "A Tractable Approach to Coverage and Rate in Cellular Networks," *IEEE Trans. Commun.*, vol. 59, no. 11, pp. 3122–3134, 2011.

[34] M. Di Renzo, W. Lu, and P. Guan, "The Intensity Matching Approach: A Tractable Stochastic Geometry Approximation to System-Level Analysis of Cellular Networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 5963–5983, 2016.

[35] J. G. Andrews, A. K. Gupta, and H. S. Dhillon, "A primer on cellular network analysis using stochastic geometry," *arXiv preprint arXiv:1604.03183*, 2016.

[36] H. S. Dhillon and J. G. Andrews, "Downlink rate distribution in heterogeneous cellular networks under generalized cell selection," *IEEE Commun. Lett.*, vol. 3, no. 1, pp. 42–45, 2014.

[37] 3GPP TS 38.214 V.16.5.0, "NR; Physical layer procedures for data (Release 16)," Mar. 2021.

[38] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel Coding Rate in the Finite Blocklength Regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.

[39] Shirvanimoghaddam *et al.*, "Short Block-Length Codes for Ultra-Reliable Low Latency Communications," *IEEE Commun. Mag.*, vol. 57, no. 2, pp. 130–137, 2019.

[40] P. Muñoz, O. Adamuz-Hinojosa, J. Navarro-Ortiz, O. Sallent, and J. Pérez-Romero, "Radio Access Network Slicing Strategies at Spectrum Planning Level in 5G and Beyond," *IEEE Access*, vol. 8, pp. 79604–79618, 2020.