Characterizing Resource Allocation Trade-offs in 5G NR Serving Multicast and Unicast Traffic

Andrey Samuylov, Dmitri Moltchanov, Roman Kovalchukov, Rustam Pirmagomedov, Yuliya Gaidamaka, Sergey Andreev, *Member, IEEE*, Yevgeni Koucheryavy, *Senior Member, IEEE*, and Konstantin Samouylov

***Abstract*—The use of highly directional antenna radiation patterns for both the access point (AP) and the user equipment (UE) in the emerging millimeter-wave (mmWave)-based New Radio (NR) systems is inherently beneficial for unicast transmissions by providing an extension of the coverage range and eventually resulting in lower required NR AP densities. On the other hand, efficient resource utilization for serving multicast sessions demands narrower antenna directivities, which yields a trade-off between these two types of traffic that eventually affects the system deployment choices. In this work, with the tools from queuing theory and stochastic geometry, we develop an analytical framework capturing both the distance- and traffic-related aspects of the NR AP serving a mixture of multicast and unicast traffic. Our numerical results indicate that the service process of unicast sessions is severely compromised when (i) the fraction of unicast sessions is significant, (ii) the spatial session arrival intensity is high, or (iii) the service time of the multicast sessions is longer than that of the unicast sessions. To balance the multicast and unicast session drop probabilities, an explicit prioritization is required. Furthermore, for a given fraction of multicast sessions, lower antenna directivity at the NR AP characterized by a smaller NR AP inter-site distance (ISD) leads to a better performance in terms of multicast and unicast session drop probabilities. Aiming to increase the ISD, while also maintaining the drop probability at the target level, the serving of multicast sessions is possible over the unicast mechanisms, but it results in worse performance for the practical NR AP antenna configurations. However, this approach may become feasible as arrays with higher numbers of antenna elements begin to be available. Our developed mathematical framework can be employed to estimate the parameters of the NR AP when handling a mixture of multicast and unicast sessions as well as drive a lower bound on the density of the NR APs, which is needed to serve a certain mixture of multicast and unicast traffic types with their target performance requirements.**

I. Introduction

Millimeter-wave (mmWave) radio technology is expected to construct a comprehensive foundation for the fifth generation (5G) of mobile systems by providing extremely high data rates and low latencies at the radio interface [1]. The first and second phases of the mmWave-based New Radio (NR) standardization have been completed by 3GPP as part of Release 15 in December 2017 and June 2018, respectively, by ratifying both LTE-anchored and standalone NR implementations. While the specification of NR systems continues – expected to be finalized by 2020 – the research community is currently focused on enabling more advanced networking functionality for mobile broadband access. One of the crucial directions along these lines is to enable the coexistence of multicast and unicast types of traffic in NR systems having directional antenna radiation patterns [2].

Reliance on multicast sessions in networking systems allows to efficiently utilize the available radio resources by serving multiple user sessions with a single transmission, thus increasing the overall utility of the network. To enable multicast capabilities in cellular systems, such as LTE, where user equipment (UE) devices may experience dissimilar propagation conditions, the access point (AP) may employ the modulation and coding scheme (MCS) associated with the UE that experiences the worst propagation conditions, which decreases the efficiency of multicasting. High directionality of the antenna radiation patterns is considered to be one of the key advantages of the emerging NR systems, by allowing for planar directivity of under 1° with linear arrays of 128×4 or more antenna elements [3], [4].

The effect is in a significant extension of coverage from a single NR AP [5] as well as a possibility to operate closer to the noise-limited mode [6]. While being essential for unicast sessions, the use of extreme antenna directivities may however result in inefficient resource utilization when serving multicast traffic. Particularly, the smaller the half-power beamwidth (HPBW) of the antenna array is, the fewer the number of multicast UE nodes becomes, which can be served simultaneously by a single antenna configuration over a single transmission. Hence, several multicast transmissions disseminating the same content need to be supported at the NR APs, which results in less efficient use of radio resources. Hence, in this work, we concentrate on answering the following questions: (i) whether multicasting needs to still be supported in 3GPP NR systems and, if so, (ii) what are the principal trade-offs associated with serving a mixture of both multicast and unicast sessions at the NR APs?

The problem of multicasting in mmWave systems has been of interest in several recent studies. The authors in [7] addressed the matter of rate adaptation in 60GHz IEEE 802.11ad systems by optimizing delay performance of user sessions.

Yu. Gaidamaka, K. Samouylov are with Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation. Email: {gaydamaka-yuv,samuylov-ke}@rudn.ru

D. Moltchanov, R. Kovalchukov, A. Samuylov, R. Pirmagomedov, S. Andreev, and Y. Koucheryavy are with Tampere University, Korkeakoulunkatu 1, 33720, Tampere, Finland. Email: firstname.lastname@tuni.fi

Yu. Gaidamaka, K. Samouylov are also with Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences (FRC CSC RAS), 44-2 Vavilov St., Moscow, 119333, Russian Federation.

The publication has been prepared with the support of the "RUDN University Program 5-100". The reported study was funded by RFBR, project numbers 19-07-00933 and 20-07-01064, Business Finland 5G-FORCE project, and the Academy of Finland, project RADIANT. This work has been developed within the framework of the COST Action CA15104, Inclusive Radio Communication Networks for 5G and beyond (IRACON).

Their solution is based on a max-min problem formulation and leads to a convex programming problem. In [8], the issue of grouping the UEs based on their proximity has been tackled. The corresponding heuristic algorithm is based on a consecutive testing of different HPBWs that maximize the sum-rate of the system. Among other conclusions, the authors demonstrated that the use of fixed HPBW might lead to non-optimal resource usage. A similar approach was proposed in [9]. The optimization framework developed in [10] not only operates with HPBW but also accounts for unequal power-sharing among beams.

A complex multicast multiplexing scheme based on non-orthogonal multiples access (NOMA) was proposed and analyzed in [11]. However, the utilization of NOMA-based access in NR deployments is still under discussion by 3GPP. Finally, the problem of multicast transmissions in systems with directional antennas was recently addressed by [12]. In that study, the authors proposed and analyzed several transmission schemes that target delay minimization during packet delivery. This literature review indicates that the research community recognizes the challenge of multicasting using directional antennas. However, to the best of our knowledge, none of the works completed so far addressed simultaneous support of both multicast and unicast traffic types in mmWave-based NR layouts. Accordingly, the matter of optimized NR system configuration for serving a mixture of multicast and unicast sessions requires a more detailed investigation.

In this contribution, we characterize the key trade-offs associated with the service process of both multicast and unicast traffic in 5G NR. To achieve this goal, we unify the tools of stochastic geometry and queuing theory by formulating a mathematical framework that captures mmWave propagation, NR system details, and the service features of multicast and unicast types of traffic at the NR AP. Our metrics of interest are related to multicast and unicast session drop probabilities as well as system resource utilization. The proposed model is then used to quantify the trade-offs between the NR AP deployment density and the performance delivered to the considered traffic types under various environmental and system conditions. These useful dependencies are then employed to yield a lower bound on the NR AP densities required to provide the desired performance levels.

The main findings of our work are as follows:

- The service of unicast sessions in terms of their drop probability is severely compromised by the presence of multicast traffic. This effect aggravates when (i) the fraction of unicast sessions increases, (ii) the spatial session arrival intensity grows, or (iii) the ratio between the service times of multicast vs. unicast sessions increases. To balance out the multicast and unicast session drop probabilities, an explicit prioritization scheme at the NR APs is required, e.g., bandwidth reservation or connection admission control.
- For a given proportion of multicast sessions in the spatial session arrival intensity, narrower antenna directivities at the NR APs characterized by smaller inter-site distance (ISD) lead to lower multicast and unicast session drop probabilities. This is due to the need of performing fewer multicast transmissions for disseminating the same content.
- An attempt to expand the ISD by enabling multicasting via the unicast service leads to significantly lower user-level performance in terms of the session drop probability for the practical ranges of the NR AP antenna directivities, i.e., higher than 1°. Reducing the HPBW further by increasing the number of antenna elements that form the NR AP radiation pattern allows to decrease the performance gap between unicast-only and mixed unicast/multicast deployments.

The rest of this text is organized as follows. In Section II, we introduce our system model. We parametrize it and assess this system for the performance metrics of interest in Section III. Our numerical results and discussion are presented in Section IV. Conclusions are drawn in the final section.

II. System Model

In this section, we first clarify and emphasize the key trade-offs involved into serving a mixture of multicast and unicast sessions at the NR APs with directional antenna radiation patterns. We then introduce our system model by formulating its core components including the deployment, propagation, antenna, and multicast vs. unicast traffic models. Finally, we specify the metrics of interest. The main notation employed in this paper is collected in Table I.

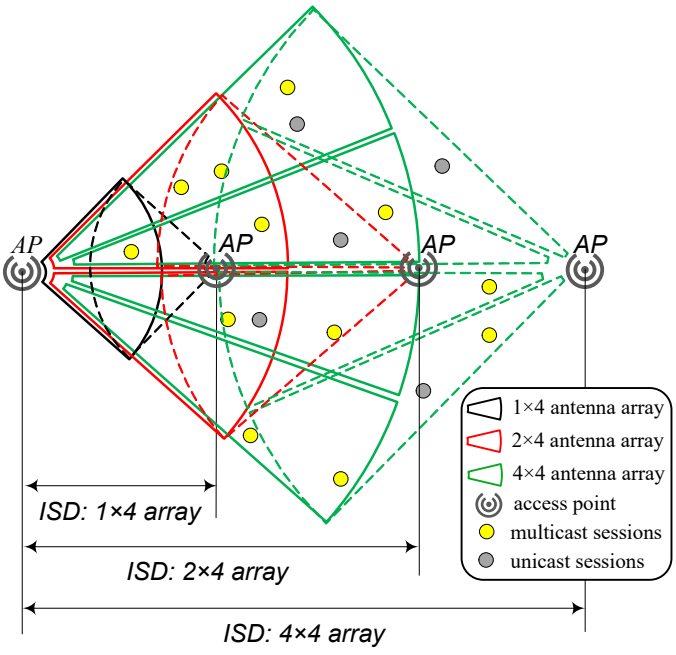


Fig. 1. Trade-offs when serving multicast traffic at NR AP.

A. Problem at a Glance

Consider two NR APs each equipped with three-sector antenna arrays as illustrated in Fig. 1. We concentrate on a certain sector covered by a single antenna array. Assuming a linear antenna array at the NR AP, linear transmit gain can be approximated by the number of antenna elements, N_A, which form the radiation pattern. Further, antenna directivity, α_A,

TABLE I
NOTATION USED IN THIS WORK.

Parameter	Definition
f_c	Carrier frequency, GHz
W	Available bandwidth, MHz
$L(y), L_{dB}(y)$	Path loss in linear and decibel scales
λ_B	Spatial pedestrian UE density, UE/m^2
h_A	Height of NR APs, m
h_U	Height of UEs, m
h_B	Height of blockers, m
r_B	Radius of blockers, m
d_E	Effective coverage range of NR APs, m
x	Two-dimensional distance between UE and NR AP, m
y	Three-dimensional distance between UE and NR AP, m
P_A	Transmit power, W
G_A, G_U	Antenna array gains at NR AP and UE ends, dBi
N_0	Power spectral density of noise, dBi/Hz
A_i, ζ_i, C_i	Propagation coefficients
α_A, α_U	Antenna array directivities at NR AP and UE, rad
N_A, N_U	Number of planar antenna array elements at NR AP and UE
$\theta^{\pm}_{3db}$	Upper and lower 3-dB points of antenna array, $^\circ$
θ_m	Location of array maximum, $^\circ$
β	Antenna array orientation, $^\circ$
$M_{S,nB}, M_{S,B}$	Shadow fading margins in non-blocked and blocked states
M_I	Interference margin
$p_B(x), p_B$	Distance-dependent and independent blockage probabilities
S_B, S_{nB}, S	SNR in LoS blocked/non-blocked states and weighed SNR, dB
S_{area}	Area covered by a single NR AP array configuration, m^2
M	Number of multicast session classes
K	Number of unicast session classes
$R_{M,m}$	Rate of class m multicast sessions, Mbps
$R_{U,k}$	Rate of class k unicast sessions, Mbps
γ_m	Offered load of class m multicast sessions, sess./s
ρ_m	Normalized offered load of class m multicast sessions, sess./m^2
a_k	Offered load of class k unicast sessions, sess./s
$p_k^{(u)}$	Probability that an arriving session is of unicast class k
$p_m^{(m)}$	Probability that an arriving session is of multicast class m
Λ	Session arrival intensity from UE side, sess./s
λ_S	Spatial session arrival intensity from antenna configuration, sess./s
$\lambda_m^{(m)}$	Arrival intensity of class m multicast sessions from UE, sess./s
$\mu_m^{(m)}$	Service intensity of class m multicast sessions, 1/s
$b_k^{(m)}$	Number of PRBs requested by a multicast session of class m
$\lambda_k^{(u)}$	Arrival intensity of class k unicast sessions, sess./s
$\mu_k^{(u)}$	Service intensity of class k unicast sessions, 1/s
$b_k^{(u)}$	Number of PRBs requested by a unicast session of class k
C	Number of servers in queuing system that model NR AP
$\widetilde{\mathcal{Z}}, \mathcal{Z}$	State spaces of infinite and finite systems
I_m	Indicator of class m multicast session in the system
$\pi_m(I_m)$	Stationary state probabilities of multicast sessions
$p_k(n_k)$	Stationary state probabilities of unicast sessions
$\vec{\pi}$	Joint stationary state probability vector
$\tilde{G}(\widetilde{\mathcal{Z}}), G(\mathcal{Z})$	Normalization constants for infinite and finite systems
$h(n), f_m(i,n)$	Auxiliary functions
s_A	Size of PRB, MHz
Δ	Subcarrier spacing, MHz
S_{th}	SNR threshold, dB
ϵ_j	Probability of CQI/MCS j
s_j	SNR thresholds, dB
$q_{M,m}$	Session drop probability of class m multicast sessions
$q_{U,k}$	Session drop probability of class k unicast sessions
u	Mean resource utilization
$F_X(x), f_X(x)$	CDF and pdf of random variable X

can be closely approximated by the HPBW, which is about $102^\circ/N_A$ [13]. As one may deduce, when using fewer antenna elements at the NR AP, the ISD between the NR APs, which ensures no coverage gaps, becomes smaller but the number of transmissions required to support a multicast service reduces.

Conversely, increasing the number of antenna elements decreases the HPBW, thus resulting in higher NR AP transmit gain and ISD distance, which reduces the cost of deployment. However, at the same time, a higher number of transmissions might be needed to serve all of the UEs involved in a multicast service, which results in ineffective utilization of system resources. This trade-off depends on the spatial session arrival intensity, the fraction of multicast sessions, the HPBW of the array, which is further complicated by the presence of unicast sessions. For a given set of system parameters and environmental characteristics, there exists an optimized ISD that yields complete coverage for the area of interest with the target multicast and unicast session drop probabilities.

B. Network Deployment

We concentrate on a tagged NR AP as part of the cellular deployment with a certain density λ_B of pedestrians as illustrated in Fig. 1. Since the methodology developed in what follows does not depend on the assumed coverage of a single NR AP and rather accounts for the fraction of space covered by a single antenna array, one may apply these results to any practical coverage obtained by using, e.g., field measurements.

All of the pedestrians carry their UEs equipped with mmWave NR modules. The heights of the NR AP and the UEs are assumed to be fixed and set to h_A and h_U, respectively. Pedestrians are modeled as cylinders with height h_B and radius r_B. The line-of-sight (LoS) propagation path between the UE and the NR AP might be occluded by pedestrians.

The NR AP has a circular coverage range, which is achieved by using three physical antennas each covering a 120°-sector. We focus on the coverage of a single antenna and define d_E as the effective coverage radius, such that no UEs inside it experience outage conditions when their LoS link is blocked, that is, there is a feasible modulation and coding scheme (MCS) for users at the distance of d_E [14]. The radius d_E is computed in Section III by using the propagation, blockage, and antenna beamforming models as detailed below.

C. Propagation, Interference, Blockage, and Beamforming

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

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

where P_A is the NR AP transmit power, G_A and G_U are the antenna array gains at the NR AP and the UE ends, respectively, N_0 is the power spectral density of noise, W is the operating bandwidth, $L(y)$ is the linear path loss, M_I is the interference margin, and M_S is the shadow fading margin.

We capture any interference from the adjacent NR APs via an interference margin M_I in (1). For a given NR AP deployment density, one may estimate it by employing stochastic geometry based models [15], [16], [17]. Similarly, the effect of shadow fading is accounted for by using the shadow fading margins, $M_{S,B}$ and $M_{S,nB}$, for the LoS blocked and non-blocked states as provided in [18].

Following [18], the path loss measured in dB is given by

$$L_{dB}(y) = \begin{cases} 32.4 + 21\log_{10} y + 20\log_{10} f_c, \text{ non-blocked}, \\ 47.4 + 21\log_{10} y + 20\log_{10} f_c, \text{ blocked}, \end{cases} \quad (2)$$

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

The path loss in the form of (2) can be represented in the linear scale by utilizing the model in the form of $A_i y^{-\zeta_i}$, where A_i and ζ_i are the propagation coefficients. Introducing the coefficients (A_1, ζ_1) and (A_2, ζ_2) that correspond to LoS non-blocked and blocked conditions, we have

$$A_1 = 10^{2\log_{10} f_c + 3.24} M_{S,nB} M_I,\ \zeta_1 = 2.1,$$
$$A_2 = 10^{2\log_{10} f_c + 4.74} M_{S,B} M_I,\ \zeta_2 = 2.1. \quad (3)$$

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

$$S(y) = \frac{P_A G_A G_U}{N_0 W} \left(\frac{y^{-\zeta}}{A_1}[1 - p_B(y)] + \frac{y^{-\zeta}}{A_2} p_B(y) \right), \quad (4)$$

where $p_B(y)$ is the blockage probability at the 3D distance y.

Introducing the coefficients

$$C_i = P_A G_A G_U / (N_0 W A_i),\ i = 1, 2, \quad (5)$$

the propagation model finally reads as

$$S(y) = C_1 y^{-\zeta}[1 - p_B(y)] + C_2 y^{-\zeta} p_B(y). \quad (6)$$

We assume linear antenna arrays at both transmit and receive sides. Following [17], [19], we consider a cone antenna model where the radiation pattern is represented as a conical zone with an angle of α coinciding with the HPBW of the antenna array. Recall that the HPBW of a linear antenna array, α, is proportional to the number of elements in the appropriate plane and is given by [13] as

$$\alpha = 2|\theta_m - \theta_{3db}|, \quad (7)$$

where θ_{3db} is the angle at which the value of the radiated power is 3dB below the maximum and θ_m is the location of the array maximum. The latter is $\theta_m = \arccos(-\beta/\pi)$, where β is the array orientation, i.e., the azimuth angle representing the physical orientation of the array. Note that $\theta_m = \pi/2$ for $\beta = 0$.

The average antenna gain over the HPBW can be found as [13]

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

where the upper and the lower 3-dB points are

$$\theta^{\pm}_{3db} = \arccos[-\beta \pm 2.782/(N\pi)], \quad (9)$$

and N is the number of antenna elements.

In our study, we assume that at most one beam is available in the subject system at a time. Note that for massive antenna arrays there might be more than one beam generated simultaneously, each of which may be steered in a different direction. The HPBW of these beams depends only on the number of the involved antenna elements and not on the total number of elements in the array. Although the developed model can be extended to capture the performance of such systems, it may require further assumptions in the system model.

D. Traffic Patterns and Service Process

Let Λ be the session arrival intensity for a single pedestrian. At an arbitrary instant of time, each of the pedestrians may initiate either a multicast or a unicast session with the corresponding probabilities p_M and p_U, $p_M + p_U = 1$. There are M and K different classes of multicast and unicast sessions: an arriving session belongs to the corresponding class m with the probability of $p^{(m)}_m$, $m = 1, 2, \ldots, M$, and to class k with the probability of $p^{(u)}_k$, $k = 1, 2, \ldots, K$, respectively. We have

$$\sum_{m=1}^{M} p^{(m)}_m = p_M, \quad \sum_{k=1}^{K} p^{(u)}_k = p_U. \quad (10)$$

Employing the superposition property of point processes, we observe that the spatial session arrival process at the NR AP serving a 120°-sector is Poisson with the intensity of $\Lambda\lambda_B\pi d_E^2 2\pi/3$ sessions per time unit [20]. We also define $\lambda^{(m)}_i = p^{(m)}_m \Lambda\lambda_B\pi d_E^2 2\pi/3$ as the spatial arrival intensity of multicast sessions of class m from a single UE and $\lambda^{(u)}_k = p^{(u)}_k \Lambda\lambda_B\pi d_E^2 2\pi/3$ as the spatial arrival intensity of unicast sessions of class k from all the UEs within coverage of the NR AP. Note that the methodology developed in what follows can be applied to any number of antennas used at the AP side by appropriately modifying the multiplier $2\pi/3$ in the spatial session arrival rates.

The choice of the UE that initiates a session is random. Hence, the geometric locations of users associated with a session are distributed uniformly within the NR AP coverage [21]. Class m of multicast and class k of unicast sessions are characterized by the exponentially distributed service times with the parameters $\mu^{(m)}_m$ and $\mu^{(u)}_k$, respectively. The corresponding session data rates are assumed to be constant and equal to $R_{M,m}$ Mbps and $R_{U,k}$ Mbps. The amounts of resources requested by a multicast and a unicast session, $b^{(m)}_m$ and $b^{(u)}_k$, depend on the size of the physical resource block (PRB), s_A, and are computed in Section III. The NR AP is assumed to operate over the bandwidth of W Hz.

A multicast session of class m is initiated by the first arrival of class m session during the so-called "off-period", i.e., the time period with no session of this class residing in the system. This session is accepted by the system if there are sufficient radio resources to serve it, whereas it is dropped otherwise. All of the sessions of class m that observe at least one session of this class currently in the system are accepted without allocating any additional resources. From the resource utilization point of view, the accepted requests overlap with each other [22].

According to the considered service discipline, each session of class m that arrives into the system during the "on-period" may increase its duration. The "on-period" ends when the last request of class m completes its service in the system. This service discipline corresponds to the conventional multicast streaming service. Resource allocation for the unicast sessions is also conventional, that is, each arrival requires a new set of resources. A unicast arrival of class k is dropped if there are under $b^{(u)}_k$ PRBs available. The metrics of interest are: (i) session drop probability of multicast session of class m, $q_{M,m}$,

$$p_B = \int_0^{d_E} p_B(x)\frac{2x}{d_E^2}dx = 1 + \frac{(h_A - h_U)e^{-2\lambda_B r_B^2}\left[e^{\frac{2d_E\lambda_B r_B(h_U-h_B)}{h_A-h_U}}(2d_E\lambda_B r_B(h_B-h_U)+h_A-h_U)-h_A+h_U\right]}{2d_E^2\lambda_B^2 r_B^2(h_B-h_U)^2}. \quad (11)$$

(ii) session drop probability of unicast session of class k, $q_{U,k}$, and (iii) system resource utilization, u.

III. Performance Evaluation Framework

In this section, we formulate our performance evaluation framework. First, we determine the number of resources requested by multicast and unicast sessions while accounting for possible blockage situations. Using it as an input parameter, we then develop a queuing model that captures the system evolution in the presence of M and K classes of multicast and unicast sessions.

A. Resource Request Characterization

First, we determine the effective coverage range of a single NR AP, d_E. Recall that the distance in question is defined as the maximum separation between the UE and the NR AP, such that the UE in the LoS blocked conditions does not reside in outage. According to our propagation model, the SNR at the maximum 2D distance d_E is equal to

$$S = C_2\left(d_E^2 + (h_A - h_U)^2\right)^{-\frac{\zeta}{2}} = S_{th}, \quad (12)$$

where S_{th} is the SNR corresponding to the lowest feasible NR MCS [14]. Solving this equation for d_E, we obtain

$$d_E = \sqrt{(C_2/S_{th})^{\frac{2}{\zeta}} - (h_A - h_U)^2}. \quad (13)$$

Note that d_E depends on C_2 from (5), which, in its turn, depends on the sector angle α according to (7)–(8). Accounting for the propagation model, we proceed with deriving the probability mass function (pmf) of the number of requested resources. Particularly, we determine the sought pmf by first establishing the pmf of the number of requested resources in the LoS non-blocked and blocked states, and then weighing them with the corresponding probabilities. Our approach is similar for multicast and unicast sessions alike. Furthermore, recall that the SNR values in the LoS non-blocked and blocked conditions differ only by a constant attenuation factor. Hence, we provide a detailed derivation of the pmf for the LoS non-blocked conditions as an example.

Let S_{nB} be a random variable (RV) denoting SNR in the LoS non-blocked conditions and $F_{S_{nB}}(s)$, $s > 0$, be its cumulative distribution function (CDF). Let x be the 2D separation distance between the NR AP and the UE. Taking into account the heights of AP, h_A, and UE, h_U, the 3D propagation path distance y is

$$y = \sqrt{x^2 + (h_A - h_U)^2}, \quad (14)$$

which leads to the following SNR at the distance of x:

$$S_{nB} = C_1 y^{-\zeta} = C_1\left(x^2 + (h_A - h_U)^2\right)^{-\frac{\zeta}{2}}. \quad (15)$$

Let us now tag an arbitrary UE within the coverage area of the NR AP. Observe that due to the assumed nature of the UE process on the landscape, the UEs are uniformly distributed within the coverage area of the AP. Hence, two-dimensional distance to the NR AP follows $f_X(x) = 2x/d_E^2$, $0 < x < d_E$, where d_E is the effective coverage range of the AP. Observe that an upper and lower bound of the 3D distance between the NR AP and the UE are $|h_A - h_U|$ and $A = \sqrt{d_E^2 + (h_A - h_U)^2}$. Therefore, the pdf of the 3D distance is provided by

$$f_Y(y) = \begin{cases} \frac{2y}{d_E^2}, y \in (|h_A - h_U|, A), \\ 0, \quad y \notin (|h_A - h_U|, A), \end{cases} \quad (16)$$

which leads to the CDF $F_Y(y)$ in the form of

$$F_Y(y) = \begin{cases} 0, & y < |h_A - h_U|, \\ \frac{y^2 - (h_A - h_U)^2}{d_E^2}, & y \in [|h_A - h_U|, A], \\ 1, & y > \sqrt{d_E^2 + (h_A - h_U)^2}. \end{cases} \quad (17)$$

Since the SNR is a monotonously decreasing function of y, its distribution can be expressed in terms of the distribution of the distance Y. Hence, we have

$$F_{S_{nB}}(s) = Pr\left\{C_1 y^{-\zeta} < s\right\} = 1 - F_Y\left(\sqrt[\zeta]{C_1/s}\right). \quad (18)$$

From (17) and (18), the SNR CDF is given by

$$F_{S_{nB}}(s) = \begin{cases} 0, & s < C_1 A^{-\zeta}, \\ \frac{A^2 - \left(\frac{C_1}{s}\right)^{2/\zeta}}{d_E^2}, & \frac{C_1}{A^\zeta} \le s < \frac{C_1}{(h_A - h_U)^\zeta}, \\ 1, & s \ge C_1(h_A - h_U)^{-\zeta}. \end{cases} \quad (19)$$

Likewise, the CDF $F_{S_B}(s)$ of the RV S_B denoting the SNR in the LoS blocked conditions has the form (19) with C_2 from (5). To determine the overall SNR CDF, we need to establish the probability of blockage, p_B. The blockage probability at a fixed 2D distance x from the NR AP is immediately available from [23], [24]

$$p_B(x) = 1 - e^{-2\lambda_B r_B\left(x\frac{h_B - h_U}{h_A - h_U} + r_B\right)}, \quad (20)$$

where r_B and h_B are the blocker radius and height, respectively, h_U is the UE height, and h_A is the NR AP height. The blockage probability can then be calculated as in (11).

Using p_B, we may now determine the averaged SNR CDF $F_S(s)$ by weighing the branches corresponding to the LoS non-blocked and blocked conditions. Let $S_j, j = 1, 2, \ldots, J$, be the SNR thresholds. Also, let ϵ_j be the probability that the UE connection at hand is assigned the Channel Quality Indicator (CQI) and the MCS j, thus requiring r_j resource units, $j = 1, 2, \ldots, J$. Using the SNR CDF $F_S(s)$, we write

$$\begin{cases} \epsilon_0 = F_S(S_1), \\ \epsilon_j = F_S(S_{j+1}) - F_S(S_j),\ j = 1, 2, \ldots, J-1, \\ \epsilon_J = 1 - F_S(S_J). \end{cases} \quad (21)$$

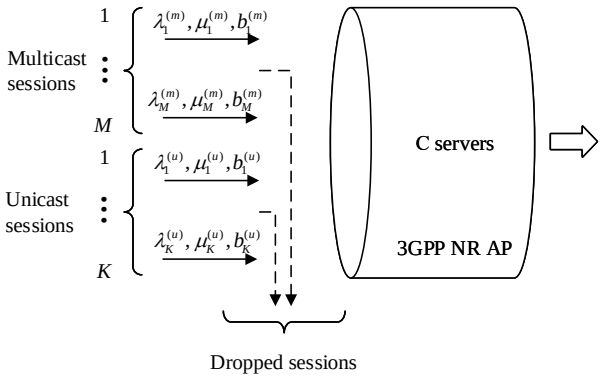


Fig. 2. Illustration of our queuing model.

The probability ϵ_j that a session requests r_j PRBs can now be used to determine the resource requirements of class m of multicast and class k of unicast sessions, $b_m^{(m)}$ and $b_k^{(u)}$.

B. Queuing Model and Analysis

Once the amount of requested resources is obtained, the task of assessing the system performance is reduced to formalizing the NR AP service process in the presence of M and K classes of multicast and unicast sessions. We approach this problem by utilizing the framework of pure loss queuing systems [25], [26] with zero waiting positions and C servers, where the number of servers is defined as $W/(s_A + \Delta)$, W is the available bandwidth at the NR AP, s_A is the size of one PRB, and Δ is the subcarrier spacing, see Fig. 2.

1) Stationary State Probabilities: Define the state of a the considered system as a vector $(\vec{\Psi}, \vec{\Phi})$, where $\vec{\Psi} = (I_1, I_2, \ldots, I_M)$ contains indicators reflecting the presence of multicast session of class m in the system, i.e., $I_m = 1$ when a multicast session of class m is present in the system, $\vec{\Phi} = (n_1, n_1, \ldots, n_K)$, where n_k is the number of unicast sessions of class k in this system. Further, let $\mathcal{Z}$ be the state space of the system,

$$\mathcal{Z} = \Big\{(\vec{\Psi}, \vec{\Phi}) : I_m \in \{0,1\},\, m = 1,2,\ldots,M, \\ n_k \in \{0,1,...\}, k = 1,2,\ldots,K\Big\}. \quad (22)$$

As one may observe, the evolution of the number of multicast and unicast sessions over (22) form a homogeneous Markov chain, $\{(\vec{\Psi}(t), \vec{\Phi}(t)), t \geq 0\}$. According to [26], this model is a generalization of the multi-class Erlang loss system [25] that can be solved for the stationary probabilities, $\vec{\pi}$, by using the state-space reduction technique [27].

To obtain the stationary state probabilities, first consider the case of an infinite number of servers, $C = \infty$. For this case, the state space $\widetilde{\mathcal{Z}}$ of the system can be expressed as

$$\widetilde{\mathcal{Z}} = \Big\{ \Big(\vec{\Psi}, \vec{\Phi}\Big) : I_m \in \{0,1\},\, m = 1,2,\ldots,M; \\ 0 \leq n_k \leq \left\lfloor \frac{C}{b_k^{(u)}} \right\rfloor, k = 1,2,\ldots,K; \\ \sum_{m=1}^{M} I_m b_m^{(m)} + \sum_{k=1}^{K} n_k b_k^{(u)} \leq C\Big\}. \quad (23)$$

When all of the sessions are admitted into the system, the components of the stochastic process describing the state evolution of the system, $\vec{\Psi}(t) \in \{0,1\}^M$ and $\vec{\Phi}(t) \in \{0,1,...\}^K$, do not affect each other. Hence, the components of the stationary state distribution $\vec{\Psi}(t), t \geq 0$ are available from [26]

$$\pi_m(I_m) = \lim_{t\to\infty} \mathrm{P}\{\Psi_m(t) = I_m\} = \frac{\gamma_m^{I_m}}{1+\gamma_m}, \quad (24)$$

where $I_m \in \{0,1\}$ and

$$\gamma_m = \mathrm{e}^{\rho_m} - 1, \quad (25)$$

where ρ_m is computed as follows

$$\rho_m = \left(1 + \frac{\lambda_m^{(m)}}{\mu_m^{(m)}}\right)^{p_m^{(m)} \lambda_B S_{\text{area}}} - 1. \quad (26)$$

Note that γ_m is the offered traffic load of multicast sessions of class m at the NR AP and $\lambda_m^{(m)}$ is the intensity of multicast requests of class m from a single UE. The exponent in (26) reflects the number of UEs initiating their multicast requests of class m. Its value depends on the coverage area $S_{\text{area}} = \pi d_E^2 \alpha/3$ of an antenna configuration that, in its turn, depends on the NR AP antenna array directivity α_A via the number of array elements. Similarly, for the unicast component of the model $\{\vec{\Phi}(t), t \geq 0\}$, the stationary state probabilities under the infinite server assumption are given by

$$p_k(n_k) = \lim_{t\to\infty} \mathrm{P}\{\Psi_k(t) = n_k\} = \frac{a_k^{n_k} \mathrm{e}^{-a_k}}{n_k!}, \quad (27)$$

which is defined over $n_k = 0,1,\ldots,$ where $a_k = \lambda_k^{(u)}/\mu_k^{(u)}$ is the offered load of class k unicast sessions at the NR AP. Here, $\lambda_k^{(u)} = p_k^{(u)} \Lambda \lambda_B S_{\text{area}}$ is the intensity of class k unicast requests at the NR AP.

Since there is no competition for the radio resources, the stationary state distribution of the composite stochastic process $(\vec{\Psi}(t), \vec{\Phi}(t))$ is provided as

$$\tilde{\pi}(\vec{\Psi}, \vec{\Phi}) = \tilde{G}^{-1}(\widetilde{\mathcal{Z}}) \prod_{m=1}^{M} \gamma_m^{I_m} \prod_{k=1}^{K} \frac{a_k^{n_k}}{n_k!}, (\vec{\Psi}, \vec{\Phi}) \in \tilde{\mathcal{Z}}, \quad (28)$$

where $\tilde{G}(\widetilde{\mathcal{Z}})$ is a normalization constant given by

$$\tilde{G}(\widetilde{\mathcal{Z}}) = e^{\sum_{k=1}^{K} a_k} \prod_{m=1}^{M} (1+\gamma_m). \quad (29)$$

Consider now the loss system with a reduced state space, i.e., $C < \infty$. The state transition diagram illustrating the case of $M = 2$ and $K = 1$ is displayed in Fig. 3. The stationary distribution for a system with the finite state space $\mathcal{Z}$ can be obtained from (28) and (29) by reducing the space $\widetilde{\mathcal{Z}}$ (22) to the space $\mathcal{Z}$ (23) and then normalizing the corresponding stationary state probabilities as

$$\pi(\vec{\Psi}, \vec{\Phi}) = G^{-1}(\mathcal{Z}) \prod_{m=1}^{M} \gamma_m^{I_m} \prod_{k=1}^{K} \frac{a_k^{n_k}}{n_k!}, (\vec{\Psi}, \vec{\Phi}) \in \mathcal{Z}, \quad (30)$$

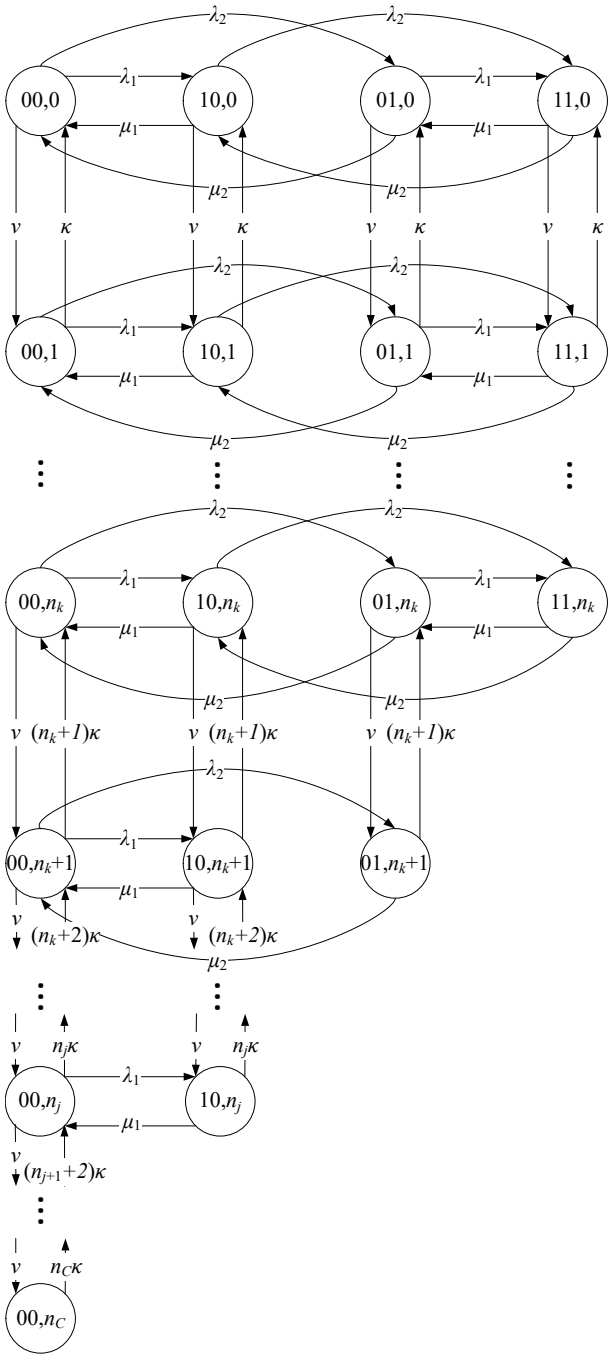


Fig. 3. State transition diagram for $C < \infty$, $M = 2$, $K = 1$.

where a normalization constant has the form of

$$G(\mathcal{Z}) = \sum_{(\vec{\Psi},\vec{\Phi})\in\mathcal{Z}} \prod_{m=1}^{M} \gamma_m^{I_m} \prod_{k=1}^{K} \frac{a_k^{n_k}}{n_k!}, \tag{31}$$

with γ_m defined in (25).

Observing (30) and (31), one may deduce that the ratio between the offered traffic loads of the multicast and unicast sessions should affect the trade-off between the multicast and unicast session drop probabilities. Furthermore, as the structure of (30) and (31) suggests, the load of the unicast session is included into the expression conventionally, i.e., the number of sessions is regulated by the factorial in the denominator. On the other hand, the effect of the multicast session rate is unrestrained. This implies that the number of multicast sessions in the system with a finite capacity should grow faster as compared to the unicast sessions, and there might be a resource capture effect leading to higher session drop probabilities of the unicast sessions and a lower fraction of resources utilized by these sessions in the system. Below, we investigate these effects numerically.

2) Performance Indicators: Once the stationary state probability vector, $\pi(\vec{\Psi},\vec{\Phi})$, is available, we can proceed by calculating the performance metrics associated with the system, which include the drop probability of class m multicast and class k unicast sessions, as well as the mean system resource utilization. However, direct calculation of the stationary state probabilities according to (30) and (31) is cumbersome due to the practical values of the numbers of servers C. To alleviate this obstacle, we develop a computationally efficient recursive algorithm that directly yields the performance indicators of interest. To this aim, we introduce an auxiliary function

$$h(n) = \begin{cases} 1, & n < 0, \\ 0, & n = 0, \\ \frac{1}{n}\sum_{k=1}^{K} b_k^{(u)} a_k h(n - b_k^{(u)}), & n = 1, 2, \ldots, C. \end{cases} \tag{32}$$

The non-normalized probabilities denoting that there are no sessions of class m and all of the multicast sessions of first i classes as well as all of the unicast sessions occupy exactly n servers are given by

$$f_m(i,n) = \begin{cases} 0, & i = 0, ..., M, n < 0, \\ h(n), & i = 0, n = 0, ..., C, \\ f_m(i-1,n) + \frac{f_m(i-1,n-b_i^{(m)})}{((1-\delta_{im})\gamma_i)^{-1}}, & \\ & i = 1, ..., M, n = 0, ..., C. \end{cases} \tag{33}$$

Note that $f_0(M,n)$ corresponds to the case where all of the unicast sessions of K classes and all of the multicast sessions of M classes occupy exactly n servers. The sought metrics can then be expressed in terms of $f_m(i,n)$ as

- drop probability of class m multicast session:

$$q_{M,m} = \frac{\sum_{n=C-b_m^{(m)}+1}^{C} f_m(M,n)}{\sum_{n=0}^{C} f_0(M,n)}, m = 1, 2, \ldots, M; \tag{34}$$

- drop probability of class k unicast session:

$$q_{U,k} = \frac{\sum_{n=C-b_k^{(u)}+1}^{C} f_0(M,n)}{\sum_{n=0}^{C} f_0(M,n)}, k = 1, 2 \ldots, K; \tag{35}$$

- mean system resource utilization:

$$u = \frac{\sum_{n=1}^{C} n f_0(M,n)}{\sum_{n=0}^{C} f_0(M,n)}. \tag{36}$$

IV. Main Numerical Results

In this section, we assess the performance of a mixture of multicast and unicast traffic service process at the NR AP. First, we validate our model by comparing its results with those obtained through computer simulations. Then, we

TABLE II
DEFAULT PARAMETERS OF NUMERICAL ASSESSMENT.

Parameter	Value
Operating frequency	28 GHz
Bandwidth, W	400 MHz
PRB size, s_A	1.44 MHz
Subcarrier spacing, Δ	0.015 MHz
Height of AP, h_A	4 m
Height of blocker, h_B	1.7 m
Height of UE, h_U	1.5 m
Blocker radius, r_B	0.4 m
Density of blockers, λ_B	0.5 bl./m^2
SNR threshold, S_{th}	-9.47 dB
Transmit power, P_A	2 W
Path loss exponent, ζ	2.1
Power spectral density of noise, N_0	-174 dBm/Hz
Blockage attenuation, B	15 dB
Fading margins, $M_{S,nB}, M_{S,B}$	4/8.2 dB
Interference margin, M_I	3 dB
UE planar antenna elements, N_U	4 el.
UE receive gain, G_U	5.57 dBi
Session data rate, $R_U = R_M$	{20,50} Mbps
Default service intensities, $\mu^{(u)}, \mu^{(m)}$	30 s
AP antenna array, N_A	{4, 8, 16, 32}$\times$ 4 el.
AP transmit gain, G_A	{5.57, 8.57, 11.57, 14.58} dBi
AP coverage range, d_E	{107, 149, 207, 288} m
Inter-site distance, ISD	$3d_E$ m
Number of unicast classes, K	1 cl./cell

continue by investigating the effect of multicast and unicast session parameters on the performance metrics including multicast and unicast session drop probabilities as well as system resource utilization. Further, we identify the maximum ISD for the typical hexagonal deployment of the NR APs to deliver the target performance guarantees over multicast and unicast sessions. Finally, we study the performance of an NR system where multicast service is implemented via unicast service.

The core system parameters are provided in Table II. Table III clarifies the mapping between the SNR and the spectral efficiency, while Table IV reflects the pre-computed relationship between the number of antenna elements at the NR AP, the effective coverage range, d_E, and the amount of resources required to maintain 20 Mbps and 50 Mbps data rates.

To conduct our performance evaluation campaign, we rely upon the following approach. We parametrize the developed queuing model by using M multicast session classes and one

TABLE III
CQI, MCS, AND SNR MAPPING FOR 5G NR.

CQI	MCS	Spectral efficiency	SNR in dB
0	out of range		
1	QPSK, 78/1024	0,15237	-9,478
2	QPSK, 120/1024	0,2344	-6,658
3	QPSK, 193/1024	0,377	-4,098
4	QPSK, 308/1024	0,6016	-1,798
5	QPSK, 449/1024	0,877	0,399
6	QPSK, 602/1024	1,1758	2,424
7	16QAM, 378/1024	1,4766	4,489
8	16QAM, 490/1024	1,9141	6,367
9	16QAM, 616/1024	2,4063	8,456
10	64QAM, 466/1024	2,7305	10,266
11	64QAM, 567/1024	3,3223	12,218
12	64QAM, 666/1024	3,9023	14,122
13	64QAM, 772/1024	4,5234	15,849
14	64QAM, 873/1024	5,1152	17,786
15	64QAM, 948/1024	5,5547	19,809

TABLE IV
SYSTEM PARAMETERS INDUCED BY NR AP ANTENNA ARRAY.

Array	Gain, dBi	HPBW, $^\circ$	d_E, m	PRBs for (20,50) Mbps
32x4	14.58	3.18	288	(7,16)
16x4	11.57	6.37	207	(6,14)
8x4	8.57	12.75	149	(5,12)
4x4	5.57	25.50	107	(5,11)

unicast session class, where M corresponds to the number of the NR AP antenna configurations (sub-sectors with the directivity angle of α_A) needed to cover a 120° sector served by a single array. Throughout this section, the number of classes corresponds to the potential number of transmissions required to disseminate the same content to all of the multicast users. We also introduce the spatial session arrival intensity, λ_S, defined as the spatial arrival intensity of all sessions in a sector covered by a single configuration of the NR AP antenna array, i.e., $\lambda_S = \lambda_B \pi d_E^2 \alpha_A / 3$, where α_A is measured in radians. The fraction of multicast sessions of all classes is then $\lambda_S \sum_{m=1}^{M} p_m^{(m)}$. Also, observe that $p_{M,i} = p_{M,j}$, $\forall i, j$, which induces $q_{M,i} = q_{M,j} = q_M$ for all the multicast classes.

A. Model Validation

We start by validating the proposed analytical framework. To this aim, we develop a single-purpose simulation environment that accepts the input parameters listed in Table II, together with the propagation and service sub-models, and returns the considered metrics of interest. To construct our simulator, we rely upon a discrete-event modeling framework (DES, [28]). The beginning of the stationary state period is determined by using an exponentially-weighted moving average test with a smoothing constant of 0.05 [29]. The statistics were collected only during the stationary state period by using the method of batch means [30] and sampling the state of the system each second of the simulation time.

A comparison of multicast and unicast session drop probabilities obtained with the developed mathematical model and the computer simulations is demonstrated in Fig. 4 as a function of the spatial session arrival intensity for the session data rates of $R_U = R_M = 20$ Mbps. Here, we specifically

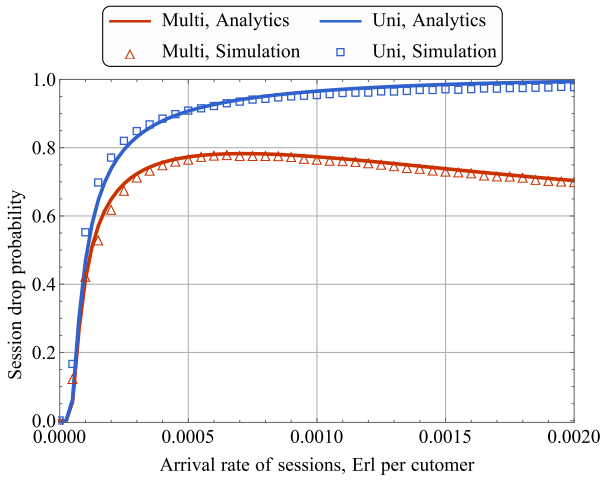
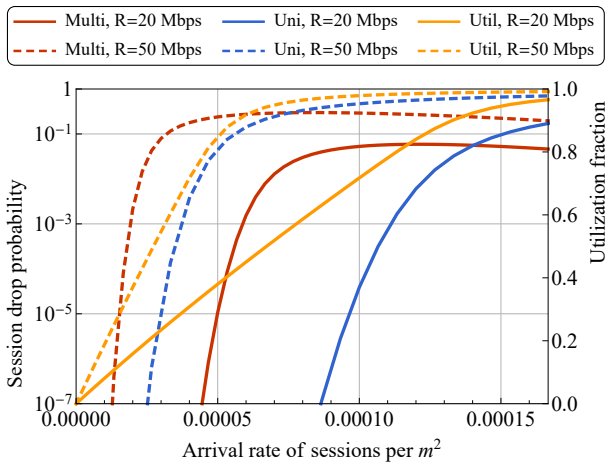
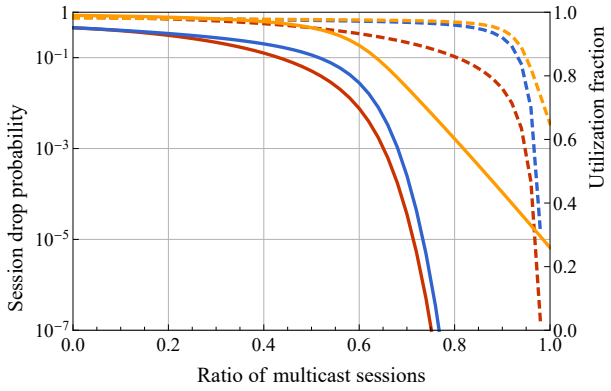


Fig. 4. Comparison of analytical and simulation results.



(a)



(b)

Fig. 5. Session drop probability as a function of arrival process parameters.

address an extreme case of 128×4 NR AP antenna array corresponding to the gain of 20.59 dBi and unit mean service times. The fraction of multicast sessions, p_M, is set to 0.5. As one may observe, the simulation data agrees closely with the theoretical results. A similar match has been noted for other input parameters as well as in case of the mean resource utilization. Therefore, we rely upon our developed analytical model to deliver the target assessment of the joint service process of multicast and unicast sessions at the NR AP.

B. Effects of Arrival and Service Characteristics

In our model, there are two types of traffic with fundamentally different service types that may have a profound effect on each other's performance at the NR APs. Therefore, we first analyze the impact of multicast and unicast arrival and service process characteristics on the user- and system-centric performance indicators, which includes the session drop probability and the system resource utilization. Throughout this section, we employ 32×4 antenna array at the NR AP that corresponds to the effective coverage distance of $d_E = 288$ m and $3.18°$ HPBW. Observe that with this antenna array, there are overall 32 classes of multicast sessions in the system.

Fig. 5 reports on the multicast and unicast session drop probability as well as the system resource utilization as a function of the spatial arrival intensity of sessions and the proportion of multicast sessions for the two data rates of multicast and unicast sessions, 20 Mbps and 50 Mbps, and 30 s of the mean service time for both types of traffic. Particularly, in Fig. 5a we keep the fraction of multicast sessions constant at $p_M = 0.5$ and vary the spatial session arrival intensity λ_S, while in Fig. 5b the latter is constant (set to 0.005 sessions per square meter) and we vary the fraction of multicast sessions.

Analyzing the effect of the spatial session arrival intensity as illustrated in Fig. 5a, we learn that for both of the considered session rates, an increase in the spatial session arrival intensity grows the multicast session drop probability faster as compared to the unicast case. Particularly, for the intensity of $7.5E-5$ and $R_M = R_U = 20$ Mbps, the multicast session drop probability is already higher than 0.01, while the unicast session drop probability is far below 10^{-7}. However, a further increase in λ_S does not impact the multicast session drop probability negatively, and for higher values of the spatial session arrival intensity it begins to decrease. In contrast, the unicast session drop probability grows exponentially for the considered range of spatial session arrival intensities.

The reason is in the special service that multicast connections receive, i.e., if there is an ongoing multicast session of class i in the system, all the additional multicast sessions of this class join this ongoing service. Such a behavior produces an implicit priority for the multicast traffic. When the intensity of multicast sessions is rather high, there is an ongoing multicast session of class i nearly at all times; hence, we observe a drastic decrease in q_M and a significant increase in q_U. The aforementioned trends also hold true for the session data rate of 50 Mbps. The difference here is that the system saturates faster, which implies that the impact of the session data rate is only quantitative. This observation is also confirmed by the behavior of the system resource utilization demonstrated in Fig. 5a.

Consider now the metrics of interest as a function of the fraction of multicast sessions, p_M, as illustrated in Fig. 5b. First, we note that the multicast session drop probability is always below its unicast counterpart. Furthermore, this difference becomes larger as the fraction of multicast sessions grows. Both observations are a direct consequence of the above "resource capture" effect. Even for the moderate values of p_M, i.e., $p_M > 0.4$, the system always has resources allocated to all the multicast classes associated with the NR AP antenna array, while only the remaining resources are available for the unicast sessions, thus inducing high unicast session drop probabilities. However, the impact of this effect is reduced when the number of the NR AP antenna array elements increases. The reason is that the area of the sector served by a particular configuration decreases, which implies that mulitcast sessions do not always exist in the system.

Analyzing these results further, observe that both drop probabilities as well as the resource utilization decrease as the fraction of multicast sessions grows. This behavior is also explained by the specific resource allocation process for multicast sessions. Hence, increasing p_M effectively reduces the loading of the system. Finally, one may notice that the rate of the multicast and unicast sessions significantly increases the gap between the corresponding drop probabilities. In terms

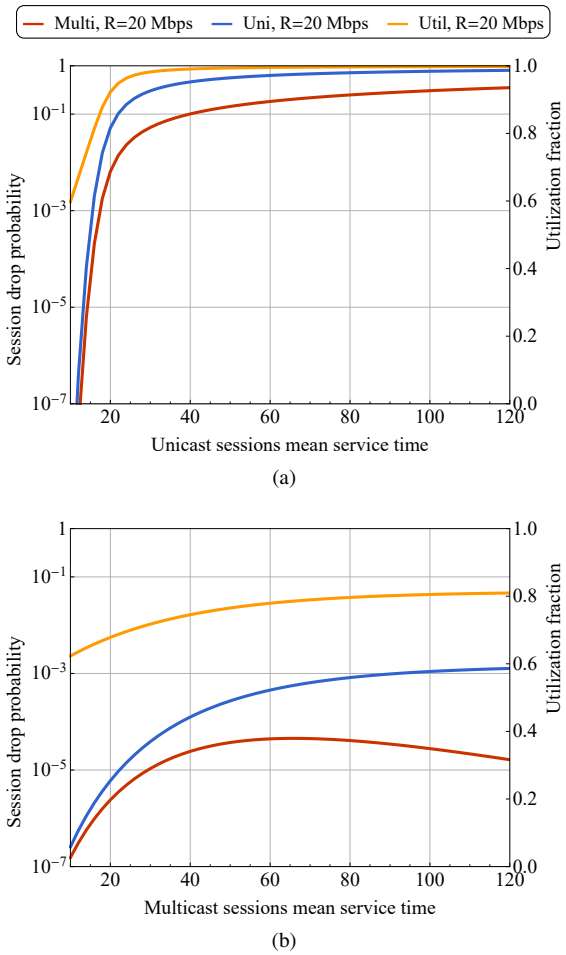


Fig. 6. Session drop probability as a function of service process parameters.

of absolute numbers, this difference may reach an order of magnitude for $R_U = R_M = 50$ Mbps, which is significantly higher as compared to the case of $R_U = R_M = 20$ Mbps.

Consider now the effects of the mean service time of multicast and unicast sessions on the corresponding drop probabilities and resource utilization as shown in Fig. 6. Recall that the service intensity is related to the mean service time as $1/\mu^{(m)}$, $1/\mu^{(u)}$. Particularly, Fig. 6a displays the mean unicast service time for $\lambda_S = 10^{-5}$, $p_M = 0.2$ and the mean multicast session intensity of $1/\mu^{(m)} = 30$ s for the two session data rates of 20 Mbps. When the mean service time of the unicast sessions, $1/\mu^{(u)}$, increases, both multicast and unicast drop probabilities grow. However, as the unicast session drop probability approaches 1 already for approximately $1/\mu^{(u)} = 80$ s, the multicast session drop probability remains well below 1, which confirms the implicit priority given to multicast sessions as a result of their specific service process even for a rather small values of $p_M = 0.2$.

The effects of the mean service time of the multicast sessions, $1/\mu^{(m)}$, are further highlighted in Fig. 6b. As one may observe, higher values of $1/\mu^{(m)}$ lead to better multicast session performance and produce a negative impact on the unicast session drop probability. In the limit, when $\mu^{(m)} \to 0$ by yielding exceptionally long multicast session durations, q_M approaches 0, while q_U is close to 1. In this case, the system is always busy with providing multicast service (32 classes for the 32×4 NR AP antenna array), thus leaving almost no room for the unicast connections. Observe that this behavior does not negatively impact the system-centric performance – the fraction of the utilized system resources – since it gradually increases as a function of both $1/\mu^{(u)}$ and $1/\mu^{(m)}$.

C. Cellular NR AP Deployments

The effective coverage distance of the NR AP, d_E, and thus the ISD depend on the number of antenna elements employed at the AP. To decrease the density of NR APs required to cover a certain area of interest, one needs to utilize all of the available antenna elements by increasing the NR AP transmit gain. However, it may not always be feasible, since the performance provided to either multicast or unicast sessions might not be satisfactory. Particularly, by increasing the number of antenna elements at the NR AP, one also decreases the HPBW, and thus grows the number of multicast classes in the system. The latter results in a higher load imposed at the NR AP, which may negatively affect both multicast and unicast session drop probabilities.

We continue by characterizing the trade-off between the number of antenna elements utilized at the NR AP and the performance metrics related to multicast and unicast sessions. The following discussion also illustrates the procedure for determining the maximum number of antenna elements at

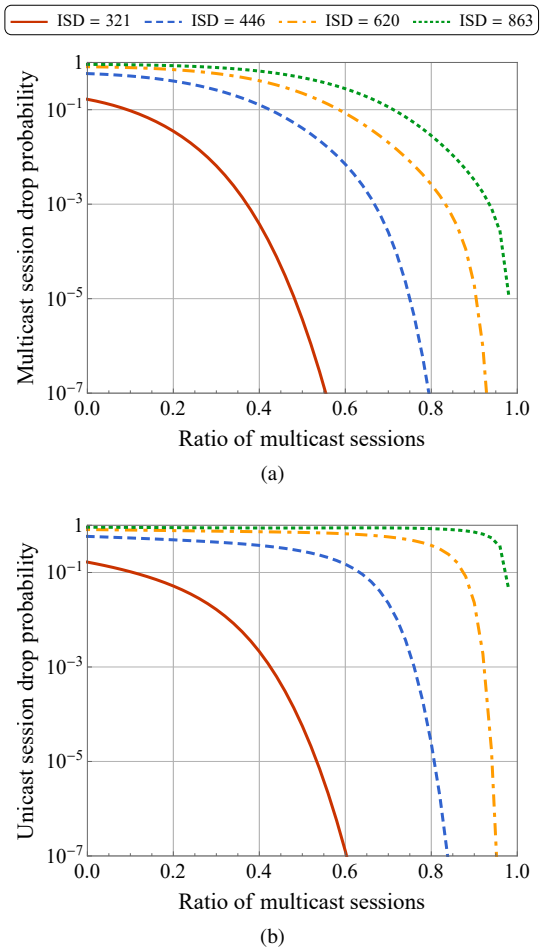
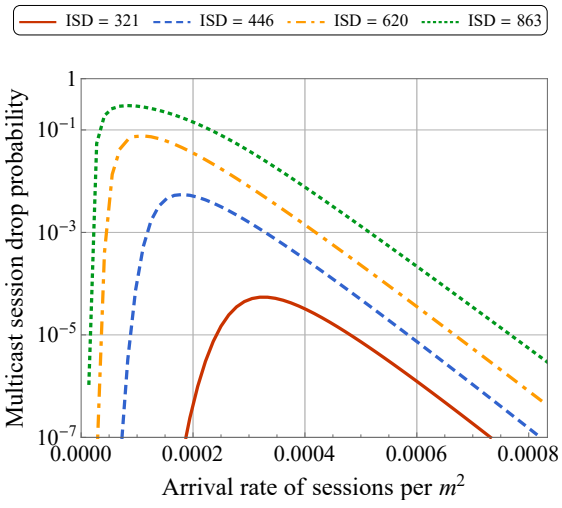
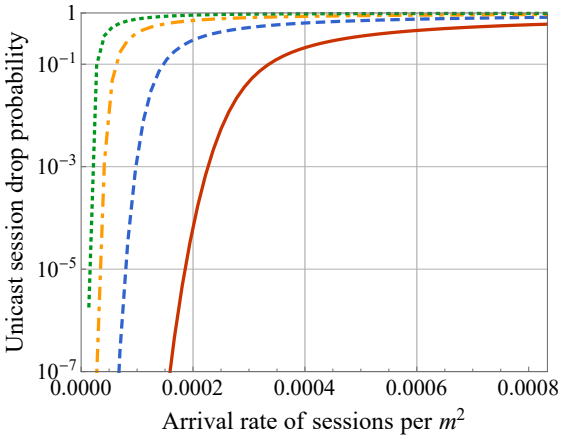


Fig. 7. Multicast and unicast session drop probabilities as a function of ISD.



(a)



(b)

Fig. 8. Multicast and unicast session drop probabilities as a function of ISD.

the NR AP, and thus the maximum ISD between the NR APs, which can be used to provide the target multicast and unicast session drop probabilities. The multicast and unicast session drop probabilities as functions of the fraction of multicast sessions, p_M, for different ISDs are displayed in Fig. 7 for the mean durations of both multicast and unicast sessions set to 30 s and the spatial session arrival intensity of 0.0015 sessions per square meter for the session data rates of $R_U = R_M = 20$ Mbps. Recall that the ISD in a cellular deployment is related to the effective coverage, d_E, as $3d_E$ [31], while the correspondence between the ISD and the antenna configuration at the NR AP is shown in Table IV.

As one may deduce by comparing the results of Fig. 7a and Fig. 7b, for the fixed values of spatial session arrival intensity and ISD, an increase in the fraction of multicast sessions leads to lower drop probabilities for both types of sessions. The reason is that higher values of p_M reduce the actual traffic load imposed on the system. Furthermore, for a small number of utilized antenna elements, e.g., 4×4 NR AP array, multicast and unicast probabilities almost coincide. However, for a higher number of antenna elements, and thus smaller HPBW, the difference becomes dramatic, especially for moderate-to-high values of p_M. Note that this difference can be rather substantial in absolute numbers. Particularly, for $p_M = 0.8$ and 16×4 NR AP antenna array corresponding to the ISD of 620 m, the multicast session drop probability is only $q_M = 0.005$, while its unicast counterpart is as high as 0.6, thus making the unicast service virtually unusable. The underlying reason is in the number of multicast classes that need to be maintained at the NR AP. Particularly, the use of 32×4 antenna array corresponding to the ISD of 863 m requires 32 separate multicast classes. The service discipline for the multicast traffic forces the NR AP to process its multicast sessions almost exclusively, thus leaving only the remaining resources for the unicast sessions.

The multicast and unicast session drop probabilities as functions of the spatial session arrival intensity are demonstrated in Fig. 8 for different ISDs, multicast session proportion of $p_M = 0.5$, session data rates of $R_U = R_M = 20$ Mbps, and mean durations of both multicast and unicast sessions of $1/\mu^{(u)} = 1/\mu^{(m)} = 30$ s. The property of multicast sessions to capture the resources is exemplified in Fig. 8a, where there is a visible peak in the multicast session drop probability for all the considered numbers of antenna array elements at the NR AP. The rationale is that for a fixed value of p_M, an increase in the spatial session arrival intensity leads to a higher number of multicast session arrivals. Due to specific service process, this intensity grows the probability that upon arrival of a new session its multicast class already exists at the NR AP, thereby reducing the multicast session drop probability.

Note that the smaller the number of antenna elements is the lower the maximum multicast session drop probability becomes. Eventually, when the spatial session arrival intensity is such that $\lambda_S \to \infty$, all the multicast classes always exist in the system and the multicast session drop probability is virtually zero. We also notice that the unicast session drop probability increases as λ_S grows for all of the considered ISDs. However, it happens faster as compared to the system with only unicast sessions, since the increased data rate is negatively affected by the decreased multicast session drop probability.

Summarizing our cellular deployment analysis, we note that although the use of higher numbers of antenna elements at the NR APs allows for extending the coverage and potentially reducing the interference [6], [17], especially in 3D deployments [15], [16], it also dramatically increases the loading of the NR AP. With power control techniques, one may reduce the offered load of the unicast traffic by limiting the coverage of the NR AP but the number of multicast classes that need to be maintained remains high due to the use of smaller HPBW.

At the same time, one may want to avoid the use of smaller numbers of antenna elements at the NR AP to form antenna radiation patterns as this leaves part of the deployment area uncovered for the multicast service and thus requires further densification of the layout. Another interesting phenomenon is that an increase in the fraction of multicast sessions decreases the multicast session drop probability. Multicast traffic implicitly receives priority over unicast traffic due to its service properties. Hence, in practical deployments under high multicast loads the operators might consider providing certain resource reservation for the unicast traffic [32] or explicit

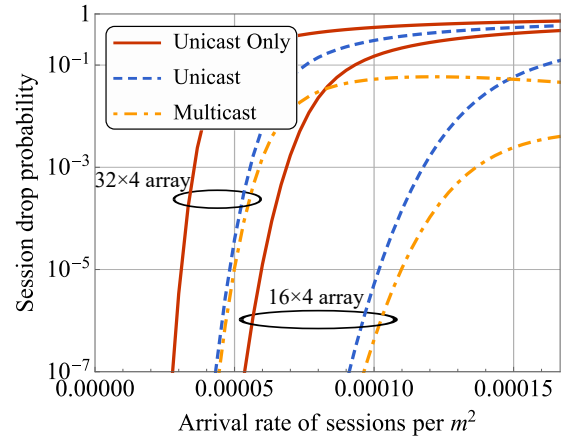


Fig. 9. Drop probabilities for multicast/unicast and unicast-only systems.

connection admission control (CAC) functions [33].

Our developed methodology allows to evaluate the parameters of a cellular NR AP deployment, such that the target user- and system-centric performance targets are met simultaneously. We emphasize that in practice these results provide a lower bound on the deployment density and have to be refined by accounting for specific propagation environments.

D. Unicast-Only Service

The use of the antenna arrays with a large numbers of elements allows to considerably improve the potential coverage of the prospective NR APs, thus significantly reducing the deployment costs. However, these massive arrays yield extremely small HPBWs that increase the number of multicast classes to be maintained at the NR AP, which causes inefficient resource utilization. Taking into account the implicit priority of the multicast sessions as discussed above, the service levels provided to the unicast sessions degrade drastically. The question we explore here is whether multicast service remains useful when antenna arrays with large numbers of elements are employed at the NR AP, or one may solely rely upon the unicast service. We address this by contrasting the system with both multicast and unicast sessions against the one, where multicast sessions are served over the unicast service.

We start by comparing the drop probabilities for the hybrid multicast/unicast system with those for the unicast-only system as illustrated in Fig. 9 being a function of the spatial session arrival intensity for the session data rates of $R_U = R_M = 20$ Mbps and the session durations of $1/\mu^{(u)} = 1/\mu^{(m)} = 30$ s, as well as two antenna array configurations, 32×4 with HPBW 3.19° and 16×4 with HPBW 3.19°, see Table IV. Particularly, Fig. 9a displays the dependence on the spatial session arrival intensity with a fixed share of multicast sessions, $p_M = 0.5$.

As one may observe, for the considered number of antenna elements the drop probability in the unicast-only system is always higher as compared to the multicast and unicast drop probabilities in the hybrid multicast/unicast system. However, notice that the gap between the two systems decreases as the number of antenna elements employed at the NR AP grows. The reason is that the overall number of multicast classes that need to be maintained at the NR AP increases. However, even for 128×4 array at the NR AP, the unicast system performs significantly worse across the entire considered range of the spatial session arrival intensities.

Finally, Fig. 9b illustrates the role of the fraction of the multicast sessions to confirm that the above conclusions hold for any mixture of multicast and unicast sessions. It is also important to note that the difference between the unicast-only and the hybrid unicast/multicast sessions grows as p_M increases. Hence, we may conclude that from the user-centric performance viewpoint, for today's antenna arrays the hybrid unicast/multicast system is by far superior to the unicast-only system. However, this conclusion may change in the future, when antenna arrays featuring more than 128×4 elements will appear. Such systems may potentially implement multicast service by using unicast sessions and thus should not compromise the ISD. The unicast-only system is much easier to control and dimension, since all sessions are treated similarly as compared to the hybrid multicast/unicast system, where the multicast sessions are implicitly prioritized.

V. Conclusions

To extend the coverage range and reduce the impact of interference, future 5G NR systems operating in millimeter-wave frequency bands are expected to employ highly directional antennas at the NR AP side. However, multicast and unicast types of traffic are characterized by drastically different service processes and may impose conflicting requirements on the antenna radiation patterns. In this work, by relying upon the tools of stochastic geometry and queuing theory, we developed a detailed mathematical framework for the performance assessment of the NR AP service for a mixture of multicast and unicast traffic types. Our framework permits the evaluation of both user- and system-centric metrics of interest, which include multicast and unicast session drop probabilities and system resource utilization.

Within the developed framework, we thoroughly investigated the effects of multicast and unicast traffic properties on their mutual performance at the NR AP. Specifically, we demonstrated that the particularities of the multicast service process implicitly introduce priority for the multicast sessions

by significantly harming the unicast session drop probability. As the offered load of the multicast sessions grows, the system becomes almost fully occupied with them, thus leaving only slim residual resources for the unicast sessions. To balance out the drop probabilities, one needs to explicitly prioritize the unicast traffic with, e.g., bandwidth reservation techniques and/or connection admission control algorithms.

Analyzing the effects of antenna arrays at the NR AP, we concluded that for a given spatial session arrival intensity from both traffic types the user-centric performance indicators improve when the number of antenna elements forming a radiation pattern decreases. Finally, we compared the considered hybrid multicast/unicast system with the one where multicast sessions are handled over the unicast service. We showed that even for higher antenna directivities this leads to a worse performance for the multicast service. However, the unicast-only system is much better in terms of fairness as it does not prioritize any of the traffic types by contrast to the hybrid multicast/unicast system.

As a result, our proposed framework allows to predict the density of the NR APs that is required to fully cover a certain deployment area with the particular target multicast/unicast traffic loads. However, we emphasize that in practice this estimate may yield a lower bound on the actual deployment density, which has to be refined with a specific propagation environment in mind.

REFERENCES

[1] "M.2083: IMT vision - Framework and overall objectives of the future development of IMT for 2020 and beyond," ITU-R technical recommendation, 2015.

[2] A. Biason and M. Zorzi, "Multicast via point to multipoint transmissions in directional 5G mmWave communications," *IEEE Communications Magazine*, vol. 57, no. 2, pp. 88–94, 2019.

[3] N. S. Jeong, Y.-C. Ou, A. Tassoudji, J. Dunworth, O. Koymen, and V. Raghavan, "A recent development of antenna-in-package for 5G millimeter-wave applications," in *2018 IEEE 19th Wireless and Microwave Technology Conference (WAMICON)*, pp. 1–3, IEEE, 2018.

[4] B. Sadhu, A. Paidimarri, M. Ferriss, M. Yeck, X. Gu, and A. Valdes-Garcia, "A 128-element dual-polarized software-defined phased array radio for mm-wave 5G experimentation," in *Proceedings of the 2nd ACM Workshop on Millimeter Wave Networks and Sensing Systems*, pp. 21–25, ACM, 2018.

[5] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!," *IEEE Access*, vol. 1, pp. 335–349, 2013.

[6] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, "What will 5G be?," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, 2014.

[7] W. Kim, T. Song, and S. Pack, "Rate adaptation for directional multicast in IEEE 802.11 ad networks," in *Consumer Electronics (ICCE), 2012 IEEE International Conference on*, pp. 364–365, IEEE, 2012.

[8] H. Park, S. Park, T. Song, and S. Pack, "An incremental multicast grouping scheme for mmwave networks with directional antennas," *IEEE Communications Letters*, vol. 17, no. 3, pp. 616–619, 2013.

[9] W. Feng, Y. Li, Y. Niu, L. Su, and D. Jin, "Multicast spatial reuse scheduling over millimeter-wave networks," in *Wireless Communications and Mobile Computing Conference (IWCMC), 2017 13th International*, pp. 317–322, IEEE, 2017.

[10] K. Sundaresan, K. Ramachandran, and S. Rangarajan, "Optimal beam scheduling for multicasting in wireless networks," in *Proceedings of the 15th annual international conference on Mobile computing and networking*, pp. 205–216, ACM, 2009.

[11] Z. Zhang, Z. Ma, Y. Xiao, M. Xiao, G. K. Karagiannidis, and P. Fan, "Non-orthogonal multiple access for cooperative multicast millimeter wave wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 8, pp. 1794–1808, 2017.

[12] A. Biason and M. Zorzi, "Multicast transmissions in directional mmwave communications," in *European Wireless 2017; 23th European Wireless Conference; Proceedings of*, pp. 1–7, VDE, 2017.

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

[14] 3GPP, "NR; Physical channels and modulation (Release 15)," 3GPP TR 38.211, Dec 2017.

[15] R. Kovalchukov, D. Moltchanov, A. Samuylov, A. Ometov, S. Andreev, Y. Koucheryavy, and K. Samouylov, "Evaluating SIR in 3D Millimeter-Wave Deployments: Direct Modeling and Feasible Approximations," *IEEE Transactions on Wireless Communications*, vol. 18, no. 2, pp. 879–896, 2019.

[16] R. Kovalchukov, D. Moltchanov, A. Samuylov, A. Ometov, S. Andreev, Y. Koucheryavy, and K. Samouylov, "Analyzing effects of directionality and random heights in drone-based mmwave communication," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 10, pp. 10064–10069, 2018.

[17] V. Petrov, M. Komarov, D. Moltchanov, J. M. Jornet, and Y. Koucheryavy, "Interference and SINR in Millimeter Wave and Terahertz Communication Systems With Blocking and Directional Antennas," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1791–1808, 2017.

[18] 3GPP, "Study on channel model for frequencies from 0.5 to 100 GHz (Release 14)," 3GPP TR 38.901 V14.1.1, July 2017.

[19] S. Singh, R. Mudumbai, and U. Madhow, "Interference analysis for highly directional 60-GHz mesh networks: The case for rethinking medium access control," *IEEE/ACM Transactions on Networking (TON)*, vol. 19, no. 5, pp. 1513–1527, 2011.

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

[21] D. Moltchanov, "Distance distributions in random networks," *Elsevier Ad Hoc Networks*, vol. 10, pp. 1146–1166, Aug. 2012.

[22] K. Samouylov and Y. Gaidamaka, "Analysis of loss systems with overlapping resource requirements," *Statistical Papers*, vol. 59, no. 4, pp. 1463–1470, 2018.

[23] 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.

[24] 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.

[25] F. P. Kelly, "Loss networks," *The annals of applied probability*, pp. 319–378, 1991.

[26] G. Basharin, Y. Gaidamaka, and K. Samouylov, "Mathematical theory of teletraffic and its application to the analysis of multiservice communication of next generation networks," *Automatic Control and Computer Sciences*, vol. 47, pp. 62–69, 2013.

[27] V. B. Iversen *et al.*, "Teletraffic engineering handbook," *ITU-D SG*, vol. 2, p. 16, 2005.

[28] B. P. Zeigler, T. G. Kim, and H. Praehofer, *Theory of modeling and simulation*. Academic press, 2000.

[29] H. G. Perros, "Computer simulation techniques: The definitive introduction!," 2009.

[30] G. S. Fishman and L. S. Yarberry, "An implementation of the batch means method," *INFORMS Journal on Computing*, vol. 9, no. 3, pp. 296–310, 1997.

[31] T. S. Rappaport *et al.*, *Wireless communications: principles and practice*, vol. 2. prentice hall PTR New Jersey, 1996.

[32] D. Moltchanov, A. Samuylov, V. Petrov, M. Gapeyenko, N. Himayat, S. Andreev, and Y. Koucheryavy, "Improving session continuity with bandwidth reservation in mmwave communications," *IEEE Wireless Communications Letters*, vol. 8, pp. 105–108, Feb 2019.

[33] S. Al-Rubaye, A. Al-Dulaimi, J. Cosmas, and A. Anpalagan, "Call admission control for non-standalone 5G ultra-dense networks," *IEEE Communications Letters*, vol. 22, no. 5, pp. 1058–1061, 2018.