

Joint Use of Guard Capacity and Multiconnectivity for Improved Session Continuity in Millimeter-Wave 5G NR Systems

Vyacheslav Begishev, Eduard Sopin, Dmitri Moltchanov, Roman Kovalchukov, Andrey Samuylov, Sergey Andreev, *Senior Member, IEEE*, Yevgeni Koucheryavy, *Senior Member, IEEE*, and Konstantin Samouylov

Abstract—The intermittent nature of millimeter wave (mmWave) links caused by human-body blockage is an intrinsic property of the 5G New Radio (NR) technology that may cause drops of sessions already accepted for service. To improve the session continuity, multiconnectivity and guard capacity mechanisms have been proposed recently. Multiconnectivity enables dynamic handover between multiple pre-established spatially-diverse links, while guard capacity reserves a fraction of radio resources for the already accepted sessions by ensuring that they will have sufficient provisions in case of link blockage. In this study, we combine the tools of queuing theory and stochastic geometry to develop a mathematical framework for capturing the joint operation of these two schemes as well as the features of mmWave radio propagation. The metrics are related to user- and system-centric performance including the system resource utilization and the new and ongoing session drop probabilities. Our results show that multiconnectivity benefits all of the considered parameters. However, the range of performance boost remains limited by the deployment density and the maximum supported degree of multiconnectivity. In its turn, guard capacity allows to further decrease the ongoing session drop probability at the expense of the new session drop probability and the system resource utilization. When implemented jointly with multiconnectivity, guard capacity does not produce noticeable negative effects on the system resource utilization as compared to its standalone use. Hence, one may prefer a joint implementation of these mechanisms for preserving the session continuity of users without compromising the resource utilization.

Index Terms—5G cellular systems, New Radio technology, multiconnectivity, guard capacity, session continuity, blockage mitigation.

Copyright (c) 2015 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

V. Begishev, E. Sopin, A. Samuylov, and K. Samouylov are with Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation. Email: {begishev-vo, sopin-es, samuylov-ak, samuylov-ke}@rudn.ru. K. Samouylov and E. Sopin are also with Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences (FRC CSC RAS), 44-2 Vavilov St, Moscow, 119333, Russian Federation. D. Moltchanov, R. Kovalchukov, S. Andreev, and Y. Koucheryavy are with Tampere University, Finland. Email: {firstname.lastname}@tuni.fi. Y. Koucheryavy is also with Higher School of Economics, Moscow, Russian Federation.

This work was supported by Business Finland (Project 5G-FORCE) and by the Academy of Finland (Projects RADIANT and IDEA-MILL). This paper has been supported by the RUDN University Strategic Academic Leadership Program (recipient K. Samouylov, supervision, project administration). The reported study was funded by RFBR, project number 19-07-00933 (recipient E. Sopin, mathematical model development) and project number 20-07-01064 (recipient V. Begishev, visualization).

I. INTRODUCTION

Millimeter-wave (mmWave) radio access is expected to become an essential part of the emerging 5G systems by providing low latencies and high data rates to the users at the last mile [1]. While the standardization of the corresponding 5G New Radio (NR) technology is to conclude soon within 3GPP, researchers continue to explore the challenges related to system design [2].

The need for reliable connectivity over mmWave-based 5G NR systems brings new challenges to system designers. The notorious problem of blockage of the propagation paths by various dynamic objects in the channel, including vehicles and human bodies, results in a rapid degradation of the received signal quality or even causes outage of the ongoing sessions. Recalling the quality of user experience metrics, it is often advantageous to drop a session at the time of its arrival rather than discontinue the service after accepting a session [3]. Acknowledging this problem, 3GPP has recently proposed multiconnectivity mechanisms. Accordingly, a user equipment (UE) is allowed to support more than a single link to the nearby NR base stations (BSs) by dynamically switching between them in the case of blockage [4].

Performance of multiconnectivity operation for 5G NR has been deeply investigated recently. Using computer simulations, the authors in [5] showed that both UE access rate and outage probability can be drastically improved with multiconnectivity operation. The capacity gains were further reported in [6]. The authors in [7] used stochastic geometry to derive a closed-form upper bound on the achievable UE rate. The gains of multiconnectivity for the UE access rate and outage probability were quantified in [8], thus showing that four links allow to achieve up to 76% capacity gains and up to 95% outage gains.

The model accounting for non-zero beamforming gains was formulated and solved in [9], where the authors demonstrated that in some conditions supporting more than two links may have a detrimental effect on the UE access rate. The authors in [10] assessed the performance of a system with various dynamic switching strategies between NR BSs, while the support of mission-critical traffic in dense street deployments of NR BSs was considered in [11].

The above studies concluded that multiconnectivity may improve session continuity by ensuring that the sessions already accepted for service will complete successfully. However, in practice, the benefits of multiconnectivity might be limited. First, the number of simultaneously supported links will

TABLE I
NOTATION OF THIS PAPER.

Parameter	Definition
f_c	Carrier frequency
r_A	Effective coverage radius
λ_B	User density
Λ	Session arrival intensity from a single UE
λ	Aggregate session arrival intensity
μ	Session service rate
h_A	NR BS height
h_U	UE height
h_B	Blocker height
r_B	Blocker radius
v	Blocker velocity
τ	Blocker run length during mobility
ζ	Path loss exponent
N_0	Thermal noise
$L_{dB}(x)$	Path loss, decibels
L_B	Blockage losses
$\beta_{i,j}(n_1, n_2, r)$	Conditional probabilities of released resources
θ_{3dB}^\pm, ν	Parameters of linear antenna array
P_T	NR BS transmit power
C_L	Cable losses
M_I	Interference margin
S_F	Shadow fading
$M_{S,nB}, M_{S,B}$	Shadow fading margins in non-blocked and blocked states
$\sigma_{S_F,B}, \sigma_{S_F,nB}$	STD of fading in LoS blocked and non-blocked states
p_C	Cell-edge coverage probability
T	Session data rate
$R_{k,0}, R_{k,1}$	Available radio resources for new and rerouted sessions
$\pi_N, \pi_{N,k}$	New session drop probability
$\pi_T, \pi_{T,k}$	Probability that rerouting leads to session drop
π_O	Ongoing session drop probability
U, U_k	Resource utilization coefficients
K_B, K_U	Number of planar antenna elements at NR BS and UE
ω_B, ω_U	Antenna directivities at NR BS and UE
G_B, G_U	NR BS transmit and UE receive antenna gains
$F_X(x), f_X(x)$	CDF and pdf of random variable X
K	Number of NR BSs
N_k	Maximum number of sessions at NR BS k
$\lambda_k = \lambda/K$	Intensity of new session arrivals to NR BS k
φ_k	Intensity of rerouted session arrivals to NR BS k
$p_{0,r}$	Probability that a new session requires r PRBs
$p_{1,r}$	Probability that an ongoing session requires r PRBs
$p_{s,r}^{(n)}$	Probability that n sessions of type s require r PRBs
γ	Guard capacity fraction
$u_i(t)$	Amount of resources occupied at NR BS at time t
Ψ	State space of our queuing model
Ψ_{n_1, n_2}	State associated with n_1 new and n_2 rerouted sessions
$\xi_1(t)$	Number of primary sessions in the system at time t
$\xi_2(t)$	Number of rerouted sessions in the system at time t
$\eta(t)$	Total amount of radio resources occupied at time t
$q_{n_1, n_2}(r)$	Stationary probabilities of our queuing model
α_k	Intensity of UE state changes with NR BS k
S_i, S	SNR with NR BS i and the overall SNR
$S_{B,i}, S_{nB,i}$	SNR at NR BS i in blocked and non-blocked states
D_i	Link distance projection to i -th nearest NR BS
$\epsilon_i(x)$	Intensity of blockers crossing the LoS blockage zone
$\eta_j(x, y)$	Probability that a blocker crosses the LoS blockage zone
s_j	SNR margins
m_j	Probability of choosing MCS j
$\text{erfc}(\cdot)$	Complementary error function

direction mobility (RDM) [14]. The speed is assumed to be constant v , while the run length is distributed exponentially with the mean value of τ . We represent the pedestrians by cylinders with constant height and radius, h_B and r_B , respectively. Each pedestrian is associated with a single UE. The height of the UE is also constant and is given by h_U . The

considered RDM model enables simpler expressions for the key metrics of interest while at the same time capturing the stochastic nature of human movement [15].

The rationale behind selecting a symmetric circular topology is manifold. First, since the framework developed in this paper allows to capture not only joint performance of the guard capacity and the multiconnectivity mechanisms but also their effects in isolation, the presented topology permits to perform an unbiased comparison between the impact of guard capacity, multiconnectivity, and their combined operation. Second, the considered topology may reflect practical deployments of NR BSs in squares seeing densely populated conditions, where mmWave-based NR technology is expected to be utilized. Finally, as we demonstrate in what follows, the use of circular deployments similar to the one shown in Fig. 1 enables tractable analysis of various parameters. The latter allows to formulate the proposed framework while omitting the over-complicating details whenever possible.

B. Propagation and Antenna Models

We consider the downlink direction from the NR BS to the UE. The signal-to-noise ratio (SNR) at the UE is written as

$$P_R(y) = \frac{P_T G_B G_U L(y) S_F C_L}{N_0 + I}, \quad (1)$$

where P_T is the emitted power, G_U and G_B are the array gains at the UE and the NR BS, y is the distance between the NR BS and the UE, $L(y)$ and S_F are the path loss and the shadow fading, respectively, C_L is the cable losses, N_0 and I are the thermal noise and interference at the UE.

The LoS path between the NR BSs and the UE can be occluded by moving pedestrians. The UE reacts to these changes by selecting a suitable MCS at the air interface as specified in TR 38.211 [13]. To represent the path loss, we utilize the urban micro (UMi) model as outlined by 3GPP [16], which explicitly accounts for blockage, i.e.,

$$L_{dB}(y) = \begin{cases} 32.4 + 21 \log(y) + 20 \log f_c, & \text{non-blocked,} \\ 52.4 + 21 \log(y) + 20 \log f_c, & \text{blocked,} \end{cases} \quad (2)$$

where f_c is the frequency in GHz, y is the distance.

At both NR BS and UE sides, we assume linear antenna arrays. We represent the antenna radiation patterns by using cone models [17], [18]. Accordingly, the radiation pattern is modeled via a cone with the angle ω centered at the transmitter or the receiver. The angle ω characterizes the half-power beamwidth (HPBW) of the antenna array. Following [19], the mean gain of a linear antenna array is

$$G = \frac{1}{\theta_{3dB}^+ - \theta_{3dB}^-} \int_{\theta_{3dB}^-}^{\theta_{3dB}^+} \frac{\sin(K\pi \cos(\theta)/2)}{\sin(\pi \cos(\theta)/2)} d\theta, \quad (3)$$

where K is the number of antenna elements, θ_{3dB}^+ and θ_{3dB}^- are the 3-dB points given by

$$\theta_{3dB}^\pm = \arccos[-\nu \pm 2.782/(K\pi)], \quad (4)$$

where ν is the array orientation measured in radians.

C. Traffic, Association, and Service

Let Λ be the arrival intensity of sessions from a single UE. We assume that each of the UEs may initiate a new session at an arbitrary instant of time. Recalling the superposition property [20], one may observe that the session arrival intensity from all of the UEs follows a Poisson process with the parameter $\lambda = \Lambda \lambda_B \pi r_A^2$. The proposed framework may assume more complex arrival patterns, including a Markovian arrival process (MAP) with correlated phase-type distributed inter-arrival times [21]. However, as we observe below, this would lead to a significantly enlarged state-space of the model, which prohibits the practical application of this framework.

The induced error associated with Poisson arrivals (typically, within 5-7% [22]) can be accounted for at the system design phase. We further assume that the session service time is distributed exponentially with the mean μ^{-1} , while the session data rate is kept constant, T Mbps. The amount of radio resources required to maintain the rate of T depends on the current MCS, which is described by the probability mass function (pmf) $p_{0,r}$, $r = 1, 2, \dots$ that is obtained in Section IV.

The amount of bandwidth at each NR BS is $R_{k,1}$. The degree of multiconnectivity, i.e., the number of links to different NR BSs that can be supported by one UE, is not limited. At each NR BS, a fraction of bandwidth, $R_{k,1}(1 - \gamma)$, is accessible for new sessions, where $\gamma \in (0, 1)$ is referred to as the *guard capacity*. On the other hand, the total set of resources $R_{k,1}$ is accessible by the sessions that are already accepted into the system. This guard capacity mechanism ensures that the sessions associated with the UEs changing their states from non-blocked to blocked are prioritized over the new sessions.

The resultant system with guard capacity, multiconnectivity, and dynamic NR BS selection capabilities operates as follows. Let $u_i(t)$, $i = 1, 2, \dots, K$, denote the radio resources currently occupied at the NR BSs. Once a new session arrives, the respective UE associates with the first NR BS discovered by employing a beamsearch procedure. If the amount of the available resources at this NR BS, $[R_{k,1}(1 - \gamma) - u_i(t)]$, is higher than the amount of the requested resources, the session is accepted for service. Otherwise, the session is dropped.

Whenever an UE – during its active session – changes its state from blocked to non-blocked or vice versa, the amount of radio resources that needs to be provided for supporting the target data rate is recalculated. In the former case, the target data rate is guaranteed to be supported and the current NR BS continues to serve the session. However, when the state changes from non-blocked to blocked, the NR BS checks whether there remain sufficient radio resources, i.e., $[R_{k,1} - u_i(t)]$, to maintain the current session. If that holds, the service of the session in question continues at the current NR BS. Otherwise, the UE searches for another NR BS by using a beamsearch procedure. If there are insufficient resources at this new NR BS, the ongoing session is considered dropped.

D. Metrics of Interest

We address both user- and system-side metrics. The first one is the new session drop probability, which is interpreted as the probability that a newly arrived session is dropped due

to insufficient radio resources available. Further, we consider the ongoing session drop probability, which captures the probability that an active session is dropped during service as a result of insufficient radio resources available at those time instants when the associated UE changes its state from non-blocked to blocked. The system-centric parameter is the mean resource utilization at the NR BS, which represents the amount of radio resources occupied at the NR BS, averaged across all of the BSs, i.e.,

$$E[U] = \frac{1}{K} \sum_{i=1}^K \frac{1}{TR_{k,i}} \int_0^T tu_i(t) dt. \quad (5)$$

E. Summary of Methodology

There are several important challenges related to modeling the service process at the NR BSs with both guard capacity and multiconnectivity operation in a dynamic blockage environment. In system-level simulations of the considered deployment, a major bottleneck is in capturing the dynamic blockage phenomena due to blocker mobility, as discussed at length in [23]. Assessing the multiconnectivity operation alone with analytical methods results in complex considerations as demonstrated in [10], [11]. The model proposed by the authors in [12] to characterize the behavior of the guard capacity mechanism also requires a set of simplifying assumptions.

To tackle the challenge of complexity, we in what follows intentionally separate the core of our model from its parametrization. Particularly, we design an analytical framework by relying upon queuing theory as our main tool in Section III. The developed framework delivers the sought metrics of interest depending on the number of NR BSs, K , the available radio resources at each NR BS, $R_{k,1}$, the guard capacity fraction, γ , the resource request pmf, $p_{0,r}$, and the intensities of the UE state changes with the NR BS k , α_k . The latter two characterize the type of deployment and are

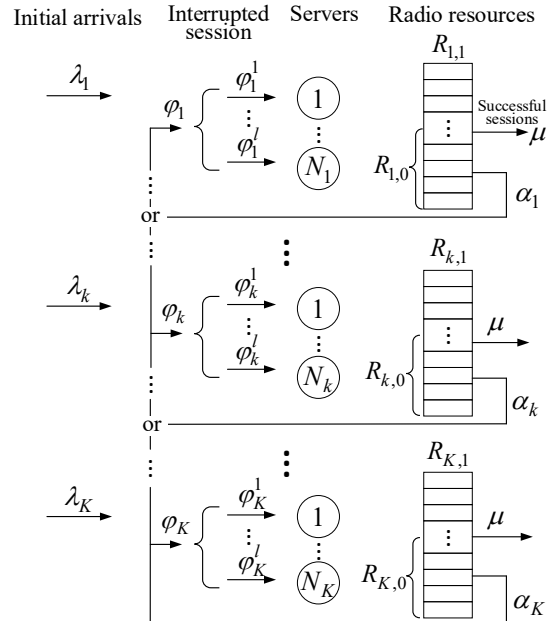


Fig. 2. Considered queuing model.

produced in Section IV. The modular structure of our proposed framework as per the below allows for a potential reuse of its core part for studying alternative 5G NR systems.

III. PERFORMANCE EVALUATION FRAMEWORK

In this section, we formulate our proposed framework. We first formalize the system model by using a queuing network with guard capacity, dynamic resource reallocation, routing, and random resource requirements. We proceed further by introducing a multidimensional Markov process characterizing the state evolution of the system at hand. The metrics of interest are then obtained. Finally, we present our numerical solution algorithm.

A. Model Formalization and Decomposition

Consider a queuing network that comprises K nodes, where k -th node is associated with N_k servers and $R_{k,1}$ resources. Here, N_k is the maximum number of sessions supported at the NR BS, while the resources are interpreted in terms of the physical resource blocks (PRBs) available at the NB BS. The arrival process of sessions at the NR BS k is homogeneous Poisson with the intensity of $\lambda_k = \lambda/K$, $k = 1, 2, \dots, K$, $\lambda = \sum_{k=1}^K \lambda_k$, where λ is the spatial session arrival intensity. Each arriving session of type k requires a discrete random amount of resources captured by the pmf $p_{0,r}$, $r > 0$. For the new sessions, only $R_{k,0}$, $R_{k,0} = R_{k,1}(1 - \gamma)$ resources are available.

Sessions that remain in service at the NR BS k are associated with a homogeneous Poisson process of transitions from the LoS non-blocked to the LoS blocked state and vice versa, having the parameter α_k , $k = 1, 2, \dots, K$. A session experiencing a state change generates a new amount of resources according to the pmf $p_{1,r}$, $r > 0$. If the amount of resources is sufficient to satisfy these new requirements, the current NR BS continues to serve the session at hand. Otherwise, the session leaves the NR BS k , releases the previously occupied resources, generates a new requirement according to the pmf $p_{1,r}$, $r > 0$, and attempts to reserve the resources at the NR BS k with probability $1/K$. Note that in the special case of $K = 1$, the session is dropped.

Such sessions are named rerouted sessions, where v denotes the number of reroutes – the so-called “level” of a session. These rerouted sessions form a secondary flow of sessions to each node with the intensity of ϕ_k , $k = 1, 2, \dots, K$. The rerouted sessions are allowed to compete for the full volume of the radio resources at the NR BS k . If the amount of free resources is insufficient at the server, or there are no free servers, then the rerouted session is lost. We also note that due to the memoryless property of the service time distribution, the residual service time after a state change from the LoS non-blocked to the LoS blocked has the same distribution as the initial service time. An illustration of the queuing model with the associated flows is offered in Fig. 2.

To tackle the outlined model, we use the network decomposition approach, which is a conventional methodology for complex queuing networks [24], [25]. The core assumption is that the service process at each node of a network is independent from the service processes at other nodes. Inter-dependencies

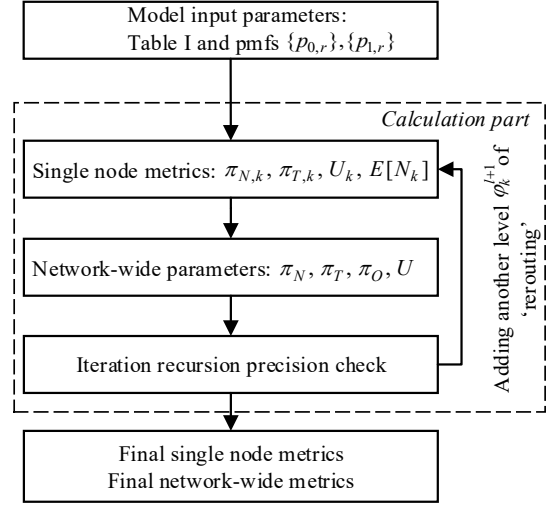


Fig. 3. Iterative nature of our solution algorithm.

between the NR BS service processes are incorporated into the numerical solution algorithm, where the characteristics of the entire system are recalculated recursively at each step until the procedure converges to a stable solution. The stability properties of this class of models were analyzed in [26].

Following the network decomposition approach, the evaluation procedure is given as follows, see Fig. 3. Initially, we start with zero intensity of the secondary flows, ϕ_k , $k = 1, 2, \dots, K$, to calculate the service characteristics at each NR BS and merge them into a product-form to characterize the overall system state. Then, we add a new level of rerouted sessions, recalculate the arrival flows to each NR BS, and repeat the procedure. Once an absolute difference in the stationary-state probabilities at steps i and $i - 1$ is less than a certain threshold value, the solution is considered to be found.

B. Operation of Single NR BS

Here, we consider the NR BS k in isolation. Due to the properties of the exponential distribution, the sojourn time of a session at this NR BS follows an exponential distribution with parameter $\mu + \alpha_k$, $i = 1, 2, \dots, K$. Hence, k -th NR BS serves two arrival processes: the Poisson arrival process of new sessions with intensity λ_k as well as the Poisson arrival process of rerouted sessions with the total intensity ϕ_k . To simplify the notation, in what follows, we omit the subscript k , e.g., $R_{k,1}$, and $R_{k,0}$ are written as R_1 and R_0 , respectively, whenever there is no ambiguity.

Observe that a complete description of the system requires an infinite-dimensional stochastic process. The reason is that one needs to track not only the types of sessions in the system but also the amounts of resources requested and received by each session. The latter is needed to appropriately release the resources upon a service completion or rerouting to another NR BS. To simplify the state description, we apply the well-known state aggregation technique (see, e.g., [27], [28]) and track the aggregate numbers of sessions from both arrival processes as well as the aggregate amount of occupied resources. Accordingly, the NR BS service process dynamics

$$\begin{aligned}
\beta_{0,j}(n_1, n_2, r) &= \left[\sum_{k=n_1}^{n_1+n_2} \frac{\binom{k-1}{n_1-1}}{\binom{n_1+n_2}{n_1}} \sum_{i=0}^{\min(r, R_0)} p_{1,r-i}^{(n_2+n_1-k)} \sum_{s=0}^i p_{0,s}^{(n_1)} p_{1,i-s}^{(k-n_1)} \right]^{-1} \left[\sum_{k=n_1}^{n_1+n_2} \frac{\binom{k-1}{n_1-1}}{\binom{n_1+n_2}{n_1}} \sum_{i=0}^{\min(r, R_0)} p_{1,r-i}^{(n_2+n_1-k)} \sum_{s=j}^i p_{0,s}^{(n_1-1)} p_{1,i-s}^{(k-n_1)} \right], \\
\beta_{1,j}(n_1, n_2, r) &= \left[\sum_{k=n_1}^{n_1+n_2} \frac{\binom{k-1}{n_1-1}}{\binom{n_1+n_2}{n_1}} \sum_{i=0}^{\min(r, R_0)} p_{1,r-i}^{(n_2+n_1-k)} \sum_{s=0}^i p_{0,s}^{(n_1)} p_{1,i-s}^{(k-n_1)} \right]^{-1} \times \left[\sum_{k=n_1}^{n_1+n_2} \frac{k-n_1}{n_2} \frac{\binom{k-1}{n_1-1}}{\binom{n_1+n_2}{n_1}} \sum_{i=0}^{\min(r, R_0)} p_{1,r-i}^{(n_2+n_1-k)} \right. \\
&\quad \left. \times \sum_{s=0}^{i-j} p_{1,j}^{(n_1)} p_{0,s}^{(n_1)} p_{1,i-s-j}^{(k-n_1-1)} + \sum_{k=n_1}^{n_1+n_2} \frac{n_1+n_2-k}{n_2} \frac{\binom{k-1}{n_1-1}}{\binom{n_1+n_2}{n_1}} \sum_{i=0}^{\min(r-j, R_0)} p_{1,r-i-j}^{(n_2+n_1-k-1)} p_{1,j} \sum_{s=0}^i p_{0,s}^{(n_1)} p_{1,i-s}^{(k-n_1)} \right]. \quad (10)
\end{aligned}$$

is represented via a three-dimensional stochastic process

$$\{\xi_1(t), \xi_2(t), \eta(t), t > 0\}, \quad (6)$$

where $\xi_1(t)$ is the number of new sessions in the system at time t , $\xi_2(t)$ is the number of rerouted sessions at time t , and $\eta(t)$ is the total volume of occupied resources.

A state of the process in (6) is defined as

$$\begin{aligned}
\Psi &= \bigcup_{0 \leq n_1+n_2 \leq N} \Psi_{n_1, n_2}, \\
\Psi_{n_1, n_2} &= \left\{ (n_1, n_2, r) : 0 \leq r \leq R_1, \sum_{i=0}^{\min(r, R_0)} p_{0,i}^{(n_1)} p_{1,r-i}^{(n_2)} > 0 \right\}, \quad (7)
\end{aligned}$$

where pmfs $\{p_{s,r}^{(n)}\}$, $r = 0, 1, \dots$, $s = \{0, 1\}$ are n -fold convolutions of pmfs $\{p_{s,r}\}$, $r = 0, 1, \dots$, and can be interpreted as the probability that n sessions of type s occupy r resources, where $s = 0$ denotes a new session, $s = 1$ refers to a rerouted session. Probabilities $p_{s,r}^{(n)}$ can be evaluated iteratively, i.e.,

$$p_{s,r}^{(n)} = \sum_{j=0}^r p_{s,j} p_{s,r-j}^{(n-1)}, s = \{0, 1\}. \quad (8)$$

The stationary probabilities $q_{n_1, n_2}(r)$ for all $(n_1, n_2, r) \in \Psi$ of the process in (6) are defined as

$$q_{n_1, n_2}(r) = \lim_{t \rightarrow \infty} P\{\xi_1(t) = n_1, \xi_2(t) = n_2, \eta(t) = r\}. \quad (9)$$

Consider now the amount of resources released by a session upon its departure. Let $\beta_{0,j}(n_1, n_2, r)$ denote the probability that, as a result of the departure of a new session, j resources are released, provided that the system resides in state (n_1, n_2, r) . For $r \leq R_0$, it can readily be obtained by applying the Bayes' law as follows

$$\beta_{0,j}(n_1, n_2, r) = \frac{p_{0,j} \sum_{i=0}^{r-j} p_{0,i}^{(n_1-1)} p_{1,r-j-i}^{(n_2)}}{\sum_{i=0}^r p_{0,i}^{(n_1)} p_{1,r-i}^{(n_2)}}. \quad (11)$$

Similarly, we determine the probability that, as a result of the departure of a rerouted session, j resources are released, provided that the system resides in state (n_1, n_2, r) , $\beta_{1,j}(n_1, n_2, r)$, i.e.,

$$\beta_{1,j}(n_1, n_2, r) = \frac{p_{1,j} \sum_{i=0}^{r-j} p_{0,i}^{(n_1)} p_{1,r-j-i}^{(n_2-1)}}{\sum_{i=0}^r p_{0,i}^{(n_1)} p_{1,r-i}^{(n_2)}}. \quad (12)$$

Estimation of $\beta_{0,j}(n_1, n_2, r)$ and $\beta_{1,j}(n_1, n_2, r)$ for $r > R_0$ is a more involved procedure. Observe that the probabilities $\beta_{0,j}(n_1, n_2, r)$ and $\beta_{1,j}(n_1, n_2, r)$ depend on the arrival order of new and rerouted sessions, which is unknown from the state description in (6). However, since the arrivals of new and rerouted sessions are independent of each other, and given the numbers of sessions of each type, any permutation of new and rerouted sessions is equally probable. The probability that the last arrived new session occupies k -th position is

$$\binom{k-1}{n_1-1} / \binom{n_1+n_2}{n_1}, \quad (13)$$

where $\binom{k-1}{n_1-1}$ is the number of ways to assign $n_1 - 1$ new sessions to the first $k - 1$ positions (the last session of this type occupies k -th position), while $\binom{n_1+n_2}{n_1}$ is the total number of ways to arrange n_1 new sessions. The probability that n_1 new sessions and $k - n_1$ rerouted sessions occupy $i \leq R_0$ resources is given by the following convolution

$$\sum_{s=0}^i p_{0,s}^{(n_1)} p_{1,i-s}^{(k-n_1)}, \quad (14)$$

while the probability that all of n_1 new and n_2 rerouted sessions occupy r resources, given that the last new session is at the position k , is also obtained via a convolution

$$\sum_{i=0}^{\min(r, R_0)} p_{1,r-i}^{(n_2+n_1-k)} \sum_{s=0}^i p_{0,s}^{(n_1)} p_{1,i-s}^{(k-n_1)}. \quad (15)$$

Finally, the probability that n_1 new and n_2 rerouted sessions fully occupy r resources is calculated as

$$\sum_{k=n_1}^{n_1+n_2} \frac{\binom{k-1}{n_1-1}}{\binom{n_1+n_2}{n_1}} \sum_{i=0}^{\min(r, R_0)} p_{1,r-i}^{(n_2+n_1-k)} \sum_{s=0}^i p_{0,s}^{(n_1)} p_{1,i-s}^{(k-n_1)}. \quad (16)$$

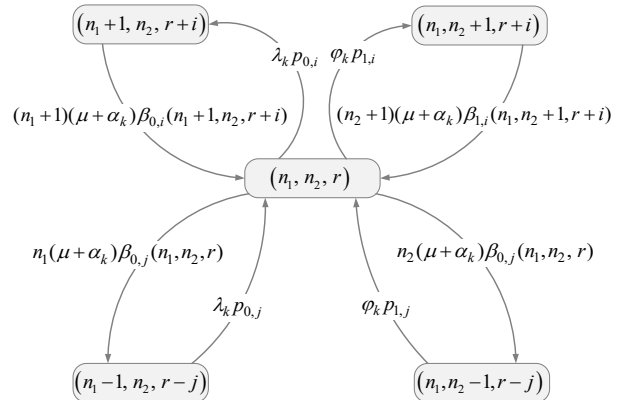


Fig. 4. Illustration of system state and associated transitions.

Applying the conditional probability rule, one can derive the probabilities $\beta_{0,j}(n_1, n_2, r)$, $j \leq r \leq R_0$, and $\beta_{1,j}(n_1, n_2, r)$, $j \leq r \leq R_1$, as in (10). Observe that the numerator of $\beta_{1,j}(n_1, n_2, r)$ in (10) comprises two terms. The first one corresponds to the case where a departing rerouted session arrived before the last new session, and the second term reflects the case where a departing session arrived after all of the new sessions.

After characterizing the probabilities in (12) and (10), we proceed by specifying a set of equilibrium equations for the transition probabilities $q_{n_1, n_2}(r)$ that the system is in state (n_1, n_2, r) and q_0 that the system is in state $(0, 0, 0)$, in the form of (17). These equations are obtained as follows:

- i) At the left-hand side (LHS), $\lambda_k \sum_{j=0}^{R_0} p_{0,j} + \phi_k \sum_{j=0}^{R_1} p_{1,j}$ is the exit intensity from state $(0, 0, 0)$. At the right-hand side (RHS), $(\mu + \alpha_k)$ is the intensity of transitions from all the states $(1, 0, j)$ and $(0, 1, j)$ to the state $(0, 0, 0)$, which represents the servicing and rerouting intensities, respectively.
- ii) This balance equation is valid under the conditions $n_1 + n_2 < N$, $r \leq R_0$. At the LHS, the term $\lambda_k \sum_{j=0}^{R_0-r} p_{0,j} + \phi_k \sum_{j=0}^{R_1-r} p_{1,j} + (n_1 + n_2)(\mu + \alpha_k)$ is the intensity of exits from the state (n_1, n_2, r) . Further, at the RHS, the first term $\lambda_k \sum_{j:(n_1-1, n_2, r-j) \in \Psi_{n_1-1, n_2}} p_{0,j} q_{n_1-1, n_2}(r-j)$ is the intensity of transitions from the state $(n_1-1, n_2, r-j)$ to the state (n_1, n_2, r) induced by a new session arrival at the NR BS k . The second term at the RHS is the transition intensities to the state (n_1, n_2, r) associated with the rerouted session arrivals at the NR BS k . The following two terms capture the transitions to the state (n_1, n_2, r) corresponding to the transition from the LoS non-blocked to the LoS blocked state. The probability $\beta_{0,j}(n_1+1, n_2, r+j)$ accounts for releasing j resources by a new session, which is leaving the system provided that the latter resides in the state $(n_1+1, n_2, r+j)$. Similarly, the probability $\beta_{1,j}(n_1, n_2+1, r+j)$ corresponds to the release of j resources upon a departure of a rerouted session given that the system state is $(n_1, n_2+1, r+j)$.
- iii) This equation is satisfied under $n_1 + n_2 < N$ and $r > R_0$. The logic is similar to the case of the previous equation.
- iv) This equation corresponds to the case $n_1 + n_2 = N$, $r \leq R_0$. At the LHS, the first term $N(n_1 + n_2)(\mu + \alpha_k)$ is the intensity of transitions from the state (n_1, n_2, r) to the state $n_1-1, n_2, r-j$ or $(n_1, n_2-1, r-j)$ induced by the service completions or by the state changes from the LoS non-blocked to the LoS blocked, respectively. At the RHS, the term $\lambda_k \sum_{j:(n_1-1, n_2, r-j) \in \Psi_{n_1-1, n_2}} p_{0,j} q_{n_1-1, n_2}(r-j)$ is the intensity of transitions from the state $(n_1-1, n_2, r-j)$ to the state (n_1, n_2, r) associated with an arrival to the NR BS k . The second term at the RHS is the transition intensities to the state (n_1, n_2, r) associated with the rerouted session arrivals.
- v) This case is similar to that of the previous equation.

The system state transition diagram is illustrated in Fig. 4. We note that, as a result of random session resource requirements, arrival and departure events lead to a transition not to a single state but to a subset of states. The destination state depends on the pmf of the session resource request.

Hence, Fig. 4 does not reflect all the states from a subset but only one typical representative state from each subset.

C. Solution and Metrics of Interest

The system of equilibrium equations specified in (17) is solved numerically. Since the number of equations in the system can be as high as $N(N+1)R_1/2$, the infinitesimal generator is stored using a sparse matrix scheme [29]. To solve this system, we employ iterative Gauss-Seidel approach [30]. The algorithm converges for diagonally dominant matrices, which always holds in our case as the absolute value of diagonal elements is at all times equal to the sum of the absolute values of non-diagonal elements.

The solution is iterative as it adds another level of rerouted sessions at every iteration. The procedure is terminated once the required accuracy is reached. At the first iteration, there are no rerouted sessions, and thus $\phi_k = 0$, $k = 0, 1, \dots, K$. Then, the algorithm proceeds as follows:

- i) using λ_k , μ , α_k , ϕ_k , and pmfs $\{p_{0,j}\}$, $\{p_{1,j}\}$, the performance metrics for a single NR BS are evaluated: the new session drop probability at the NR BS k $\pi_{N,k}$, the probability that a rerouting from the NR BS k leads to a session drop $\pi_{T,k}$, the resource utilization coefficient U_k , and the mean number of sessions at the NR BS k , $E[N_k]$;
- ii) network-wide parameters are evaluated: the new session drop probability π_N , the probability that a rerouting from an arbitrary NR BS leads to a session drop π_T , as well as the individual and the overall resource utilization coefficients, U_k and U ;
- iii) if the desired precision is achieved, the algorithm stops; otherwise, the arrival intensity of sessions at a higher reroute level ϕ_k^{k+1} is evaluated and another iteration starts.

After deriving the stationary state probabilities at the NR BS, we proceed with evaluating the new session drop probability at the NR BS k , $\pi_{N,k}$, and the probability that a rerouting from the NR BS k leads to a session drop, $\pi_{T,k}$, as

$$\begin{aligned} \pi_{N,k} &= 1 - \sum_{0 \leq n_1 + n_2 \leq N-1} \sum_{r \leq R_{0,k}: (n_1, n_2, r) \in \Psi_{n_1, n_2}} q_{n_1, n_2}(r) \sum_{j=0}^{R_{0,k}-r} p_{0,j}, \\ \pi_{T,k} &= 1 - \sum_{0 \leq n_1 + n_2 \leq N-1} \sum_{r: (n_1, n_2, r) \in \Psi_{n_1, n_2}} q_{n_1, n_2}(r) \sum_{j=0}^{R_{1,k}-r} p_{1,j}. \end{aligned} \quad (18)$$

The total intensity of the rerouted sessions is given by

$$\phi_k = \sum_{v=1}^{\infty} \phi_k^v, \quad (19)$$

$$\phi_k^1 = \sum_{i=1}^K \lambda_i (1 - \pi_{N,i}) \frac{\alpha_i}{\mu + \alpha_i} \phi_{i,k}^0, \quad (20)$$

$$\phi_k^v = \sum_{i=1}^K \phi_i^{v-1} (1 - \pi_{T,i}) \frac{\alpha_i}{\mu + \alpha_i} \phi_{i,k}^{v-1}, v > 1, \quad (21)$$

where v is the number of rerouting events.

$$\begin{aligned}
\text{(i)} \quad q_0 & \left[\lambda_k \sum_{j=0}^{R_0} p_{0,j} + \Phi_k \sum_{j=0}^{R_1} p_{1,j} \right] = (\mu + \alpha_k) \left[\sum_{j:(1,0,j) \in \Psi_{1,0}} q_{1,0}(j) + \sum_{j:(0,1,j) \in \Psi_{0,1}} q_{0,1}(j) \right], \\
\text{(ii)} \quad q_{n_1, n_2}(r) & \left[\lambda_k \sum_{j=0}^{R_0-r} p_{0,j} + \Phi_k \sum_{j=0}^{R_1-r} p_{1,j} + (n_1 + n_2)(\mu + \alpha_k) \right] = \lambda_k \sum_{j:(n_1-1, n_2, r-j) \in \Psi_{n_1-1, n_2}} p_{0,j} q_{n_1-1, n_2}(r-j) + \Phi_k \sum_{j:(n_1, n_2-1, r-j) \in \Psi_{n_1, n_2-1}} p_{1,j} q_{n_1, n_2-1}(r-j) \\
& + (n_1 + 1)(\mu + \alpha_k) \sum_{j:(n_1+1, n_2, r+j) \in \Psi_{n_1+1, n_2}} q_{n_1+1, n_2}(r+j) \beta_{0,j}(n_1 + 1, n_2, r+j) + \\
& + (n_2 + 1)(\mu + \alpha_k) \sum_{j:(n_1, n_2+1, r+j) \in \Psi_{n_1, n_2+1}} q_{n_1, n_2+1}(r+j) \beta_{1,j}(n_1, n_2 + 1, r+j), \quad 0 < n_1 + n_2 < N, r \leq R_0, \\
\text{(iii)} \quad q_{n_1, n_2}(r) & \left[\Phi_k \sum_{j=0}^{R_1-r} p_{1,j} + (n_1 + n_2)(\mu + \alpha_k) \right] = \Phi_k \sum_{j:(n_1, n_2-1, r-j) \in \Psi_{n_1, n_2-1}} p_{1,j} q_{n_1, n_2-1}(r-j) + (n_1 + 1)(\mu + \alpha_k) \sum_{j:(n_1+1, n_2, r+j) \in \Psi_{n_1+1, n_2}} q_{n_1+1, n_2}(r+j) \times \\
& \times \beta_{0,j}(n_1 + 1, n_2, r+j) + (n_2 + 1)(\mu + \alpha_k) \sum_{j:(n_1, n_2+1, r+j) \in \Psi_{n_1, n_2+1}} q_{n_1, n_2+1}(r+j) \beta_{1,j}(n_1, n_2 + 1, r+j), \quad 0 < n_1 + n_2 < N, r > R_0, \\
\text{(iv)} \quad N(\mu + \alpha_k) q_{n_1, n_2}(r) & = \lambda_k \sum_{j:(n_1-1, n_2, r-j) \in \Psi_{n_1-1, n_2}} p_{0,j} q_{n_1-1, n_2}(r-j) + \Phi_k \sum_{j:(n_1, n_2-1, r-j) \in \Psi_{n_1, n_2-1}} p_{1,j} q_{n_1, n_2-1}(r-j), \quad n_1 + n_2 = N, r \leq R_0, \\
\text{(v)} \quad N(\mu + \alpha_k) q_{n_1, n_2}(r) & = \Phi_k \sum_{j:(n_1, n_2-1, r-j) \in \Psi_{n_1, n_2-1}} p_{1,j} q_{n_1, n_2-1}(r-j), \quad n_1 + n_2 = N, r > R_0.
\end{aligned} \tag{17}$$

Further, the new session drop probability and the probability that a rerouting from an arbitrary NR BS leads to a session drop, π_N and π_T , respectively, are calculated as

$$\pi_N = \sum_{k=1}^K \frac{\lambda_k}{\lambda} \pi_{N,k}, \quad \pi_T = \sum_{k=1}^K \frac{\Phi_k}{\varphi} \pi_{T,k}, \tag{22}$$

where the unknowns are provided by

$$\lambda = \sum_{k=1}^K \lambda_k, \quad \varphi = \sum_{k=1}^K \Phi_k. \tag{23}$$

Another important parameter of interest is the ongoing session drop probability π_O , i.e., the probability that the service process of an arbitrary accepted session is interrupted, and that session is dropped. This probability is obtained as

$$\pi_O = \lim_{t \rightarrow \infty} \frac{\varphi \pi_T t}{\lambda(1 - \pi_N)t} = \frac{\varphi \pi_T}{\lambda(1 - \pi_N)}, \tag{24}$$

where the numerator represents the number of ongoing sessions dropped during time interval t , while the denominator stands for the number of accepted sessions during time t .

Finally, the resource utilization coefficient is given by

$$U = \frac{\sum_{k=1}^K U_k R_{1,k}}{\sum_{k=1}^K R_{1,k}}, \tag{25}$$

where the individual utilization coefficients are

$$U_k = \frac{1}{R_{1,k}} \sum_{1 \leq n_1 + n_2 \leq N} \sum_{r \leq R_{1,k}: (n_1, n_2, r) \in \Psi_{n_1, n_2}} r q_{n_1, n_2}(r). \tag{26}$$

IV. PARAMETRIZATION OF DEVELOPED FRAMEWORK

The developed mathematical framework accepts two parameters at its input, which depend on the spatial locations of the UEs: (i) the pmf of the amount of radio resources needed for the session $p_{0,r}$, $r = 1, 2, \dots$, and (ii) the intensities of the states changes at the UE, α_i , $i = 1, 2, \dots, K$. We derive these parameters in this section.

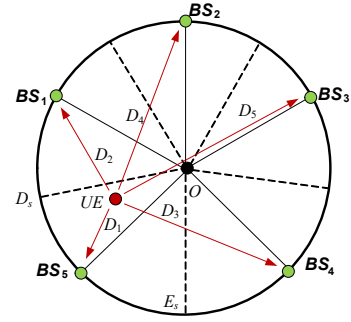


Fig. 5. Link projections to nearest NR BS for $K = 5$.

A. Characterizing Resource Requirements

We proceed with deriving the pmf of the radio resources requested by a session, $p_{0,r}$. Using the properties of the RDM model [14], we observe that the UE locations are distributed uniformly across the deployment area. Further, recall that upon a new session arrival, its UE is associated with the first NR BS found by a beamsearch procedure. This is equivalent to choosing a random NR BS located on the circumference. Similarly, in case of UE state change, a randomly chosen NR BS is attempted.

Hence, the pmf of the session resource requirements, $p_{0,r}$, can be obtained as a sum of the pmfs corresponding to the NR BSs weighed with the coefficient $1/K$. Further, notice that the pmf of the resource requirements with every NR BS is a weighed sum of the pmfs corresponding to the blocked and non-blocked conditions. In what follows, we determine the pmf of the radio resource requirements in the non-blocked state. The rest of the pmfs are obtained similarly.

To obtain the pmf of the radio resource requirements in the non-blocked state at the NR BS i , we first determine the link projections to these NR BSs from the UE being uniformly distributed over the deployment area. The procedure for characterizing the link projections is similar for all of the

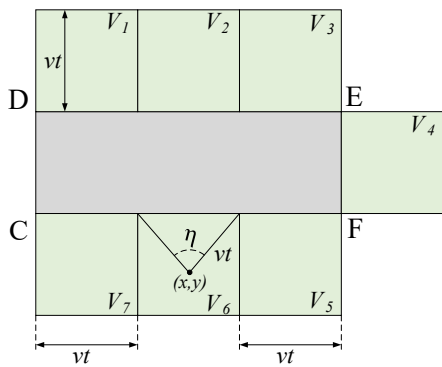


Fig. 7. Evaluation of temporal intensity of blockers.

B. Intensity of State Changes

The intensity of UE transitions from non-blocked to blocked state and back characterizes the temporal properties of the blockage process that was studied in detail by [23], [34], [35]. Therefore, below we briefly summarize the general procedure by referring to the previously published material as necessary.

To determine the sought parameter, the following steps are required: (i) characterize the nature of the temporal pedestrian arrival process into the UE blockage zone by using the UE mobility model, (ii) determine the parameters of the resultant process, (iii) using the UE blockage zone geometry, obtain the distribution of time that a pedestrian spends in the UE blockage zone, (iv) use the arrival process of pedestrians into the UE blockage zone and the UE blockage zone sojourn time distribution to determine the mean times when the LoS is blocked and non-blocked, and (v) obtain the intensity of UE state changes.

In [23], the authors demonstrated that for the RDM model the process of pedestrians entering the LoS blockage zone is approximately Poisson in nature. The authors also derived an integral expression for characterizing the temporal intensity of this process for the UE located at the distance of x , i.e.,

$$\varepsilon(x) = \frac{\lambda_B e^{-1/\tau}}{2\pi} \sum_{j=1}^7 \iint_{V_j} \eta_j(x, y) dx dy, \quad (38)$$

where $\eta_j(x, y)$ are provided by

$$\begin{aligned} \eta_1(x, y) &= ([x_D - x]/vt), j = 1, 3, 5, 7, \\ \eta_2(x, y) &= 2 \cos^{-1}([x_E - x]/vt), j = 2, 6, \\ \eta_4(x, y) &= 2 \tan^{-1}([x - x_E]/[y - y_E]), \end{aligned} \quad (39)$$

where x_D, x_E, y_E are given in Fig. 7.

The distribution of time that a pedestrian spends in the LoS blockage zone was also reported in [23]. Similarly to [34], one may observe that the LoS blockage and non-blockage periods coincide with the busy and empty periods in the M/G/1 queuing system, where the arrival process is Poisson with the intensity of $\varepsilon(x)$, while the service time distribution is given by the distribution of time when a pedestrian occludes the LoS propagation path. These metrics of interest are provided in, e.g., [36]. Once the mean blockage and non-blockage periods are available, the intensity in question is inversely proportional to their sum. Finally, the intensity of the UE state changes with the NR BS k , α_k , is obtained by averaging.

TABLE II
NUMERICAL EVALUATION PARAMETERS.

Parameter	Value
Number of NR BSs, K	$\{1, 2, \dots, 5\}$
Carrier frequency, f_c	28 GHz
Radio resources at each NR BS, $R_{k,1}$	400 Hz
Transmit power, P_T	2 W
Planar antenna elements at NR BS, K_B	32
Planar antenna elements at UE, K_U	4
Effective coverage radius, r_A	100 m
LoS blockage loss, L_B	20 dB
NR BS height, h_A	4 m
UE height, h_U	1.5 m
Blocker height, h_B	1.7 m
Blocker radius, r_B	0.3 m
Pedestrian density, λ_B	0.03; 0.15; 2 users/m ²
Blocker velocity, v	1 m/s
Session data rate, T	3; 6 Mbps
Mean session service time, $1/\mu$	20 s
Session arrival intensity from a single UE, Λ	$3.14E - 4$ sess./s
Guard capacity range, γ	$\{0.0, 0.01, \dots, 0.1\}$
Cable losses, C_L	2 dB
Interference margin, M_I	3 dB
Cell-edge coverage probability, p_C	0.1
STD of shadow fading, $\sigma_{Sf,B}$	3 dB

V. NUMERICAL PERFORMANCE ASSESSMENT

In this section, we evaluate the operation of a joint guard capacity and multiconnectivity implementation in a dynamic blockage environment. Particularly, we start by reporting their individual performance and then proceed with evaluating a combined performance. The parameters used in what follows are summarized in Table II.

An appropriate comparison methodology in the presence of multiconnectivity requires an additional clarification. To offer insights into the operating regimes of the NR BSs, for the considered session data rates, $R = \{3, 5\}$ Mbps, we first identify the session arrival intensity from a single user, Λ . This results in both new and ongoing session drop probabilities being bounded within 0.1% and 5% – by having a single NR BS in the area of interest. Then, to provide a fair comparison with an increasing number of NR BSs, we upscale the session arrival intensity as $K\Lambda$ and assess the resultant gains of the multiconnectivity operation.

For the selected system parameters provided in Table II, the session arrival intensity from a single user is $\Lambda = 3.14E - 4$ sessions per second. We also note that the coverage radius is set to be 100 m. The rationale behind utilizing this value in our numerical campaign is that at this distance blockage events do not lead to outage. To apply the developed methodology to other deployment cases, one needs to parametrize the framework with appropriate distance by ensuring that it does not yield outage.

A. Accuracy of Developed Model

We begin with evaluating the accuracy of the developed mathematical framework by providing a comparison with the simulation data. To perform this, we develop a simulation environment that accounts for the radio-specific properties as

well as for the dynamic blockage, UE associations, and resource allocation with guard bandwidth and multiconnectivity mechanisms as specified in Section II, mindful of both system- and user-centric metrics. In our simulation framework, we relax the two major modeling assumptions: (i) state aggregation technique employed by the queuing model and (ii) weighing of SNR to compute the session resource requirements.

The simulation engine follows the widely accepted discrete-event simulation (DES) framework [37] with additional enhancements for parallel execution and data analysis. The simulation data are collected during the stationary period only, which is detected via exponentially weighed statistics according to [38]. To eliminate the effect of residual correlation and ensure that the conventional statistical methods can be applied, we sample the state of the system every second, and further apply the method of replications [39]. Due to low simulation system complexity, we are able to obtain a sufficient number of independent and identically distributed (i.i.d.) observations, such that the interval estimates are smaller than $\pm 0.05x$, where x is the point estimate of a parameter. Hence, only the point estimates are shown.

Fig. 8 reports on the user-centric performance metrics as estimated with the proposed mathematical framework and the simulation approach for the session rate $T = 3$ Mbps, $K = 3$ NR BSs, and $\gamma = 0.1$. The analytical data approximate the simulation results fairly accurately. The observed deviations are attributed to the modeling assumptions. Equivalent results were observed for other input parameters including the resource utilization coefficient. This implies that the assumptions adopted as part of our mathematical framework do not produce substantial distortions to the investigated parameters. Below, we therefore use our mathematical framework to assess the key performance indicators.

B. Performance of Individual Mechanisms

The guard capacity and the multiconnectivity mechanisms can be utilized in isolation. That is, in the case of guard capacity only regime, whenever a blockage occurs, a session may attempt to reserve a new set of resources at the same BS, as it now has access to the entire pool of resources. However, if these resources are insufficient, the session is dropped. In the

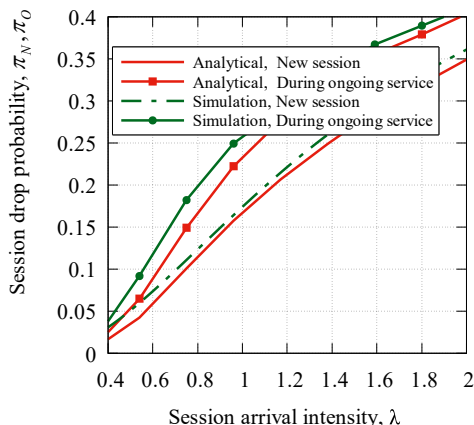
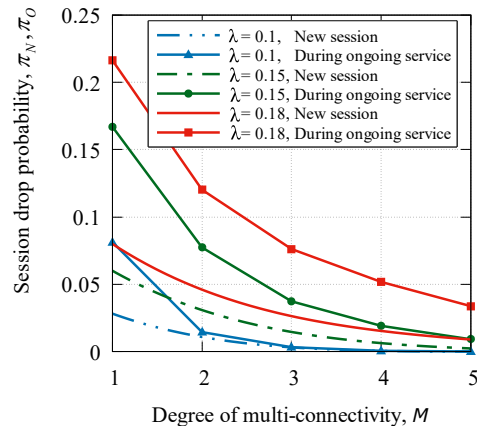
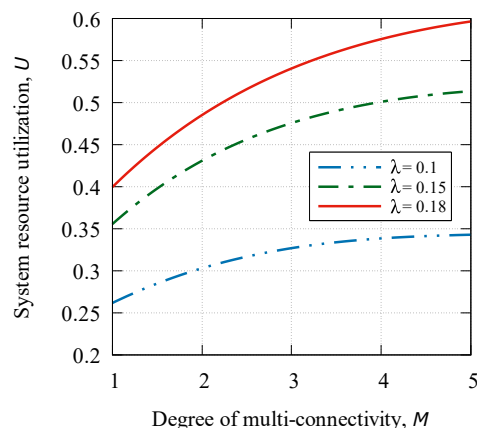


Fig. 8. Comparing simulation and analytical results.



(a) Session drop probabilities



(b) NR BS resource utilization

Fig. 9. User- and system-centric metrics in presence of multiconnectivity.

case of multiconnectivity, no resource reservation is made, and new as well as ongoing sessions have access to the full set of resources. Should blockage happen, a session is allowed to be rerouted, which may positively affect the session continuity. To represent the multiconnectivity only feature, the value of guard capacity has to be set to zero. To model the guard capacity mechanism in isolation, one needs to limit the scenario to a single cell, thus reducing the rerouting options to the same cell. These regimes are the special cases of the proposed performance evaluation framework. Hence, before we proceed with quantifying the joint effects of the two mechanisms, we evaluate the cases where they are used in isolation.

We first assess the performance of our system in the presence of multiconnectivity mechanism only. To this aim, Fig. 9(a) and Fig. 9(b) study the user- and system-centric parameters for various numbers of NR BSs available within the considered service area, K , for the session data rate of $T = 3$ Mbps and $\gamma = 0$. Analyzing the presented data, one may notice that an increased degree of multiconnectivity results in higher resource utilization. Particularly, the latter approaches 0.6 for $K = 5$ and $\lambda = 0.18$ sessions per second per NR BS. These system-centric gains are also associated with a decrease in both ongoing and new session drop probabilities.

The rationale behind the above behavior is that the higher

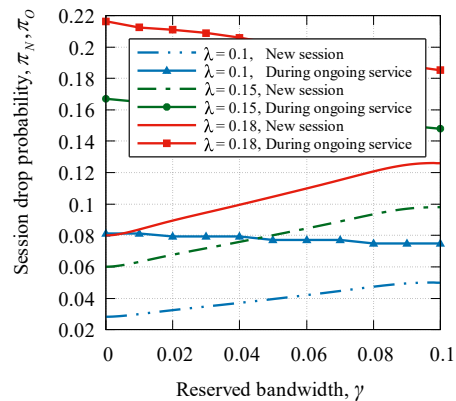
values of K increase the diversity of connectivity options available to the UEs when they experience state changes from the LoS non-blocked to the LoS blocked. As a result, the load becomes more evenly distributed among the NR BSs, thus decreasing the ongoing session drop probability. However, depending on the input system parameters (particularly, on the session arrival rate), the ongoing session drop probability might still be much higher than the new session drop probability. To decrease it, one needs to significantly reduce the offered traffic load into the system (via e.g., explicit connection admission control mechanisms), which will also decrease the system resource utilization. Despite significant gains in all three considered metrics of interest, multiconnectivity alone does not provide the means to balance the new session drop probability and the drop probability of accepted sessions.

We now proceed with understanding the performance of the guard capacity mechanism separately. Fig. 10 illustrates the ongoing and the new session drop probabilities together with the system resource utilization coefficient for $K = 1$ and $T = 3$ Mbps. As one may observe, the use of guard capacity does not drastically deteriorate the utilization of the system resources for the considered range of γ and the three assumed session arrival intensities. The new session drop probability clearly increases as γ grows, thus reducing the fraction of radio resources for the new sessions. Here, constant utilization of system resources is explained by dropping fewer sessions that are already accepted by the system, which also prohibits new sessions from being accepted for service.

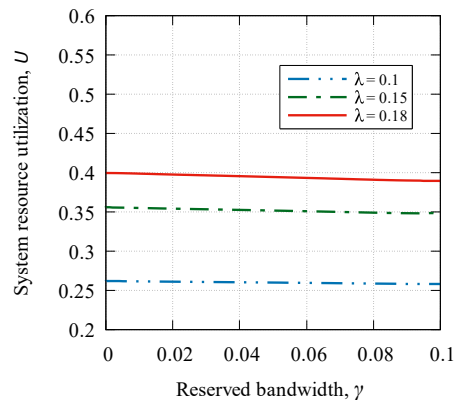
Summarizing the behavior of the two mechanisms in isolation, we note that guard capacity may complement multiconnectivity by demonstrating further positive effects for the session continuity in 5G NR systems. In their joint implementation, multiconnectivity is expected to compensate for the reduced probability of accepting new sessions and further increase the utilization of radio resources at the NR BSs. From the practical standpoint, such joint operation may improve the degrees of freedom in controlling the 5G NR service processes, thus allowing the operators to achieve the desired trade-off between the considered user-centric performance metrics, while at the same time ensuring that the usage of system resources remains adequate.

C. Joint Use of Two Mechanisms

We continue with analyzing a joint implementation of the two mechanisms by assessing the system response to various degrees of multiconnectivity. Fig. 11 shows the considered user- and system-centric parameters as a function of K for different values of guard capacity γ , $\lambda = 0.15$, and the session data rate of $T = 3$ Mbps. Assessing the collected data, one may notice that both the ongoing and the new session drop probabilities decrease as the degree of multiconnectivity grows for all the values of γ . It is also evident that guard capacity positively affects the ongoing session drop probability, while its effect on the new session drop probability is, not surprisingly, negative. As one may establish, this improvement – in absolute values – is smaller than the corresponding degradation in terms of the fraction of new sessions accepted for service, as seen in Fig. 11(a).



(a) Session drop probabilities



(b) NR BS resource utilization

Fig. 10. User- and system-centric metrics in presence of guard capacity.

Understanding the system resource utilization displayed in Fig. 11(c), one may deduce that the positive effect of the multiconnectivity mechanism on this parameter also translates into having the system with a joint implementation of the mechanisms under evaluation. More importantly, the values of guard capacity do not affect this parameter significantly. The rationale behind this behavior is in that more sessions accepted by the NR BSs remain in the system up to their service completion and utilize the system resources. From the practical point of view, the system spends less resources by serving sessions that will never be completed. Summarizing, a joint implementation of the considered mechanisms does offer an additional degree of freedom for the operators in controlling the performance of their deployment. Furthermore, while the value of γ remains within $(0, 0.1)$, the potential trade-off does not negatively affect the resource utilization at the NR BSs.

We now study the response of our system to other input parameters, including the requested session data rate, the blocker density in the environment, and the session arrival intensity. We first focus on the impact of the data rate of sessions as shown in Fig. 12(a) for $\gamma = 0.05$ and $T = 3$ Mbps. Interpretation of the presented results requires a careful treatment. Recall that the traffic load into the system is provided by $\rho = \lambda TK / \mu$, where λ is the spatial session arrival intensity, K is the number of NR BSs, and μ is the session service intensity. To isolate the effect of the mean session data

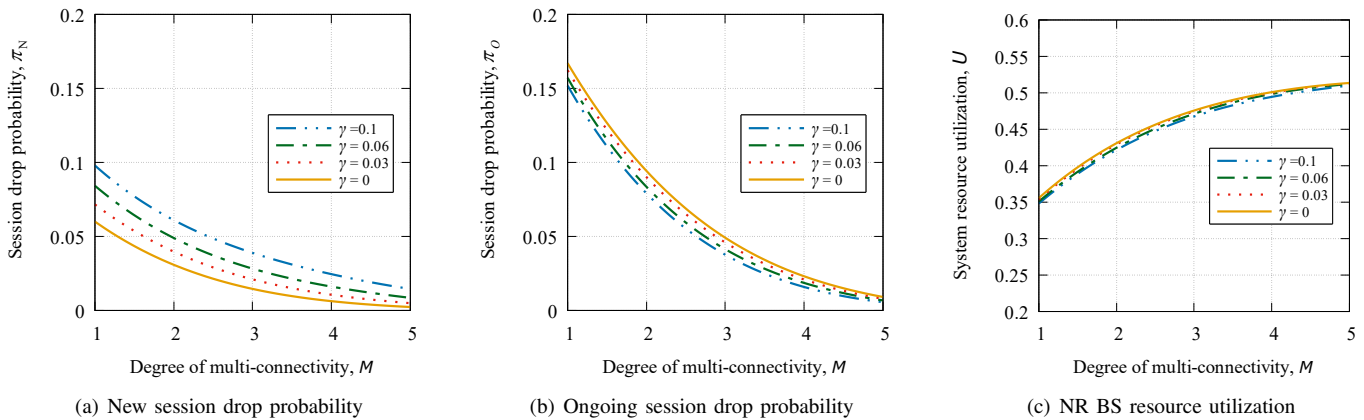


Fig. 11. User- and system-centric metrics in presence of guard capacity and multicommunity.

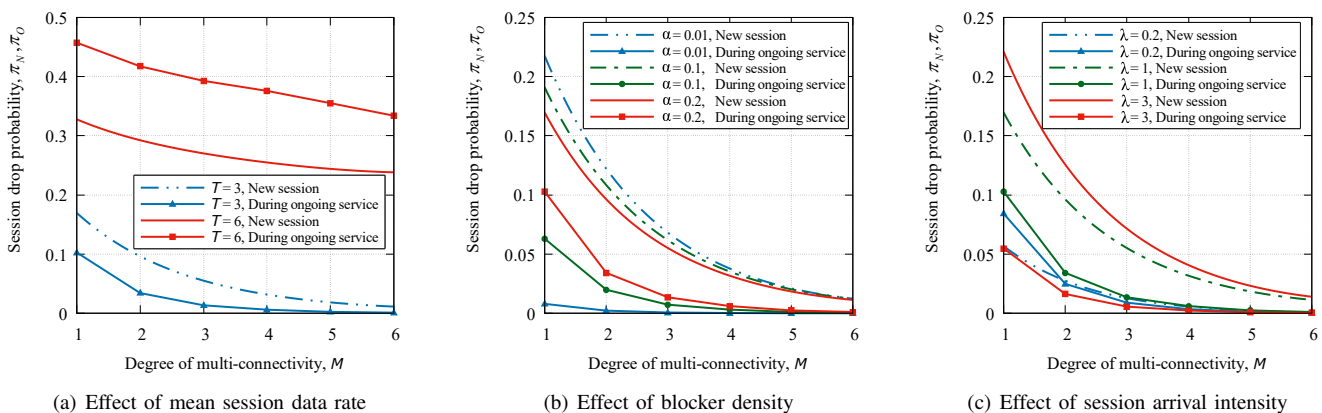


Fig. 12. Impact of system parameters on user- and system-centric performance metrics.

rate, we decrease λ as T grows, such that ρ is kept constant.

As one may notice while analyzing the user-centric performance metrics in Fig. 12(a), the data rate of the sessions affects both drop probabilities not only quantitatively but also qualitatively. Particularly, for smaller mean data rates, the new session drop probability is always higher than the probability of dropping a session during its ongoing service for all the considered values of the degree of multicommunity. However, when the mean data rate increases from $T = 3$ Mbps to $T = 6$ Mbps, the ongoing session drop probability rises above the new session drop probability. The reason is that an increase in T leads to higher mean resource requirements, where reserving only 5% of the radio resources is not sufficient to lower the ongoing session drop probability. Therefore, one may conclude that the mean session data rate affects the choice of the values of K and γ that satisfy the desired session drop probabilities.

The blocker density, λ_B , is another important variable that may drastically affect the choice of K and γ . Indeed, for a given UE speed, v , it is translated into the intensity of changes between the LoS blocked and the LoS non-blocked states, α_k , $k = 1, 2, \dots, K$, as described in Section IV. Since the mean number of UE state changes is $\sum_{k=1}^K \alpha_k / K\mu$, where μ is the service intensity, an increase in the blocker density should lead to a higher probability of dropping the ongoing sessions and a lower probability of rejecting the new sessions.

Accordingly, Fig. 12(b) displays the ongoing and the new

session drop probabilities over a range of UE state change intensities α , $T = 3$ Mbps, $\gamma = 0.05$, $\lambda = 0.15$, and the default values of $\nu = 1$ and $\mu = 1/20$. As one may learn, the above hypothesis is confirmed. It is important to note that for any given K and an extremely small crowd density of $\lambda_B = 0.01$, the ongoing session drop probability remains low. Once the density increases, the difference between the two probabilities shrinks, and eventually the probability of dropping a session already accepted by the system may exceed the one for the new sessions in dense crowds, thus rendering the system unusable. One may return the system back to its operating regime by increasing the value of guard capacity.

Finally, Fig. 12(c) illustrates the effect of the session arrival intensity on the session drop probabilities for $\gamma = 0.05$ and $T = 3$ Mbps. Here, a growth in λ yields higher offered traffic load ρ , and thus increases the probability of dropping a new session. Similarly, another considered user-centric parameter also grows as λ increases from 0.2 to 3. However, for a much higher session arrival intensity of 3, it decreases drastically. The explanation here is that, for such values of λ and $\gamma = 0.05$, the system operates at its capacity limit: it is busy with many sessions whose mean resource requirements are lower than the mean requested resource requirements. Hence, when the UE state changes, the probability that its session will remain in the system grows.

VI. CONCLUSIONS

Session continuity is a crucial user-centric performance indicator that may be severely deteriorated by dynamic link blockage in dense deployments of mmWave-based NR systems. To mitigate these negative effects, the guard capacity and the multiconnectivity mechanisms were recently proposed. In this study, we developed an accurate mathematical framework that is capable of simultaneously capturing the radio propagation and the session service processes at the NR BSs when implementing these mechanisms jointly. Using the developed framework, we investigated the system- and the user-centric performance metrics in dense deployments of 5G NR systems.

The findings of the performed numerical evaluation suggest that an integration of the considered mechanisms does improve the degrees of freedom in selecting the operating point of the 5G NR systems. More specifically, the use of guard capacity allows to further decrease the drop probability of sessions already accepted for service, thus improving the session continuity as compared to the multiconnectivity only case. However, the gains in the ongoing session drop probability are smaller than the associated degradation in the new session drop probability. At the same time, when implemented jointly with multiconnectivity, guard capacity does not produce any notable negative effects on the system resource utilization as compared to its use in isolation.

As a result, one may recommend a combined utilization of the considered mechanisms for mobile operators that prioritize session continuity of their subscribers, and remain unwilling to compromise the system performance. The said joint implementation of the two schemes does not require additional signaling at the 5G NR air interface, or between the NR BSs and their served UEs. However, assuming such extra signaling, one may develop even more comprehensive inter-BS switching strategies (e.g., based on the current loading of the NR BSs), which may further improve the user- and system-centric performance indicators.

REFERENCES

- [1] "M.2376: Technical feasibility of IMT in bands above 6 GHz," ITU-R technical report, 2015.
- [2] Y. Niu, Y. Li, D. Jin, L. Su, and A. Vasilakos V, "A survey of millimeter wave communications (mmWave) for 5G: opportunities and challenges," *Wireless Networks*, vol. 21, pp. 2657–2676, November 2015.
- [3] N. Seitz, "ITU-T QoS standards for IP-based networks," *IEEE Communications Magazine*, vol. 41, no. 6, pp. 82–89, 2003.
- [4] 3GPP, "NR; Multi-connectivity; Overall description; Stage-2 (Release 15)," 3GPP TS 37.340, December 2017.
- [5] D. S. Michalopoulos, I. Viering, and L. Du, "User-plane multi-connectivity aspects in 5G," in *2016 23rd International Conference on Telecommunications (ICT)*, pp. 1–5, May 2016.
- [6] F. B. Tesema, A. Awada, I. Viering, M. Simsek, and G. P. Fettweis, "Mobility modeling and performance evaluation of multi-connectivity in 5G intra-frequency networks," in *IEEE Globecom Workshops (GC Wkshps)*, pp. 1–6, December 2015.
- [7] D. Moltchanov, A. Ometov, S. Andreev, and Y. Koucheryavy, "Upper bound on capacity of 5G mmwave cellular with multi-connectivity capabilities," *Electronics Letters*, vol. 54, no. 11, pp. 724–726, 2018.
- [8] M. Gapeyenko, V. Petrov, D. Moltchanov, M. R. Akdeniz, S. Andreev, N. Himayat, and Y. Koucheryavy, "On the degree of multi-connectivity in 5g millimeter-wave cellular urban deployments," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 2, pp. 1973–1978, 2019.
- [9] M. Gerasimenko, D. Moltchanov, M. Gapeyenko, S. Andreev, and Y. Koucheryavy, "Capacity of multi-connectivity mmWave systems with dynamic blockage and directional antennas," *IEEE Transactions on Vehicular Technology*, 2019.

- [10] V. Petrov, D. Solomitckii, A. Samuylov, M. A. Lema, M. Gapeyenko, D. Moltchanov, S. Andreev, V. Naumov, K. Samouylov, M. Dohler, *et al.*, "Dynamic multi-connectivity performance in ultra-dense urban mmWave deployments," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 9, pp. 2038–2055, 2017.
- [11] V. Petrov, M. A. Lema, M. Gapeyenko, K. Antonakoglou, D. Moltchanov, F. Sardis, A. Samuylov, S. Andreev, Y. Koucheryavy, and M. Dohler, "Achieving end-to-end reliability of mission-critical traffic in softwarized 5g networks," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 3, pp. 485–501, 2018.
- [12] D. Moltchanov, A. Samuylov, V. Petrov, M. Gapeyenko, N. Himayat, S. Andreev, and Y. Koucheryavy, "Improving session continuity with bandwidth reservation in mmWave communications," *IEEE Wireless Communications Letters*, vol. 8, no. 1, pp. 105–108, 2018.
- [13] 3GPP, "NR; Physical channels and modulation (Release 15)," 3GPP TR 38.211, Dec 2017.
- [14] P. Nain, D. Towsley, B. Liu, and Z. Liu, "Properties of random direction models," in *IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 3, pp. 1897–1907, March 2005.
- [15] A. Orsino, D. Moltchanov, M. Gapeyenko, A. Samuylov, S. Andreev, L. Militano, G. Araniti, and Y. Koucheryavy, "Direct connection on the move: Characterization of user mobility in cellular-assisted d2d systems," *IEEE Vehicular Technology Magazine*, vol. 11, no. 3, pp. 38–48, 2016.
- [16] 3GPP, "Study on channel model for frequencies from 0.5 to 100 GHz (Release 14)," 3GPP TR 38.901 V14.1.1, July 2017.
- [17] V. Petrov, M. Komarov, D. Moltchanov, J. M. Jornet, and Y. Koucheryavy, "Interference and SINR in millimeter wave and terahertz communication systems with blocking and directional antennas," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1791–1808, 2017.
- [18] S. Singh, R. Mudumbai, and U. Madhow, "Interference analysis for highly directional 60-ghz mesh networks: The case for rethinking medium access control," *IEEE/ACM Transactions on Networking (TON)*, vol. 19, no. 5, pp. 1513–1527, 2011.
- [19] A. B. Constantine *et al.*, "Antenna theory: analysis and design," *Microstrip Antennas*, John Wiley & Sons, 2005.
- [20] J. F. C. Kingman, *Poisson processes*. Wiley Online Library, 1993.
- [21] D. M. Lucantoni, "New results on the single server queue with a batch markovian arrival process," *Communications in Statistics. Stochastic Models*, vol. 7, no. 1, pp. 1–46, 1991.
- [22] M. Pióro and D. Medhi, *Routing, flow, and capacity design in communication and computer networks*. Elsevier, 2004.
- [23] M. Gapeyenko, A. Samuylov, M. Gerasimenko, D. Moltchanov, S. Singh, M. R. Akdeniz, E. Aryafar, N. Himayat, S. Andreev, and Y. Koucheryavy, "On the temporal effects of mobile blockers in urban millimeter-wave cellular scenarios," *IEEE Transactions on Vehicular Technology*, available online, 2017.
- [24] P. Kuehn, "Approximate analysis of general queuing networks by decomposition," *IEEE Transactions on Communications*, vol. 27, no. 1, pp. 113–126, 1979.
- [25] G. R. Bitran and D. Tirupati, "Multiproduct queueing networks with deterministic routing: Decomposition approach and the notion of interference," *Management Science*, vol. 34, no. 1, pp. 75–100, 1988.
- [26] E. Sopin, K. Ageev, and K. Samouylov, "Approximate analysis of the limited resources queuing system with signals," in *Proceedings of 33rd European Conference for Modeling and Simulation (ECMS)*, ECMS, 2019.
- [27] A. Bobbio and K. S. Trivedi, "An aggregation technique for the transient analysis of stiff markov chains," *IEEE Transactions on computers*, no. 9, pp. 803–814, 1986.
- [28] G. Ciardo and E. Smirni, "Etaqa: an efficient technique for the analysis of qbd-processes by aggregation," *Performance Evaluation*, vol. 36, pp. 71–93, 1999.
- [29] D. Langr and P. Tvrdik, "Evaluation criteria for sparse matrix storage formats," *IEEE Transactions on parallel and distributed systems*, vol. 27, no. 2, pp. 428–440, 2016.
- [30] J. Xu, "Iterative methods by space decomposition and subspace correction," *SIAM Review*, vol. 34, pp. 581–613, 1992.
- [31] R. Kovalchukov, D. Moltchanov, A. Samuylov, A. Ometov, S. Andreev, Y. Koucheryavy, and K. Samouylov, "Evaluating SIR in 3D millimeter-wave deployments: Direct modeling and feasible approximations," *IEEE Transactions on Wireless Communications*, vol. 18, no. 2, pp. 879–896, 2019.

- [32] R. Kovalchukov, D. Moltchanov, A. Samuylov, A. Ometov, S. Andreev, Y. Koucheryavy, and K. Samouylov, "Analyzing effects of directionality and random heights in drone-based mmWave communication," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 10, pp. 10064–10069, 2018.
- [33] S. Ross, *Introduction to probability models*. Academic Press, 2010.
- [34] M. Gapeyenko, A. Samuylov, M. Gerasimenko, D. Moltchanov, S. Singh, M. R. Akdeniz, E. Aryafar, S. Andreev, N. Himayat, and Y. Koucheryavy, "Spatially-consistent human body blockage modeling: a state generation procedure," *IEEE Transactions on Mobile Computing*, vol. 9, no. 17, p. 20, 2019.
- [35] M. Gapeyenko, V. Petrov, D. Moltchanov, S. Andreev, N. Himayat, and Y. Koucheryavy, "Flexible and reliable UAV-assisted backhaul operation in 5G mmWave cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 11, pp. 2486–2496, 2018.
- [36] D. J. Daley, "The busy period of the $M/GI/\infty$ queue," *Queueing Systems*, vol. 38, no. 2, pp. 195–204, 2001.
- [37] B. P. Zeigler, T. G. Kim, and H. Praehofer, *Theory of modeling and simulation*. Academic press, 2000.
- [38] H. G. Perros, "Computer simulation techniques: The definitive introduction!," 2009.
- [39] G. S. Fishman and L. S. Yarbberly, "An implementation of the batch means method," *INFORMS Journal on Computing*, vol. 9, no. 3, pp. 296–310, 1997.