User Persistence in 5G/6G mmWave/sub-THz Systems with Blockage: Does it Pay Off?

E. Sopin, D. Moltchanov, A. Maslov, V. Begishev, A. Samuylov,
K. Samuyilov, Y. Koucheryavy

***Abstract*—5G New Radio (NR) millimeter wave (mmWave) and future sub-terahertz (sub-THz) 6G systems are designed for rate-greedy multimedia applications. Such applications often utilize rate adaptation allowing to vary the rate according to the network conditions. Unreliable nature of these bands subject to blockage may induce persistent user behavior resulting in repeated attempts (retrials) to continue the service. These effects may lead to the loss of sessions during the service and service provisioning at reduced rate. In this paper, by using the stochastic geometry and queuing theory, we characterize performance of adaptive rate-greedy applications at mmWave/sub-THz base stations. We consider user- and system-centric performance including new and ongoing session drop probabilities, resource utilization, and the amount of wasted resources. Our results show that the user persistence improves the new and ongoing session drop probabilities: making two retrials brings both down by $20-70\%$ while for fully persistent behavior the gain is 300%. The impact is more pronounced in underloaded system conditions. User persistence also increases the usage of system resources by $20-40\%$ and decreases the percentage of wasted resources by two/three times. However, these gains come at the expense of decreased successful session completion probability without retrials negatively affecting the overall user experience.**

***Index Terms*—5G/6G, mmWave, terahertz, performance, user persistence, retrials, queuing theory**

I. Introduction

The recently standardized millimeter wave band (28-100 GHz) version of 5G New Radio (NR) technology is developed to satisfy the needs to rate-greedy multimedia applications including augmented/virtual/extended reality (AR/VR/XR), holographic communications, etc [1]. In the future 6G cellular systems, it will be complemented with sub-terahertz (sub-THz) radio access technologies (RAT) operating in the lower part of the terahertz band, 100-300 GHz [2].

The loss of connectivity due to the human body blockage events is currently one of the main obstacles towards the use of mmWave 5G NR systems and this problem will persist in what follows with introduction of sub-THz 6G systems. As demonstrated recently [3], [4], system level solutions proposed in the past including the use of multi-connectivity option [5], network densification (conventional and based on Integrated Access and Backhaul Technology, [6]) does not allow to fully eliminate the impact of outages caused by blockage. The research towards physical-layer blockage avoidance solutions such as utilization of diffraction-free and self-reconstructing beams (e.g., Bessel beamforming) has just recently started [7] and the community does not expect it to commence producing operational antennas in the coming years. Thus, in the coming years, the use of mmWave/sub-THz systems will be subject to occasional session drops caused by blockages.

Rate-greedy multimedia applications often require constant bit rate (CBR) service. However, even these applications are equipped with rate adaptation capabilities at the application layer [8]. Specifically, the application may track the state of the connection and react to the reduced network throughput and/or the temporal loss of connectivity by reducing the required throughput. The most common example is video streaming, live or stored, that allows dynamically switching between bitrate levels depending on the current network conditions. When the rate variability is high and/or the connectivity is frequently lost, users perceive service at the degraded quality, e.g., in lower resolution or with a lower frame-per-second rate.

Outages caused by blockage naturally lead to persistent user behaviour in terms of repeated session re-establishment attempts (called retrials thereafter). It may result in degraded quality of user experience (QoE) as the session will be served at a reduced rate. Furthermore, the repeated attempts may increase the offered traffic load preventing new sessions from entering the system and also increase the wastage of resources. However, this feature might be advantageous to mmWave/sub-THz systems from the system-centric point of view (resource utilization) and user-centric point of view as well (eventual session drop probability), where the session might be accepted at the degraded quality at the second or third session establishment attempts. This realistic behaviour has not been considered so far for mmWave/sub-THz systems.

The aim of this paper is to characterize the effects of persistent user behavior in 5G/6G mmWave/sub-THz RATs with rate adaptation and dynamic blockage conditions. To this aim, we use the tools of stochastic geometry to capture channel specifics, renewal processes to characterize blockage dynamics and queuing theory for capturing service process at mmWave/sub-THz base stations (BS). Our metrics of interest include not only the user-centric performance measures such as new and ongoing session drop probabilities, but the efficiency of resource utilization and service delivery including the percentage of resources wasted for serving sessions that are not completed and probability that the service is delivered with reduced rate.

The main contributions of our study are:

- Mathematical framework to evaluate user- and system-

D. Moltchanov, A. Samuylov, and Y. Koucheryavy are with Tampere University, Finland. Email: {firstname.lastname}@tuni.fi.

E. Sopin, A. Maslov, V. Begishev, and K. Samuyilov are independent authors. Email: {eduard.sopin,alexmaslov16,vobegishev, ksamuylov}@gmail.com.

centric performance with persistent user behavior and application rate adaptation in 5G/6G mmWave/sub-THz systems, where the modeling errors are limited to 5-7%.
- Observations that the user persistence and rate adaptation: (i) increase the usage of resources by around 20-40% and allow to use them efficiently decreasing the percentage of wasted resources by two/three times, (ii) decrease the ongoing session drop probability but also decrease the successful session completion probability without retrials, (iii) the impact is more evident in underloaded conditions.
- A way to improve the successful session completion probability without retrials and decrease the amount of wasted resources is to increase the BS antenna gain – this allows to improve both metrics by 100% and more depending on the system parameters.

The rest of the paper is organized as follows. In Section II we review the related work. Next, in Section III we introduce our system model. The performance evaluation framework is developed in Section IV. Numerical results are presented in Section V. Conclusions are drawn in the last section.

II. Related work

Modern 5G NR mmWave and future sub-THz 6G systems are subject to principally similar impairments (albeit with different magnitudes) -– high propagation losses and path blockage by dynamic objects such as humans, vehicles, and buildings. The efficient use of mmWave and sub-THz bands requires large antenna arrays capable of creating directional steerable antenna radiation patterns with the beamwidth of just a few degrees or even less. This feature is vital for 5G/6G systems not only allowing to overcome severe path losses but ensuring almost interference-free environment [9], [10]. However, blockage still leads to unreliable service often resulting in outage conditions.

To assess performance of applications in mmWave/sub-THz systems dynamic blockage models have been developed [11], [12]. The stochastic structure of these models are generally similar and they differ mainly in the attenuation caused by blockage. For mmWave systems in outdoor conditions the human body blockage of 15-25 dB has been reported [13]–[15]. While for sub-THz systems, these measurements are still scarce, few studies reported up to 15 dB losses [16], [17]. Equipped with these models, researchers have characterized performance of elastic and non-adaptive applications in different deployment options. The former type of traffic is generated by applications utilizing TCP at the transport layer while the latter presumes CBR non-adaptive applications.

A number of studies assessed radio-related performance of sub-THz systems including signal-to-noise plus interference ratio (SINR), coverage, outage, and spectral efficiency. Specifically, the authors in [9], [10] assessed interference and SINR in systems with highly-directional antennas highlighting that sub-THz systems would potentially allow for extremely dense deployments without notable link quality degradation. The authors in [18] assessed coverage of such systems. These models, however, accounted for basic sub-THz channel specifics including atmospheric absorption, blockage, and directional nature of the antenna radiation patterns. Recently, more comprehensive propagation models specifying advanced factors such as fast and shadow fading started to appear, see, e.g. [19] for a comprehensive survey. Equipped with these models, the authors in [20] provided closed-form expression for outage probability showing that it is highly sensitive to the fading parameters. However, we note that these models neglect the applications and user-behavior specifics concentrating solely on radio part performance.

To study performance of elastic traffic in mmWave/sub-THz systems either simulation-based approach or stochastic geometry have been utilized [21], [22]. Several studies in the mmWave domain [21]–[25] indicated that the relationship between the size of the data-link buffer, the round-trip time, and the bandwidth-delay product may heavily affect the efficiency of resource utilization in links with highly varying capacity. Some studies joined the dynamic blockage models with well-known PFTK and Mathis TCP models and characterized the attained rate analytically, see, e.g. [26]. Specifically, they showed that the blockage does not prevent TCP from reaching its optimal rate at the mmWave interface. The rationale is that the blockage and non-blockage interval durations are large enough and thus TCP may efficiently adapt its rate. Finally, to optimize performance of TCP in mmWave/sub-THz systems, few TCP extensions have been proposed [23], [27], [28].

For non-adaptive CBR applications researchers mainly concentrated on improving reliability of session service process by utilizing various techniques including bandwidth reservation and 3GPP multiconnectivity functionality, including inter- and intra-RAT options. Specifically, it has been shown that temporal offloading of "heavy-weight" mmWave NR sessions onto microwave LTE and sub-6 GHz technologies does not increase service reliability but also leads to drastic performance degradation of single-RAT user equipment (UE) utilizing LTE/sub-6 GHz systems only [29]. Alternatively, utilization of intra-RAT multiconnectivity within mmWave- or sub-THz systems still leads to non-negligible active session drops while requiring extreme power consumption from handheld devices and is only feasible in very dense deployments of these systems [3], [30]. Finally, offloading 5G NR sessions onto IEEE 802.11ad/ay analyzed in [31], may lead to rate reduction due to the use of random access mechanism.

Being characterized by unreliable nature with blockage leading to the loss of connectivity, 5G/6G mmWave/sub-THz systems will naturally lead to persistent user behaviour with users attempting to re-establish interrupted sessions. For rate-greedy CBR sessions with application layer rate adaptation this behaviour may lead to a number of undesirable effects such as service provisioning at degraded rate and inefficient use and even wastage of system resources.

III. System model

In this paper, we define the system model as a combination of deployment, radio part, traffic, and service models. Then, we introduce the model for user persistence. Finally, we define the metrics of interest.

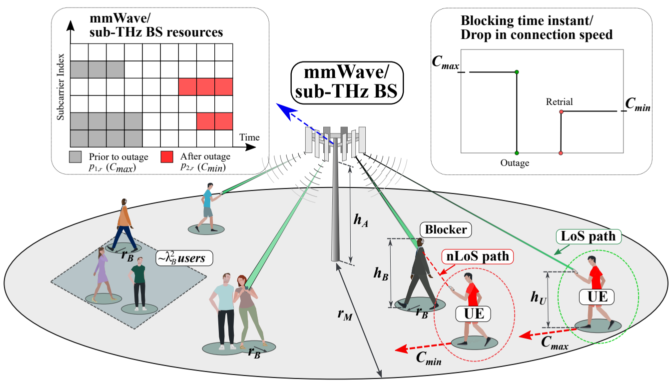


Fig. 1. The considered deployment scenario.

A. Deployment and User Mobility Models

We consider a single mmWave/sub-THz BS of circular form with the coverage radius of r_M as illustrated in Fig. 1, providing service to the users in the crowded environment, where there is high traffic demands, e.g., city square. The radius of the cell, r_M, depends on the utilized modulation and coding scheme (MCS) and is computed using the propagation model defined below. The available bandwidth is B MHz corresponding to R primary resource blocks (PRBs). For a 5G NR system, R can be estimated using B and a certain numerology specifying the size of PRB. The height of BS and UE are assumed to be h_A and h_U, respectively. The height of pedestrians is h_P and we also assume that $h_P > h_U$.

The service area is filled with pedestrians. The initial distribution of pedestrians is assumed to follow a homogeneous Poisson point process (PPP) in $\Re^2$ with density λ_B units/km^2. Pedestrians move according to a random direction model (RDM, [32]). That is, each pedestrian, independently of others, chooses the direction of movement uniformly in $(0, 2\pi)$ and moves in the selected direction for exponentially distributed time. The speed and the mean run duration are the two parameters of the model denoted by v_B m/s and an τ_B m, respectively. Due to the independence properties of the RDM model, the process of pedestrian locations is PPP at all the times as any random translation of PPP is PPP again [33]. The considered mmWave/sub-THz BS is assumed to serve UEs associated with pedestrians. The limiting distribution of the RDM mobility model is uniform over the cell area. Assuming that the flux of users over the cell boundary is balanced, the geometric location of a randomly chosen UE with an active session is uniformly distributed in the cell area.

In our paper, we concentrate on a single-cell environment as we study a phenomenon, user impatience, that is in most cases localized within a single cell. To this aim, we additionally assume that the session duration is smaller than the cell residence time. However, we note that if one wants to introduce a functionality that would require to consider multiple cells operating together when serving users, e.g., multi-connectivity functionality specified by 3GPP in [5], one may extend the proposed framework to the multi-cell scenario. This can be done by following the approach proposed in [3], where we investigated the impact of the multi-connectivity on the session drop probability. In this case, instead of utilizing a single resource-based queuing system as we do in what follows, one needs a network of resource-based queuing systems.

B. Radio Part Models

To reflect the wireless part in the system model, we account for several mmWave-specific radio part design including propagation, antenna, and blockage models.

1) Blockage Models: We assume that pedestrians may dynamically block the propagation path between mmWave/sub-THz BS and UE. Blockers are represented by cylinders with base radius r_B and height coinciding the the height of pedestrians, h_P. Using the results of [11], for RDM mobility model the time-averaged blockage probability is given by [34]

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

h_B is the height of BS, y is the 3D distance between BS and UE, λ_B units/m^2 is the density of pedestrians, r_B and h_P are the radius and height of pedestrians.

Let further $\Psi_B(x)$ and $\Phi_B(x)$ be the random blocked and non-blocked periods. As shown in [11], these intervals can be approximated by exponential distributions with means

$$E[\Psi(x)] = \frac{e^{-\sigma(x) v_B/2r_B} - 1}{\sigma(x)},\ E[\Phi_B(x)] = \frac{1}{\sigma(x)}, \quad (2)$$

where $\sigma(x)$ is the the temporal intensity of blockers entering the blockage zone associated with UE,

$$\sigma(x) = \frac{2}{5}\lambda_B v_B[4r_B + 2x(h_P - h_U)/(h_A - h_U)], \quad (3)$$

where x is 2D distance between UE and BS.

2) Propagation Model: We consider the 3GPP standardized Urban Micro (UMi) propagation model. According to it, the path loss in non-blocked state at the 3D distance y is

$$L_{dB}(y) = 32.4 + 10\zeta\log(y) + 20\log f_C, \quad (4)$$

where f_C is the carrier frequency in GHz, $\zeta = 2.1$ is the propagation exponent [35].

This model can be written in the linear form as

$$L(y) = 10^{2\log_{10} f_C + 3.24} y^{-2.1}. \quad (5)$$

3) Antenna Model: At both BS and UE we assume planar antenna arrays. By following to [9], [36], we use the cone antenna model, where the width of the beam coincides with the radiation pattern's half-power beamwidth (HPBW). The mean gain over HPBW is known to be [37]

$$G = \frac{1}{\theta^+_{3db} - \theta^-_{3db}} \int_{\theta^-_{3db}}^{\theta^+_{3db}} \frac{\sin(N_{(\cdot)}\pi\cos(\theta)/2)}{\sin(\pi\cos(\theta)/2)} d\theta, \quad (6)$$

where $N_{(\cdot)}$ is the number of antenna elements in the appropriate plane, θ^+_{3db} and θ^-_{3db} are 3-dB points. The HPBW can be approximated by $102^\circ/N_{(\cdot)}$ [37].

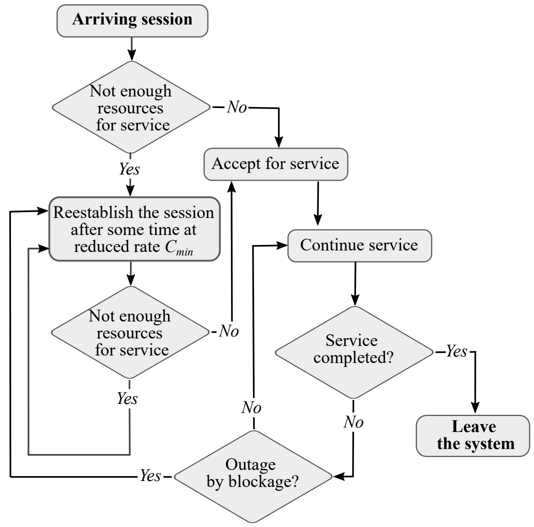


Fig. 2. The service procedure with persistent user behavior.

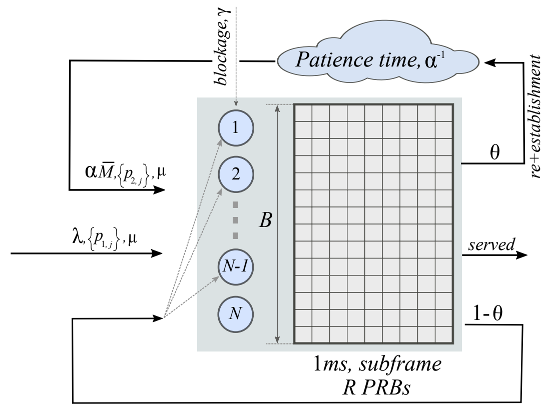


Fig. 3. The schematic representation of the model.

C. Traffic and Rate Adaptation Models

In the considered system sessions are generated by UEs associated with pedestrians. Let λ_U be the session arrival intensity from a single UE and assume that pedestrians generate sessions independently of each other. Then, by superposition property, the arrival process is Poisson in nature with the session arrival intensity given by

$$\lambda = \lambda_U \pi r_M^2. \tag{7}$$

We assume that a single UE at any given time instant may have no more than one active session. The rationale is that the applications we consider usually fully occupy attention from the users such that they do not run other applications in the foreground. Applications running in the background are usually not quality-of-service (QoS) sensitive and thus can be classified to a separate 5G "best effort" slice. That is why we assumed only one active session per user.

We assume that there are no priorities in service between users and arriving sessions are served on a first come first served (FCFS) basis. We consider a heavy bitrate service such as 4kHD video, AR/VR/XR, or holographic communications characterized by the non-elastic but rate adaptive traffic nature. That is, each arriving session is assumed to request constant bitrate $C_{\max}$. In case of service interruption caused by blockage the application layer rate adaptation mechanism reduces the rate to $C_{\min}$, $C_{\min} < C_{\max}$. Note that due to the random location of UEs in the service area the actual amount of resources associated with bitrates $C_{\min}$ and $C_{\max}$ is random. The session service time follows exponential distribution with mean μ^{-1}.

D. Session Service and User Persistence

We consider persistent user behavior with the overall service procedure sketched in Fig. 2. Specifically, the following joint admission and user behavior procedure is considered. Upon arrival, a new session requests a random amount of PRBs with probability mass function (pmf) $p_{1,j}$, $1 \leq j \leq R$, corresponding to the rate $C_{\max}$.[1] If these resources are available, the session is accepted to the system. If not, the user waits for exponentially distributed time with mean α^{-1} with probability θ reflecting the persistent user behavior and then re-initiates the session. This behavior is modeled by the queuing system with random resources requirements and orbit as sketched in Fig. 3. When user re-initiates the session, it returns from the orbit having reduced resource requirements with pmf $p_{2,j}$, $1 \leq j \leq R$ corresponding to the rate $C_{\min}$. Handling of this returning sessions is similar to the new sessions.

Any session that is currently in the system is associated with external Poisson process with intensity γ representing blockage events. The blockage intensity is computed by utilizing blocked and non-blocked intervals in (2). Upon these events the BS attempts to re-allocate resources required to continue service. If there are no sufficient amount of resources available, the user decides to wait for exponentially distributed time with mean α^{-1} with probability θ, i.e., the session leaves for orbit. The rest of the procedure is similar to the one described above.

In the considered system, a session can be successfully served if its service time completes when it is in the service. The session can also be successfully served at the initial rate $C_{\max}$ when it never experiences blockage event or lack of resources upon arrival. Alternatively, a session can be completely lost at the moment of arrival if the user decides to not re-initialize it with probability $1-\theta$. A session whose service has already started can be dropped during the service leading to the wastage of system resources.

E. Metrics of Interest

We are interested in implications of persistent user behavior in 5G/6G mmWave/sub-THz cellular systems with application layer rate adaptation. We consider both user- and system-centric metrics. The former metrics include the new and ongoing session drop probabilities. Of interest are also the probabilities of successful session completion with and without retrials. Note that the former metric implies that the

[1]Note that the randomness in the amount of requested resources for a fixed bitrate is caused by random locations of UEs in the cell coverage.

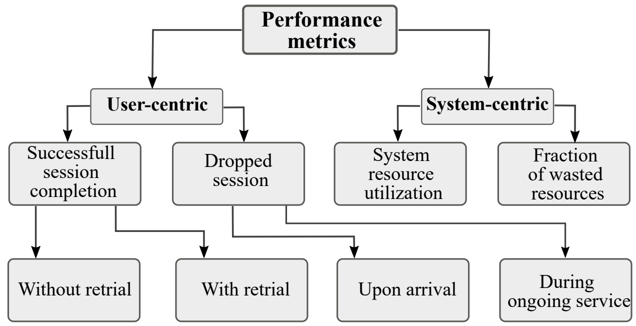


Fig. 4. The relation between the considered performance metrics.

user has to re-establish the session and thus experiences rate degradation. For the system-centric measures we consider the resource utilization and the fraction of wasted resources. The latter represents the share of resources utilized by the system to serve sessions whose service is not completed.

The relation between the considered metrics is shown in Fig.4. They allow to characterize the QoE and also track the reason why it degrades.

IV. Performance Evaluation Framework

In this section, we present the developed analytical framework in terms resource queuing system with losses, signals, and retrials. Here, signals correspond to the changes in blocked/non-blocked state of a radio channel and the retrials reflect the persistent user behaviour. We characterize this system by deriving all the considered performance metrics. Finally, we parameterize the model by using the propagation, channel, and blockage models.

A. Model Solution

The behaviour of the mmWave/sub-THz BS service model specified in Section III can be described by a three-dimensional stochastic process $X(t) = (\xi(t), \delta(t), \varphi(t))$, where $\xi(t)$ is the number of sessions in service, $\delta(t)$ is the total number of resources occupied by them, and $\varphi(t)$ is the number of sessions on the orbit, i.e., waiting for re-establishment. Each component of the stochastic process $X(t)$ is assumed to be limited by a finite value. The number of occupied resources $\delta(t)$ cannot be more than the total number of PRBs R. The number of sessions in service $\xi(t)$ is limited by N, which can be evaluated as the fraction of total number of PRBs R to the minimal PRB requirements of a session. On the other hand, the number of sessions on the orbit $\varphi(t)$ should not be limited, naturally. However, this leads to infinite state space of $X(t)$, and the numerical solution of the system of balance equations will be practically infeasible. To avoid this, we introduce the upper limit on the maximum number of sessions in the orbit M, and choose it large enough to ensure that the session drop probability caused by the finite orbit size is negligibly small.

Also note that keeping track of only total amount of resources occupied by all sessions instead of a vector of occupied resources by each session is typical in the analysis of the ReLSs [3], [4], [38]. In this case, one cannot exactly determine the amount of released resources upon the session departure, but it can be approximated by well-known Bayes's formula. It was shown in [39] that this approach provides accurate approximation of the mmWave/sub-THz BS service process. In what follows, we will refer to the new arriving session as *primary sessions*, and to those that re-enter the system as *secondary sessions*.

The state space of the stochastic process $X(t)$ is given by

$$S = \bigcup_{n=0}^{N} S_n, \tag{8}$$

$$S_n = \left\{ (n, r, m) : 0 \le m \le M, \sum_{i=0}^{n} \sum_{j=0}^{r} p_{1,j}^{(i)} p_{2,r-j}^{(n-i)} > 0 \right\},$$

where the pmf $\{p_{s,j}^{(i)}\}$, $1 \le j \le R$, $s = \{1, 2\}$ is an i-fold convolution of the pmf $\{p_{s,j}\}, 1 \le j \le R$ and $p_{s,j}^{(i)}$ denotes the probability that i sessions of type s, $s = \{1, 2\}$ totally occupy j resources.

The discrete convolutions involved in (8) can be efficiently evaluated recurrently as

$$\begin{cases} p_{s,0}^{(0)} = 1, & p_{s,j}^{(0)} = 0,\ 1 \le j \le R, \\ p_{s,j}^{(1)} = p_{s,j}, & 1 \le j \le R, \\ p_{s,j}^{(i)} = \sum_{r=1}^{j-1} p_{s,r} p_{s,j-r}^{(i-1)}, & 1 \le j \le R,\ i \ge 2. \end{cases} \tag{9}$$

To reduce the state space of the system we do not distinguish primary and secondary sessions in the state description. To this aim, we assume that the resource requirements of sessions in the system are distributed according to pmf $p_{3,j}, 1 \le j \le R$, which is a mixture of $p_{1,j}$ and $p_{2,j}$,

$$p_{3,j} = \bar{N}_1 p_{1,j} + \bar{N}_2 p_{2,j}, \tag{10}$$

where $\bar{N}_1$ and $\bar{N}_2$, $\bar{N}_1 + \bar{N}_2 = 1$, are the fractions of primary and secondary sessions in the system, respectively. We will derive $\bar{N}_1$ and $\bar{N}_2$ in what follows.

There are four types of events that may cause the state transition of $X(t)$. We summarize them together with the corresponding probabilities, presuming that the system is initially in the state (n, r, m):

1) *Arrival of a primary session that requires j PRBs for service.* If $j \le R - r$ and $n < N$, then this session is accepted for service and the system jumps to the state $(n+1, r+j, m)$. Alternatively, if $j > R-r$ or $n = N$, and $m < M$ then this session either leaves for the orbit with probability θ changing the system state to $(n, r, m+1)$ or leaves the system with probability $1 - \theta$ with no change in the system state. Finally, if there is no place on the orbit, $m = M$, the session leaves the system.
2) *Service completion of a session.* In this case, the resources occupied by this session are released. Recall, that since it is impossible to exactly determine the number of the released PRBs, we employ the probabilistic approach. We assume that a session releases j PRBs with probability $p_{3,j} p_{3,r-j}^{(n-1)} / p_{3,r}^{(n)}$ and the system moves to the state $(n - 1, r - j, m)$.
3) *Signal arrival for a session.* Upon blockage event, we need to re-calculate the amount of resources required to maintain a given rate. Thus, we assume that the session

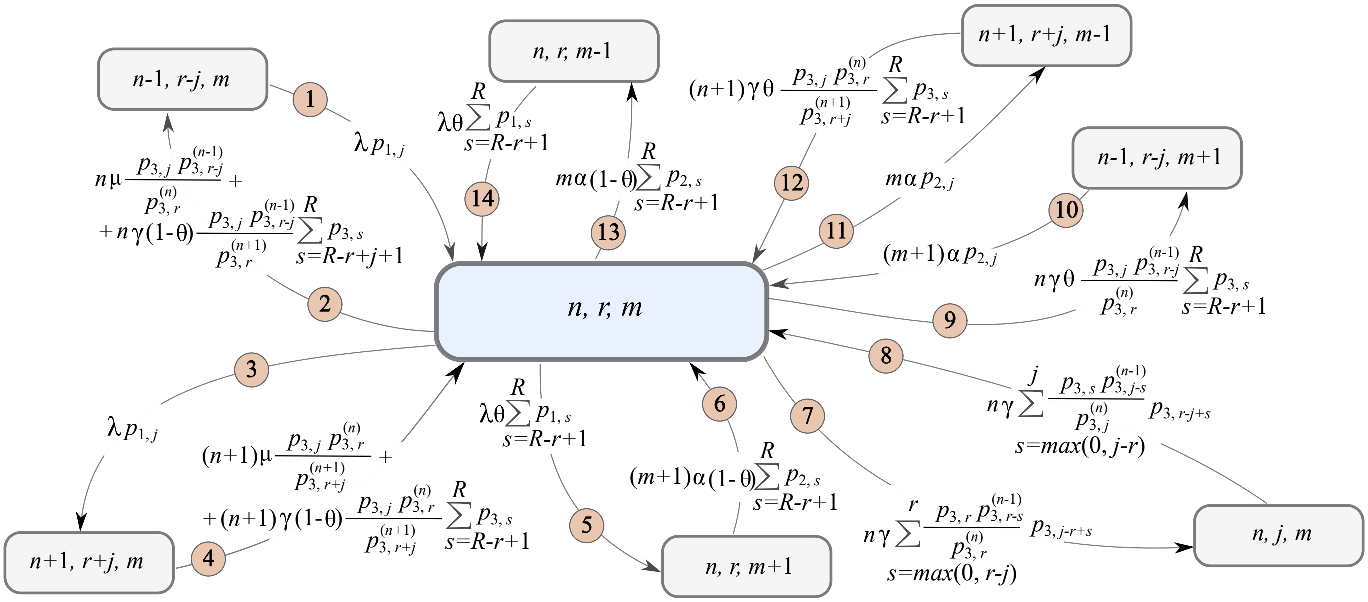


Fig. 5. Generic state of the system with transitions to the neighboring states: 1. Arrival and acceptance of a primary session; 2. Successful completion of the service or loss of a session due to arrival of a signal; 3. Arrival and acceptance of a primary session; 4. Successful completion of the service or loss of a session due to arrival of a signal; 5. Arrival of a primary session and redirection to the orbit due to the lack of resources; 6. Unsuccessful attempt of a secondary session to continue its service due to the lack of resources; 7. Successful change of resource requirements of a session upon a signal arrival (service is continued); 8. Successful change of resource requirements of a session upon a signal arrival (service is continued); 9. Unsuccessful change of resource requirements of a session upon a signal arrival (session is dropped); 10. Successful attempt of a secondary session to continue its service; 11. Successful attempt of a secondary session to continue its service; 12. Unsuccessful change of resource requirements of a session (service is continued); 13. Arrival of a primary session and its re-direction to the orbit due to the lack of resources; 14. Unsuccessful attempt of a secondary session to continue its service (session is dropped).

releases j resource units with probability $p_{3,j}p^{(n-1)}_{3,r-j}/p^{(n)}_{3,r}$, and attempts to occupy the newly generated resource requirement i having pmf $p_{3,i}$. If $i \leq R-r+j$, then the session continues its service, and the system switches to the state $(n, r-j+i, m)$. Otherwise, if $i > R-r+j$, the session either leaves for the orbit with probability θ or leaves the system with probability $1-\theta$.

4) *Arrival of a secondary session that requires j PRBs.* Once the waiting time of a session on the orbit is over, it tries to re-start its service with the reduced resource requirements having pmf $p_{2,j}, 1 \leq j \leq R$. If $j \leq R-r$ and $n < N$, then the session is accepted for service, and the system switches to the state $(n+1, r+j, m-1)$. Otherwise, the session may either return to the orbit with probability θ or leave the system with probability $1-\theta$.

These four types of events cause eight main types of system state changes: (i) arrival and acceptance of a primary session, (ii) arrival and re-routing to the orbit of a primary session, (iii) service completion of a session, (iv) resource re-allocation caused by a signal arrival, (v) re-routing a session from the system to the orbit as a result of a insufficient amount of resources in case of signal arrival, (vi) session leaving the system as a result of a signal arrival and user refusing to re-establish it, (vii) acceptance of a secondary session (returning from the orbit as a results of session re-establishment attempt) for service, and, finally, (viii) secondary session leaving the system as a result of observing insufficient amount of resources upon arrival. A generic state of the system and associated transitions are illustrated in Fig. 5. Note that the actual number of transitions is greater than listed above due to symmetry of events associated with blockage.

Let $q(n, r, m)$ be the stationary state probability of $X(t)$,

$$q(n, r, m) = \lim_{t\to\infty} P\{\xi(t) = n, \delta(t) = r, \varphi(t) = m\}. \quad (12)$$

Based on the state transitions described above, one may derive the system of equilibrium equations as in (11), where $u(\cdot)$ is Heaviside function, i.e., $u(x) = 0$, $x \leq 0$, $u(x) = 1$, $x > 0$. Note that (11) is a system of linear equations that can be solved numerically for stationary state probabilities by utilizing the standard techniques. We recommend the iterative Gauss-Seidel method [40].

B. Performance Metrics

Having obtained the stationary state probabilities, we can proceed deriving the user- and system-centric metrics. Observe that the life-cycle of a session may contain numerous "service" and "suspension" rounds, see Fig. 6. We define the service round as the time interval during which a session is continuously served. At the end of a service round, a session is either dropped, re-routed to the orbit, or finishes the service and leaves the system. Similarly, a suspension round is the time interval that a session continuously spends on the orbit. At the end of a suspension round, a session is either dropped or accepted for service again. To derive the metrics of interest, we need auxiliary probabilities that are summarized Table I along with the sought metrics.

$$
\begin{aligned}
&(\lambda + m\alpha)q(0,0,m) = \mu \sum_{j:(1,j,m)\in S_1} q(1,j,m), \quad 0 \le m \le M,\\
&q(n,r,m)\left[\lambda\left(\sum_{j=0}^{R-r} p_{1,j} + u(M-m)\theta \sum_{j=R-r+1}^{R} p_{1,j}\right) + n\mu + n\gamma\left(1 - \sum_{j=1}^{r} \frac{p_{3,j}p_{3,r-j}^{(n-1)}}{p_{3,r}^{(n)}} p_{3,j}\right) + \right.\\
&\quad \left. + m\alpha\left(\sum_{j=1}^{R-r} p_{2,j} + (1-\theta)\sum_{j=R-r+1}^{R} p_{2,j}\right)\right] = \lambda \sum_{j<r,(n-1,j,m)\in S_{n-1}} q(n-1,j,m)p_{1,r-j} + u(m)\lambda\theta q(n,r,m-1)\sum_{j=R-r+1}^{R} p_{1,j} +\\
&\quad + (n+1)\mu \sum_{j>r,(n+1,j,m)\in S_{n+1}} q(n+1,j,m)\frac{p_{3,j-r}p_{3,r}^{(n)}}{p_{3,j}^{(n+1)}} + n\gamma \sum_{(n,j,m)\in S_n, j\ne r} q(n,j,m) \sum_{s=max(0,j-r)}^{j} \frac{p_{3,s}p_{3,j-s}^{(n-1)}}{p_{3,j}^{(n)}} p_{3,r-j+s} +\\
&\quad + (n+1)\gamma\left(\sum_{j>r,(n+1,j,m)\in S_{n+1}} (u(m)\theta q(n+1,j,m-1) + (1-\theta)q(n+1,j,m))\frac{p_{3,j-r}p_{3,r}^{(n)}}{p_{3,j}^{(n+1)}} \sum_{s=R-r+1}^{R} p_{3,s}\right) +\\
&\quad + u(M-m)(m+1)\alpha\left((1-\theta)q(n,r,m+1)\sum_{j=R-r+1}^{R} p_{2,j} + \sum_{j<r,(n-1,j,m+1)\in S_{n-1}} q(n-1,j,m+1)p_{2,r-j}\right),\\
&\quad 0 < n < N, \quad 0 \le m \le M, \quad (n,r,m) \in S,\\
&q(N,r,m)\left[u(M-m)\theta\lambda + N\mu + N\gamma\left(1 - \sum_{j=1}^{r}\frac{p_{3,j}p_{3,r-j}^{(N-1)}}{p_{3,r}^{(N)}}p_{3,j}\right) + m\alpha(1-\theta)\right] = \lambda \sum_{j<r,(N-1,j,m)\in S_{N-1}} q(N-1,j,m)p_{1,r-j} +\\
&\quad + u(m)\lambda\theta q(N,r,m-1) + N\gamma \sum_{(N,j,m)\in S_N, j\ne r} q(N,j,m) \sum_{s=max(0,j-r)}^{j} \frac{p_{3,s}p_{3,j-s}^{(N-1)}}{p_{3,j}^{(N)}} p_{3,r-j+s} + u(M-m)(m+1)\alpha \times\\
&\quad \times\left((1-\theta)q(N,r,m+1) + \sum_{j<r,(N-1,j,m+1)\in S_{N-1}} q(N-1,j,m+1)p_{2,r-j}\right), 0 \le m \le M, \quad (N,r,m) \in S_N. \qquad (11)
\end{aligned}
$$

1) New Session Drop Probability: We start with the new session drop probability π_N, which is the probability that a session does not receive any service during its life-cycle. This may happen as a result of the following two events: (i) an arriving primary session is immediately dropped, (ii) an arriving primary session is re-routed to the orbit and at the end of the suspension round is dropped. The probability π_1 of the first event can be found directly from the stationary distribution of the system,

$$
\begin{aligned}
\pi_1 = &\sum_{r=1}^{R}\left(q(N,r,M) + (1-\theta)\sum_{m=0}^{M-1} q(N,r,m)\right) +\\
&+ \sum_{n=1}^{N-1}\sum_{r=1}^{R}\left(q(n,r,M) + (1-\theta)\sum_{m=0}^{M-1} q(n,r,m)\right) \times\\
&\times \sum_{j=R-r+1}^{R} p_{1,j}. \qquad (13)
\end{aligned}
$$

To obtain the probability of the second event, we need the probability π_2 that an arriving primary session is re-routed to the orbit, which is also evaluated directly from the stationary state probabilities as follows

$$
\begin{aligned}
\pi_2 = &\theta \sum_{n=1}^{N-1}\sum_{r=1}^{R}\sum_{m=0}^{M-1} q(n,r,m) \sum_{j=R-r+1}^{R} p_{1,j} +\\
&+ \theta \sum_{r=1}^{R}\sum_{m=0}^{M-1} q(N,r,m), \qquad (14)
\end{aligned}
$$

where the complementary probability $(1 - \pi_1 - \pi_2)$ is the probability that a primary session is accepted for service.

We also need probabilities π_3 and π_4 that a secondary session is either dropped or returns to the orbit upon an attempt to re-establish it. These probabilities can be found as the fractions of corresponding session arrival intensities. The average intensity of secondary session arrivals is $\alpha\bar{M}$, where $\bar{M}$ is the average number of sessions on the orbit provided by

$$
\bar{M} = \sum_{n=0}^{N}\sum_{r=0}^{R}\sum_{m=1}^{M} mq(n,r,m). \qquad (15)
$$

Therefore, probability π_3 is the fraction of the average intensity ν_3 of session dropped upon their return from the orbit to the average intensity of all secondary session arrivals,

$$
\pi_3 = \nu_3/\alpha\bar{M}, \qquad (16)
$$

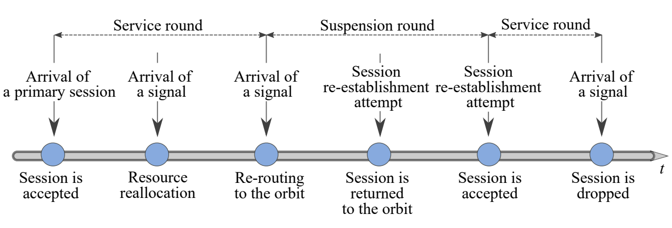


Fig. 6. Life-cycle of a session.

TABLE I
AUXILIARY AND FINAL METRICS

Metric	Probability that
π_1	arriving primary session is dropped
π_2	arriving primary session is routed to the orbit
π_3	secondary session is dropped on an attempt to start service
π_4	secondary session returns to the orbit on attempt to start service
π_5	session will be dropped at the end of the current service round
π_6	session will be routed to the orbit at the end of the service round
π_7	secondary session is dropped at the end of the suspension round
π_N	new session drop
π_O	ongoing session drop
π_S	session is served without visiting the orbit
π_W	session is served with at least one visit to the orbit
π_{orb}	session visits the orbit least once

where ν_3 is evaluated as

$$\nu_3 = (1-\theta)\alpha \sum_{n=1}^{N-1}\sum_{r=1}^{R}\sum_{m=1}^{M} mq(n,r,m) \sum_{j=R-r+1}^{R} p_{2,j} + \\ + (1-\theta)\alpha \sum_{r=1}^{R}\sum_{m=1}^{M} mq(N,r,m). \quad (17)$$

Similarly, probability π_4 that a secondary session returns to the orbit upon an attempt to re-start its service is the fraction of the average intensity ν_4 of returns to the orbit to the average intensity of the secondary session arrivals. Since ν_4 differs from ν_3 by a multiplier only, we have

$$\pi_4 = \frac{\theta}{1-\theta}\pi_3, \quad (18)$$

where the probability $(1-\pi_3-\pi 4)$ is the probability that an attempt to re-establish the secondary session is successful.

Finally, the new session drop probability π_N can be written in terms of the introduced auxiliary variables as follows

$$\pi_N = \pi_1 + \pi_2 \left(\sum_{k=0}^{\infty} \pi_4^k\right) \pi_3 = \pi_1 + \frac{\pi_2\pi_3}{1-\pi_4}. \quad (19)$$

2) Ongoing Session Drop Probability: We refer to a session as "ongoing" if its service has already started. Observe that an ongoing session may be dropped upon the following two events: (i) when a signal specifying the blockage event arrives provided that the new resource requirements of a session are greater than the unoccupied amount of resources in the system, (ii) when an attempt to return to the service from the orbit fails, i.e., there are no sufficient resources to re-start the service and the user decides not to re-establish the session.

Consider a single service round of a session. At the end of it, there are three possible outcomes: (i) the session is dropped, (ii) the session is re-routed to the orbit, and (iii) the session successfully finishes its service and leaves the system. Denote the corresponding probabilities by π_5, π_6, and $1-\pi_5-\pi_6$.

The probabilities, π_5 and π_6, can be evaluated as fractions of intensities. The probability π_5 that a session is dropped at the end of the current service round is a fraction of the average intensity ν_5 of session drops caused by signal arrivals to the average intensity of primary and secondary sessions accepted for service, that is,

$$\pi_5 = \frac{\nu_5}{\lambda(1-\pi_1-\pi_2) + \alpha\bar{M}(1-\pi_3-\pi_4)}, \quad (20)$$

where $\lambda(1-\pi_1-\pi_2)$ is the primary sessions acceptance intensity, $\alpha\bar{M}(1-\pi_3-\pi_4)$ is the average secondary sessions acceptance intensity, and ν_5 is evaluated as

$$\nu_5 = \gamma \sum_{n=2}^{N}\sum_{r=1}^{R} n \left(q(n,r,M) + \sum_{m=0}^{M-1} (1-\theta) q(n,r,m) \right) \times \\ \times \sum_{j=1}^{r} \frac{p_{3,j} p_{3,r-j}^{(n-1)}}{p_r^{(n)}} \sum_{s=R-r+j+1}^{R} p_{3,s}. \quad (21)$$

Similarly, π_6 is a fraction of the average intensity ν_6 of sessions leaving for orbit as a result of signal arrivals to the average intensity of primary and secondary sessions accepted for service, i.e.,

$$\pi_6 = \frac{\nu_6}{\lambda(1-\pi_1-\pi_2) + \alpha\bar{M}(1-\pi_3-\pi_4)}, \quad (22)$$

where ν_6 is computed as

$$\nu_6 = \theta\gamma \sum_{n=2}^{N}\sum_{r=1}^{R} n \sum_{m=0}^{M-1} q(n,r,m) \times \\ \times \sum_{j=1}^{r} \frac{p_{3,j} p_{3,r-j}^{(n-1)}}{p_r^{(n)}} \sum_{s=R-r+j+1}^{R} p_{3,s}. \quad (23)$$

Now, consider a suspension round of a session. At the end of it, there are only two possible outcomes: (i) the session is dropped due to insufficient amount of resources and user decides not to re-establish it and (ii) the session is accepted for service initiating a new service round. Denote these probabilities by π_7 and $1-\pi_7$, respectively. Observe that the session returning from the orbit can be dropped due to insufficient amount of available resources and user deciding not to re-establish the session. Accounting for the possibility of dropping the session at any return from the orbit, we have the following for π_7

$$\pi_7 = \sum_{k=0}^{\infty} \pi_4^k \pi_3 = \frac{\pi_3}{1-\pi_4}. \quad (24)$$

Finally, by combining the terms, we can write the ongoing session drop probability π_O as follows

$$\pi_O = (1-\pi_N)\Big(\sum_{k=0}^{\infty} \left(\pi_6(1-\pi_7)\right)^k \pi_5 + \\ + \sum_{k=0}^{\infty} \left(\pi_6(1-\pi_7)\right)^k \pi_6\pi_7 \Big) = \\ = (1-\pi_N)\frac{\pi_5 + \pi_6\pi_7}{1-\pi_6(1-\pi_7)}, \quad (25)$$

where $1-\pi_N$ is the probability that a session receives at least one service round, then proceeds through $k \geq 0$ suspension and service rounds, and then either is dropped at the end of the last service round with probability π_5 or returns to the orbit for the last time and is finally dropped at the end of the suspension round with probability $\pi_6\pi_7$.

3) Service With and Without Visiting an Orbit (Retrial): Critical measures of interest are successful service completions with and without retrial. The latter implies that the session rate is degraded by the application layer rate adaptation. The probability π_S that a primary session is accepted for service, and its service is finished during the first service round is

$$\pi_S = (1 - \pi_1 - \pi_2)(1 - \pi_5 - \pi_6). \quad (26)$$

The successful service completion probability of a session provided that it visited an orbit and experienced rate degradation π_W can be evaluated as

$$\pi_W = 1 - \pi_N - \pi_O - \pi_S. \quad (27)$$

Finally, we are also interested in the probability π_{orb} that a session takes at least one suspension round which is given by

$$\pi_{orb} = 1 - \pi_1 - \pi_S. \quad (28)$$

4) System-Centric Metrics: In addition to user-centric performance metrics we are interested in operator-centric ones with the efficiency of resource utilization and fraction of wasted resources being the most important ones. The system resource utilization is the average fraction of occupied resources and can be computed directly by utilizing the stationary state probabilities of the system as

$$U = \frac{1}{R} \sum_{n=1}^{N} \sum_{r=1}^{R} \sum_{m=0}^{M} r q(n, r, m). \quad (29)$$

A critical metric for 5G/6G mmWave/sub-THz systems with unreliable service induced by blockage is the fraction of wasted resources U_W. This metric is interpreted as the time averaged fraction of resources provided to sessions that are eventually dropped. It can be found as

$$U_W = \frac{\pi_O}{\pi_O + \pi_S + \pi_W} U. \quad (30)$$

where $\pi_O + \pi_S + \pi_W$ is a fraction of sessions that received at least one service round, and π_O is a fraction of sessions that were dropped after receiving at least one service round.

C. Computational Algorithm

To determine the stationary state distribution of $\{q(n, r, m)\}$ by solving the system of equations in (11), we utilize the weighted average resource requirements distribution $p_{3,j}$, $1 \leq j \leq R$. However, the weights themselves cannot be determined before the system in (11) is solved.

To determine the weights and eventually obtain the solution for $\{q(n, r, m)\}$ we propose the following iterative algorithm, see Fig. 7. The procedure starts with $\bar{N}_1 = 1$ and $\bar{N}_2 = 0$ that are substituted to (10). After that, the system (11) is solved, and the obtained stationary state distribution is utilized to evaluate auxiliary probabilities π_1, π_2, π_3 and π_4. The latter are then used to calculate the updated values of weights as

$$\bar{N}_1 = \frac{\lambda(1 - \pi_1 - \pi_2)}{\lambda(1 - \pi_1 - \pi_2) + \alpha \bar{M}(1 - \pi_3 - \pi_4)};$$
$$\bar{N}_2 = \frac{\alpha \bar{M}(1 - \pi_3 - \pi_4)}{\lambda(1 - \pi_1 - \pi_2) + \alpha \bar{M}(1 - \pi_3 - \pi_4)}, \quad (31)$$

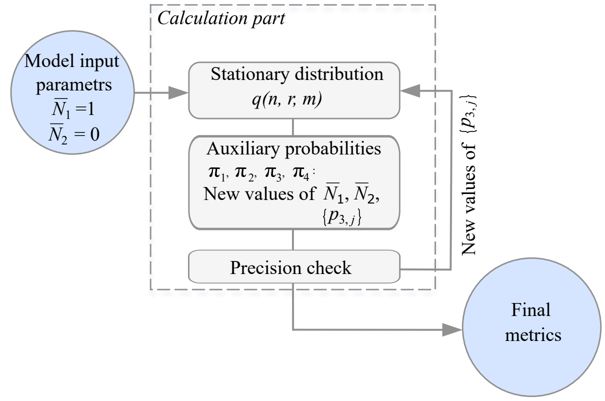


Fig. 7. Iterative algorithm for calculating performance metrics.

and the new weighted average distribution $p_{3,j}$, $1 \leq j \leq R$.

The procedure proceeds until the difference between the previous and the current value of $\bar{N}_1$ reaches the desired accuracy level. Note that we make the code for solving the considered system for performance metrics of interest freely available at GitHub [41].

D. System Parameterization

The formulated and solved system accounts for the major specifics of the 5G/6G mmWave/sub-THz systems – outage caused by blockage situations and random resource requirements. The intensity of blockage events is provided in Section III while $p_{3,j}$ is expressed in terms of $p_{1,j}$ and $p_{2,j}$ in (10). Thus, to complete parameterization of the system we need to calculate the resource requirements of primary and secondary sessions, $p_{1,j}$, $p_{2,j}$, $1 \leq j \leq R$. Recall, that the principal difference between these pmfs is that $\{p_{2,j}\}$ characterizes the resource requirements when the application rate is reduced. Thus, in what follows, we show how to derive $p_{1,j}$ only.

Since our goal in this paper was mainly to characterize the impact of user persistence on the service performance in 5G/6G mmWave/sub-THz systems our choice was to account for the main factors before and after rate reduction. These include shadow fading, fast fading, and interference. Since it is the difference between resources requested before and after rate reduction that mainly affects service performance, we do that consistently by utilizing their margins, as it is often done in practice. The chosen parameters for fast and shadow fading margin are taken from the 3GPP TR 38.901.

The value of SINR at the UE in the non-blocked state can be approximated by

$$S(y) = \frac{C y^{-\zeta} \Omega \Upsilon}{(N_0 B + I)}, \quad (32)$$

where $C = P_T G_A G_U 10^{-2 \log_{10} f_C - 3.24}$, see (5), P_T is BS emitted power, G_A and G_U are the BS and UE gains calculated by using (6), I is the interference, Υ and Ω are the fast and shadow fading, y is 3D distance between UE and BS, B is the bandwidth, and $\zeta = 2.1$ [35].

The coverage radius of BS, r_M can be determined based on the SINR threshold, $S_{\min}$, representing the minimum SINR value for MCS schemes by solving (32) with respect to y

$$r_M = \sqrt{\left(\frac{Cy^{-\zeta}\Omega\Upsilon}{(N_0B+I)S_{\min}}\right)^{2/\zeta}}, \tag{33}$$

where one may replace Ω, Υ, and I by their margins, Ω_M, Υ_M, and I_M, as it is often done in practical coverage assessment, leading to

$$r_M = \sqrt{\left(\frac{Cy^{-\zeta}}{(N_0B+I_M+\Omega_M+\Upsilon_M)S_{\min}}\right)^{2/\zeta}}, \tag{34}$$

where I_M can be computed using models in [9], [18] while Ω_M and Υ_M are provided in [35].

One may now proceed determining the sought pmf $p_{1,j}$. Recall, that geometric locations of users generating sessions are assumed to be uniformly distributed in the cell coverage. Thus, the probability density function (pdf) of the distance to a randomly selected UE is given by $f_D(x) = 2x/r_M^2$, $0 < x < r_M$ [42]. Therefore, the pdf of the 3D distance, Y, is

$$f_Y(y) = 2y/r_M^2,\ y \in (|h_A - h_U|, Q), \tag{35}$$

where $Q = \sqrt{r_M^2 + (h_A - h_U)^2}$, h_A is the height of BS, h_U is the height of UE, leading to the cumulative distribution function (CDF) $F_Y(y)$ in the following form

$$F_Y(y) = \frac{y^2 - (h_A - h_U)^2}{r_M^2},\ y \in [|h_A - h_U|, Q]. \tag{36}$$

The SINR CDF can be written by using Y as

$$F_S(s) = Pr\left\{Ay^{-\zeta} < s\right\} = 1 - F_Y\left(\sqrt[\zeta]{A/s}\right), \tag{37}$$

where $A = C/(N_0B + I_M + \Omega_M + \Upsilon_M)$.

Now, the SINR CDF is given by

$$F_S(s) = \frac{Q^2 - \left(\frac{A}{s}\right)^{\frac{2}{\zeta}}}{r_M^2},\ \frac{A}{Q^\zeta} \le s < \frac{A}{(h_A - h_U)^\zeta}. \tag{38}$$

Finally, by defining SINR boundaries for MCS-SINR mapping [43] as S_j, $j = 1, 2, \ldots, J$ and also letting ϵ_j be the probability that the session is associated with MCS j and requires r_j PRBs, we have

$$\epsilon_j = F_S(S_{j+1}) - F_S(S_j),\ j = 1, 2, \ldots, J, \tag{39}$$

and the probability ϵ_j that a session requests r_j PRBs can now be used to obtain the resource requirements pmf $p_{1,r}$.

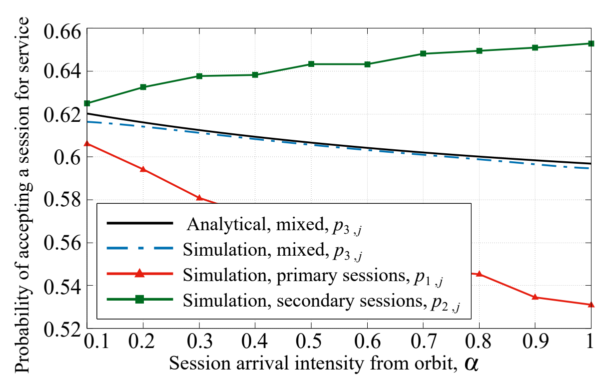
Note that the framework can be extended to account for the case of more than one active session per user. This can be done by changing the session arrival intensity from a single user, λ_U, introduced in Section III-C and utilized to compute the aggregated session arrival rate λ. The limiting factors, however, are that: (i) no UEs should dominate in the aggregated traffic arrival rate λ and (ii) the ratio $lambda_B/lambda_U$ should be more than approximately 10-20. These two limitations ensure that there is enough spatial diversity in the arrival session process that is required when calculating the pmf of the resource requirements $p_{1,j}$ and $p_{2,j}$. If these requirements are not satisfied, one needs to further modify the way how these pmfs are estimated by explicitly accounting for spatial correlation in the session arrival process.

TABLE II
DEFAULT PARAMETERS FOR NUMERICAL ASSESSMENT

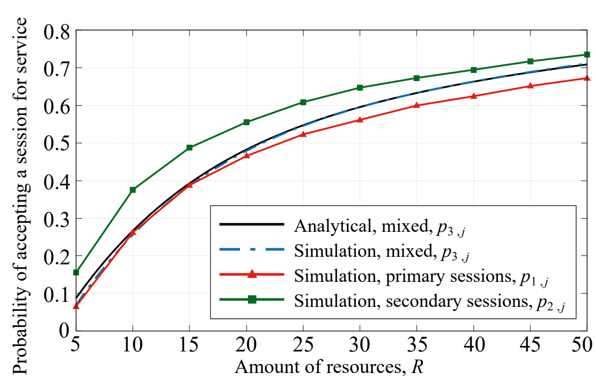
Parameter	Value
Carrier frequency, f_c	28 GHz
Bandwidth, B	100 MHz
Height of BS, h_A	10 m
Height of blocker, h_B	1.7 m
Height of UE, h_U	1.5 m
Blocker radius, r_B	0.4 m
Blocker velocity, v	1 m/s
Mean blocker movement duration, $1/\tau$	5 m
SINR threshold, S_B	-9.47 dB
Transmit power, P_T	2 W
Path loss exponent, ζ	2.1
Interference margin, I_M	3 dB
Fast and shadow fading margins, Υ_M, Ω_M	3 dB, 3 dB
Blocker intensity, λ_B	0.05 units/m^2
Thermal noise, N_0	-174 dBm/Hz
Blockage attenuation, L_B	15 dB
UE antenna array	4×4 el.
Maximum session rate, $C_{\max}$	10 Mbps
Minimum session rate, $C_{\min}$	5 Mbps
Mean session service time, $1/\mu$	10 s
Session arrival intensity, λ	0.3, sess./s
Utilized 5G NR numerology	3
BS antenna array	16×16 el.

V. Numerical Results

In this section, we will numerically study our system addressing the impact of unreliable wireless medium, user persistence, and rate adaptation. Specifically, we start with assessing the errors introduced by our modeling assumptions and verifying the intermediate metrics by using the simulation

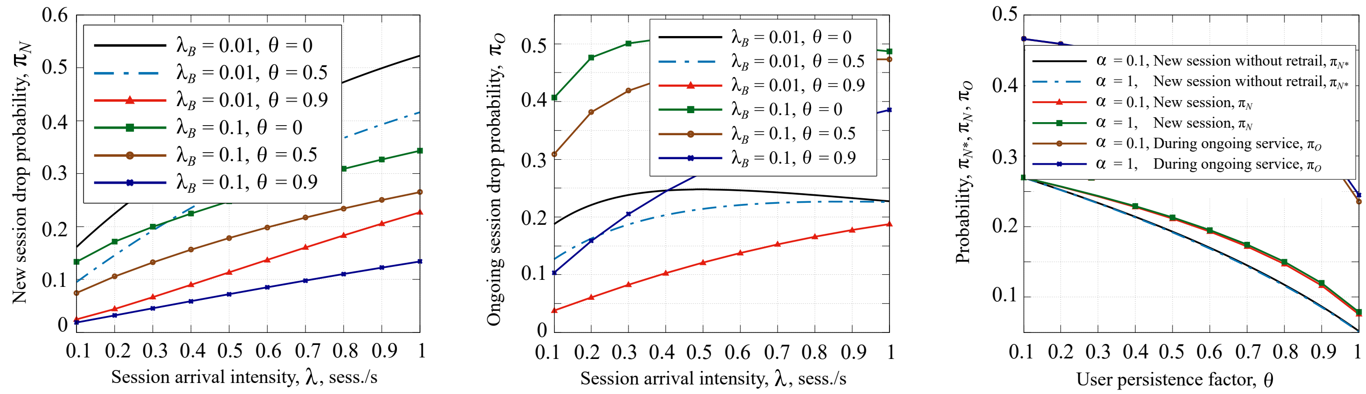


(a) Mean values of pmfs of resource requirements



(b) Ongoing session drop probability

Fig. 8. Assessment of intermediate parameters.



(a) New session drop probability (b) Ongoing session drop probability (c) Impact of persistence factor

Fig. 9. Main user-centric performance measures of the system.

studies. Next, we consider the main user-centric performance measures – new and ongoing session drop probabilities – as a function of environmental characteristics and persistent user behavior. Owing to the complex nature of these final performance measures, we then proceed explaining the rationale for their behavior by analyzing complementary probabilities, such as session completion probability with and without retrials. Finally, we study the impact of radio part specifics including the antenna arrays, etc. For numerical examples we utilize 5G NR system with numerology 3 characterized by the PRB size of 1.44 MHz. The default system parameters utilized in this section are provided in Table II.

In our study, we utilize the RDM mobility model as specified in Section III. In fact, we need RDM analytical properties to re-utilize the results on the duration of the blockage and non-blockage intervals reported in [11] to further establish the intensity of blockages. For different mobility models these results might slightly deviate. The derivation of these metrics for other models are much more complex as compared to RDM. We would like to note that we utilize only two properties of the mobility model in our study: (i) limiting distribution which should be uniform in the cell coverage (with balanced flux across the cell boundary) and (ii) local behavior as in [11], when estimating the probability that a blocker will cross the LoS blockage zone. The former one is intrinsic for a number of models utilized in the literature (e.g., Brownian motion, Brownian motion with jumps, Levy flights [44]), while the latter behavior could serve as a reason for deviation. That is, by following [45] RDM and random waypoint model (RWP) may result in higher blockage intensity due to straight motion, while Brownian motion should be characterized with lower blockage intensity. Levy flights and Brownian motion with jumps are in-between these two cases.

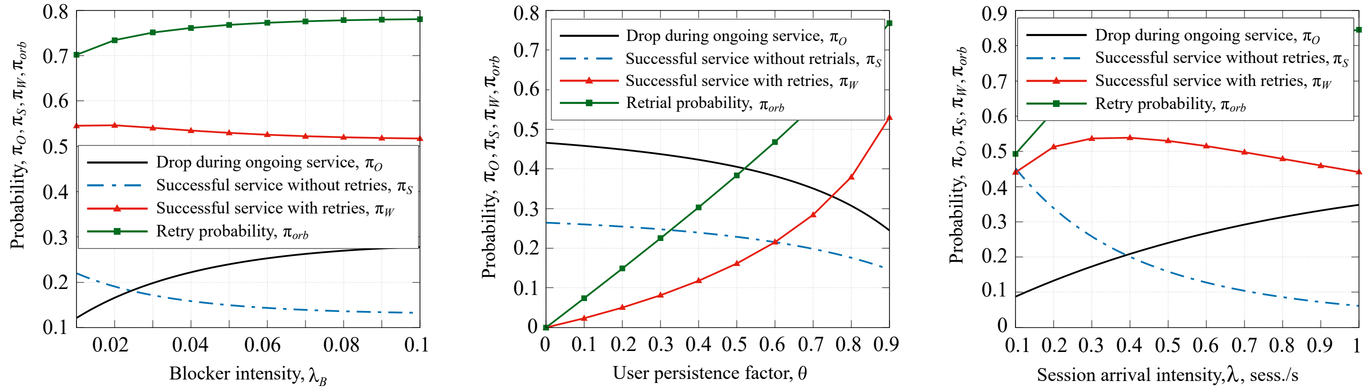
A. Assessment of Utilized Assumptions

Recall, that the critical assumption in our study allowing to build a fully tractable model is the pmf of resource requirements utilized by sessions upon user retrials, $p_{3,j}, 1 \leq j \leq R$, which is computed as a weighted sum of primary and secondary sessions in the system in (10). Thus, in Fig. 8(a) we validate this assumption by comparing the probability of accepting a session upon retrial with that obtained using computer simulations (black and blue lines) as a function of the intensity of retrials α. Here, we also show the results for probability of accepting primary and secondary sessions obtained using the computer simulations (red and green lines) with exact pmfs $p_{1,j}$ and $p_{2,j}$. First of all, we see that the considered metric obtained using the simulations and the proposed model closely coincide for the whole considered range of α. We further see that when the mean time between retrials is large (e.g., $1/\alpha = 10$ s), this approximation results in overestimation of the actual probability of accepting a session upon retrial. However, even for large values of α, the gap is only $0.05 - 0.07$.

To understand the ultimate impact of the utilized approximation, Fig. 8(b) shows the probability of accepting a session upon retrial as a function of the number of PRBs obtained using the proposed model and computer simulations (black and blue lines). Similarly to the dependence on α, the proposed approximation closely coincides with the probability obtained using computer simulations and these curves are located in-between two other probabilities. We also observe that the approximation error becomes smaller as the system capacity increases. Thus, for the practical range of the amount of system resources ($R > 100/1.44 \approx 70$ PRBs) *the error we introduce by utilizing pmf $p_{3,j}$ is upper bounded by $5 - 7\%$.*

B. Main Performance Measures

Now, we proceed with user-centric performance measures in Fig. 9, where the new and ongoing session drop probabilities are shown as a function of session arrival intensity in Fig. 9(a) and Fig. 9(b) and user persistence factor in Fig. 9(c), where the mean retrial time is $1/\alpha = 1$ s. Starting with the new session drop probability, we observe that it logically increases as the arrival intensity of sessions increases. Here, we may also observe that the higher blockage density of $\lambda_B = 0.1$ bl./m^2 leads to lower new session drop probability as compared to $\lambda_B = 0.01$ for all the values of the user persistence factor θ. Note that this is typical for systems allowing for session interruptions, see, e.g., [3], [29], where the authors have shown that without user persistence the blockage events leading to session interruptions allows to free up resources



(a) As a function of the blockers' density

(b) As a function of user persistence

(c) As a function of arrival intensity

Fig. 10. Advanced user-centric performance metrics of the system.

of the system for new sessions. Notably, in our case, the user persistence drastically improves the new session drop probability. Specifically, for $\lambda_B = 0.01$ bl./m^2 and no retrials ($\theta = 0$), the considered metric is 0.17 and decreases down to 0.1 for two retrials on average, ($\theta = 0.5$).

Analyzing the ongoing session drop probability, shown in Fig. 9(b), we observe a drastic difference between curves corresponding to $\lambda_B = 0.1$ bl./m^2 and $\lambda_B = 0.01$ bl./m^2 reaching $0.2 - 0.3$, i.e., the higher blockers density leads to higher ongoing session drop probability compensating for the lower new session drop probability shown in Fig. 9(a). In general, for the considered values of the session arrival intensities, a large fraction of the sessions accepted for service are eventually lost. However, user persistence may drastically improve the successful session completion probability. Specifically, for $\lambda = 0.1$ sess./s, $\lambda_B = 0.1$ bl./m^2 user persistence drives the ongoing session drop probability down from 0.4 to 0.1. We note that the impact of user persistence is more profound for underloaded system, e.g., when λ is small. The rationale is that in these conditions, session interruptions caused by blockage is the only impairment affecting session continuity.

Fig. 9(c) highlights the impact of retrial intensity, α, on the new and ongoing session drop probability for $\lambda = 0.3$ sess./s., $\lambda_B = 0.05$ bl./m^2. We observe that the considered drop probabilities are independent of the mean time to re-initiate the session, $1/\alpha$, across the whole range of the user persistence factor θ. In other words, the service process at 5G/6G mmWave/sub-THz BS is "stable" and the re-established session sees similar conditions irrespective of the time it took for a user to re-establish the session. Of interest is also the metric's behavior as a function of θ. We observe a linear behavior for small values of θ and faster decline for larger values. It implies that *the user has to be very persistent for the service to be successfully completed*.

C. Detailed Behaviour and Advanced Metrics

To understand the impact of user persistence in detail, we now take a closer look at the system behavior. To this aim, Fig. 10 shows ongoing session drop probability, π_O, probability of successful service completion with and without retrials, π_S and π_W, and retrial probability, π_{orb}, as a function of blockers density in Fig. 10(a), user persistence factor in Fig. 10(b), and session arrival intensity in Fig. 10(c). Here, $\theta = 0.9$, $1/\alpha = 1$ s, $\lambda = 0.3$ sess./s. We note that the successful service completion with retrials implies that the user has to wait to receive service and also experiences rate degradation which is important QoE parameter.

Starting with blockers density in Fig. 10(a), we observe that the retrial probability increases with λ_B and so does the ongoing session drop probability. This eventually leads to the decrease in both successful service completion probabilities. Due to the very persistent user behavior, $\theta = 0.9$, a large fraction of sessions still successfully completed even for high blockers density of $\lambda_B = 0.3$ bl./m^2. Nevertheless, the users need to be persistent and wait most of the time as the probability of service completion with retrials is 3-4 times higher than that without retrials.

We now proceed with the impact of user persistence factor shown in Fig. 10(b), where $\lambda_B = 0.05$ bl./m^2. Here, we see that as θ increases, the retrial probability also increases and so does the successful session completion probability with retrials. This happens at the expense of the decreased session completion probability without retrials. The rationale is that the higher user persistence factor allows more sessions that experience interruption due to blockage to continue their service, thus, limiting the amount of resources available for new sessions. That is, *persistence of some users forces other users to be persistent too*. From this point of view, persistence provides negative impact on QoE.

Finally, Fig. 10(c) shows the dependence on session arrival intensity λ, were $\theta = 0.9$, $1/\alpha = 1$ s, $\lambda_B = 0.05$ bl./m^2. We see that higher values of λ lead to higher ongoing session drop probability. The retrial probability also increases due to higher on average load as sessions arriving upon retrial find insufficient amount of resources and, thus, require another retrial. However, there are clearly two regimes in the system, the one until $\lambda = 0.3$ and another one after that point. In underloaded system with persistent users, the increase in the arrival traffic improves session completion probability with retrials at the expense of session completion probability without retrials. However, higher overall load negatively affects both considered session completion probabilities.

D. System-Centric Performance Measures

We now proceed with system-centric performance measures – the system resource utilization and fraction of wasted resources shown in Fig. 11 as a function of user persistence factor θ and arrival traffic intensity λ. Recall, that the latter metric is unique for systems, where session service may not complete successfully. By analyzing these data, where $\lambda = 0.3$ sess./s., $1/\alpha = 1$ s, one may observe that *the user persistence positively affects both metrics allowing to utilize more system resources and do that more efficiently*. Specifically, for $\lambda = 0.1$ bl./m^2 and $\theta = 0$, around 40% of resources out of 55% occupied are wasted. When $\theta = 0.9$, not only the system utilizes 80% of resources but only 15% of those are spent for sessions that will not be served successfully. These gains hold for other blocker densities. The difference is that for smaller blockers density the resource utilization is generally higher by $10-15\%$.

E. The Impact of Radio Part

Finally, we observe the impact of radio part parameters, specifically, the number of planar antenna elements on the user- and system-centric performance measures. Fig. 12 shows these metrics for different values of user persistence factor, $\lambda = 0.3$ sess./s., $\lambda_B = 0.1$ bl./m^2, $1/\alpha = 1$ s. In both graphs, we see a very interesting behavior – the metrics plateau as the number of planar antenna elements increases. The rationale is that the increase in the number of antenna elements leads to higher gains and thus better SINR at the UEs. This eventually decreases the mean value and variance of the resource request pmfs as discussed in [46]. Thus, for a given coverage radius, there is an upper bound of the number of utilized elements that may affect user- and system-centric performance measures.

By analyzing Fig. 12(a) further, we observe that the retrial probability decreases as a result of decreased resource requirements. As a result, successful service completion probability without retrials increases by two times, from around 0.25 to 0.5. The associated increase in the completion probability with retrials is much milder as most of the sessions are served without retrials. This implies that *using large arrays at the BS side, is an effective way to improve QoE*. Similar observations are made with respect to the system resource utilization and fraction of wasted resources shown in Fig. 12(b).

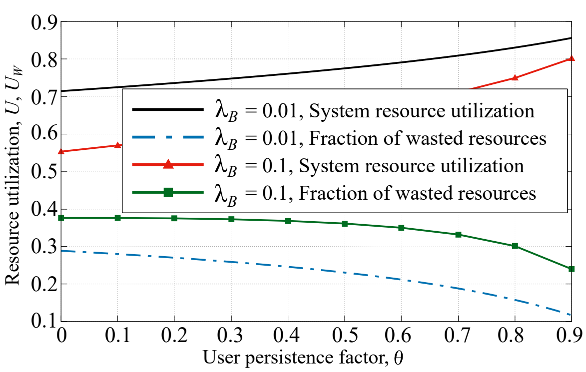
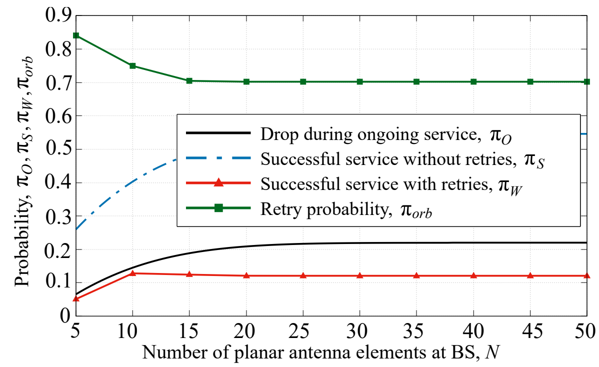
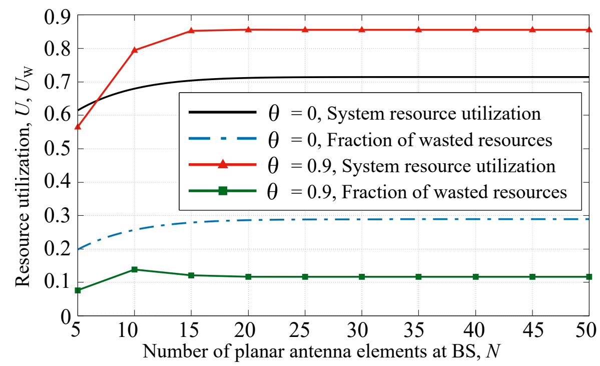


Fig. 11. System utilization as a function of user persistence factor.



(a) Drop probabilities



(b) Resource utilization

Fig. 12. The impact of radio part parameters.

VI. Conclusions

In this paper, motivated by the inherently unreliable nature of mmWave/sub-THz transmission media that are prone to connection interruptions caused by blockage, we proposed a model of service process at BS with persistent user behavior. The model also captures specifics of mmWave/sub-THz propagation conditions, intermittent connectivity, session rate adaptation, and resource allocation at BS. We then characterized user- and system-centric performance measures including not only new and ongoing session drop probabilities and system resource utilization but the amount of wasted resources and successfully session completion probability with and without retrials. The comparison with computer simulations show that the model captures qualitative behavior of the metrics, while quantitatively, the modeling errors are limited to 5-7%.

Our numerical results show that the user persistence allows to improve the new session and ongoing session drop probability: making two retrials brings both down by $20-70\%$ while for fully persistent users the gain increases by two-three times. The effect of persistence is more evident in underloaded system, where the blockage is the only impairment affecting session continuity. The achieved gains are independent of retrial times, i.e., the time user spends prior to session re-establishment. However, higher user persistence not only decreases the ongoing session drop probability but also decreases successful session completion probability without retrials. This can be considered as a negative impact.

In general, despite persistence is stressful for the user and directly leads to QoE degradation, it is an important behavior that may positively affect the system's performance. Specifically, user persistence not only increases the usage of system resources by approximately $20-40\%$ but allows to use those efficiently decreasing the percentage of wasted resources by two/three times. We also identified that utilizing large arrays at BS having high gain in transmit direction improves successful session completion probability without retrials by two 100% and more depending on the system parameters.

References

[1] R. Vannithamby and S. Talwar, *Towards 5G: Applications, requirements and candidate technologies*. John Wiley & Sons, 2017.

[2] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan, and M. Zorzi, "Toward 6G networks: Use cases and technologies," *IEEE Communications Magazine*, vol. 58, no. 3, pp. 55–61, 2020.

[3] V. Begishev, E. Sopin, D. Moltchanov, R. Kovalchukov, A. Samuylov, S. Andreev, Y. Koucheryavy, and K. Samouylov, "Joint use of guard capacity and multiconnectivity for improved session continuity in millimeter-wave 5g nr systems," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 3, pp. 2657–2672, 2021.

[4] E. Sopin, D. Moltchanov, A. Daraseliya, Y. Koucheryavy, and Y. Gaidamaka, "User association and multi-connectivity strategies in joint terahertz and millimeter wave 6G systems," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 12, pp. 12765–12781, 2022.

[5] 3GPP, "NR; Multi-connectivity; Stage 2 (Release 16)," 3GPP TS 37.340 V16.0.0, 3GPP, December 2019.

[6] Y. Sadovaya, D. Moltchanov, W. Mao, O. Orhan, S.-p. Yeh, H. Nikopour, S. Talwar, and S. Andreev, "Integrated access and backhaul in millimeter-wave cellular: Benefits and challenges," *IEEE Communications Magazine*, vol. 60, no. 9, pp. 81–86, 2022.

[7] I. V. Reddy, D. Bodet, A. Singh, V. Petrov, C. Liberale, and J. M. Jornet, "Ultrabroadband terahertz-band communications with self-healing bessel beams," *Communications Engineering*, vol. 2, no. 1, p. 70, 2023.

[8] C. Liu, I. Bouazizi, and M. Gabbouj, "Rate adaptation for adaptive HTTP streaming," in *Proceedings of the second annual ACM conference on Multimedia systems*, pp. 169–174, 2011.

[9] 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.

[10] 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 Trans. on Wir. Comm.*, vol. 18, no. 2, pp. 879–896, 2018.

[11] 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*, vol. 66, no. 11, pp. 10124–10138, 2017.

[12] V. Raghavan, V. Podshivalov, J. Hulten, M. A. Tassoudji, A. Sampath, O. H. Koymen, and J. Li, "Spatio-temporal impact of hand and body blockage for millimeter-wave user equipment design at 28 GHz," *IEEE Communications Magazine*, vol. 56, no. 12, pp. 46–52, 2018.

[13] W. Qi, J. Huang, J. Sun, Y. Tan, C.-X. Wang, and X. Ge, "Measurements and modeling of human blockage effects for multiple millimeter wave bands," in *2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC)*, pp. 1604–1609, IEEE, 2017.

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

[15] G. R. MacCartney, S. Deng, S. Sun, and T. S. Rappaport, "Millimeter-wave human blockage at 73 ghz with a simple double knife-edge diffraction model and extension for directional antennas," in *2016 IEEE 84th Vehicular Technology Conference (VTC-Fall)*, pp. 1–6, IEEE, 2016.

[16] A. Shurakov, D. Moltchanov, A. Prikhodko, A. Khakimov, E. Mokrov, V. Begishev, I. Belikov, Y. Koucheryavy, and G. Gol'tsman, "Empirical blockage characterization and detection in indoor sub-thz communications," *Computer Communications*, vol. 201, pp. 48–58, 2023.

[17] A. Shurakov, P. Rozhkova, A. Khakimov, E. Mokrov, A. Prikhodko, V. Begishev, Y. Koucheryavy, M. Komarov, and G. Gol'tsman, "Dynamic blockage in indoor reflection-aided sub-terahertz wireless communications," *IEEE Access*, vol. 11, pp. 134677–134689, 2023.

[18] A. Shafie, N. Yang, S. Durrani, X. Zhou, C. Han, and M. Juntti, "Coverage analysis for 3D terahertz communication systems," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 6, pp. 1817–1832, 2021.

[19] D. Serghiou, M. Khalily, T. W. Brown, and R. Tafazolli, "Terahertz channel propagation phenomena, measurement techniques and modeling for 6g wireless communication applications: A survey, open challenges and future research directions," *IEEE Comm. Surveys & Tutorials*, 2022.

[20] J. Ye, S. Dang, G. Ma, O. Amin, B. Shihada, and M.-S. Alouini, "On outage performance of terahertz wireless communication systems," *IEEE Transactions on Communications*, vol. 70, no. 1, pp. 649–663, 2021.

[21] M. Polese, R. Jana, and M. Zorzi, "TCP and MP-TCP in 5G mmWave networks," *IEEE Internet Computing*, vol. 21, no. 5, pp. 12–19, 2017.

[22] M. Zhang, M. Polese, M. Mezzavilla, J. Zhu, S. Rangan, S. Panwar, and M. Zorzi, "Will TCP work in mmWave 5G cellular networks?," *IEEE Communications Magazine*, vol. 57, no. 1, pp. 65–71, 2019.

[23] M. Polese, M. Mezzavilla, M. Zhang, J. Zhu, S. Rangan, S. Panwar, and M. Zorzi, "MilliProxy: A TCP proxy architecture for 5G mmWave cellular systems," in *2017 51st Asilomar Conference on Signals, Systems, and Computers*, pp. 951–957, IEEE, 2017.

[24] M. Pieska and A. Kassler, "TCP performance over 5G mmWave links—Tradeoff between capacity and latency," in *2017 IEEE 13th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, pp. 385–394, IEEE, 2017.

[25] D. A. Hayes, D. Ros, and O. Alay, "On the importance of TCP splitting proxies for future 5G mmWave communications," in *IEEE 44th LCN Symposium on Emerging Topics in Networking (LCN Symposium)*, pp. 108–116, 2019.

[26] D. Moltchanov, A. Ometov, P. Kustarev, O. Evsutin, J. Hosek, and Y. Koucheryavy, "Analytical TCP Model for Millimeter-Wave 5G NR Systems in Dynamic Human Body Blockage Environment," *Sensors*, vol. 20, no. 14, p. 3880, 2020.

[27] T. Azzino, M. Drago, M. Polese, A. Zanella, and M. Zorzi, "X-TCP: a cross layer approach for TCP uplink flows in mmwave networks," in *2017 16th Annual Mediterranean Ad Hoc Networking Workshop (Med-Hoc-Net)*, pp. 1–6, IEEE, 2017.

[28] W. Na, B. Bae, S. Cho, and N. Kim, "DL-TCP: Deep learning-based transmission control protocol for disaster 5G mmWave networks," *IEEE Access*, vol. 7, pp. 145134–145144, 2019.

[29] V. Begishev, E. Sopin, D. Moltchanov, R. Pirmagomedov, A. Samuylov, S. Andreev, Y. Koucheryavy, and K. Samouylov, "Performance analysis of multi-band microwave and millimeter-wave operation in 5G NR systems," *IEEE Transactions on Wireless Communications*, vol. 20, no. 6, pp. 3475–3490, 2021.

[30] V. Beschastnyi, D. Ostrikova, D. Moltchanov, Y. Gaidamaka, Y. Koucheryavy, and K. Samouylov, "Balancing Latency and Energy Efficiency in mmWave 5G NR Systems with Multiconnectivity," *IEEE Communications Letters*, 2022.

[31] A. Daraseliya, M. Korshykov, E. Sopin, D. Moltchanov, Y. Koucheryavy, and K. Samouylov, "Handling overflow traffic in millimeter wave 5g nr deployments using nr-u technology," in *2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 1–7, IEEE, 2020.

[32] P. Nain, D. Towsley, B. Liu, and Z. Liu, "Properties of random direction models," in *Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies.*, vol. 3, pp. 1897–1907, IEEE, 2005.

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

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

[35] 3GPP, "Study on channel model for frequencies from 0.5 to 100 GHz (Release 14)," 3GPP TR 38.901 V14.1.1, 3GPP, July 2017.

[36] 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*, vol. 19, pp. 1513–1527, October 2011.

[37] A. B. Constantine *et al.*, "Antenna theory: analysis and design," *Microstrip Antennas, John Wiley & Sons*, 2005.

[38] V. Begishev, D. Moltchanov, E. Sopin, A. Samuylov, S. Andreev, Y. Koucheryavy, and K. Samouylov, "Quantifying the impact of guard capacity on session continuity in 3GPP new radio systems," *IEEE Trans. on Veh. Tech.*, vol. 68, no. 12, pp. 12345–12359, 2019.

[39] D. Moltchanov, E. Sopin, V. Begishev, A. Samuylov, Y. Koucheryavy, and K. Samouylov, “A Tutorial on Mathematical Modeling of 5G/6G Millimeter Wave and Terahertz Cellular Systems,” *IEEE Communications Surveys and Tutorials*, vol. 24, no. 2, pp. 1072–1116, 2022.

[40] J. P. Milaszewicz, “Improving jacobi and gauss-seidel iterations,” *Linear Algebra and Its Applications*, vol. 93, pp. 161–170, 1987.

[41] A. Maslov, “Source code.” https://github.com/AlexKeeper1506/ScienceCode. Accessed: 2024-5-26.

[42] D. Moltchanov, “Distance distributions in random networks,” *Ad Hoc Networks*, vol. 10, no. 6, pp. 1146–1166, 2012.

[43] 3GPP, “NR; Physical Channels and Modulation (Release 15),” 3GPP TR 38.211, 3GPP, Dec 2017.

[44] 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 Veh. Tech. Mag.*, vol. 11, no. 3, pp. 38–48, 2016.

[45] S. Redner, *A guide to first-passage processes*. Cambridge university press, 2001.

[46] R. Kovalchukov, D. Moltchanov, Y. Gaidamaka, and E. Bobrikova, “An accurate approximation of resource request distributions in millimeter wave 3GPP New Radio systems,” in *Internet of Things, Smart Spaces, and Next Generation Networks and Systems*, pp. 572–585, Springer, 2019.