IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 4 July 2023, accepted 9 August 2023, date of publication 21 August 2023, date of current version 29 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3307021

RESEARCH ARTICLE

Packet Level Performance of 5G NR System Under Blockage and Micromobility Impairments

EMIL KHAYROV AND YEVGENI KOUCHERYAVY
Higher School of Economics, National Research University, 101000 Moscow, Russia

Corresponding author: Emil Khayrov (ekhayrov@hse.ru)

This work was supported by the Grant for Research Centers in the field of AI provided by the Analytical Center for the Government of the Russian Federation (ACRF) in accordance with the agreement on the provision of subsidies (identifier of the agreement 000000D730321P5Q0002) and the agreement with HSE University No. 70-2021-00139.

ABSTRACT With the introduction of the Integrated Access and Backhaul (IAB) technology by 3GPP, the last mile in beyond 5G cellular systems will be characterized by multiple wireless backhaul links between the user equipment and base station. In such multi-hop configurations, packet latency becomes a dominant factor characterizing the quality of service (QoS) delivered to the users. At the same time, directional communications at millimeter wave (mmWave) frequencies are prone to frequent outages caused by blockage and micromobility leading to periods of silence and subsequent rate compensation. In this paper, we are interested in assessing the impact of mmWave propagation specifics on the delay and medium access control (MAC) buffer overflow performance at a 5G NR IAB network node. To this aim, we first propose a continuous-time Markov chain model for the bitrate provided by such a system that explicitly accounts for mmWave-specific propagation. Then, by formulating a queuing model of packet transmission, we characterize the mean packet delay experienced by a traffic source. Our results illustrate that the mean delay and MAC buffer overflow probability at the NR BS interface is a step function that is not strictly decreasing due to the packet encapsulation specifics of the modulation and coding (MCS) adaptation mechanism. Further, these metrics are non-negligible even for the system having a sufficient amount of resources. For IAB architecture, where the access link may not be available the mean delay increases exponentially even when the arrival rate is adjusted to the available resources.

INDEX TERMS 5G, blockage, IAB, micromobility, millimeter wave, packet model, queuing theory.

I. INTRODUCTION

Nowadays, as the standardization of the 5G New Radio (NR) technology operating in both millimeter wave (mmWave) and microwave (μWave) is over, the network operators start to deploy these systems. However, the rollouts of mmWave-based 5G NR technology are hampered by unique propagation properties limiting the coverage area of base stations (BS) to just a few hundreds of meters.

The mmWave band poses unique challenges for reliable system design. In addition to extremely high path losses, the blockage phenomenon, which has been already studied in detail in literature [1], [2], makes these communications subject to frequent temporal loss of connectivity. Micromobility effects, occurring as a result of fast displacement or rotation of user equipment (UE) in the hands of a user, have also been shown to produce frequent outages in systems with highly directional antennas [3], [4]. In spite, recently, methods of machine learning (ML) and artificial intelligence (AI) have been proposed to predict blockage and micromobility events, they may still severely affect the data transmission process. Thus, efficient utilization of these technologies naturally requires particularly dense deployments to ensure uninterrupted service when advanced functionalities such as inter and intra multi-connectivity are utilized [5], [6].

To reduce capital expenditures of network operators, 3GPP has recently proposed the Integrated Access and Backhaul (IAB) technology as a key enabler for beyond 5G cellular systems [7], [8]. The IAB architecture relies upon the utilization of relay nodes called IAB nodes to deliver user traffic to/from

The associate editor coordinating the review of this manuscript and approving it for publication was Bilal Khawaja.

the BS via wireless multi-hop communication. In such systems, in addition to conventional performance metrics such as spectral efficiency and the fraction of time in the outage, the packet latency becomes of crucial interest. To dimension IAB systems, new performance evaluation models at the packet level are needed.

Even though 5G and beyond systems are inherently packet-switched, most available performance evaluation models are concerned with data sessions and operate at the corresponding relatively large time scales, see Section II for a review. Specifically, little is known about the packet-level delay characteristics at the mmWave air interface, where packet transmission performance may be affected not only by application specifics (e.g., bursty traffic) but also by wireless channel conditions. For example, outages of various duration due to blockage or micromobility may result in backlogs at UEs requiring a rate compensation when the channel becomes available. This may lead to significant bitrate variability at the air interface even when the application generates a constant bitrate pattern. Rate variability is known to drastically affect the delay performance of traffic transmission towards BS [9], [10], [11]. Thus, understanding the packet-level performance of traffic transmission in mmWave systems is crucial for accurate system dimensioning for beyond 5G communications.

In this paper, we characterize the mean packet delay and medium access control (MAC) buffer overflow probability at the mmWave 5G NR air interface by explicitly accounting for blockage and outage periods caused by micromobility impairments inherent for directional communications in these bands. In doing so, we rely on the propagation properties and utilize the continuous-time Markov chain (CTMC) framework. Then, by applying the apparatus of the queuing theory we develop a performance evaluation model of the packet transmission process at an mmWave IAB node and employ this model for assessing packet delay and other performance metrics. To generalize our model we will also account for IAB specifics.

The main contributions of our study are:

- a mathematical model to investigate the realistic IP packet delay and MAC buffer overflow probability at the 3GPP NR radio interface according to specifics of the protocol parameters and IAB architecture;
- the mean delay and MAC buffer overflow probability are not monotonously decreasing functions if the channel quality and are non-negligible even for the system having a sufficient amount of resources;
- in IAB-enabled systems, where the access link may not always be available, the mean delay increases exponentially with the linear reduction of allocated transmission time even when the traffic rate is adjusted accordingly.

The rest of the paper is organized as follows. First, in Section II we overview the few studies endeavoring to characterize beyond 5G systems' performance at the packet level and the corresponding models. Our system model is then introduced in Section III. In Section IV, we develop the packet service rate model. The performance evaluation model is then developed in Section V. Numerical results are provided and discussed in Section VI. Conclusions are drawn in the last section.

II. RELATED WORK

The mmWave band offers a large bandwidth allowing to achieve the access bitrates required for the 5G enhanced Mobile Broadband (eMBB) services. However, the use of this band is associated with a number of challenges including high propagation losses and the blockage and micromobility phenomena. Whereas the former can be partially compensated by utilizing massive antenna arrays with high transmit and receive gains, the latter two can only be partially compensated and may lead to drastic signal-to-interference-plus-noise (SINR) degradation or even outages.

The dynamic blockage phenomenon has been fairly well investigated in the literature. Specifically, the authors of [2] consider an outdoor scenario where blockers move according to the random direction model (RDM, [12]), while users are stationary. They have utilized the renewal theory to derive the cumulative distribution functions (CDF) of the blocked and non-blocked conditions' duration. An inverse model where blockers are stationary while the user is moving is developed in [13], and CDF of blocked and non-blocked periods are also reported. Finally, in [14] a model where blockers and users are both mobile is addressed. The authors have demonstrated that only the mean values of blocked and non-blocked intervals can be obtained.

The micromobility effect consists of quick rotation of the UE about the transverse and longitudinal axes and manifests itself in the frequent loss of beam alignment between UE and BS [15]. The authors of [4] propose to capture this phenomenon by utilizing a two-dimensional Brownian motion process. Statistical properties of micromobility processes for different applications including phone calling, gaming, virtual reality, and video watching have been reported in [3]. The authors of [16] compare several candidate models and demonstrate that a two-dimensional Markov model provides the best accuracy.

By utilizing dynamic blockage and micromobility models, several different deployment scenarios have been evaluated. A number of studies assessed the impact of inter- and intra-RAT (radio access technology) dual- and multi-connectivity functionality in the presence of dynamic blockage, see [6] and [17]. In general, most authors agree that multi-connectivity alone cannot guarantee outage-free service for UEs. As for the micromobility, the authors of [4] have derived the capacity of the channel under micromobility impairments showing that on-demand realignment leads to better spectral efficiency. The joint impact of dynamic blockage and micromobility is investigated in [18], where the authors notice that for the practical antenna configuration switching time, only periodic beam alignment can provide outage-free performance.

The studies above mainly focus on session-level performance characteristics such as spectral efficiency, capacity, outage probability, etc. However, in multi-hop IAB topologies, latency becomes another critical metric of interest. Inherently, this metric is defined at the network layer characterizing the time a packet spends in the network. The authors of [19] analyze the performance of a mmWave backhaul link by utilizing the framework of discrete-time queuing systems with correlated arrival flows. Among other conclusions, they highlight that the correlation in the input flow may severely affect the mean and CDF of the latency experienced by a packet.

Only a few packet layer models are available for mmWave systems with blockage impairments [20], [21]. In [20], the human blockage is predicted by detecting changes in the received signal power when the blockage is about to occur. This method allows the scheduler to adapt resource allocation and mitigate losses before the blockage happens. The authors show that the performance gains of the proposed scheduler as compared to the conventional ones (e.g., proportional fair (PF) and max-min) increase when the arrival rate and duration of blockage are higher. The study in [21] proposes further modifications to the proportional fair scheduler that not only improve the priority of UEs in blockage conditions but also reduce the response time to any kind of fluctuations, primarily blockages. This improvement is achieved by introducing an exponential scaling that depends on the modulation and coding scheme (MCS) to the classification of UE weights [22].

Summarizing, we observe that the session-level performance of mmWave 5G NR systems with blockage and micromobility impairments is fairly well investigated so far. Contrarily, only little is known with respect to the packet delay performance in these systems. In the following sections, we will proceed to develop a model characterizing the mean delay experienced by packets traversing the mmWave 5G NR air interface. To generalize the model we will also account for IAB specifics.

III. SYSTEM MODEL

In this section, we specify our system model by defining its components, namely the deployment, propagation, blockage, micromobility, antenna, and beam-searching models.

A. DEPLOYMENT MODEL

We consider a single mmWave 5G NR BS mounted at height h_A and having a circular coverage area of radius r, see Fig. 1. We assume that UEs are randomly and uniformly distributed in the coverage area and concentrate on a single UE at a distance x from the BS. The UE height is assumed constant and denoted by h_U. We consider the uplink direction.

We also note that an IAB node can be considered in place of the BS. Moreover, the packet transmission process model formulated and solved in Section V captures the case of communication between UE and an IAB node or donor operating in the half-duplex regime.

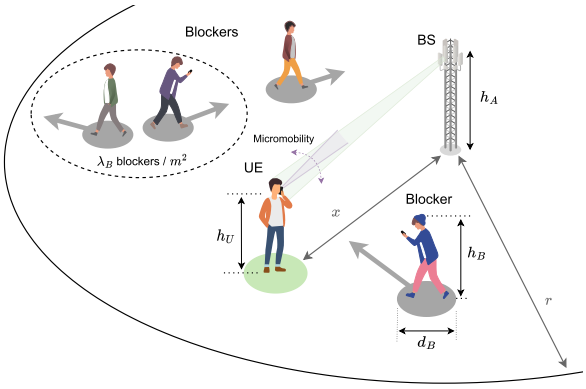


FIGURE 1. The considered 5G mmWave NR deployment.

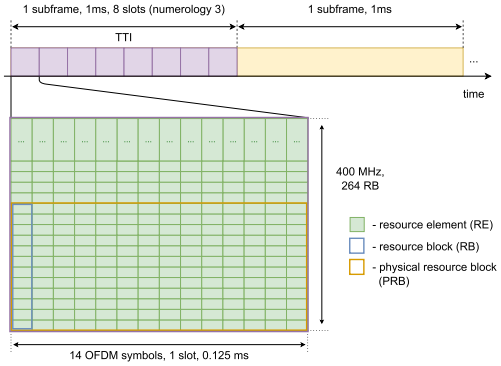


FIGURE 2. The OFDMA "resource grid" of the 5G NR radio interface.

B. NR RADIO INTERFACE

The physical layer of the 5G NR interface defines frames that are 10 ms long, each containing 10 sub-frames of 1 ms in duration. Depending on numerology, each sub-frame is divided into several slots. The number of slots in numerology n is 2^n, with the duration of $1/2^n$ ms each [23].

By allocating resources to a UE, the BS schedules available time slots and a range of frequencies during which the UE can transmit data. This distribution over time and frequency forms the so-called OFDMA (Orthogonal Frequency-Division Multiple Access) "resource grid" shown in Fig. 2. Here, the X-axis represents time, where each slot is also divided into 14 OFDM symbols. The Y-axis represents the entire available frequency channel ($\leq$ 400 MHz) utilized by the BS.

After initializing the connection, the UE and the BS acquire the channel quality indicator (CQI). Knowing the CQI of a specific UE, the BS allocates the required amount of resources to provide the requested bitrate. Then, the UE starts passing IP packets from different applications through 3GPP's UE-side stack of protocols: SDAP (Service Data Adaptation Protocol), PDCP (Packet Data Compression Protocol), and RLC (Radio Link Protocol). Each of them provides different instructions for different types of traffic, which forms a logical channel, called radio bearer.

TABLE 1. CQI, MCS and SINR mapping for 5G NR.

CQI	Modulation	Q_m	Code rate, R_c	SINR in dB
0	outage			
1	QPSK	2	78/1024	-9.478
2	QPSK	2	120/1024	-6.658
3	QPSK	2	193/1024	-4.098
4	QPSK	2	308/1024	-1.798
5	QPSK	2	449/1024	0.399
6	QPSK	2	602/1024	2.424
7	16QAM	4	378/1024	4.489
8	16QAM	4	490/1024	6.367
9	16QAM	4	616/1024	8.456
10	64QAM	6	466/1024	10.266
11	64QAM	6	567/1024	12.218
12	64QAM	6	666/1024	14.122
13	64QAM	6	772/1024	15.849
14	64QAM	6	873/1024	17.786
15	64QAM	6	948/1024	19.809

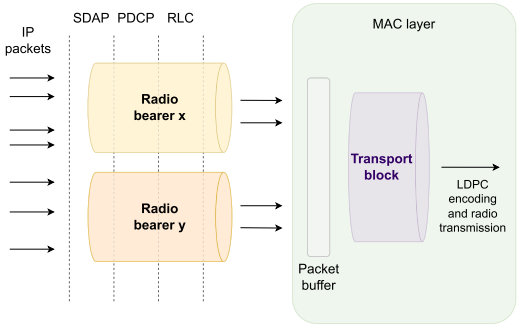


FIGURE 3. Packet transmission process in 3GPP 5G NR.

After SDAP/PDCP/RLC, the packet is dispatched to the MAC layer to be included in a transport block for further transmission. Each transport block contains several packets to be transmitted at the radio interface. Although the transport block's size is constant for some allocated number of resource blocks, the number of packets therein may be different depending on the channel state. In our paper, we assume that the average packet size S_P is 512 bytes, which is typical for video applications. Occasional SINR degradation entails different MCSs to preserve the block error rate (BLER) at the default 10 %. Since the use of different MCSs leads to different rules for processing and duplicating bits of information, we define the useful size of the transport block, S_{TB}, as the number of bits allocated to carry data packets per ms. This parameter depends on the SINR, say S, through the corresponding MCS and more specifically the code rate R_C and the modulation order Q_m (see Table 1), and also on the total number of physical resource blocks (PRB) allocated for transmission, N_{PRB}. Thus, $S_{TB} = S_{TB}(S, N_{PRB})$.

We assume the number of PRBs, N_{PRB}, to be chosen so that the value $1000 \times S_{TB}$ [bps] meets the constant bitrate requirement. Let $D_{[bps]}$ represent the demanded bitrate, and $N^{\max_{PRB}}$ denote the maximum available number of PRBs. Then, the required number of PRBs to provide D is given by

$$\tilde{N}_{PRB}(S, D) = \min\{n \in \mathbb{N} : S_{TB}(S, n) \times 10^3 \geq D\}, \quad (1)$$

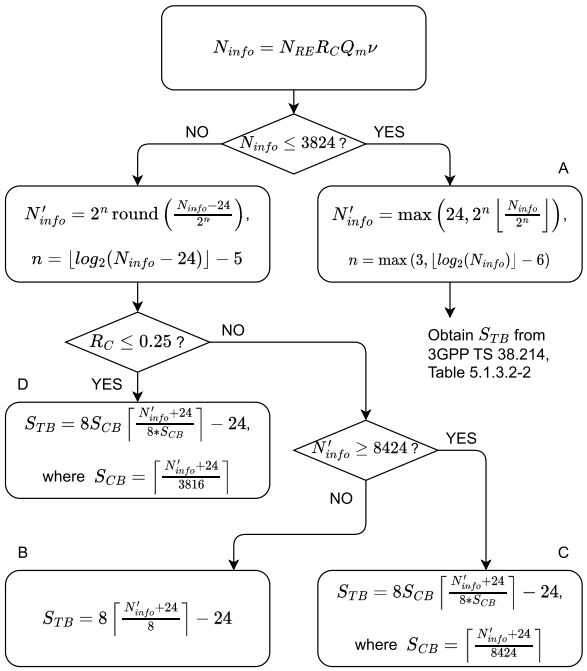


FIGURE 4. The diagram to determine the transport block size S_{TB}.

and the provided number of PRBs is $\min\{\tilde{N}_{PRB}(S, D), N^{\max}_{PRB}\}$.

The algorithm for determining the transport block size (TBS) S_{TB} is specified in 3GPP TR 38.211 [24]. First, the maximum data rate supported by the channel is computed as

$$N_{info} = R_C Q_m N_{RE} \nu, \quad (2)$$

where ν is the number of MIMO layers and

$$N_{RE} = \min\{156, N_{sc} N_{symb} - N_{DMRS} - N_{oh}\} N_{PRB}, \quad (3)$$

is the total number of resource elements (REs) allocated for PUSCH, where $N_{sc} = 12$ is the number of subcarriers in a PRB, N_{symb} is the number of scheduled OFDM symbols within the slot, N_{DMRS} is the number of REs for DM-RS per PRB, and N_{oh}, which we assume 0.

Once N_{info} is computed, S_{TB} is determined according to the diagram in Fig 4, or directly from Table 5.1.3.2-2 of [24] if N_{info} is relatively small.

Note that in the considered protocol stack buffering may happen at multiple layers. In this work, similarly to [25] and [26], we consider packet processing at the MAC layer, just prior to the actual transmission over the physical medium.

C. BLOCKAGE MODEL

Pedestrians act as blockers for mmWave propagation. To capture dynamic human blockage we represent humans by cylinders of diameter d_B and height h_B corresponding to an average person's height. Note that $h_U < h_B < h_A$. Blockers are assumed to move in $\mathbb{R}^2$ according to the RDM. In RDM, a blocker first chooses a direction of movement uniformly in $[0, 2\pi)$ and then moves in this direction for a time exponentially distributed with parameter α_B at a constant

speed v_B. The density of blockers is assumed constant, λ_B bl/m^2.

Depending on the current link state (blockage/non-blockage) and the distance between the BS and UE, an appropriate MCS is utilized to maintain reliable data transmission with 10% BLER. We assume that the BS attempts to maintain the requested bitrate of the UE on average. If a blockage occurs, more PRBs are allocated to the UE (if available) to compensate for the degraded MCS.

D. PROPAGATION MODEL

The received SINR at the UE can be written as

$$S^{(c)}(x) = \frac{P_{\mathrm{A}} G_{\mathrm{A}} G_{\mathrm{U}}}{L_{PL}^{(c)}(x) N_0 R_b M_{\mathrm{SF}} L_a}, \tag{4}$$

where P_{A} is the transmit power, G_{A} and G_{U} are the antenna gains at the BS and the UE, $L_{PL}(x)$ represents the propagation losses at a distance x, N_0 is the thermal noise, R_b is the physical resource block size, $L_a = N_F + C_L + M_I$ is the aggregated losses' coefficient, which accumulates the losses of interference, noise figure and cable losses, c defines the channel state, which is non-blocked (nb) or blocked (b), and M_{SF} is the slow fading margin given by [27]

$$M_{\mathrm{SF[dB]}} = \sqrt{2}\mathrm{erfc}^{-1}(2p_{\mathrm{out}})\sigma_{\mathrm{SF}}, \tag{5}$$

where $p_{\mathrm{out}} = 0.05$ is the target fraction of time in outage conditions at the cell boundary, $\mathrm{erfc}^{-1}(\cdot)$ is the inverse complementary error function, and σ_{SF} is the standard deviation of shadow fading in dB.

Following 3GPP [28], the path loss measured in dB is

$$\begin{aligned} L_{PL}^{(nb)}(x) &= L^{(NB)} + 10\zeta \log(\sqrt{x^2 + \Delta h^2}) + 20 \log f_c, \\ L_{PL}^{(b)}(x) &= L^{(B)} + 10\zeta \log(\sqrt{x^2 + \Delta h^2}) + 20 \log f_c, \end{aligned} \tag{6}$$

where $L^{(NB)} = 32.4$ dB and $L^{(B)} = 52.4$ dB are the human blockage attenuation coefficients with no-blockage and blockage correspondingly, f_c is the carrier frequency in GHz, x is the 2D distance between BS and UE measured in meters, $\zeta = 2.1$ is the path loss coefficient, and $\Delta h = h_A - h_U$.

Given the propagation and antenna models, we can define the effective maximum BS coverage radius r where a blockage does not lead to an outage as [29]

$$r = \sqrt{\left(\frac{P_{\mathrm{A}} G_{\mathrm{A}} G_{\mathrm{U}}}{L^{(B)} f_c^2 N_0 R_b S_{\mathrm{th}} M_{\mathrm{SF}} L_a}\right)^{2/\zeta} - \Delta h^2}, \tag{7}$$

where S_{th} is the SINR threshold.

E. MICROMOBILITY AND BEAM ALIGNMENT MODELS

In addition to blockage, the link between UE and BS is subject to micromobility impairments [3], [15], [30]. By micromobility we understand quick rotations of the UE in the hands of the user. This phenomenon leads to repeated loss of connectivity due to beam misalignment even when the UE is in good propagation conditions. To capture micromobility specifics we utilize the model from [4], which provides the probability density function (pdf) of the time to an outage. We denote the mean time to an outage obtained from this pdf by β.

We account for delays caused by the beam alignment (BA) procedure and assume the on-demand exhaustive search procedure, according to which one side of a communication link is put in omnidirectional mode, while the other switches between configurations searching for the best one. Then the roles are reversed. Note however that the model developed further in this work may account for other beam alignment procedures, e.g., hierarchical beam alignment.

Micromobility always results in an outage. The average time the UE spends in outage conditions is denoted by T_{BA}, and once the link is up again the BS compensates for the lost connectivity by increasing the available bitrate by a factor of 2 during, on average, the same time period T_{BA}.

F. ANTENNA MODEL

Planar antenna arrays are assumed at both the UE and BS. To model the radiation patterns, similarly to [31], [32], and [33], we utilize cone models with constant gain over the main lobe. Following [34], the half-power beamwidth (HPBW) of the main lobe is approximated by $\alpha = 2|\theta_m - \theta_{3db}^{\pm}|$, where $\theta_{3db}^{\pm}$ are the upper and low bounds of the HPBW, and θ_m is the lobe's maximum. The latter parameter is calculated as $\theta_m = \arccos(-1/\pi)$, while the former can be estimated as $\theta_{3db}^{\pm} = \arccos[\pm 2.782/(N_{(\cdot)}\pi)]$.

The mean gain over the main lobe is given by

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

where $N_{(\cdot)}$ is the number of rows or columns [34].

G. TRAFFIC MODEL

In this work, we assume that the applications utilizing the NR radio interface generate packets according to the homogeneous Poisson process with rate λ packets/ms. This model corresponds to the ON state of the ftp3 model defined in 3GPP specifications. We define the parameter λ based on the available bitrate D and the packet size S_p in bits as $\lambda = D/S_p$.

Packets generated by applications are handled as explained in Section III-B.

H. METRICS OF INTEREST

For the described model, we are primarily interested in the mean latency W experienced by a packet. We define packet latency as the time from its arrival to the MAC layer's buffer to the end of its successful transmission. To evaluate W, we first formalize the packet rate model in Section IV and the corresponding packet transmission model in Section V, where we solve the latter for the metric of interest.

IV. PACKET SERVICE RATE MODEL

In this section, to reflect variations in the useful transport block's size, S_{TB}, we develop a tractable packet service rate model that accounts for blockage and micromobility impairments. The model comprises a governing continuous-time

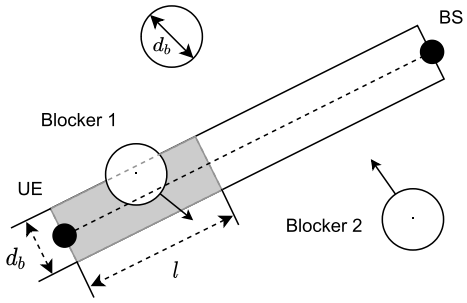


FIGURE 5. Blockers occlude the line-of-sight when entering the gray rectangle.

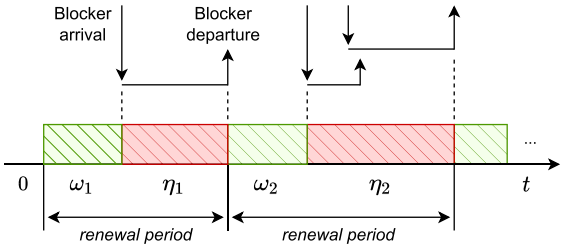


FIGURE 6. An illustration of the blockage renewal process.

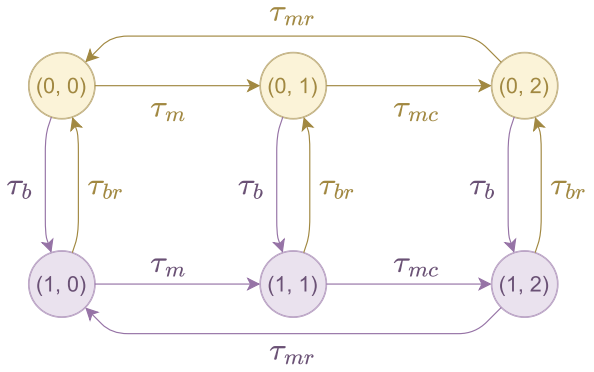


FIGURE 7. Transition diagram for the TBS model.

TABLE 2. Description of states and TBS values.

State	TBS	Description
(0,0)	$S_{TB}(S^{(nb)}, \min\{\tilde{N}^{(nb)}_{PRB}, N^{\max}_{PRB}\})$	Non-blocked, connected
(0,1)	0	Non-blocked, BA outage
(0,2)	$S_{TB}(S^{(nb)}, \min\{2\tilde{N}^{(nb)}_{PRB}, N^{\max}_{PRB}\})$	Non-blocked, BA compens.
(1,0)	$S_{TB}(S^{(b)}, \min\{\tilde{N}^{(b)}_{PRB}, N^{\max}_{PRB}\})$	Blocked, connected
(1,1)	0	Blocked, BA outage
(1,2)	$S_{TB}(S^{(b)}, \min\{2\tilde{N}^{(b)}_{PRB}, N^{\max}_{PRB}\})$	Blocked, BA compens.

Markov chain (CTMC), each state of which corresponds to certain channel conditions (blockage, micromobility), and a mapping to the values of the S_{TB} parameter.

A. BLOCKAGE PROCESS

We define the blockage area as a rectangle of width d_b and length $l = x(h_B - h_U)/(h_A - h_U)$ (see Fig. 5), where x is the 2D distance between UE and BS. A communication channel is considered blocked if a blocker enters the rectangle and therefore occludes the communication link. We follow [35] and model the blockage process as a renewal process with two alternating states: non-blocked and blocked (see Fig. 6).

We denote the sojourn time in the non-blocked state by ω. The authors in [35] demonstrated that it follows the exponential distribution with mean

$$\mu_\omega(x) = \gamma_B^{-1}(x), \tag{9}$$

where x is the 2D distance between BS and UE, $\gamma_B(x)$ is the rate at which blockers enter the blockage area associated with the UE given by [36]

$$\gamma_B(x) = \frac{2}{5}\lambda_B v_B \left(2\, d_B + 2\, x \frac{h_B - h_U}{h_A - h_U}\right). \tag{10}$$

The sojourn time in the blocked state, which we denote by η, corresponds to the time between a blocker's arrival into an empty blockage area and the departure of the last blocker among those who entered the area while blockage was occurring. the distribution of this time coincides with the busy period in M/G/∞ queuing system [35] that is available in the integral form only [37]. By utilizing M/M/∞ approximation, the mean blockage period is given by

$$\mu_\eta(x) = \frac{e^{\gamma_B(x) v_B / d_B} - 1}{\gamma_B(x)}. \tag{11}$$

B. THE GOVERNING CTMC

Let $\psi_1(t)$ equal 1 if at time $t > 0$ the UE is in the blockage state and 0 otherwise. Let further $\psi_2(t)$ equal 1 if at time $t > 0$ the UE is in a BA-related outage, 2 if BA-related compensation occurs, and 0 otherwise. Now, consider a bivariate stochastic process $\{\psi(t) = (\psi_1(t), \psi_2(t)), t > 0\}$ defined over a state space $\Psi = \{0, 1\} \times \{0, 1, 2\}$. In view of the assumptions made previously and also further in this subsection, the process is a CTMC by construction.

The transition diagram of $\{\psi(t), t > 0\}$ with the corresponding notation is depicted in Fig. 7. The transition intensities are obtained as follows. Let τ_b be the transition intensity from non-blocked to blocked state. It is given by $\tau_b = 1/\mu_\omega$, where $\mu_\omega = \mu_\omega(x)$ is the average time in the non-blocked state at a given distance $0 < x \leq r$. Similarly, the reverse transition from blocked to non-blocked state is provided by $\tau_{br} = 1/\mu_\eta$, where $\mu_\eta = \mu_\eta(x)$ is the average time in the non-blocked state.

Since micromobility events are independent of the current blockage process state, the transition intensities in both cases are the same and we have $\tau_m = 1/\beta$. We recall that the mean duration of the beam alignment procedure is T_{BA} and assume this value as the mean of the exponentially distributed sojourn times in the corresponding states. Thus, the transition intensities from the outage state to the compensation state and from compensation to blocked/non-blocked states are given by $\tau_{mc} = \tau_{mr} = 1/T_{BA}$.

Let the states of $\{\psi(t), t > 0\}$ be ordered lexicographically and $\mathbf{I}$ denote the identity matrix. Then the infinitesimal generator of $\{\psi(t), t > 0\}$ is of the block form

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_0 - \mu_\omega^{-1}\mathbf{I} & \mu_\omega^{-1}\mathbf{I} \\ \mu_\eta^{-1}\mathbf{I} & \mathbf{A}_0 - \mu_\eta^{-1}\mathbf{I} \end{pmatrix} \tag{12}$$

with

$$\mathbf{A}_0 = \begin{pmatrix} -\beta^{-1} & \beta^{-1} & 0 \\ 0 & -T_{BA}^{-1} & T_{BA}^{-1} \\ T_{BA}^{-1} & 0 & -T_{BA}^{-1} \end{pmatrix}. \tag{13}$$

The row vector $\mathbf{q} = (q_\psi)_{\psi\in\Psi}$ representing the stationary distribution of $\{\psi(t), t > 0\}$, i.e,

$$q_\psi = lim_{t\to\infty}\mathbb{P}\{\psi(t) = \psi\}, \ \psi \in \Psi. \tag{14}$$

can now be found from

$$\begin{aligned} \mathbf{qA} &= \mathbf{0}, \\ \mathbf{q1} &= 0, \end{aligned} \tag{15}$$

where $\mathbf{0}$ and $\mathbf{1}$ denote a row vector of zeros and a column vector of ones, respectively.

C. TBS VALUES

The TBS, S_{TB}, is determined as described in Section III-B with the SINR value $S^{(c)}(x)$, $c \in \{b, nb\}$, at a given distance $0 < x \le r$ and the number of PRBs given by (1). Let us denote the TBS when there is no blockage by

$$\gamma_{(0,0)} = S_{TB}(S^{(nb)}(x), \min\{\tilde{N}_{PRB}(S^{(nb)}(x), D), N_{PRB}^{\max}\}). \tag{16}$$

In what follows, whenever possible, we omit parameters x and D for brevity. Also, we denote $\tilde{N}_{PRB}^{(c)} = \tilde{N}_{PRB}(S^{(c)})$.

Now, to account for micromobility, we define $\gamma_{(0,1)} = 0$ as the TBS when the UE and BS perform beam alignment and

$$\gamma_{(0,2)} = S_{TB}(S^{(nb)}, \min\{2\tilde{N}_{PRB}^{(nb)}, N_{PRB}^{\max}\}) \tag{17}$$

as the TBS when an outage caused by beam misalignment is compensated during the non-blockage time.

In the blocked state we have

$$\gamma_{(1,0)} = S_{TB}(S^{(b)}, \min\{\tilde{N}_{PRB}^{(b)}, N_{PRB}^{\max}\}). \tag{18}$$

Note that micromobility can also happen in this state. We define the TBS in the beam alignment and its compensation states as follows

$$\gamma_{(1,1)} = 0, \ \gamma_{(1,2)} = S_{TB}(S^{(b)}, \min\{2\tilde{N}_{PRB}^{(b)}, N_{PRB}^{\max}\}). \tag{19}$$

The CTMC $\{\psi(t), t > 0\}$ along with the mapping γ_ψ, $\psi \in \Psi$, represents the packet service rate model. The states of the process and the corresponding TBS are summarized in Table 2. The mean TBS at a given distance $0 < x \le r$ is

$$\mu_{S_{TB}} = \sum_{\psi\in\Psi} \gamma_\psi q_\psi. \tag{20}$$

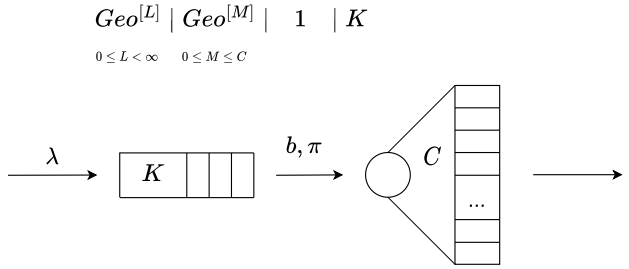


FIGURE 8. Queuing system model.

V. PACKET TRANSMISSION MODEL

In this section, we develop a queuing model of the packet transmission process at the 5G NR radio interface. We start with a formalization, then proceed to derive the steady-state distribution of the resulting discrete-time Markov chain (DTMC), and finally determine the metrics of interest.

A. MODEL FORMALIZATION

We model the packet transmission process at the MAC layer as a single-server queuing system with batch arrivals and batch service, see Fig. 8. The system has a queue of a finite size K representing the limited buffer space at the MAC layer [25], [26]. The service discipline is first come first served (FCFS). Packets are served in batches whose size is bounded by the server's capacity $0 < C \le K$, which corresponds to the mean useful size of the transport block, specifically

$$C = \lceil S_p^{-1}\mu_{S_{TB}} \rceil. \tag{21}$$

Although the server's capacity is defined, the server does not provide additional space to accommodate the packets, and the system's capacity is determined by the queue's size K. A packet scheduled for service keeps its place in the queue and vacates it only upon service completion. Served packets depart the queue and the system simultaneously.

The system time is assumed discrete and indexed by $n \in \{0, 1, 2, \dots\}$, with the time slot duration corresponding to the scheduling time interval in 5G NR, namely 1 ms. Changes in the system (arrivals and service completions) occur at the time slot boundaries as shown in Fig. 9.

According to the assumption made in Section III-G and the properties of the Poisson arrival process, the number of packets arriving in each time slot follows Poisson law, i.e., the probability that the arriving batch is of size $i \ge 0$ is

$$l_i = \frac{\lambda^i}{i!}e^{-\lambda}, \tag{22}$$

where λ is the packet arrival rate.

The service/transmission time of a batch of packets is assumed to be a single time slot. However, we account for imperfect transmission with non-zero BLER and hybrid automatic repeat request procedure (HARQ). To this aim, we consider the probability of re-transmission denoted by π. Accordingly, $1 - \pi$ is the probability of the correct reception

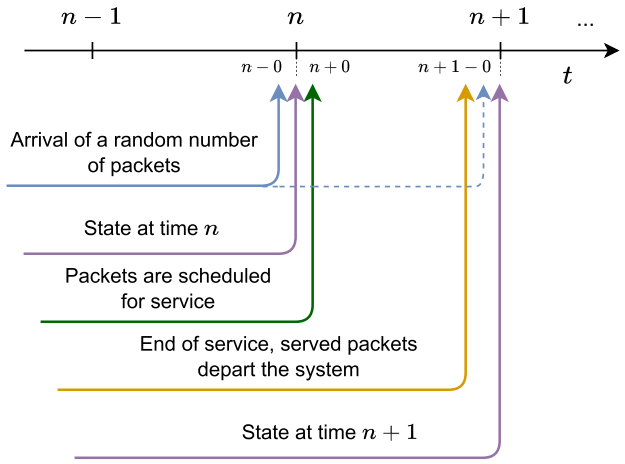


FIGURE 9. Timing of events.

of a packet. We also assume that packet transmissions are independent of each other, i.e., any packet is incorrectly received with probability π independently of other packets in the same or other batches.

The number of packets scheduled for transmission in each time slot depends on the number of packets in the queue: if they are fewer than C, only those in the queue are scheduled. Among the packets scheduled for service, only those transmitted correctly with probability $1-\pi$ are considered served and depart the queue (and the system) by the end of the time slot. The packets transmitted incorrectly remain in the queue to be rescheduled for transmission in the next time slot.

Finally, to capture the IAB specifics, we assume that at the end of each time slot, the scheduled batch can be served and depart the system with a probability b. In other words, b is the probability that the server is available in the time slot. This parameter stems from a resource allocation algorithm implemented at the IAB donor and $b = 1$ in the stand-alone BS scenario.

Thus, a given packet is served and departs the system in the current time slot if (a) it is scheduled for transmission in this time slot (the number of packets in the queue before the given packet is smaller than C), (b) the server is available, and (c) the packet is transmitted correctly. Denote by $\{m_{j,i}, 0 \le i \le C\}$ the conditional probability distribution of the number of packets served in a time slot given that the server is available and there are j packets in the queue, $0 \le j \le K$. Observe that the number of served packets can be zero if the system is empty at the beginning of the time slot or if all the scheduled packets have been received incorrectly. The distribution $\{m_{j,i}, 0 \le i \le C\}$ is binomial with parameters $\min\{j, C\}$ and $1-\pi$, i.e.,

$$m_{j,i} = \binom{\min(j,\,C)}{i}(1-\pi)^i \pi^{\min(j,C)-i},\ 0 \le i \le C. \quad (23)$$

B. STEADY-STATE DISTRIBUTION

Let X_n represent the number of packets in the system at time n. Since the next state of the system depends only on the current state and current arrivals and departures, the process $\{X_n, n = 0, 1, \dots\}$ is a Markov chain with a finite state space $\mathcal{X} = \{0, 1, 2, \dots, K\}$ where all states communicate.

The general structure of its transition matrix is

$$P = \begin{pmatrix} p_{0,0} & p_{0,1} & \cdots & p_{0,C} & \cdots & p_{0,K} \\ p_{1,0} & p_{1,1} & \cdots & p_{1,C} & \cdots & p_{1,K} \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ p_{C,0} & p_{C,1} & \cdots & p_{C,C} & \cdots & p_{C,K} \\ 0 & p_{C+1,1} & \cdots & p_{C+1,C} & \cdots & p_{C+1,K} \\ \vdots & & & \vdots & & \vdots \\ 0 & 0 & \dots & \dots & \dots & p_{K,K} \end{pmatrix}. \quad (24)$$

Having the distributions $\{l_i, i \le 0\}$ and $\{m_{j,i}, 0 \le i \le C\}$, $0 \le j \le K$, we can obtain the entries of the transition matrix. First of all, we observe that

$$p_{0,0} = l_0 \quad (25)$$

is the probability that there will be no packets in the system at time $n+1$ if there are no packets at time n, i.e., the probability to remain in state 0. The probability of moving from state 0 to state $1 \le j \le K-1$ is given by

$$p_{0,j} = l_j, \quad (26)$$

whereas the probability to transit from 0 to K is

$$p_{0,K} = \sum_{j=K}^{\infty} l_j = 1 - \sum_{j=0}^{K-1} l_j. \quad (27)$$

The reverse transitions are provided by

$$p_{i,0} = l_0 bm_{i,i},\ i = \overline{1, C}. \quad (28)$$

Since there is an upper bound of the number of packets that can be served in a time slot, C, the following probabilities are zeros:

$$\begin{aligned} &p_{i,0} = 0,\ \ C < i \le K, \\ &p_{i,j} = 0,\ \ C + j < i \le K,\ 0 \le j < K - C. \end{aligned} \quad (29)$$

The rest of the transition probabilities are

$$\begin{aligned} p_{j,0} &= l_0 bm_{j,j} + \sum_{i=j+1}^{C} l_{i-j} bm_{j,i},\ 1 \le j < C, \\ p_{i,i} &= l_0 \bar{b} + l_0 bm_{i,0} + \sum_{k=1}^{\min(i,C)} l_k bm_{i,k},\ 1 \le i < K, \\ p_{i,j} &= l_0 bm_{i,1} + \sum_{k=2}^{\min(i,C)} l_{k-1} bm_{i,k},\ i = j+1, j > 0, \\ p_{i,j} &= l_0 bm_{i,i-j} + \sum_{k=i-j+1}^{\min(i,C)} l_{j-i+k} bm_{i,k},\ j+1 \le i \le j+C, \\ p_{i,j} &= l_{j-i}\bar{b} + l_{j-i} bm_{i,0} + \sum_{k=1}^{\min(i,C)} l_{j-i+k} bm_{i,k},\ i < j < K, \\ p_{i,K} &= 1 - \sum_{j=0}^{K-1} p_{i,j},\ 0 \le i \le K. \end{aligned} \quad (30)$$

In (30), $\overline{b} = 1 - b$ and the term $l_0\overline{b}$ captures the event that there have been no arrivals and no service completions due to the server's unavailability in the time slot. The term $l_0 b m_{i,0}$ corresponds to the event that there have been no arrivals and no packets in the scheduled batch have been transmitted correctly. Finally, the term $l_j b m_{i,k}$ gives the probability that j packets have arrived and k packets have been transmitted correctly and departed the system out of i originally in the queue.

Now that we have the transition matrix (24), the steady-state probabilities p_i that there are i packets in the system, $0 \le i \le K$, can be obtained from the equilibrium equations with the normalization condition, i.e., from

$$\begin{cases} p_0 = p_{0,0}p_0 + p_{1,0}p_1 + \ldots + p_{C,0}p_C, \\ p_1 = p_{0,1}p_0 + p_{1,1}p_1 + \ldots + p_{C+1,1}p_{C+1}, \\ p_2 = p_{0,2}p_0 + p_{1,2}p_1 + \ldots + p_{C+2,2}p_{C+2}, \\ \ldots \\ p_K = p_{0,K}p_0 + p_{1,K}p_1 + \ldots + p_{K,K}p_K, \\ p_0 + p_1 + \ldots + p_K = 1. \end{cases} \tag{31}$$

C. PERFORMANCE MEASURES

The ultimate metric of interest is the mean packet delay, corresponding to the average total time a packet spends in the system. To determine it, we first need to obtain the packet loss probability, B, representing the probability of the MAC buffer overflow. It can be found as the ratio of the expected number of lost packets to the arrival rate. Since packets arrive in batches, we represent this probability as the weighted sum

$$B = \sum_{i=1}^{\infty} \frac{i l_i}{\lambda} B^{(i)} \tag{32}$$

of conditional loss probabilities $B^{(i)}$ given that the packet arrives in a batch of $i > 0$ packets. These are given by

$$\begin{aligned} B^{(1)} &= p_K^*, \\ B^{(2)} &= p_K^* + \frac{1}{2}p_{K-1}^*, \\ B^{(3)} &= p_K^* + \frac{2}{3}p_{K-1}^* + \frac{1}{3}p_{K-2}^*, \\ B^{(k)} &= \sum_{j=0}^{\min(k,K)-1} \frac{k-j}{k} p_{K-j}^*, \ k > 3, \end{aligned} \tag{33}$$

where p_n^* is the probability that an arriving batch sees n packets in the system

$$p_n^* = p_n\overline{b} + b \sum_{j=0}^{\min(C,K-n)} p_{n+j} m_{n+j,j}, \ 0 \le n \le K. \tag{34}$$

Now, to determine the mean packet delay, we apply Little's law in the following form

$$W = \frac{\overline{K}}{\lambda(1-B)}, \tag{35}$$

where $\overline{K} = \sum_{i=1}^{K} i p_i$ is the mean number of packets in system.

TABLE 3. Parameters for numerical analysis.

Notation	Value	Description
f_c	28 GHz	Operating frequency
G_{A}	11.57 dBm	BS antenna gain
G_{U}	5.57 dBm	UE antenna gain
R_b	1.44 MHz	PRB size
N_{symb}	14	Number of scheduled OFDM symbols
N_{DMRS}	12	REs for DM-RS (per PRB)
h_B, h_A, h_U	1.7, 4, 1.5 m	Height of the blockers, BS and UE
C_L	2 dB	Cable losses
M_I	3 dB	Interference margin
N_F	7 dB	Noise figure
N_0	– 174 dBm/Hz	Noise power spectral density
P_{A}	2 W (33 dBm)	Transmitting BS power
$\mathrm{S_{th}}$	– 9.478 dB	SNR threshold
σ_{SF}	8.2 dB	Shadow fading STD
ν	2	Number of MIMO layers
D	20, 30, 50 Mbps	Available bitrate
v_B	1.66 m/s	Blockers' speed
λ_B	0.01 bl/m^2	Default blockers' density
β	1 s	Mean time to outage (micromobility)
b	1	Default IAB access probability
δ	2 μs	Antenna array switching time
N_A	16×4	Number of BS antenna configurations
N_U	4×4	Number of UE antenna configurations
K	28 pkts (14 KB)	Max pkts in queue (MAC buffer size)
r	$\approx$ 137.3 m	Radius of circular coverage area
x	$2r/3 \approx 91.5$ m	Default distance between UE and BS

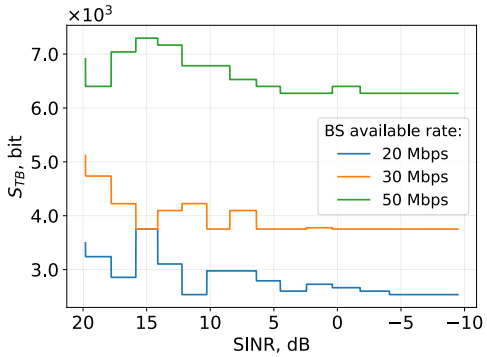
VI. NUMERICAL RESULTS

In this section, we report our numerical results. To better understand the impact of environmental characteristics on the delay performance of the NR air interface, we first assess the dependency on these parameters of the protocol stack's variables which determine the bitrate experienced by the UE. Next, we proceed with the analysis of the packet delay and MAC buffer overflow probability. Finally, the IAB scenario is addressed by quantifying the impact of the fraction of time the access link is operational.

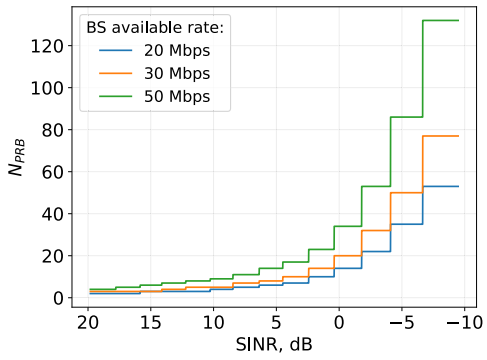
The default system parameter values for the analysis are provided in Table 3. We assume that the UE utilizes an application requiring a high-speed connection with bitrates D of 20, 30, and 50 Mbps. These correspond to packet arrival rates λ of approximately 4.88, 7.32, and 12.21 pkts/ms. The resources required to deliver these bitrates are estimated by accounting for the channel state information as discussed in Section III. Furthermore, we consider the cases of unlimited and limited BS resources. In the latter we assume $N_{PRB}^{\max} = 30$. The value is chosen so that, for a UE at distance $x = 2r/3$ from the BS, resources suffice to provide the three considered bitrates, namely 20, 30, and 50 Mbps, in non-blocked conditions but are not enough to provide $D = 30$ Mbps in the blockage state. Note that we consider the case of high micromobility with mean time to outage of 1 s corresponding to VR and gaming applications [3].

A. PROTOCOL STACK VARIABLES

We start our discussion on the impact of environmental parameters on the protocol stack's variables with Fig. 10 showing the behavior of the useful transport block size, S_{TB}, and the number of allocated PRBs, N_{PRB}, as functions of



(a) S_{TB} as a function of SINR



(b) N_{PRB} as a function of SINR

FIGURE 10. Dependency of S_{TB} and N_{PRB} on SINR.

SINR assuming unlimited BS resources. As one may observe in Fig. 10 there are sharp changes in the values of S_{TB}. These changes implicitly depend on N_{PRB}: with SINR degradation the MCS deteriorates, and the BS increases the amount of allocated PRBs to provide the same target bitrate. Specifically, the SINR values at which S_{TB} drops correspond to such MCS degradation instances where the target bitrate is still supported with the same amount of PRBs. In contrast, rises of S_{TB} indicate an MCS degradation to the point where the target bitrate is no longer satisfied and additional PRBs are allocated. The latter can be seen from Fig. 10(b), which depicts the dependency of the required amount of resources on the SINR. It also demonstrates that in the worst considered channel conditions, the number of PRBs required to maintain the target bitrates is extremely high. Such amounts of PRBs may not be available at the BS, which motivates us, in what follows, to also study the case of limited resources.

The blockers' density λ_B largely determines how much time the UE is in blockage conditions. As λ_B increases from 0 to 0.1, the fraction of time the UE is blocked goes from 0 to approximately 95 %. Fig. 11 further shows the dependency of the average useful transport block size S_{TB} on the blockers' density λ_B in this range. Here, dashed lines represent the system with limited resources and solid lines, with unlimited.

As one may observe in Fig. 11, in the case of unlimited resources S_{TB} does not decrease significantly because the BS timely allocates additional resources to fulfill

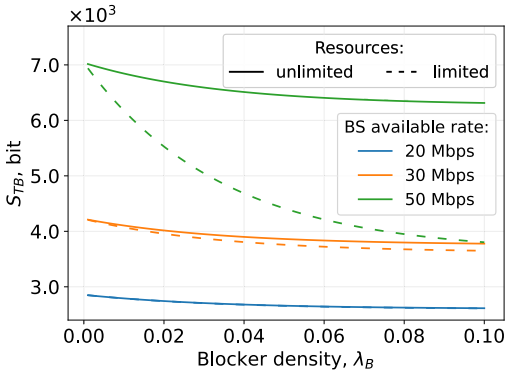


FIGURE 11. Dependency of the average S_{TB} on the blockers' density.

the requirements. A slight decrease in the value of S_{TB}, especially noticeable for $D = 50$ Mbps, is explained by the fact that as the UE goes between non-blocked and blocked states, its SINR switches between approx. 16.13 and −3.86 dB. As it can be seen from Fig. 10, these yield respectively the maximum and the minimum S_{TB} for $D = 50$ Mbps. So, as the probability for the UE to be in the blockage state increases with λ_B, its average S_{TB} decreases from its maximum to its minimum.

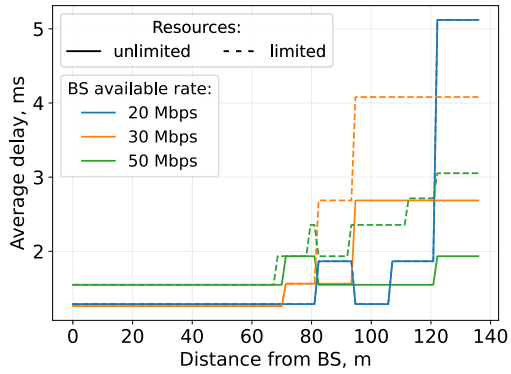
For the system with limited resources, we observe in Fig. 11 that for $D = 50$ Mbps the considered metric drops sharply. The rationale is that for a higher blockers' density, the UE spends more and more time in the blockage state, in which the available number of PRBs is insufficient to maintain the target bitrate. We see that this degradation is insignificant for $D = 30$ Mbps and does not occur at all for $D = 20$ Mbps, which is explained by the choice of $N_{PRB}^{\max}$.

B. DELAY PERFORMANCE OF NR BS

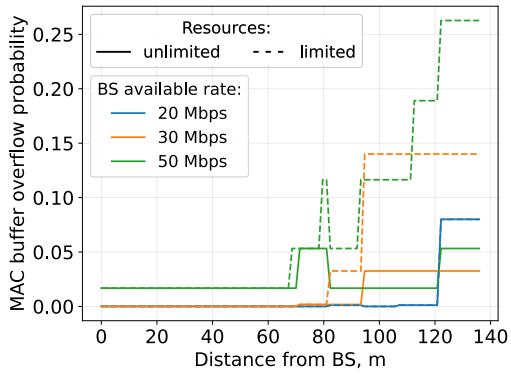
We are now in a position to evaluate the mean delay and the MAC buffer overflow probability specified in Section V-C. The average delay metric, W, measured in ms is calculated according to (35) and the MAC buffer overflow probability, B, by (32).

Fig. 12 depicts the average delay and the MAC buffer overflow probability as a function of the distance x between the UE and the BS. Here, as previously, dashed lines show the metrics for the case of limited resources available at the BS, and solid lines for unlimited. Recall that as the UE moves away from the BS, the SINR decreases, and S_{TB} changes accordingly. If the current amount of allocated PRBs is enough to provide the target bitrate D, there is no need to allocate additional resources. Otherwise, the BS increases the number of allocated PRBs if they are available. We also note that occasional down jumps of the curves correspond to the change of the MCSs.

By analyzing the data presented in Fig. 12, we observe that the mean delay has a qualitatively similar behavior for both limited and unlimited resources. The rationale is that the mechanisms of resource allocation are the same in both

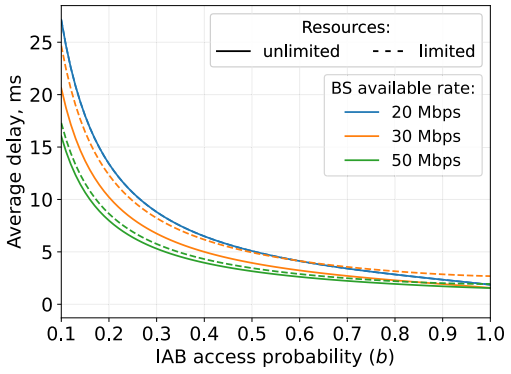


(a) Average packet delay W

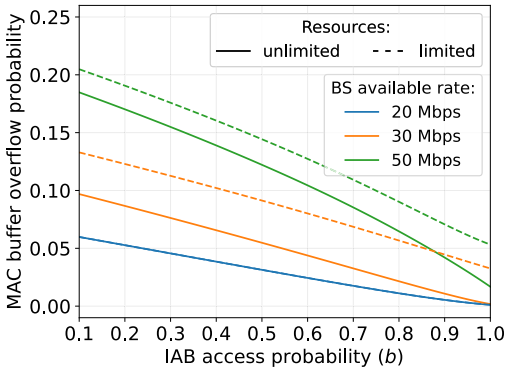


(b) MAC buffer overflow probability B

FIGURE 12. Dependency of the average packet delay, W, and the MAC buffer overflow probability, B, on the distance x between UE and BS.



(a) Average packet delay W



(b) MAC buffer overflow probability B

FIGURE 13. Dependency of the average packet delay, W, and the MAC buffer overflow probability, B, on the IAB access probability b.

cases. Still, the delays experienced in unlimited resources conditions are well below than those in limited resources conditions at larger distances to the BS. Also, at larger distances, we observe a significant difference in delay for different target bitrates. Further, as could be expected, the limited resources conditions lead to much higher MAC buffer overflow probabilities as depicted in Fig. 12(b). In contrast, in the system with unlimited resources the buffer overflow probability becomes non-negligible only at large distances between the UE and the BS.

As one may further observe for the system with limited resources in Fig. 12, at a distance of fewer than 70 meters, performance is stable as the SINR is always sufficient to ensure the minimum delay. As the distance increases further, performance starts to deteriorate because the MCS changes. At this point, the changes in the value of S_{TB} lead to an increase of both the mean packet delay and the MAC buffer overflow probability for the target bitrates of 30 and 50 Mbps. If after recalculating S_{TB}, the target bitrate is no longer provided, then the amount of allocated resources increases. At the same time, as one may observe, there might be a case when increasing the amount of resources to meet the same target bitrate requirements leads to an improvement in mean delay performance. This effect can be seen in Fig. 12 for $D = 50$ Mbps at around 80 meters and for $D = 20$ Mbps at around 90–100 meters.

C. THE IMPACT OF THE IAB ARCHITECTURE

Having gained an understanding of the performance of an NR BS that can dedicate the whole set of its resources to the UE, we complete our analysis by assessing the impact of the IAB architecture. To this end, Fig. 13 presents the average packet delay, W, and the MAC buffer overflow probability, B, as functions of the fraction of time the access link is available, b. Here again, both unlimited and limited resources are considered. We specifically note that the packet arrival rate λ here is also multiplied by b to not overflow the model in standby time slots (i.e., time slots without transmission). In this way, we keep the ratio of the arrival and service rates, i.e., the offered load to the system, constant and equal to that in the case of $b = 1$. We recall that $b = 1$ corresponds to the stand-alone BS scenario considered previously.

As one may observe in Fig. 13, qualitatively, the average packet delay decreases as b becomes higher, i.e., more time is provided to service the access link. We note that the delay increases exponentially as b becomes small reaching 17–25 ms for $b = 0.1$. The negative impact of $b < 1$ is even more pronounced when we look at the MAC buffer overflow probabilities in Fig. 12(b). Here the case of the largest target bitrate of 50 Mbps is the most affected with the overflow probability over 10 % already at $b = 0.6$ even for unlimited resources. Such an effect of discontinuous packet

service should be taken into account when dimensioning in-band IAB systems.

VII. CONCLUSION

Motivated by the lack of packet-level mathematical models for 5G NR systems, in this paper we proposed the models of the packet service rate and of the packet transmission process at the 5G NR air interface. The former model is based on the CTMC framework and captures rate variations caused by packetization specifics of the protocol stack and by the human body blockage and micromobility events. The packet transmission model is designed as a discrete-time queuing system for server vacations and batch arrivals and service. It accounts for erroneous packet transmission and a finite MAC buffer, as well as for specifics of an IAB deployment. We used these models to derive the mean packet delay and the MAC buffer overflow probability.

Our numerical results demonstrate that the transport block size and the amount of allocated PRBs have non-linear dependency on the SINR with jumps associated with MCS changes. This has a noticeable impact on the delay and overflow performance as they are not monotonously decreasing functions of the SINR and have up and down jumps. We further observed that the mean delay is non-negligible even in a system having a sufficient amount of resources. In limited resource conditions, one should provide sufficient resources to handle human body blockage as otherwise the MAC buffer overflows quickly. Finally, investigating the IAB operational regime, where the access link may not always be available at all times, we observed that the mean delay increases exponentially when link availability drops below 50 % even when the arrival traffic rate is decreased accordingly.

In our future work we plan to extend the proposed model to the case of multiple UEs. To this aim, one needs to account for dynamic bandwidth sharing between UEs based on the utilized scheduling strategy. We also plan to investigate the use of ML and AI techniques to improve the scheduling performance of packets at the 5G NR air interface.

REFERENCES

[1] T. Bai, R. Vaze, and R. W. Heath Jr., "Analysis of blockage effects on urban cellular networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 9, pp. 5070–5083, Sep. 2014.

[2] 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 *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–7.

[3] N. Stepanov, D. Moltchanov, V. Begishev, A. Turlikov, and Y. Koucheryavy, "Statistical analysis and modeling of user micromobility for THz cellular communications," *IEEE Trans. Veh. Technol.*, vol. 71, no. 1, pp. 725–738, Jan. 2022.

[4] V. Petrov, D. Moltchanov, Y. Koucheryavy, and J. M. Jornet, "Capacity and outage of terahertz communications with user micro-mobility and beam misalignment," *IEEE Trans. Veh. Technol.*, vol. 69, no. 6, pp. 6822–6827, Jun. 2020.

[5] 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 Trans. Wireless Commun.*, vol. 20, no. 6, pp. 3475–3490, Jun. 2021.

[6] 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 Trans. Veh. Technol.*, vol. 70, no. 3, pp. 2657–2672, Mar. 2021.

[7] M. Cudak, A. Ghosh, A. Ghosh, and J. Andrews, "Integrated access and backhaul: A key enabler for 5G millimeter-wave deployments," *IEEE Commun. Mag.*, vol. 59, no. 4, pp. 88–94, Apr. 2021.

[8] M. Polese, M. Giordani, T. Zugno, A. Roy, S. Goyal, D. Castor, and M. Zorzi, "Integrated access and backhaul in 5G mmWave networks: Potential and challenges," *IEEE Commun. Mag.*, vol. 58, no. 3, pp. 62–68, Mar. 2020.

[9] M. J. Neely, "Delay analysis for maximal scheduling with flow control in wireless networks with bursty traffic," *IEEE/ACM Trans. Netw.*, vol. 17, no. 4, pp. 1146–1159, Aug. 2009.

[10] D. Rajan, A. Sabharwal, and B. Aazhang, "Delay-bounded packet scheduling of bursty traffic over wireless channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 1, pp. 125–144, Jan. 2004.

[11] M. Wang, J. Liu, W. Chen, and A. Ephremides, "Joint queue-aware and channel-aware delay optimal scheduling of arbitrarily bursty traffic over multi-state time-varying channels," *IEEE Trans. Commun.*, vol. 67, no. 1, pp. 503–517, Jan. 2019.

[12] P. Nain, D. Towsley, B. Liu, and Z. Liu, "Properties of random direction models," in *Proc. IEEE 24th Annu. Joint Conf. IEEE Comput. Commun. Soc.*, vol. 3, Mar. 2005, pp. 1897–1907.

[13] 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 Trans. Mobile Comput.*, vol. 19, no. 9, pp. 2221–2233, Sep. 2020.

[14] D. Moltchanov, A. Ometov, and Y. Koucheryavy, "Analytical characterization of the blockage process in 3GPP new radio systems with trilateral mobility and multi-connectivity," *Comput. Commun.*, vol. 146, pp. 110–120, Oct. 2019.

[15] R. Singh and D. Sicker, "Parameter modeling for small-scale mobility in indoor THz communication," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.

[16] N. Stepanov, A. Turlikov, V. Begishev, Y. Koucheryavy, and D. Moltchanov, "Accuracy assessment of user micromobility models for THz cellular systems," in *Proc. 5th ACM Workshop Millim.-Wave Terahertz Netw. Sens. Syst.*, pp. 37–42, 2021.

[17] M. Polese, M. Giordani, M. Mezzavilla, S. Rangan, and M. Zorzi, "Improved handover through dual connectivity in 5G mmWave mobile networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 2069–2084, Sep. 2017.

[18] D. Moltchanov, Y. Gaidamaka, D. Ostrikova, V. Beschastnyi, Y. Koucheryavy, and K. Samouylov, "Ergodic outage and capacity of terahertz systems under micromobility and blockage impairments," *IEEE Trans. Wireless Commun.*, vol. 21, no. 5, pp. 3024–3039, May 2022.

[19] A. Ometov, D. Moltchanov, M. Komarov, S. V. Volvenko, and Y. Koucheryavy, "Packet level performance assessment of mmWave backhauling technology for 3GPP NR systems," *IEEE Access*, vol. 7, pp. 9860–9871, 2019.

[20] F. Firyaguna, A. Bonfante, J. Kibilda, and N. Marchetti, "Performance evaluation of scheduling in 5G-mmWave networks under human blockage," 2020, *arXiv:2007.13112*.

[21] J. Ma, A. Aijaz, and M. Beach, "Recent results on proportional fair scheduling for mmWave-based industrial wireless networks," in *Proc. IEEE 92nd Veh. Technol. Conf. (VTC-Fall)*, Nov. 2020, pp. 1–5.

[22] F. Capozzi, G. Piro, L. A. Grieco, G. Boggia, and P. Camarda, "Downlink packet scheduling in LTE cellular networks: Key design issues and a survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 2, pp. 678–700, 2nd Quart., 2013.

[23] *NR; Physical Channels and Modulation (Release 17)*, document TR 38.211, 3GPP, Sep. 2022.

[24] *NR; Physical Layer Procedures for Data (Release 17)*, document TS 38.214 V17.3.0, 3GPP, Sep. 2022.

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

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

[27] R. Kovalchukov, D. Moltchanov, Y. Gaidamaka, and E. Bobrikova, "An accurate approximation of resource request distributions in millimeter wave 3GPP new radio systems," in *Proc. 19th Int. Conf. Next Gener. Wired/Wireless Netw. (NEWAN), 12th Conf. ruSMART*, St. Petersburg, Russia. Cham, Switzerland: Springer, Aug. 2019, pp. 572–585.
[28] *Study on Channel Model for Frequencies From 0.5 to 100 GHz (Release 14)*, document TR 38.901 V14.1.1, 3GPP, Jul. 2017.
[29] A. Samuylov, D. Moltchanov, R. Kovalchukov, R. Pirmagomedov, Y. Gaidamaka, S. Andreev, Y. Koucheryavy, and K. Samouylov, "Characterizing resource allocation trade-offs in 5G NR serving multicast and unicast traffic," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3421–3434, May 2020.
[30] V. Petrov, D. Moltchanov, Y. Koucheryavy, and J. M. Jornet, "The effect of small-scale mobility on terahertz band communications," in *Proc. 5th ACM Int. Conf. Nanosc. Comput. Commun.*, Sep. 2018, pp. 1–2.
[31] 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 Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1791–1808, Mar. 2017.
[32] 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 Commun. Surveys Tuts.*, vol. 24, no. 2, pp. 1072–1116, 2nd Quart., 2022.
[33] 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 Trans. Netw.*, vol. 19, no. 5, pp. 1513–1527, Oct. 2011.
[34] C. A. Balanis, *Antenna Theory: Analysis and Design*, 4th ed. Hoboken, NJ, USA: Wiley, 2016.
[35] 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 Trans. Veh. Technol.*, vol. 66, no. 11, pp. 10124–10138, Nov. 2017.
[36] 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 J. Sel. Areas Commun.*, vol. 36, no. 11, pp. 2486–2496, Nov. 2018.
[37] D. J. Daley, "The busy period of the M/GI/∞ queue," *Queueing Syst.*, vol. 38, no. 2, pp. 195–204, 2001.

EMIL KHAYROV received the B.Sc. degree in applied mathematics and informatics and the M.Sc. degree in fundamental informatics and information technologies from RUDN University, Moscow, Russia, in 2019 and 2021, respectively. He is currently pursuing the Ph.D. degree in electronics, radio engineering, and communication systems with the Higher School of Economics (HSE University).

His research interests include performance analysis in 5G+ networks, which includes packet traffic analysis in both mmWave and THz, UAV-based scenarios in mmWave networks, MAC protocols for the Industrial Internet of Things in THz, and neural networks for image recognition.

YEVGENI KOUCHERYAVY received the Ph.D. degree from the Tampere University of Technology, Tampere, Finland, in 2004.

He is currently a Professor with the Moscow Institute of Electronics and Mathematics, HSE University. He has authored numerous publications in the field of advanced wired and wireless networking and communications. His current research interests include various aspects of heterogeneous wireless communication networks and systems, and emerging communication technologies for digitally augmented future beings.