Quantifying the Millimeter Wave New Radio Base Stations Density for Network Slicing with Prescribed SLAs*

Yevgeni Koucheryavy[a,c], Ekaterina Lisovskaya[b,*], Dmitri Moltchanov[a], Roman Kovalchukov[a], Andrey Samuylov[b]

[a] *Unit of Electrical Engineering, Tampere University, 33720 Tampere, Finland*
[b] *Peoples' Friendship University of Russia (RUDN University)*
6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation
[c] *Higher School of Economics, National Research University, Moscow 101000, Russia*

Abstract

Network slicing is expected to become an integral part of future 5G systems providing a simple mechanism for physical network operators to diversify their business models. New Radio (NR) technology operating in millimeter wave (mmWave) band is one of the critical bearers for this functionality, providing extraordinary capacity at the air interface. This paper provides a mathematical tool for assessing the upper and lower bounds of NR BS density needed to maintain the requested slice rate guarantees. The upper bound corresponds to the full traffic isolation between slices while the lower one – to the full mixing of traffic from the slices. To this aim, we unite the tools of stochastic geometry and queuing theory formulating a performance evaluation framework that allows assessing the rate violation metrics in a dynamic network slicing environment. The developed framework captures specifics of mmWave NR technology, including antenna directivity at the UE and NR BS sides, propagation and blockage losses, as well as the service process with location-dependent resource requirements. Our results show that for considered schemes, the operational regime of the system changes abruptly with respect to the density of NR BSs. The difference between full isolation and full mixing schemes becomes bigger in environments with high session arrival intensities that naturally require dense deployments. Thus, at the initial market penetration phase, full isolation can be used without compromising the network performance. However, at mature stages, more complex schemes are needed to reduce the capital expenditures of the operators.

Keywords: Quality assurance, Network slicing, Cellular radio, Millimeter wave communication

1. Introduction

The concept of network slicing, introduced by 3GPP in Release 14 [1] as one of the essential features of 5G cellular systems and then clarified in Release 15 and 16, is expected to drastically simplify the market entrance for mobile virtual network operators (MVNO) as well as provisioning of differentiated quality to network services [2, 3]. This functionality is a major paradigm shift in the cellular world enabling multi-layer network structures similar to that of the modern Internet and allowing resource sharing with logical isolation among multiple tenants and/or services in multi-domain context [4]. The technology relies on the set of requirements and mechanisms ensuring resource allocation across both air interface and core network and promises to virtualize resources of the physical mobile network operator (PMNO) in a secure way providing a strong degree of isolation between MVNOs [5].

According to 3GPP [1] and GSMA [6], there has to be not only logical but physical isolation between the traffic of different slices. This isolation at the data plane might be ensured using mechanisms operating at different timescales such as packet scheduling, connection admission control (CAC), network planning, etc. However, at the same time, PMNO is expected to ensure the efficient use of network resources benefiting from the statistical multiplexing of traffic demands, which inherently contradicts the requirement of physical traffic isolation. As a result, the level of physical traffic isolation is left to PMNOs.

While future 5G systems are expected to be heterogeneous in nature, allowing to use of multiple radio interfaces [7], the recently standardized New Radio (NR) technology has already become become the 5G technological enabler at the air interface. Although NR is built to operate over a wide variety of frequency bands [8], it is millimeter wave (mmWave) band implementation that matches 5G system requirements with respect to the access

*This paper has been supported by the RUDN University Strategic Academic Leadership Program (recipient E. Lisovskaya, investigation). The reported study was funded by RFBR, project number 19-07-00933 (recipient R. Kovalchukov, simulation model development), 20-07-01064 (recipient A. Samouylov, problem formulation and analysis).

*Corresponding author

Email addresses: evgeny.kucheryavy@tuni.fi (Yevgeni Koucheryavy), lisovskaya-eyu@rudn.ru (Ekaterina Lisovskaya), dmitri.moltchanov@tuni.fi (Dmitri Moltchanov), roman.kovalchukov@tuni.fi (Roman Kovalchukov), samuylov-ak@rudn.ru (Andrey Samuylov)

rate potentially enabling multi-tenancy and multi-service network slicing at the air interface [9]. However, along with promises to bring extraordinary rates to the access interface, mmWave NR systems are characterized by a set of unique challenges. High propagation losses combined with mmWave propagation's sensitivity to blockages by small objects (e.g., human bodies, cars, trees) significantly limit the coverage of mmWave systems [10]. This challenge becomes even more dramatic in the environments with dynamic blockages, e.g., crowded streets, city centers [11, 12]. The recent empirical [10] and theoretical [11, 12] studies have deeply investigated this process in various network deployments. Particularly, it has been shown that the signal-to-noise ratio (SNR) may rapidly fluctuate fading by 10-30 dB at sub-second timescales. These effects make network slicing provisioning with prescribed rate guarantees an extremely complex procedure.

Most of the prior research work related to enabling network slicing at the air interface concentrated on developing algorithms trying to satisfy a certain isolation and performance criterion (see Section 2). However, these models often have no reference theoretical bounds to benchmark their performance. Providing a mathematical tool for assessing the upper and lower bounds of NR BS density needed to maintain the requested slice rate guarantees is the main motivation for this paper. Accordingly, this study aims to characterize the lower and upper bounds on the density of mmWave NR BSs corresponding to full isolation and full mixing of slices at the NR interface required to support the prescribed slice rate guarantees. To this aim, we first characterize the rate provided by a single NR BS in dense deployment accounting for inherent features of mmWave propagation, including directional communications, path loss, dynamic blockage, and interference. Then, we formulate the slice service process of a single NR BS accounting for slice and session dynamics. The sought key performance indicators (KPI), used to dimension the mmWave NR access network, are those specified by GSMA [6] for service level agreement (SLA) and include: (i) fraction of time rate guarantees are violated, (ii) the intensity of rate violations, (iii) mean duration of slice rate violation, and (iv) network resource utilization.

The main contributions of our study are:

- mathematical framework for assessing the required density of NR BSs for given characteristics of the dynamic slice arrival process and associated slice rate requirements specified as a part of SLA;
- lower and upper bounds on the NR BS density corresponding to the full isolation and full mixing of slices at the NR interface required to support a given set of slices with prescribed rate guarantees;
- observation that the gap between full mixing and full isolation strategies increases with the increased traffic load imposed on the network and can reach 10^{-3} BS/km^2 implying that at mature phases or NR deployment more sophisticated slice isolation schemes might be required to reduce capital expenditures of network operators.

The rest of the paper is organized as follows. In Section 2, we overview the technological background of network slicing and review the related works. System model is introduced in Section 3. We develop our framework in Section 4 and Section 5. Numerical results are elaborated in Section 6. Conclusions are drawn in the last section.

2. Background and State-of-the-Art

In this section, we provide the technological background of network slicing addressing the main components standardized by 3GPP and GSMA. Then, we review the recent literature addressing resource sharing at the air interface.

2.1. 3GPP Standardization and Terminology

The concept of network slicing has been originally introduced in 3GPP in Release 14 [1] in the context of vehicular-to-everything (V2X) communications. Later, in 3GPP Releases 15 and 16, it has been extended to include MVNOs and bundling of services with similar QoS requirements. The use-cases are flexible and include service provisioning for third parties such as MVNOs, content providers, etc.

According to 3GPP TS 23.501, a network slice is a logical network providing the prescribed set of networking capabilities. The latter is defined in terms of network functions, which represent certain network capabilities related to either data or control plane [13]. More specifically, following GSMA [14], a slice instance is a set of network functions and resources, including those at the air interface and in the core network representing an existing slice. From the external point of view, slice instance is expected to be seen as a dedicated network.

The QoS parameters for a slice are assumed to be negotiated between PMNO and external entities buying the service and specified in SLA. Although 3GPP does not explicitly specify these parameters, recently, the SLA template has been proposed by GSMA as an open standard NG.116 "Generic network slice template" [6]. Technically, slices are differentiated by the slice identifier, called *single network slice selection assistance information (S-NSSAI)*, that specifies the slice/service type (SST) describing the global application of a slice and slice differentiator (SD) – an optional field used to differentiate between slices of the same type. Three values of SST have been specified: eMBB, URLLC, mMTC. Examples of slice template parameters defined in NG.116 are shown in Table 1.

2.2. Resource Sharing at the Air Interface

Over the last few years, several authors have discussed and demonstrated the advantages of network slicing for 5G systems. The authors in [15] revealed the importance of

Table 1: Basic slice parameters according to GSMA NG.116.

Parameter	**Example of values**
Region Specification	1: Full country, 2: List of regions
Delay guarantees	0: Not supported, 1: Supported
Guaranteed rate	0: Not specified, 10 Mbps
Maximum rate	100 Mbps, 20 Gbps
Guaranteed UE rate	180 Kbps – VoIP, 25 Mbps – gaming
Maximum UE rate	50/400/1000 Mbps
Isolation level	0: No isolation, 1: Physical, 2: Logical
Physical isolation	0: Processes, 1: Memory, 2: Physical
Logical isolation	0: Virtual, 1: NF, 2. Tenant
Maximum packet size	eMBB 1500, IoT 40, URLLC 160
Mission-crit. support	0: Non-critical, 1: Mission-critical
Mission-crit. capab.	1: Prioritization, 2: Preemption
Mission-critical service	1: MCPTT, 2: MCData, 3: MCVideo
MMTel support	0: Not supported, 1: Supported
Number of connection	10^5 sessions, 10^7 sessions
Number of terminals	10^5 terminals, 10^7 terminals
Performance prediction	1: Rate, 2: Latency, 3: Service success
Accuracy	+/- 1 m, +/- 0,01 m
Radio spectrum	n1, n77, n38
Investigation	0: N/A, 1: Passive, 2: Active
Service continuity	0: N/A, 1: SSC 1, 2: SSC 2
Simultaneous slice use	0: Any, 2: Same SST value
Slice QoS parameters	1: Voice, 2: Video, 3: Gaming
Resource type	0: GBR, 1: Delay-crit., 2: Non-GBR
Priority level	10 – signaling, 30 – Real time gaming
Packet delay budget	Driving: 2 ms, VR 1 ms
Packet error rate	10^{-6} – MC data, 10^{-2} – V2X
Maximum packet loss	1%; 5%
Terminal density	10^5 UE/km^2, 10^6 UE/km^2 MIoT

providing QoS guarantees at the air interface for CDN systems. The use of slicing at the air interface of 5G systems for mMTC traffic is proposed and analyzed in [16]. Network slicing for vehicle-to-everything (V2X) and vehicle-to-vehicle (V2V), as well as its coexistence with other slice types, have been discussed in [17].

Despite network slicing is defined as an end-to-end phenomenon, its performance is expected to be mostly affected by the air interface between UE and BS as well as the service process at the BSs. Thus, recent performance evaluation studies mainly concentrated on delivering QoS guarantees at radio access network. The performance of network slicing at the air interface has been addressed in [18]. The authors have proposed and evaluated the algorithm for provisioning differentiated QoS guarantees to prospective applications that have to be supported in 5G systems, including uRLLC, eMBB, and mMTC. They also emphasized the need for new advanced resource allocation and management techniques for QoS provisioning that would account for the specifics of the wireless interface. The authors in [19] utilized the game-theoretic approach to formulate and solve the task of slice service optimization with minimum rate guarantees. Contrarily to [18], the resource allocation model is much more comprehensive compared to the wireless access model. The authors in [20] considered the end-to-end slicing in the core network and at the air interface, developing the set of algorithms and protocols for QoS provisioning.

The emergence of NR technologies providing abundant resources and flexible numerology at the air interface promises to deliver the flexibility of resource allocation enabling fine-grained network slicing [21]. Using the computer simulations, the authors in [22] characterized the trade-offs between QoS provisioning to network slices and resource utilization in NR deployments. Implementation aspects are also considered in [3, 23]. However, the literature on the use of network slicing in NR systems is still limited.

To summarize, the task of delivering QoS guarantees to network slices is a complex one. Its solution has to account for both specifics of wireless access and the service process of slices at BSs. Unfortunately, most of the studies have been putting more emphasis on either of these two parts. Furthermore, most of the studies performed so far concentrated on proposing and evaluating the slicing management and control procedures without comparing them to other proposals or performance bounds. Our research aims to provide lower and upper bounds on the density of NR BSs needed to provide the performance guarantees to a dynamically varying number of slices by accounting for both wireless and resource allocation specifics at NR BSs.

3. System Model

In this section, we specify the considered system and its main components. Particularly, we first start describing the wireless part by specifying propagation, blockage, and antenna models for NR BS and UE, and then proceed describing the traffic and slice service process. The notation utilized in this paper is summarized in Table 2.

3.1. NR Deployment Model

We consider a Poisson point process (PPP) of NR BSs with BS density of ξ units/m^2, as illustrated in Fig. 1(a). In this field of BSs, we concentrate on a tagged NR BS serving pedestrians that are also assumed to form PPP with density λ_B (see Fig. 1(b)). All of the pedestrians are associated with UEs equipped with mmWave NR modules. The heights of the NR BS and the UEs are h_A and h_U, respectively. Pedestrians are modeled as cylinders with a height h_B and radius r_B. Pedestrians may dynamically occlude the line-of-sight (LoS) propagation path between the UE and the NR BS.

The NR BS is equipped with three physical antennas, each covering a 120°-sector. We focus on a single antenna sector and define r_E as the effective coverage radius. Radius is chosen such that no UEs inside it experience outage conditions when their LoS link is blocked, so there is a modulation and coding scheme (MCS) for users at the distance of r_E [24]. Note that this assumption does not qualitatively affect the analysis principles in the rest of the paper. The only affected parameter is the number of UEs covered which is explicitly accounted for in the numerical results section. The radius r_E is computed in Section 5 by using the propagation, blockage, and antenna models detailed below.

Table 2: Notation used in this paper.

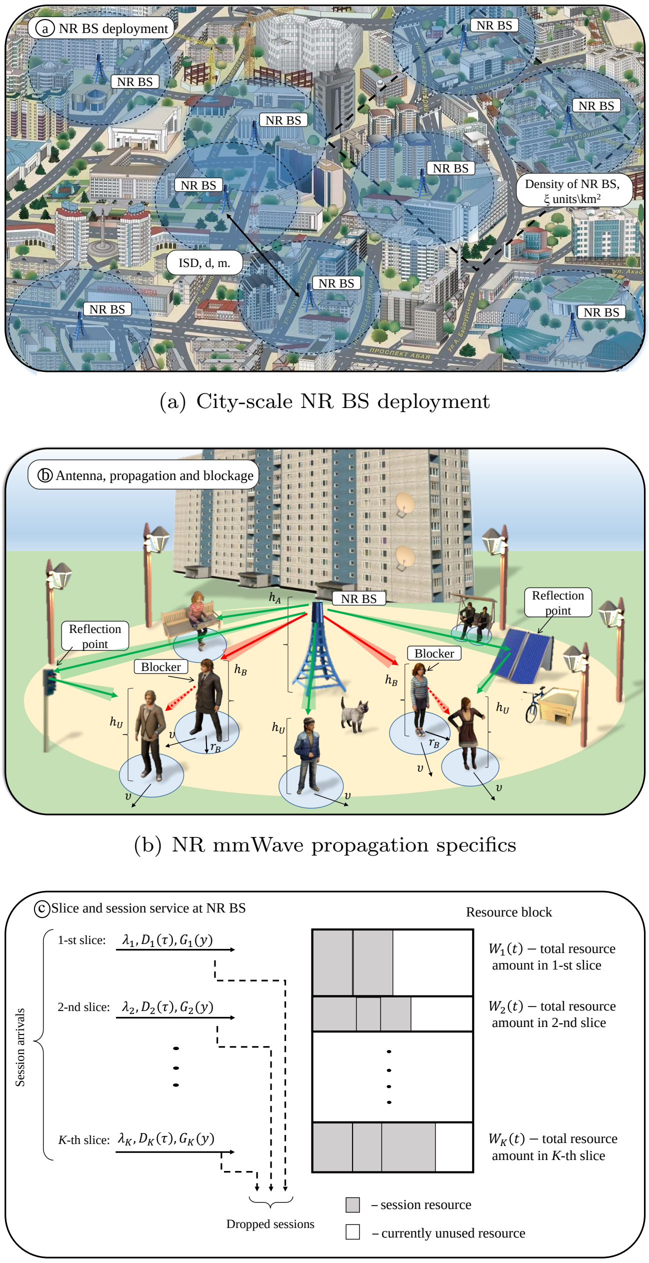
Parameter	Definition
Wireless part	
ξ	Density of NR BS (units/m^2)
λ_B	Density of pedestrians (units/m^2)
h_A, h_U	Heights of NR BS and UE
h_B, r_B	Height and radius of pedestrians
r_E	Effective coverage radius of NR BS
$S(y)$	SNR at UE located at a distance of y
P_A	NR BS transmit power
G_A, G_U	Antenna array gains at NR BS and UE
N_0	Power spectral density of noise at UE
B	Bandwidth
$L(y)$	Linear path loss to UE located at a distance of y
I	Interference
$M_{S,1}, M_{S,2}$	Fading margins in non-blocked and blocked states
f_c	Carrier frequency
A_i, ζ_i	Propagation and path loss coefficients
$p_B(y)$	Blockage probability at a distance y
α_A, α_U	HPBW of radiation patterns at NR BS and UE
$\theta_{3db}, \theta_m, \beta$	Parameters of antenna array
G	The mean antenna gain over HPBW
N_A	Number of horizontal antenna elements at NR BS
R	Rate provided by a single NR BS
P_R	Received signal strength
r_I	Interference radius
M_I	Interference margin
μ_{P_R}, μ_I	Means of P_R and I
$\sigma^2_{P_R}, \sigma^2_I$	Variances of P_R and I
$K_{P_R,I}$	The covariance between P_R and I
$p_C(x)$	Exposure probability
S_{nB}, S_B	SNR in the non-blocked and blocked conditions
S_j	The SNR thresholds for MCSs
r_j	Resource units requested by a single session
ϵ_j	The probability that a session requests r_j PRBs
Queuing system part	
B_i	Amount of resources allocated for slice i
Λ	The session arrival intensity from a pedestrian
p_i	The probability that a session belongs to slice i
b_i	The mean service time of sessions of slice i
a_i, α_i	Moments of resource request by a session in slice i
$N^+(B_i)$	Intensity of rate violations per unit time
$T^+(B_i)$	Fraction of time rate guarantees are violated
$\tau_m(B_i)$	Mean duration of rate violation
$r(B_i)$	Mean resources utilization

3.2. Propagation, Blockage, and Antenna Models

We assume that pedestrians might temporarily block the LoS path between the UE and the NR BS. Depending on the current link state (LoS non-blocked or blocked) and the distance between the NR BS and the UE, the session employs an appropriate MCS to maintain reliable data transmission. The signal-to-noise ratio (SNR) at the receiver located at the distance of y from the NR BS along the propagation path is

$$S(y) = \frac{P_A G_A G_U}{N_0 B L(y) I M_S}, \tag{1}$$

where P_A is the NR BS transmit power G_A and G_U are the antenna array gains at the NR BS, and the UE ends, respectively. N_0 is the power spectral density (psd) of noise. B is the bandwidth, $L(y)$ is the path loss in linear


ⓐ NR BS deployment
NR BS
Density of NR BS, ξ units\km²
ISD, d, m.
ⓑ Antenna, propagation and blockage
NR BS
h_A
Reflection point
Blocker
h_B
h_U
r_B
v
ⓒ Slice and session service at NR BS
Resource block
Session arrivals
1-st slice: $\lambda_1, D_1(\tau), G_1(y)$
2-nd slice: $\lambda_2, D_2(\tau), G_2(y)$
K-th slice: $\lambda_K, D_K(\tau), G_K(y)$
$W_1(t)$ – total resource amount in 1-st slice
$W_2(t)$ – total resource amount in 2-nd slice
$W_K(t)$ – total resource amount in K-th slice
Dropped sessions
– session resource
– currently unused resource

(a) City-scale NR BS deployment

(b) NR mmWave propagation specifics

(c) Abstracted slice service process at NR BS

Figure 1: The main elements of network slicing over NR deployment.

scale, and I is the interference, and M_S is the shadow fading margin.

The effect of shadow fading is accounted by using the shadow fading margins, $M_{S,1}$ and $M_{S,2}$ for the LoS non-blocked and blocked states as provided in [25]. Further, following [25], the path loss measured in dB is

$$L_{dB}(y) = \begin{cases} 32.4 + 21\log_{10} y + 20\log_{10} f_c, \text{ non-bl.,} \\ 32.4 + 31.9\log_{10} y + 20\log_{10} f_c, \text{ blocked,} \end{cases} \tag{2}$$

where f_c is the carrier frequency in GHz and y is the three-dimensional (3D) distance between the NR BS and the UE.

The path loss in (2) can be represented in the linear scale by utilizing the model in the form of $A_i y^{-\zeta_i}$, where

A_i and ζ_i are the propagation coefficients. Introducing the coefficients (A_1, ζ_1) and (A_2, ζ_2) corresponding to LoS non-blocked and blocked conditions, we have

$$A = A_1 = A_2 = 10^{2 \log_{10} f_c + 3.24}, \; \zeta_1 = 2.1, \; \zeta_2 = 3.19. \tag{3}$$

The value of SNR at the UE can then be written as

$$S(y) = P_A G_A G_U \left[\frac{y^{-\zeta_1}[1 - p_B(y)]}{A_1 M_{S,1}} + \frac{y^{-\zeta_2} p_B(y)}{A_2 M_{S,2}} \right], \tag{4}$$

where $p_B(y)$ is the blockage probability [11, 26]

$$p_B(y) = 1 - \exp^{-2\lambda_B r_B \left[\sqrt{y^2 - (h_A - h_U)^2} \frac{h_B - h_U}{h_A - h_U} + r_B \right]}, \tag{5}$$

where λ_B is the density of pedestrians, r_B and h_B are the pedestrians' radius and height, h_U is the height of NR UE, h_A is the height of NR BS.

Introducing the coefficients

$$C_i = P_A G_A G_U / (A_i M_{S,i}), \; i = 1, 2, \tag{6}$$

the propagation model finally reads as

$$S(y) = C_1 y^{-\zeta_1}[1 - p_B(y)] + C_2 y^{-\zeta_2} p_B(y). \tag{7}$$

Note that the mmWave NR technology is not reliable as blockage may lead to outage situations, where the network connectivity can be lost. This effect may eventually lead to the drop of UE sessions initially accepted for service. In our paper, we consider the NR configuration, where multi-connectivity functionality is not utilized but the system is still expected to support reliable connectivity. For this reason, in what follows, we consider only NLoS deployment, implying that the received signal strength is sufficient to support active connection even in case of blockage.

We consider planar antenna arrays at both transmit and receive sides. Following [27, 28], we utilize a cone antenna model, where the radiation pattern is represented as a conical zone with an angle of α coinciding with the half power beamwidth (HPBW) of the antenna array. Recall that the HPBW of a linear antenna array, α, is given by [29] $\alpha = 2|\theta_m - \theta_{3db}|$, where θ_{3db} is the angle at which the value of the radiated power is 3dB below the array maximum and $\theta_m = \pi/2$ is the location of the array maximum.

The antenna gain over the HPBW can be found as [29]

$$G = \frac{1}{\theta_{3db}^{+} - \theta_{3db}^{-}} \int_{\theta_{3db}^{-}}^{\theta_{3db}^{+}} \frac{\sin(N\pi \cos(\theta)/2)}{\sin(\pi \cos(\theta)/2)} d\theta, \tag{8}$$

where the upper and the lower 3-dB points are

$$\theta_{3db}^{\pm} = \arccos[\pm 2.782/(N\pi)], \tag{9}$$

and N is the number of antenna elements.

3.3. Slice Composition, Isolation, and Service

Following 3GPP and GSMA, we consider dynamic network slices, where a slice is initialized dynamically depending on the presence or absence of sessions of the slice, see Fig. 1(c). There might be overall K slices in the system. As long as there are no sessions at the NR BS that belong to this slice, the slice consumes no resources. A slice is initiated with a new session of this slice arriving at the system. Once initiated, slice is allocated B_i, $i = 1, 2, \ldots, K$, resources. When the last remaining session of a slice leaves, all the allocated resources for this slice are released.

Let Λ be the session arrival intensity from a single pedestrian. An arriving session belongs to the slice i with the probability of p_i, $i = 1, 2, \ldots, K$, $\sum_{i=1}^{K} p_i = 1$. Employing the superposition property of point processes [30], the spatial session arrival process at the NR BS is Poisson with the intensity $\Lambda \lambda_B \pi r_E^2 2\pi/3$ sessions per time unit [30]. We define $\lambda_i = p_i \Lambda \lambda_B \pi r_E^2 2\pi/3$ as the spatial arrival intensity of slice i sessions in coverage of NR BS.

The choice of UE that initiates a session is random. Hence, by utilizing the property of Poisson processes, we observe that the geometric locations of users associated with a session are distributed uniformly within the NR BS coverage. Sessions belonging to slice i are characterized by the generally distributed service times $D_i(\tau)$ with mean d_i. The amount of resources requested by a session depends on the service rate R_i, physical resource block (PRB) size, emitted power, and channel characteristics. For our model to represent a variety of potential applications, we assume that the session resource requirements ν_i follow general distribution $G_i(y)$ with mean a_i and second raw moment α_i. The associated cumulative distribution functions (CDF) of the amount of requested resources are computed in Section 5.

We consider the following slice isolation schemes:

- *Full isolation strategy.* According to this scheme, a slice of type i is initiated by the first arrival of a session that belongs to this slice during the so-called "off-period", i.e., the time with no session of this slice residing in the system. This session is accepted by the system if there are sufficient radio resources to serve it, whereas it is dropped otherwise. Each further session that belongs to slice i and arrives in the system during the slice i "on-period" may increase its duration. The slice "on-period" ends when a session that belongs to this slice leaves the system with no session of this slice remaining. At this time instant, the resources reserved for this slice are released. If there are no sufficient resources available for arriving session that belong to slice i, i.e., when $W_i(t) + \nu_i > B_i$, where $W_i(t)$ is the current amount of occupied resources of slice i, the session is dropped.
- *Full mixing strategy.* In this scheme, no isolation to slices is provided at the data-plane. Sessions that

belong to slices are not differentiated, and the decision about their acceptance is based on the full set of resources, B, upon session arrivals. This strategy allows to fully benefit from the statistical multiplexing at the data plane and is expected to provide an upper bound on traffic performance in the network with data plane traffic isolation. The best algorithms with traffic isolation should operate close to this bound.

3.4. SLA and Metrics of Interest

We assume that MVNO or service requester establishes an SLA with PMNO describing the set of QoS guarantees that need to be provided to the slice. Although 3GPP standards do not explicitly specify the potential SLA content, the set QoS metrics that need to be provisioned for a slice are logically divided into two layers: session and packet layer [31], see Table 1. At the session layer, there has to be a connection admission control (CAC) mechanism responsible for admitting slices to the systems and sessions to a slice. These decisions are based on the available resources in the system and resources requested by a slice/session. If these are satisfied, the packet scheduler can meet QoS metrics that have to be provided to individual packets.

In this work, we concentrate on the session level QoS provisioning. Thus, the main target is to ensure that the bandwidth allocated to a slice satisfies the constraints specified in SLA. Following GSMA guidelines [32], we consider the following metrics of interest: (i) fraction of time rate guarantees are violated, (ii) the intensity of rate violations, (iii) mean duration of slice rate violation, and (iv) network resource utilization. The ultimate target is to *identify the density of NR BS such that the metrics mentioned above are satisfied.*

4. The Effective NR BS Capacity

In this section, by accounting for propagation, blockage, and antenna specifics of mmWave NR systems, we characterize the effective NR BS capacity as a function of NR BS density. Note that we assume no interference cancellation techniques implying that the developed framework provides pessimistic bounds. However, it might be further extended to capture interference suppression techniques. This would require appropriate changes to the interference analysis performed below.

4.1. Taylor Approximation of Capacity Equation

To determine the NB BS capacity we obtain the capacity of a single arbitrarily chosen UE operating over the bandwidth B. Using the Shannon formula we may write

$$R{=}B\log_2\left[1{+}\frac{C_1Y_0^{-\zeta_1}[1{-}p_B(Y_0)]+C_2Y_0^{-\zeta_2}p_B(Y_0)}{BN_0+I}\right], \quad (10)$$

where C_i and ζ_i are constants specified in Section 3, $p_B(Y_0)$ is the blockage probability provided in (5), Y_0 is the distance to the serving NR BS, N_0 is the thermal noise, and I is the interference, defined as

$$I=\sum_{i=1}^{\infty}(C_1Y_i^{-\zeta_1}[1-p_B(Y_i)]+C_2Y_i^{-\zeta_2}p_B(Y_i)), \quad (11)$$

where Y_i are the distances to interfering NR BS.

Note that derivation according to (10) is a complex procedure as the achieved rate is actually a function of a random number of random variables (RV). Obtaining capacity pdf is thus a complicated task that cannot be done in RV domain [27]. To obtain the mean capacity, we thus employ Taylor series expansion of the capacity function $R=g(P_R,I)=B\log_2(1+P_R/(BN_0+I))$, where $P_R=C_1Y_0^{-\zeta_1}[1-p_B(Y_0)]+C_2Y_0^{-\zeta_2}p_B(Y_0)$ is the received signal strength in the nominator of (10). A second-order approximation is obtained by expanding $g(x_1,x_2)$ around $\vec{\mu}=(E[P_R],E[I])=(\mu_{P_R},\mu_I)$, which leads to [33]

$$\begin{aligned}E[g(\vec{\mu})]\approx g(\vec{\mu})+\frac{g''_{x_1x_1}(\vec{\mu})\sigma^2_{P_R}+2g''_{x_1x_2}(\vec{\mu})K_{P_R,I}}{2}+\\+\frac{g''_{x_2x_2}(\vec{\mu})\sigma_I^2}{2}+O(n),\end{aligned} \quad (12)$$

where $K_{P_R,I}$ is the covariance between P_R and I, while $\sigma^2_{P_R}$ and σ^2_I are the variances of P_R and I, respectively. Observing that

$$\begin{aligned}&g''_{x_1,x_1}(x_1,x_2)=g''_{x_1,x_2}(x_1,x_2)=\\&\qquad=-\frac{B}{\ln 2\,(BN_0+x_1+x_2)^2},\\&g''_{x_2,x_2}(x_1,x_2)=\\&\qquad=\frac{Bx_1\,(2BN_0+2x_2+x_1)}{\ln 2\,(BN_0+x_1+x_2)^2\,(BN_0+x_2)^2},\end{aligned} \quad (13)$$

we arrive at the following approximation

$$\begin{aligned}E[R]\approx B\log_2\left(1+\frac{\mu_{P_R}}{BN_0+\mu_I}\right)+\\+\frac{B}{2\ln 2\,(BN_0+\mu_I+\mu_{P_R})^2}\times\\\times\left[\frac{\sigma_I^2\mu_{P_R}\,(2BN_0+2\mu_I+\mu_{P_R})}{(BN_0+\mu_I)^2}-\left(2K_{P_R,I}+\sigma^2_{P_R}\right)\right],\end{aligned} \quad (14)$$

implying that we need to determine the first three moments of interference, moments of the received signal strength and covariance between them.

4.2. Derivation of Involved Components

Below, to determine interference and other components required in (14), we adopt a conventional stochastic geometry approach [34] heavily utilized in other studies (e.g., [27, 35] among others): (i) specify a circular zone around the target UE, (ii) assume that no interference is generated from the BS located outside, (iii) determine the interference coming from the BS located inside this zone.

Consider now the circle centered at the tagged UE of radius r_I. The radius r_I is computed such that the potential interference coming from NR BS located outside is less than the noise floor [27]. Recalling that NR BSs are assumed to organize PPP in $\Re^2$ we see that the number of them inside the circle of radius r_I follows a Poisson distribution with parameter $\xi\pi r_I^2$ BS/m^2 [30]. Further, observe that the distance from the tagged UE to /NR BS located in the circle, Y_0, is characterized by probability density function (pdf) $f_{Y_0}(y) = 2y/r_E^2$, $0 < y < r_E$, where r_E is the effective coverage of NR BS [36].

The effective coverage radius, r_E, is determined by the interplay between inter-BS distance, $r_{E,V}$, and the maximum coverage of NR BSs, $r_{E,S}$. Thus, we have $r_E = \min(r_{E,V}, r_{E,S})$. The latter component is defined as the maximum separation distance between the UE and the NR BS, such that the UE in the LoS blocked conditions is not in the outage. According to our propagation model, the SNR at the maximum 2D distance $r_{E,S}$ is

$$S = C_2 \left(r_{E,S}^2 + (h_A - h_U)^2\right)^{-\frac{\zeta_2}{2}} = S_{th}, \tag{15}$$

where S_{th} is the SNR corresponding to the lowest feasible NR MCS [24]. Solving for $r_{E,S}$, we obtain

$$r_{E,S} = \sqrt{(C_2/S_{th})^{\frac{2}{\zeta_2}} - (h_A - h_U)^2}. \tag{16}$$

We approximate the radius $r_{E,V}$, characterizing the half distance between NR BS locations, by a radius of a circle approximating the area of the Voronoi cell induced by NR BS locations in $\Re^2$. Since the actual area of the Voronoi cell is not known [37], we utilize computer simulations to obtain $r_{E,V}$.

Consider the received signal strength, $P_R = C_1 Y_0^{-\zeta_1}[1 - p_B(Y_0)] + C_2 Y_0^{-\zeta_2} p_B(Y_0)$. Observe, that it is a function of a single RV, Y_0, representing the distance to the nearest BS. Conditioning on m UEs in the circle of radius r_E, the CDF of the distance to the nearest NR BS is

$$F_{Y_0}(x) = 1 - (1 - F_Y(x))^m, \tag{17}$$

where the unknown term

$$F_Y(x) = \int_0^x f_Y(y)dy = \frac{x^2}{r_E^2},\ 0 < x < r_I, \tag{18}$$

is the CDF of the distance from the NR BS to UE.

Accounting for Poisson distributed number of UEs in the coverage area of NR BS we arrive at the following expression for the CDF of the distance to the NR BS

$$F_{Y_0}(x) = \frac{\sum_{m=1}^{\infty} \frac{(\xi\pi r_E^2)^m}{m!} e^{-\xi\pi r_E^2} \left[1 - \left(1 - \frac{x^2}{r_I^2}\right)^m\right]}{1 - e^{-\xi\pi r_E^2}}. \tag{19}$$

The raw moments of received signal now read as

$$E[P_R^n] = \int\limits_0^{r_E} \left[\frac{C_1}{x^{\zeta_1}}[1 - p_B(x)] + \frac{C_2}{x^{\zeta_2}} p_B(x)\right]^n dF_{Y_0}(x). \tag{20}$$

The moments of aggregated interference at a randomly selected UE from a field of interfering NR BS deployed with the density ξ can be found using the Campbell theorem [38] as follows

$$E[I^n] = \int_0^{r_I} \left[C_1 x^{-\zeta_1}(1 - p_B(x)) + C_2 x^{-\zeta_2} p_B(x)\right]^n \times \\ \times p_C(x) 2\xi\pi x dx, \tag{21}$$

where $2\xi\pi x dx$ is the probability of having an interferer in the infinitesimal increment of the circumference dx, $p_C(x)$ is the probability that the transmit antennas of the interfering nodes are oriented such that they contribute to the interference at the tagged UE (termed here the exposure probability), $p_B(x)$ is the LoS blockage probability.

Consider the exposure probability, $p_C(x)$. Assuming practical antenna arrays with only a few elements forming the radiation pattern in the vertical domain we approximate the exposure probability accounting for horizontal plane only [27], i.e.,

$$p_C(x) = p_C = \alpha_A \alpha_U / 4\pi^2, \tag{22}$$

where α_A and α_U are HPBW of antenna radiation pattern at NR BS and UE sides that can be directly estimated as a function of antenna elements as shown in Section 3.

In our case, UE is always associated with the nearest NR BS. The distance to this BS is obtained in (19). Thus, conditioned on this distance, we observe that there are no NR BS located closed than Y_0. Thus, the moments of interference are now

$$E[I^n] = \int_0^{r_I} f_{Y_0}(x) \Bigg(\int_x^{r_I} \Bigg[C_1 y^{-\zeta_1}(1 - p_B(y)) + \\ + C_2 y^{-\zeta_2} p_B(y) \Bigg]^n p_C(y) 2\xi\pi y dy \Bigg) dx. \tag{23}$$

where $f_{Y_0}(x)dx$ is the pdf of Y_0, i.e.,

$$f_{Y_0}(x) = F'_{Y_0}(x) = \frac{2x\xi\pi e^{-\xi\pi x^2}}{1 - e^{-\xi\pi r_E^2}}. \tag{24}$$

We determine the covariance of the received signal strength and interference. Expressing it as $K_{P_R,I} = E[P_R I] - E[P_R]E[I]$ we now observe that $E[P_R I]$ is the only unknown while $E[P_R]$ and $E[I]$ are provided in (20) and (25), respectively. The missing term $E[P_R I]$ can be obtained as

$$E[P_R I] = \int_0^{r_E} f_{Y_0}(x)[C_1 x^{-\zeta}[1 - p_B(x)] + C_2 x^{-\zeta} p_B(x)] \times \\ \times \left[\int\limits_x^{r_I} \left[\frac{C_1}{y^{\zeta}}(1 - p_B(y)) + \frac{C_2}{y^{\zeta}} p_B(y)\right] p_C(y) 2\xi\pi y dy \right] dx. \tag{25}$$

5. Dynamic Slice Service Process

Having characterized the service rate of NR BS, we now proceed with modeling the service process of slices at individual NR BS. To this aim, we first derive the session resource requirements and then continue specifying and solving the queuing system. Finally, we obtain the metrics of interest.

5.1. Session Resource Requirements

Accounting for the propagation model, we first derive the probability mass function (pmf) of the number of requested resources. Particularly, we determine the sought pmf by first establishing the pmf of the number of required resources in the LoS non-blocked and blocked states, and then weighing them with the corresponding probabilities. Since the SNR in the LoS non-blocked and blocked states differ only by a path loss exponent, we provide a detailed derivation of the pmf for the LoS non-blocked conditions as an example.

Similarly to the previous section, we tag an arbitrary UE within the coverage area of the NR BS. Let S_{nB} be a RV denoting SNR in the LoS non-blocked conditions and $F_{S_{nB}}(s)$, $s > 0$, be its CDFs. Since $S_{nB} = C_1 Y_0 / I$, the latter can be written as

$$F_{S_{nB}}(s) = P\left\{C_1 y^{-\zeta_1}/I < s\right\}. \qquad (26)$$

Observe that RV S_{nB} is a function of two RVs, the distance between NR BS and UE, Y_0, and interference, I. Unfortunately, the distribution of interference, I, cannot be obtained in closed-form in the RV domain, as discussed in the previous section. To provide a viable approximation, we introduce the interference margin M_I and set it equal to the mean interference estimated in the previous section. The task in (26), thus, reduces to linear transformation of RV Y_0 [39]. Distribution of SNR in the blocked state, $F_{S_B}(s)$, can be obtained similarly. To determine the overall SNR CDF, $F_S(s)$, we need to weight the these distributions with blockage probability provided in (5).

Let $S_j, j = 1, 2, \dots, J$, be the SNR thresholds and ϵ_j be the probability that the UE connection is assigned the Channel Quality Indicator (CQI) and the MCS j, thus requiring r_j resource units, $j = 1, 2, \dots, J$. Using $F_S(s)$, we write

$$\begin{cases} \epsilon_0 = F_S(S_1), \\ \epsilon_j = F_S(S_{j+1}) - F_S(S_j),\ j = 1, 2, \dots, J-1, \\ \epsilon_J = 1 - F_S(S_J). \end{cases} \qquad (27)$$

The probabilities ϵ_j and the rate R_i can now be used to determine the pmf of resource requirements of individual sessions.

5.2. Queuing Model

Both considered isolation schemes defined in Section 3 can be analyzed using the queuing system with a finite amount of resources and random session resource requirements. The rationale is that even when the requested rate of a session is the same, random locations of the UEs with respect to the NR BS lead to random resource requirements. For the full isolation scheme, one may apply it to individual slices while for the full mixing scheme, one may apply to the whole available set of resources.

There are techniques allowing to solve resource queuing systems with a finite amount of resources. However, exact analysis of such systems is rarely feasible and, thus, the authors refer to various approximations at different stages of analysis. These approximations include the state aggregation technique, Bayesian approximation for the amount of released resources by a customer leaving the system and simulations, see, e.g., [40, 41]. However, even empowered with these approximations, the derivation of the state probability distributions often requires complex numerical algorithms. To simplify computational requirements associated with obtaining the metrics of interest and, thus, provide and an easy to use network dimensioning model, in our study, we consider an approximation of the state probability distribution by firstly assuming that the resource volume is unlimited, and secondly, by constructing a Gaussian approximation of the probability distribution of the total volume of the occupied resource in the system. The advantage of the proposed approach is that it does not require any complex numerical algorithm.

The core of the approximation lies in the fact that in a system with an unlimited amount of resources, the probability that the total amount of occupied resources will take a value greater than some B is the upper limit for assessing the probability that exactly B resources will be occupied in a system with losses [42]. Although more refined loss probability estimates have been reported in literature [43, 44], those are not directly applicable due to random resource requirements associated with session arrivals.

5.2.1. Full Isolation

We first address the case of full isolation. Observe that in this case, the resource utilization of individual slices is independent. We may characterize the service process of sessions that belong to the same slice using the queuing system formulated below. Similarly to [45], we consider the queuing system with an unlimited amount of resources, see Fig. 1(c). The arrival process is assumed to be a stationary Poisson process with the parameter λ_i. Note that in practice, λ_i is estimated using the subscriber density in the area of interest intensity of their session requests. The session service time, $\eta_i \geq 0$, is assumed to be generally distributed with CDF $D_i(\tau) = P\{\eta_i < \tau\}$ with finite first moment. The amount of resources requested by each session in a slice is characterized by RV $\nu_i \geq 0$ with probability mass function (pmf) $\epsilon_k = P\{\nu_i = r_k\}$ with finite first two moments. Once the session's service is completed, the amount of resources occupied by this session is released.

The sum of all resources occupied by active slice sessions is called *total amount of the occupied resources* in the system. Let $W_i(t)$, $i = 1, 2, \dots, K$, be the total amount of the occupied resource in the system by ith slice at time t. The task is to find the stationary probability distribution of the stochastic process $W_i(t)$. The behavior of the system could be described by the stochastic process $\{W_i(t), t > t_0\}$ defined over the continual state space $\mathbf{W_i}$ denoting the amount of occupied PRBs. Observe that the process is not Markovian, as we need to take into account the fact that at the service completion moment, the session releases exactly the amount of the resource that it occupied upon arrival. Thus, the evolution of a stochastic process at an arbitrary time moment depends on the evolution preceding this moment. To solve the system, we resort to the method of dynamic screening originally introduced in [45, 46]. The main result is as follows.

Theorem 1. *The approximation of the stationary probability distribution of the total amount of the occupied resources by the i-th slice under the condition of increasing intensity of arrivals follows Gaussian distribution with the parameters $m_i = \lambda_i a_i d_i$ and $\sigma_i^2 = \lambda_i \alpha_i d_i$, where λ_i is the intensity of session arrivals, a_i and α_i are the first and second raw moments of the number of requested resources by a session ν_i, and d_i is the mean service time.*

The proof is provided in Appendix A.

5.2.2. Full Mixing

Now, consider the case when slices are not individually isolated and share the resources, i.e., $W(t) = \sum_{i=1}^{K} W_i(t)$. Using [47], we can formulate the next theorem.

Theorem 2. *The approximation of the stationary probability distribution of the total amount of the occupied resources by all of the slices under the condition of increasing intensity of arrivals follows Gaussian distribution with the parameters $m = \sum_{i=1}^{K} m_i$ and $\sigma^2 = \sum_{i=1}^{K} \sigma_i^2$.*

This theorem directly stems from Theorem 1, applying the central limit theorem and thus omitted here.

5.3. Metrics of Interest

5.3.1. Full Isolation

Having characterized the behavior of the queuing process we now proceed to obtain the metrics of interest. Recalling the results of the previous section, one may observe that the sought metrics can be expressed as characteristics of the Gaussian stochastic process with respect to bandwidth B_i specifying the rate provided to a slice in full isolation scheme or the whole available bandwidth for full mixing scheme. Recall that the trajectory of a continuous differentiable stochastic process $\{W_i(t), t > t_0\}$ intersects a fixed level B_i at some arbitrary point $t = t_s$, if in the vicinity of this point $[t_s - \Delta t/2, t_s + \Delta t/2]$ the following holds for arbitrarily small $\Delta t > 0$

$$[W_i(t_s - \Delta t/2) - B_i]\,[W_i(t_s + \Delta t/2) - B_i] < 0. \qquad (28)$$

Particularly, if the derivative is positive, i.e.,

$$W_i(t_s) = B_i, \quad W_i'(t_s) > 0, \qquad (29)$$

the intersection is from the bottom.

Having these properties in mind we now proceed to analyze the metrics of interest. The fraction of time, $T^+(B_i)$, the slice service rate guarantees are violated, coincides with the fraction of time the considered Gaussian stochastic process remains atop of B_i. That is, we have

$$T^+(B_i) = 1 - \Phi\left(\frac{B_i - m_i}{\sigma_i}\right), \qquad (30)$$

where $\Phi(x)$ is the Laplace error function

$$\Phi(x) = \frac{1}{2\pi}\int_{-\infty}^{x} e^{-\frac{y^2}{2}}\,dy. \qquad (31)$$

Considering the intensity of service violations we recall that to determine the mean number of outliers of a stochastic process per unit time one needs to determine the joint pdf (jpdf) of the stochastic process and its time derivative. For Gaussian processes, this metric can be expressed in terms of its autocorrelation function (ACF) which is derived in Appendix B. Further, applying the results of [48], the intensity of service violations is

$$N^+(B_i) = \frac{1}{2\pi}\frac{\sigma_i'}{\sigma_i}\exp\left(-\frac{(B_i - m_i)^2}{2\sigma_i^2}\right), \qquad (32)$$

where $\sigma_i' = R_i''(0)$ is the standard deviation of the derivative of $W_i(t)$.

The mean rate violation duration, $\tau_m(B_i)$, can be expressed as the ratio of the total mean duration, $T^+(B_i)$, stochastic process $W_i(t)$ remains atop B_i, to the mean number of such intervals $N^+(B_i)$, i.e.,

$$\tau_m(B_i) = T^+(B_i)/N^+(B_i). \qquad (33)$$

Finally, the resource utilization of the system is given by

$$r(B_i) = m_i/B_i. \qquad (34)$$

5.3.2. Full Mixing

Determining performance measures associated with individual slice performance in full mixing strategy is more involved. To compare considered strategies, we will consider the two measures: (i) the fraction of time rate guarantees are violated and (ii) the intensity of rate violations.

Recall that for the full isolation strategy, the fraction of the time, when rate guarantees are violated, is obtained utilizing (30). Thus, we may obtain the fraction time, when the rate guarantees for at least one of the slices are violated as

$$T^+(B) = 1 - \prod_{i=1}^{K}(1 - T^+(B_i)). \qquad (35)$$

To determine the intensity of rate violations for the full mixing strategy we observe that the service violation

Table 3: Parameters for numerical calculations.

Parameter	**Value**
Carrier frequency, f_c	28 GHz
Bandwidth, B	400 MHz
Number of slices, K	4
Slice bandwidth, B_i	100 MHz
Density of NR BS, ξ	$10^{-6}, \dots, 10^{-3}$, 1/m^2
Density of pedestrians, λ_B	0.1, 0.3, 0.5, 1.0 1/m^2
Height of the NR BS, h_A	6 m
Height of the UEs, h_U	1.5 m
Height of pedestrians, h_B	1.7 m
Radius of pedestrians, r_B	0.6 m
NR BS transmit power, P_A	2 W
Horizontal antenna elements at BS, N_A	$4, 8, 16, 32, 64, 128$
Antenna array gains at NR BS, G_A	$4 \times N_A$
Antenna array gains at UE, G_U	4×4
Power spectral density of noise, N_0	4.14×10^{-21} W/Hz
Fading margins in blocked state, $M_{S,2}$	6.61 dB
Session rates, R_i	5 Mbps
Default pedestrians density, λ_B	0.3, $1/m^2$
Default UE session intensity, λ_i	1/600, $1/m^2$/slice
Service intensity, μ	1/120 1/s

events of individual slices organize a point process in time with the aggregated intensity provided in (32), i.e.,

$$N^{+}(B) = \sum_{i=1}^{K} N_i^{+}(B). \tag{36}$$

Utilizing (37) we thus observe that the intensity of rate violations for an individual slice is proportional to the arrival intensity of this slice, i.e.,

$$N_i^{+}(B) = N^{+}(B)\frac{\lambda_i}{\sum_{i=1}^{K}\lambda_k}. \tag{37}$$

6. Numerical Results

In this section, we numerically elaborate on the results of the developed framework. We start assessing the accuracy of the developed model. Then, we proceed to characterize the wireless part reporting on the effective coverage and capacity of NR BS as well as mean resource characterization. Further, we concentrate our attention on analyzing the properties of the slicing service process with full isolation using the metrics of interest derived in the previous section. Finally, we compare full isolation and full mixing schemes. The default system parameters utilized in this section are provided in Table 3.

Note that in what follows, to keep the number of illustrations manageable, we perform our numerical analysis having the session arrival intensity fixed and varying the NR BS deployment density ξ. The reason is that the effect of the UE session arrival intensity is rather straightforward, i.e., higher values of ξ lead to worse performance in terms of service violation. Furthermore, these metrics are inter-related, and the conclusions formulated for a given session arrival intensity qualitatively hold for other intensities as well.

6.1. Accuracy Assessment

The developed simulation framework is based on discrete-event simulation (DES) technique [49] and implements the system specified in Section 3. The software is developed by using the Java programming language with multi-thread optimizations. The simulation procedure comprises two phases: DES simulations and data analysis. The rationale for providing a comparison with simulations is that the developed framework involves several approximations that are needed to retain the analytical tractability. Mainly, these are (i) approximation of the NR BS rate using Taylor series approximation, (ii) NR BS coverage approximation using circle having the area coinciding with the area of the corresponding Voronoi cell, (iii) approximation of the loss system with the equivalent system having infinite amount of resources. To assess the accuracy of the developed model, we implemented the considered deployment and isolation schemes in a computer simulator carefully accounting for all input system parameters.

The simulation data have been gathered during the steady-state period of the system, only using batch mean methods with the 10^3 batches each having 10^3 samples [50]. The beginning of the steady-state period has been determined using the exponentially weighted moving average (EWMA) statistics as outlined in [49]. Note that due to the ability to generate an exceptionally large number of samples and the use of batch means method, the confidence limits corresponding to the level of significance $\alpha = 0.05$ are minimal. Thus, in what follows, only point estimates are shown.

Fig. 2 illustrates the comparison between analytical data and data obtained using computer simulations for a fraction of time rate guarantees are violated for full isolation scheme. We specifically consider the system's practical operational regime, spanning the region from 0.1 and 10^{-5} session drop probabilities. We also illustrate the equivalent Erlang loss model, where each arriving session occupies a fixed amount of resources. Recall that the developed model provides the upper bound on the session loss probability while the equivalent Erlang model with averaged resource requirements sets the lower bound on

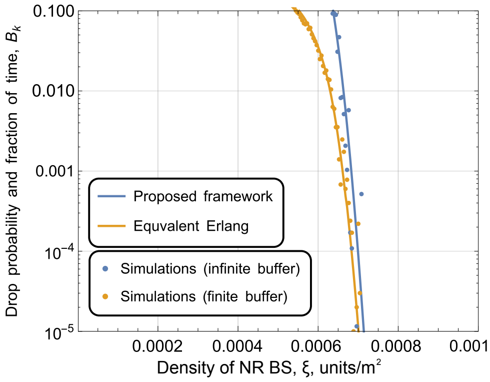
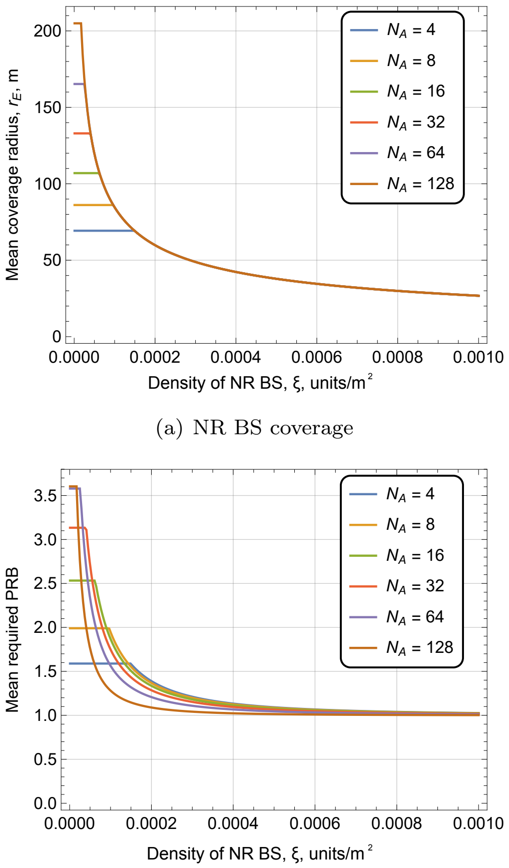


Figure 2: NR BS coverage and associated resource request.



(a) NR BS coverage

(b) Mean number of required PRBs

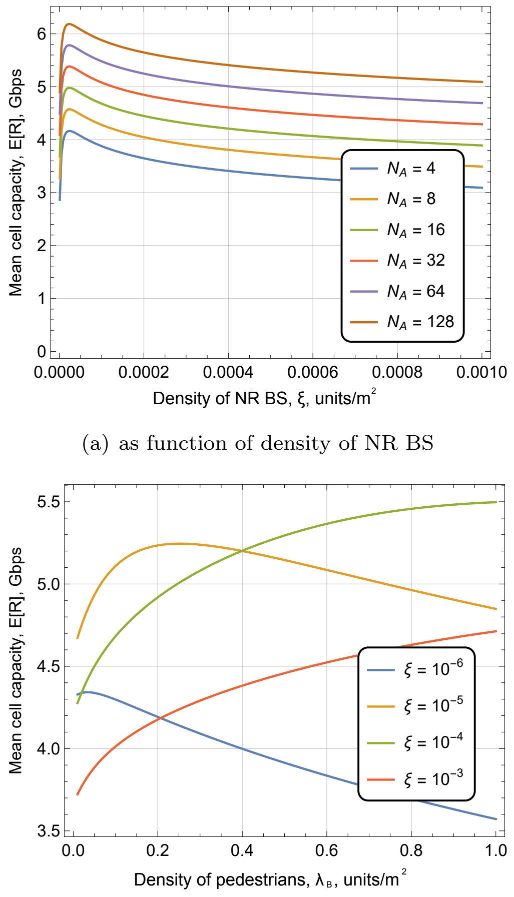
Figure 3: NR BS coverage and associated mean resource request (N_A is the number of horizontal antenna elements at NR BS).

system performance. As one may observe, the simulation data matches with the modeled one very well.

Furthermore, the equivalent Erlang model approaches the developed one as the density of NR BS increases. Similar conclusions have been observed for other metrics of interest and full mixing scheme as well. Thus, in what follows, we will use the model data only.

6.2. Network Rate and Coverage

We now proceed characterizing the intermediate parameters induced mainly by the radio part of the considered system. The coverage radius of a single NR BS is illustrated in Fig. 3(a) as a function of NR BS density, ξ, units/m^2 for several typical antenna configurations of NR BS. As one may observe, for rather small NR BS density, i.e., less than approximately 10^{-4} corresponding to the mean inter-site distance (ISD) of around 100 m, cell size is limited by the outage. In other words, NR BSs are sparsely distributed over the landscape. However, increasing ξ full coverage is achieved, and deployment starts to be limited by NR BS density. In this region, the interference begins to affect system performance. The increase in the number of antenna elements allows moving this boundary point to the region of higher values of ξ.



(a) as function of density of NR BS

(b) as function of density of pedestrians

Figure 4: NR BS aggregated cell rate as function of input parameters (N_A is the number of horizontal antenna elements at NR BS, ξ is the density of NR BSs).

Fig. 3(b) further shows the dependence of the mean resource request (expressed in terms of the mean number of PRBs) by a single UE as a function of NR BS density, ξ, $1/m^2$ for several typical antenna configurations of NR BS. These values are computed using resource request pmf used to calculate our metrics of interest in what follows. Observe that the mean resource requirements are characterized by a behavior similar to NR BS coverage in Fig. 3(a). The reason is that the amount of PRBs requested by a randomly selected UE is heavily affected by the cell coverage. Also, notice that starting from the bending point, the mean resource requirements gradually decrease along with the coverage radius. The reason is that a higher number of antenna elements are characterized by higher resource requirements, especially for low NR BS density ξ. However, this difference becomes smaller as ξ increases, and cell coverage starts to be mainly dictated by NR BS density and is also affected by interference.

The aggregated average cell spectral efficiency (that is, cell rate for $B = 1$ Hz) is illustrated in Fig. 4 as a function of input system parameters. Analyzing Fig. 4(a) one may observe that for all the considered antenna arrays at NR BS, the cell spectral efficiency peaks at rather small NR

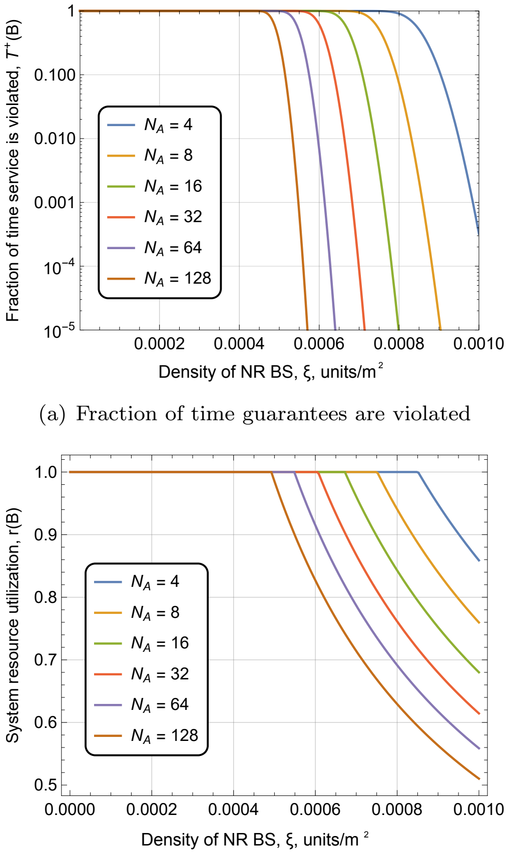
BS densities corresponding to large ISDs. The reason is that for shorter distances, the cell coverage is maximized as it is not affected by interference while at the higher NR BS densities, interference starts to play a significant role. Nevertheless, the decrease in the cell spectral efficiency is rather mild, and the system may still provide exceptional area capacity in the region of high values of NR BS density ξ. Note that the maximums in Fig. 4(a), corresponding to the number of horizontal elements at BS antenna array N_A, are attained at different x-coordinates compared to the bending points between coverage and interference constrained regimes in Fig. 3(b). The difference is attributed to the effect of the logarithmic function in the expression for spectral efficiency.

Finally, Fig. 4(b) illustrates the effect of the density of pedestrians, λ_B, on the cell capacity for several values of ξ. As one may observe, for rather sparse deployments, the effect is strictly negative. It is explained by the fact that the cell coverage is almost fully regulated by the cell edge UE outage probability and is practically unaffected by interference at these densities. When the NR BS density increases, the effect becomes more complicated. Particularly, for higher values of λ_B, the spectral efficiency still decreases. However, in the region of small values, we observe an increase in the cell spectral efficiency. This increase is explained by the transition from SNR limited regime to NR BS density limited regime. Further, an increase in the NR BS density leads to stable growth in the associated cell spectral efficiency despite higher interference experienced in the denser environment.

6.3. Full Isolation Scheme Performance

Empowered with the understanding of the radio part of the bearer service provided by NR technology, we now proceed to assess the performance of the full isolation mechanism. Particularly, in this section, we consider the coexistence of $K = 4$ slices in the system, where each slice is allocated $B_i = 100$ MHz of resources.

We start with the first-order service characteristics, including the fraction of time slice rate guarantees are violated and system resource utilization shown in Fig. 5(a) as a function of NR BS deployment density for several antenna configurations. First of all, as one may observe, the system performance in terms of the fraction of time guarantees are violated decreases when the antenna array at NR BS improves from 4×4 to 128×4. Particularly, the NR deployment with the latter antenna arrays may operate with much sparser density ξ of NR BSs. The reason is higher coverage of NR BS. Furthermore, the system changes its operational regime abruptly, and the higher the number of antenna elements makes the decline sharper. Particularly, the considered metrics drops from 1 to 10^{-5} for 128×4 array from 4.5×10^{-4} (ISD of $\sim$ 60 m) to 5.5×10^{-4} (ISD of $\sim$ 50 m) NR BS per squared meter. The similar drop for 4×4 array is observed between 7.5×10^{-4} (ISD of $\sim$ 40 m) to 9.5×10^{-4} (ISD of $\sim$ 20 m).



(a) Fraction of time guarantees are violated

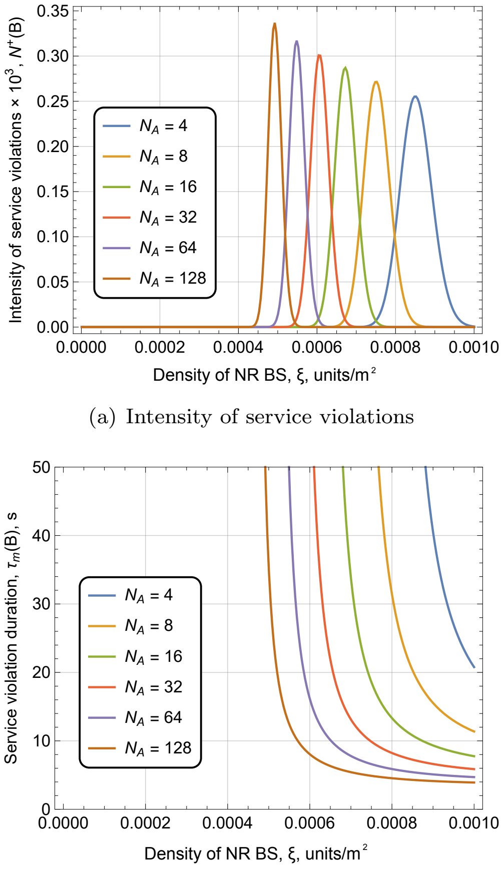
(b) System resource utilization

Figure 5: First-order service characteristics of full isolation scheme (N_A is the number of horizontal antenna elements at NR BS).

The associated resource utilization of the system is shown in Fig. 5(b). Expectedly, when in an overloaded regime, all the system resources are fully utilized. However, as the antenna array at NR BS is improved, the resource utilization decreases. A hyperbolic decrease in the considered parameter is also observed when the density of NR BS increases. The effect is different from the reported exponential decrease in [27], as the cell capacity is negatively affected by higher interference when ξ increases.

The system's practical operational regime starts when the fraction of time rate guarantees are violated close to zero. However, the service is also negatively affected by higher-order metrics of interest, including the intensity of service violations that defines how often rate guarantees are violated. This metric is shown in Fig. 6(a) as a function of NR BS deployment density for several antenna configurations. Further analysis is naturally complemented with a mean duration of rate guarantees violation illustrated in Fig. 6(b) for the same input parameters.

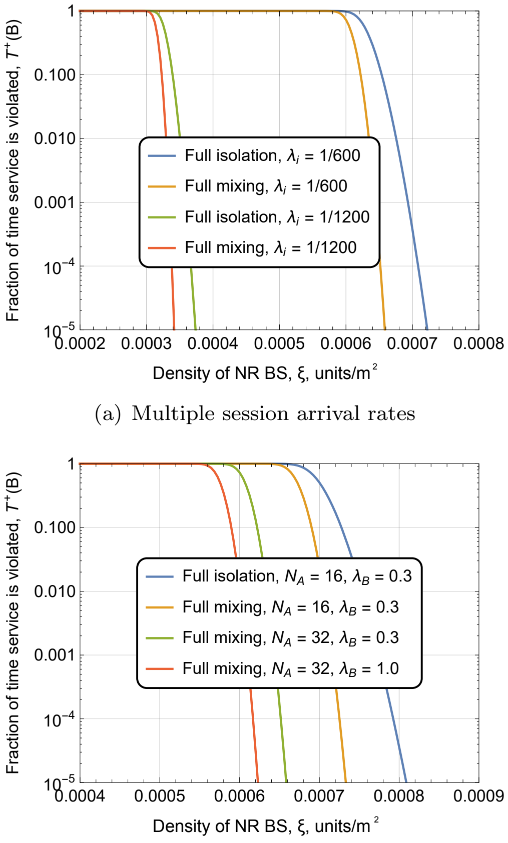
The bell-shaped structure of the intensity function is explained by the fact that overflow periods are first extremely long leading to small intensity of rate violations. The associated mean duration of rate violation is extremely large, as shown in Fig. 6(b) for this region of ξ values. Further increase in NR BS density decreases the load im-



(a) Intensity of service violations

(b) Mean duration of rate violation

Figure 6: Second-order service characteristics of full isolation scheme (N_A is the number of horizontal antenna elements at NR BS, λ_B is the pedestrians density).



(a) Multiple session arrival rates

(b) Multiple pedestrians density

Figure 7: Fraction of rate guarantees violation (N_A is the number of horizontal antenna elements at NR BS, λ_B is the pedestrians density, λ_i is the session arrival intensity of slice i).

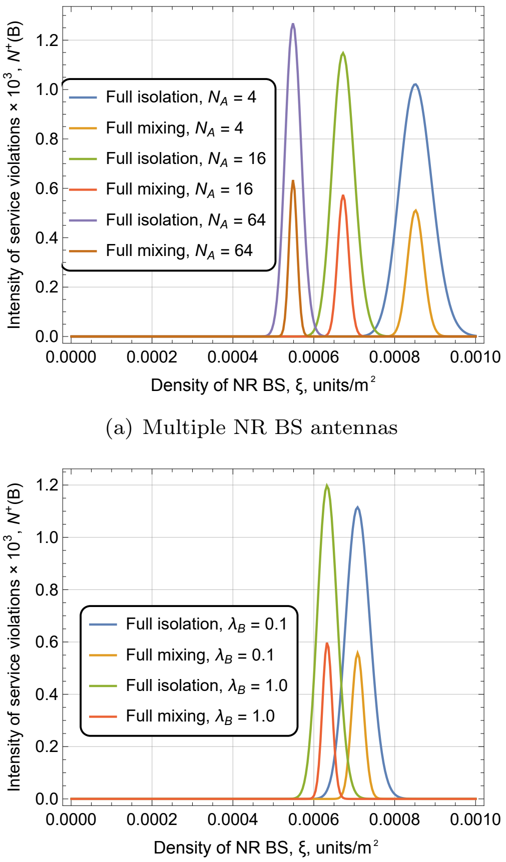
posed at a single NR BS, increasing the amount of time rate guarantees are satisfied as we have noticed analyzing Fig. 5(a). Thus, the associated rate violation intensity increases and then peaks. The further decrease in intensity is associated with a further decrease in the load at NR BS. Observe that the mean duration of rate violation related to this decrease also diminishes, see Fig. 6(b). Although the considered metric eventually reaches minimal values, it happens later than the fraction of time rate guarantees are violated approaches zero, setting up the lower bound on the number of NR BS density.

6.4. Comparison of Full Isolation and Full Mixing Schemes

Having revealed the properties of the full isolation scheme, we can compare its performance to the full mixing strategy characterizing the gap provided by statistical multiplexing of slices within a fully shared set of resources. We use two metrics of interest to perform this comparison: (i) fraction of time rate guarantees are violated and (ii) intensity of rate guarantees violations. Note that the calculation of the latter metric is different for the full mixing scheme.

The fraction of time rate guarantees are violated for full isolation, and full mixing schemes are shown in Fig. 7 for different session arrival intensities to a single slice λ_k, the number of horizontal antenna elements at NR BS, N_A, and pedestrians density in the environment, λ_B. Expectedly, the full mixing scheme outperforms the one with full isolation exposing the gap of approximately 10^{-4} between the densities of NR BS required to achieve a nearly close-to-zero fraction of time rate guarantees are violated. The revealed difference between them is rather small for small session arrival intensities and increases when the system load becomes higher, see Fig. 7(a) or when more sophisticated antenna arrays are utilized at NR BS, see Fig. 7(b). In practice, it means that the full isolation scheme can be used at the initial market penetration phase without compromising system performance. However, denser deployments, especially those associated with high traffic conditions, may require more sophisticated schemes approaching full mixing performance for cost efficiency.

Fig. 8 shows the second-order characteristic – the intensity of rate guarantees violations for considered schemes as a function of the number of horizontal antenna elements at NR BS, N_A, and pedestrians density in the environment, λ_B. As one may observe, qualitatively, the structure of this metric (i.e., bell-shaped curve) is similar for considered schemes. Quantitatively, it decreases drastically in absolute values, approximately by half, when the full mixing scheme is employed. In practice, it implies that the



(a) Multiple NR BS antennas

(b) Multiple pedestrians density

Figure 8: Intensity of rate guarantees violations.

density of NR BSs for the full isolation scheme needs to be increased to reach the same performance level provided by the full mixing scheme.

7. Conclusions

In this paper, aiming at quantifying the performance of mmWave 5G NR deployments supporting the network slicing mechanism at the air interface, we unite the tools of stochastic geometry and queuing theory to develop a detailed performance evaluation framework capable of characterizing two extreme cases, full isolation and full mixing between slices. From the practical point of view, we have provided a tool to determine the density of NR BSs needed to satisfy rate guarantees in the presence of K dynamics slices in the system with different traffic characteristics. Using the developed framework, PMNOs may evaluate QoS metrics that can be delivered to a given composition of slices in their network. The developed framework can also be utilized as a baseline for the development of more sophisticated isolation strategies required by network operators.

Our numerical evaluation campaign reveals that the full isolation and full mixing systems' operational regime changes rather abruptly with respect to the density of NR BSs. However, the system parameters may drastically affect the required density. When deploying operational systems, one may also account for second-order properties of the considered scheme, including the intensity and duration of rate violations. Even when the fraction of time when rate guarantees are violated is low, these metrics may still take on high values leading to frequent service interruptions. Furthermore, the difference between considered schemes is more significant for environments with high session arrival intensities that naturally require dense deployments. Practically, it implies that at the initial market penetration phase, the full isolation strategy can be utilized without compromising the network performance. However, at mature deployment phases, more sophisticated schemes may reduce the capital expenditures of network operators.

References

[1] 3GPP, Service requirements for V2X services (Release 14), 3GPP TS 22.185 V14.3.0 (March 2017).

[2] J. Ordonez-Lucena, P. Ameigeiras, D. Lopez, J. J. Ramos-Munoz, J. Lorca, J. Folgueira, Network slicing for 5g with sdn/nfv: Concepts, architectures, and challenges, IEEE Communications Magazine 55 (5) (2017) 80–87.

[3] K. Samdanis, X. Costa-Perez, V. Sciancalepore, From network sharing to multi-tenancy: The 5g network slice broker, IEEE Communications Magazine 54 (7) (2016) 32–39.

[4] S. O. Oladejo, O. E. Falowo, 5g network slicing: A multi-tenancy scenario, in: 2017 Global Wireless Summit (GWS), IEEE, 2017, pp. 88–92.

[5] I. Ahmad, T. Kumar, M. Liyanage, J. Okwuibe, M. Ylianttila, A. Gurtov, Overview of 5g security challenges and solutions, IEEE Communications Standards Magazine 2 (1) (2018) 36–43.

[6] GSM Association, Generic network slice template, NG.116 (May 2019).

[7] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. De Silva, F. Tufvesson, A. Benjebbour, G. Wunder, 5g: A tutorial overview of standards, trials, challenges, deployment, and practice, IEEE journal on selected areas in communications 35 (6) (2017) 1201–1221.

[8] 3GPP, NR; User Equipment (UE) radio transmission and reception; Part 1: Range 1 Standalone, 3GPP TS 38.101-1 (Jan. 2020).

[9] A. Ksentini, N. Nikaein, Toward enforcing network slicing on ran: Flexibility and resources abstraction, IEEE Communications Magazine 55 (6) (2017) 102–108.

[10] G. R. MacCartney, T. S. Rappaport, S. Rangan, Rapid fading due to human blockage in pedestrian crowds at 5g millimeter-wave frequencies, in: GLOBECOM 2017-2017 IEEE Global Communications Conference, IEEE, 2017, pp. 1–7.

[11] M. Gapeyenko, A. Samuylov, M. Gerasimenko, D. Moltchanov, S. Singh, M. R. Akdeniz, E. Aryafar, N. Himayat, S. Andreev, Y. Koucheryavy, On the temporal effects of mobile blockers in urban millimeter-wave cellular scenarios, IEEE Transactions on Vehicular Technology 66 (11) (2017) 10124–10138.

[12] A. Samuylov, M. Gapeyenko, D. Moltchanov, M. Gerasimenko, S. Singh, N. Himayat, S. Andreev, Y. Koucheryavy, Characterizing spatial correlation of blockage statistics in urban mmwave systems, in: 2016 IEEE Globecom Workshops (GC Wkshps), IEEE, 2016, pp. 1–7.

[13] 3GPP, System Architecture for the 5G System, TS 23.501 V15.2.0 (July 2018).

[14] GSM Association, Network slicing use case requirements, Tech. rep. (Apr. 2018).

[15] S. Retal, M. Bagaa, T. Taleb, H. Flinck, Content delivery network slicing: Qoe and cost awareness, in: 2017 IEEE International Conference on Communications (ICC), IEEE, 2017, pp. 1–6.

[16] J. Ni, X. Lin, X. S. Shen, Efficient and secure service-oriented authentication supporting network slicing for 5g-enabled iot, IEEE Journal on Selected Areas in Communications 36 (3) (2018) 644–657.

[17] C. Campolo, A. Molinaro, A. Iera, F. Menichella, 5g network slicing for vehicle-to-everything services, IEEE Wireless Communications 24 (6) (2017) 38–45.

[18] H. Zhang, N. Liu, X. Chu, K. Long, A.-H. Aghvami, V. C. Leung, Network slicing based 5g and future mobile networks: mobility, resource management, and challenges, IEEE communications magazine 55 (8) (2017) 138–145.

[19] P. Caballero, A. Banchs, G. De Veciana, X. Costa-Pérez, A. Azcorra, Network slicing for guaranteed rate services: Admission control and resource allocation games, IEEE Transactions on Wireless Communications 17 (10) (2018) 6419–6432.

[20] G. Garcia-Aviles, M. Gramaglia, P. Serrano, A. Banchs, Posens: A practical open source solution for end-to-end network slicing, IEEE Wireless Communications 25 (5) (2018) 30–37.

[21] S. E. Elayoubi, S. B. Jemaa, Z. Altman, A. Galindo-Serrano, 5g ran slicing for verticals: Enablers and challenges, IEEE Communications Magazine 57 (1) (2019) 28–34.

[22] C. Sexton, N. Marchetti, L. A. DaSilva, Customization and trade-offs in 5g ran slicing, IEEE Communications Magazine 57 (4) (2019) 116–122.

[23] H. Xiang, W. Zhou, M. Daneshmand, M. Peng, Network slicing in fog radio access networks: Issues and challenges, IEEE Communications Magazine 55 (12) (2017) 110–116.

[24] 3GPP, NR; Physical channels and modulation (Release 15), 3GPP TR 38.211 (Dec 2017).

[25] 3GPP, Study on channel model for frequencies from 0.5 to 100 GHz (Release 14), 3GPP TR 38.901 V14.1.1 (July 2017).

[26] M. Gapeyenko, A. Samuylov, M. Gerasimenko, D. Moltchanov, S. Singh, E. Aryafar, S.-p. Yeh, N. Himayat, S. Andreev, Y. Koucheryavy, Analysis of human-body blockage in urban millimeter-wave cellular communications, in: Communications (ICC), 2016 IEEE International Conference on, IEEE, 2016, pp. 1–7.

[27] V. Petrov, M. Komarov, D. Moltchanov, J. M. Jornet, Y. Koucheryavy, Interference and SINR in Millimeter Wave and Terahertz Communication Systems With Blocking and Directional Antennas, IEEE Trans. on Wirel. Comm. 16 (3) (2017) 1791–1808.

[28] S. Singh, R. Mudumbai, U. Madhow, Interference analysis for highly directional 60-ghz mesh networks: The case for rethinking medium access control, IEEE/ACM Transactions on Networking (TON) 19 (5) (2011) 1513–1527.

[29] A. B. Constantine, et al., Antenna theory: analysis and design, Microstrip Antennas (third edition), John Wiley & Sons.

[30] J. F. C. Kingman, Poisson processes, Wiley Online Library, 1993.

[31] 3GPP, Study on Enhancement of Network Slicing (Release 16), 3GPP TS 23.740 V16.0.0 (Dec. 2018).

[32] GSMA, Generic Network Slice Template, NG.116 V1.0, GSM Association (May 2019).

[33] M. G. Kendall, et al., The advanced theory of statistics. vols. 1., The advanced theory of statistics. Vols. 1. 1 (Ed. 4).

[34] M. Haenggi, Stochastic geometry for wireless networks, Cambridge University Press, 2012.

[35] M. Haenggi, J. G. Andrews, F. Baccelli, O. Dousse, M. Franceschetti, Stochastic geometry and random graphs for the analysis and design of wireless networks, IEEE journal on selected areas in communications 27 (7) (2009) 1029–1046.

[36] D. Moltchanov, Distance distributions in random networks, Elsevier Ad Hoc Networks 10 (6) (2012) 1146–1166.

[37] M. Tanemura, Statistical distributions of poisson voronoi cells in two and three dimensions, FORMA-TOKYO- 18 (4) (2003) 221–247.

[38] S. N. Chiu, D. Stoyan, W. S. Kendall, J. Mecke, Stochastic geometry and its applications, John Wiley & Sons, 2013.

[39] S. M. Ross, Introduction to probability models, Academic press, 2014.

[40] V. Begishev, D. Moltchanov, E. Sopin, A. Samuylov, S. Andreev, Y. Koucheryavy, K. Samouylov, Quantifying the impact of guard capacity on session continuity in 3gpp new radio systems, IEEE Transactions on Vehicular Technology.

[41] E. S. Sopin, K. A. Ageev, S. Shorgin, Simulation of the limited resources queuing system for performance analysis of wireless networks., in: ECMS, 2018, pp. 505–509.

[42] H. Sakasegawa, M. Miyazawa, G. Yamazaki, Evaluating the Overflow Probability Using the Infinite Queue, Management Science 39 (10) (1993) 1238–1245.

[43] F. N. Gouweleeuw, H. C. Tijms, Computing Loss Probabilities in Discrete-Time Queues, Operations Research 46 (1) (1998) 149–154. `doi:10.1287/opre.46.1.149`.

[44] A. Gyorgy, T. Borsos, Estimates on the packet loss ratio via queue tail probabilities, in: GLOBECOM'01. IEEE Global Telecommunications Conference (Cat. No.01CH37270), Vol. 4, 2001, pp. 2407–2411 vol.4.

[45] E. Lisovskaya, S. Moiseeva, M. Pagano, The Total Capacity of Customers in the Infinite-Server Queue with MMPP Arrivals, Communications in Computer and Information Science (2016) 110–120.

[46] A. Moiseev, A. Nazarov, Queueing network $MAP-(GI/\infty)K$ with high-rate arrivals, European Journal of Operational Research 254 (1) (2016) 161 – 168. `doi:https://doi.org/10.1016/j.ejor.2016.04.011`.

[47] B. R. Levin, Theoretical Foundations of Statistical Radio Engineering, Radio and communication, 1989, (in Russian).

[48] V. I. Tikhonov, V. I. Himenko, Ejections of trajectories of stochastic processes, Nauka, 1987.

[49] H. G. Perros, Computer simulation techniques: The definitive introduction! (2009).

[50] G. S. Fishman, L. S. Yarberry, An implementation of the batch means method, INFORMS Journal on Computing 9 (3) (1997) 296–310.

Appendix A.

We proceed to find the stationary probability distribution of the stochastic process $W_i(t)$. We fix a certain time instant $T > t_0$ and mark all the epochs of session arrivals on the upper time axis, see Fig. A.9. Further, we consider only those sessions that have not completed their service by the time T. Now, we translate the arrival time instants of the original process to the screened process (lower time axis in Fig. A.9). Consider the probability that a session that has arrived at the system at the time instant t and occupied ν_i resources will not finish its service up until the instant T. Denoting this probability by $S_i(t)$ we observe that $S_i(t) = 1 - D_i(T-t)$. On the other hand, the session finishes its service by time T and releases all the occupied resources with probability $1 - S_i(t)$. In the screened process, the latter sessions are not taken into account.

Denote the total amount of resources occupied by screened sessions by $V_i(t)$. The associated state space is $\mathbf{V_i}$. It has been shown in [45] that at the time instant $t = T$, the pmfs of RVs $W_i(T)$ and $V_i(T)$ coincide, i.e.,

$$P\{W_i(T) < v_i\} = P\{V_i(T) < v_i\},\, v_i > 0. \tag{A.1}$$

Applying the total probability law to the pdf of the stochastic process $V_i(t)$, $\partial P\{V_i(t) = v_i\}/\partial v_i = p(v_i, t)$,

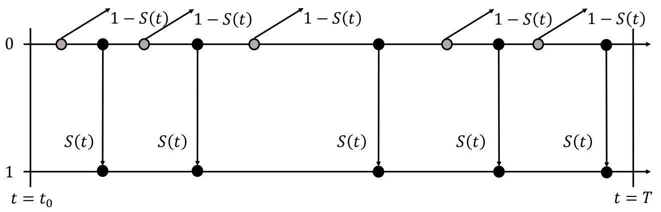


Figure A.9: Screening of arriving sessions.

$v_i > 0$, we can write the system of balance (Kolmogorov) differential equations in the following form

$$\frac{\partial p(v_i,t)}{\partial t} = \lambda_i S_i(t)\left[\sum_{k=0}^{J} p(v_i - r_k, t)\epsilon_k - p(v_i,t)\right], \quad \text{(A.2)}$$

with the initial condition

$$p(v_i, t_0) = \begin{cases} 1, & \text{if } v_i = 0, \\ 0, & \text{otherwise.} \end{cases} \quad \text{(A.3)}$$

Introduce the partial characteristic functions

$$h(z,t) = \sum_{v_i=0}^{\infty} e^{jzv_i} p(v_i,t), \quad j = \sqrt{-1}. \quad \text{(A.4)}$$

Now, using the introduced functions (A.4) we may rewrite the system of differential equations (A.2) as

$$\frac{\partial h(z,t)}{\partial t} = \lambda_i S_i(t) h(z,t)\left[G_i^*(z) - 1\right], \quad \text{(A.5)}$$

with the initial condition $h(y, t_0) = 1$ and

$$G_i^*(z) = \sum_{k=0}^{J} e^{jzr_k}\epsilon_k. \quad \text{(A.6)}$$

We now look for the solution of (A.5) using the asymptotic analysis under the condition of increasing intensity of arrivals, i.e., $\lambda_i \to \infty$. Using the following substitutions

$$\varepsilon = \frac{1}{\lambda_i}, \quad z = \varepsilon\overline{z}, \quad h(z,t) = f_1(\overline{z}, t, \varepsilon), \quad \text{(A.7)}$$

we rewrite (A.5) as

$$\varepsilon\frac{\partial f_1(\overline{z},t,\varepsilon)}{\partial t} = f_1(\overline{z},t,\varepsilon) S_i(t)(G_i^*(\varepsilon\overline{z}) - 1), \quad \text{(A.8)}$$

with the initial condition now expressed by $f_1(\overline{z}, t_0, \varepsilon) = 1$.

Thus, we look for the asymptotic solution

$$f_1(\overline{z},t) = \lim_{\varepsilon\to 0} f_1(\overline{z},t,\varepsilon),\ \varepsilon \to 0. \quad \text{(A.9)}$$

Applying the decomposition

$$e^{j\varepsilon\overline{z}} = 1 + j\varepsilon\overline{z} + O(\varepsilon^2), \quad \text{(A.10)}$$

we derive the following differential equation for $f_1(\overline{z}, t)$

$$\frac{\partial f_1(\overline{z},t)}{\partial t} = j\overline{z} a_i S_i(t) f_1(\overline{z},t), \quad \text{(A.11)}$$

where $a_i = \sum_{k=0}^{J} r_k\epsilon_k$ is the mean value of the amount of resources requested by a single session.

The solution of (A.11) has the following form

$$f_1(\overline{z},t) = \exp\left\{ j\overline{z} a_i \int_{t_0}^{t} S_i(\tau) d\tau \right\}. \quad \text{(A.12)}$$

Accounting for (A.7), we have the following asymptotic approximation

$$h(z,t) \approx \exp\left\{ jz\lambda_i a_i \int_{t_0}^{t} S_i(\tau) d\tau \right\}. \quad \text{(A.13)}$$

Let us now introduce function $h(z,t)$ in the form

$$h(z,t) = h_2(z,t)\exp\left\{ jz\lambda_i a_i \int_{t_0}^{t} S_i(\tau) d\tau \right\}. \quad \text{(A.14)}$$

Substituting (A.14) into (A.5), we obtain

$$\begin{aligned} \frac{\partial h_2(z,t)}{\partial t} &+ jz\lambda_i a_i S_i(t) h_2(z,t) = \\ &= h_2(z,t)\lambda_i S_i(t)(G_i^*(z) - 1), \end{aligned} \quad \text{(A.15)}$$

with the initial condition $h_2(z, t_0) = 1$. Using

$$\varepsilon^2 = \frac{1}{\lambda_i}, \quad z = \varepsilon\overline{z}, \quad h_2(z,t) = f_2(\overline{z},t,\varepsilon), \quad \text{(A.16)}$$

and substituting (A.7) into (A.15), we obtain

$$\begin{aligned} \varepsilon^2\frac{\partial f_2(\overline{z},t,\varepsilon)}{\partial t} &+ j\varepsilon\overline{z} a_i S_i(t) f_2(\overline{z},t,\varepsilon) = \\ &= f_2(\overline{z},t,\varepsilon) S_i(t)(G_i^*(\varepsilon\overline{z}) - 1), \end{aligned} \quad \text{(A.17)}$$

with initial condition $f_2(\overline{z}, t_0, \varepsilon) = 1$.

We now proceed determining the asymptotic solution $f_2(\overline{z},t) = \lim_{\varepsilon\to 0} f_2(\overline{z},t,\varepsilon)$ of (A.17) when $\varepsilon \to 0$. Applying the following decomposition

$$e^{j\varepsilon\overline{z}} = 1 + j\varepsilon\overline{z} + \frac{(j\varepsilon\overline{z})^2}{2} + O(\varepsilon^3), \quad \text{(A.18)}$$

we derive the following equation

$$\frac{\partial f_2(\overline{z},t)}{\partial t} = \frac{(j\overline{z})^2}{2}\alpha_i S_i(t) f_2(\overline{z},t), \quad \text{(A.19)}$$

where $\alpha_i = \sum_{k=0}^{J} r_k^2\epsilon_k$ is the second raw moment of the amount of resources requested by a single session.

The solution of (A.19) is given by

$$f_2(\overline{z},t) = \exp\left\{ \frac{(j\overline{z})^2}{2}\alpha_i \int_{t_0}^{t} S_i(\tau) d\tau \right\}. \quad \text{(A.20)}$$

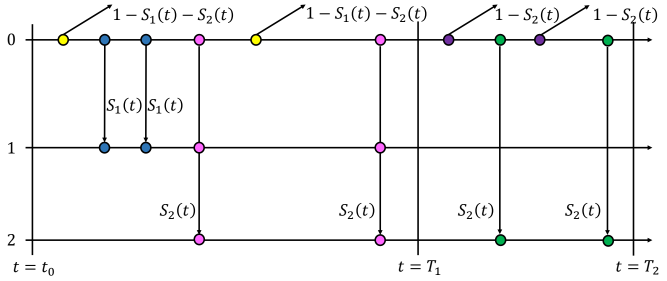


Figure B.10: Two-dimensional screening of the arrival process.

Using (A.16) and (A.14), we finally obtain the sought characteristic function

$$h(z,t) = \exp\left\{\lambda_i\left[jza_i + \frac{(jz)^2}{2}\alpha_i\right]\int_{t_0}^{t} S_i(\tau)d\tau\right\}. \quad \text{(A.21)}$$

Appendix B.

We proceed to obtain the ACF of the derivative of the Gaussian process $W_i(t)$ required to determine the intensity of service violations. To this aim, we apply the same methodology as in Appendix A. However, we consider the screening of the arrival process on two time axes for two arbitrary time instants, T_1 and T_2, as shown in Fig. B.10, where the auxiliary axis labeled "Axis 0" displays the arrival time instants. Apply the following rules to the events occurring in the time interval $t_0 \leq t < T_1$:

1. An arrival at time t is screened to Axis 1 only if its service is completed by time T_1. These arrivals are marked by blue color. Then, screening probability to Axis 1 is provided by

$$\begin{aligned} S_1(t) = P\{T_1 - t < \eta_i < T_2 - t\} = \\ = D_i(T_2 - t) - D_i(T_1 - t), \end{aligned} \quad \text{(B.1)}$$

   where η_i is a service time.
2. An arrival at time t is screened to Axis 2 only if its service is not completed by time $T_2 > T_1 > t$, i.e., it remains in the system after T_2. These arrivals are marked by pink color. The screening probability to Axis 2 is then

$$S_2(t) = P\{\eta_i > T_2 - t\} = 1 - D_i(T_2 - t). \quad \text{(B.2)}$$

3. Finally, an arrival at time t is not screened on any of the axes if its service is completed in (T_1, T_2). These arrivals are marked by yellow color. The probability of this event is readily obtained as

$$1 - S_1(t) - S_2(t) = P\{\eta_i < T_1 - t\} = D_i(T_1 - t). \quad \text{(B.3)}$$

Consider now the events in the interval $T_1 \leq t < T_2$:

1. An arrival is screened to Axis 2 only if its service is completed by time T_2. These arrivals are marked with green color and their screening probability is

$$S_2(t) = P\{\eta_i > T_2 - t\} = 1 - D_i(T_2 - t). \quad \text{(B.4)}$$

2. An arrival is not screened to any of the axes if its service is not completed by time T_2. These arrivals are marked with purple color and their screening probability is

$$1 - S_2(t) = P\{\eta_i < T_2 - t\} = D_i(T_2 - t). \quad \text{(B.5)}$$

Let now $\{V_1(t), V_2(t), t > t_0\}$ be the two-dimensional stochastic process of the total amount of resources occupied by the screened arrivals on the Axes 1 and 2 and let

$$p(v_1, v_2, t) = \partial^2/(\partial v_1 \partial v_2) P\{V_1(t) < v_1, V_2(t) < v_2\}, \quad \text{(B.6)}$$

be its jpdf.

Since the process is Markov in nature we may write the system of Kolmogorov differential equations for $p(v_1, v_2, t)$ in the following form

$$\begin{aligned} &\frac{\partial p(v_1, v_2, t)}{\partial t} = \lambda_i S_1(t) \sum_{k=1}^{J} p(v_1 - r_k, v_2, t) G_i(k) + \\ &+ \lambda_i S_2(t) \sum_{k=1}^{J} \sum_{l=1}^{J} p(v_1 - r_k, v_2 - r_l, t) G_i(k) G_i(l) - \\ &- [\lambda_i S_1(t) + \lambda_i S_2(t)]\, p(v_1, v_2, t),\ t_0 \leq t < T_1, \\ &\frac{\partial p(v_1, v_2, t)}{\partial t} = \lambda_i S_2(t) \sum_{k=1}^{J} p(v_1, v_2 - r_k, t) G_i(k) - \\ &- \lambda_i S_2(t) p(v_1, v_2, t),\ T_1 \leq t < T_2. \end{aligned} \quad \text{(B.7)}$$

Introducing the characteristic function

$$h(z_1, z_2, t) = \sum_{v_1=0}^{\infty} e^{jz_1 v_1} \sum_{v_2=0}^{\infty} e^{jz_2 v_2} p(v_1, v_2, t), \quad \text{(B.8)}$$

one may now utilize the asymptotic analysis to solve the resulting equation similarly to Section (5) to obtain the sought jpdf. Differentiating it twice with respect to u_1 and u_2, setting $t_0 \to -\infty$, which is inherent for a stationary regime, we obtain the the ACF in the following form

$$R_i(\tau) = \lambda_i a_i^2 \left[\int_{\tau}^{\infty} (1 - D_i(x))dx\right]. \quad \text{(B.9)}$$

Finally, the standard deviation required in (32) is

$$\sigma_i' = R_i''(0). \quad \text{(B.10)}$$