Optimal Multicasting in Millimeter Wave 5G NR with Multi-beam Directional Antennas

Optimal Multicasting in Millimeter Wave 5G NR with Multi-beam Directional Antennas

Nadezhda Chukhno, Olga Chukhno, Dmitri Moltchanov, Antonella Molinaro, Yuliya Gaidamaka, Konstantin Samouylov, Yevgeni Koucheryavy, and Giuseppe Araniti

Abstract—The support of multicast communications in the fifth-generation (5G) New Radio (NR) system poses unique challenges to system designers. Particularly, the highly directional antennas do not allow to serve all the user equipment devices (UEs) that belong to the same multicast session in a single transmission. However, the capability of modern antenna arrays to utilize multiple beams simultaneously, with potentially varying half-power beamwidth, adds a new degree of freedom to the UE scheduling. This work addresses the challenge of optimal multicasting in 5G millimeter wave (mmWave) systems by presenting a globally optimal solution for multi-beam antenna operation. The optimization problem is formulated as a special case of multi-period variable cost and size bin packing problem that allows to not impose any constraints on the number of the beams and their configurations. We also propose heuristic solutions having polynomial time complexity. Our results show that for small cell radii of up to 100 meters, a single beam is always utilized. For higher cell coverage and practical ranges of the number of users (5-50), the optimal number of beams is upper bounded by 3.

Index Terms—5G, New Radio, Millimeter Wave, Multicast, Multi-beam antennas, Optimization, Heuristic Algorithms.

1 Introduction

The growth in demand for mobile multimedia services poses considerable challenges in providing reliable service quality, with the support of a large number of users competing for limited radio resources in cellular networks [1]. The New Radio (NR) technology is expected to be the primary enabler of the fifth-generation (5G) cellular system's air interface. While the basic functionality of NR has already been specified in 3GPP Rel. 15 [2] and Rel. 16 [3], several advanced functionalities are still not defined. One of these critical functionalities is multicasting that has been planned for 3GPP Rel. 17 onwards [4], [5].

Multicasting is a prominent technique applied to improve bandwidth efficiency compared to unicast transmission [6], [7]. In the multicast regime, a base station (BS) can transmit the packet to many users simultaneously using the same band and modulation and coding scheme (MCS). In the microwave spectrum with typically omnidirectional transmissions, multicast is a natural scheme to implement. However, in highly directional systems, i.e., the millimeter wave (mmWave) band communications considered for NR, the use of extremely directional radiation patterns at the BS's antennas poses some challenges to the multicast operation design, which still remain unsolved or even unaddressed [8], [9].

In exchange for the promised extraordinary rates at the air interface, mmWave NR systems bring the following hurdles [10]. First of all, the use of highly directional antenna radiation patterns does not allow to serve simultaneously, in a single transmission, all the user equipment devices (UEs), which belong to the same multicast session and are located in very large regions [11]. Indeed, the signal-to-interference-plus-noise ratio (SINR) decreases with larger beams. Secondly, NR is expected to work with considerably larger antenna arrays, hence increasing the design complexity with respect to relatively simple microwave antenna configurations [12]. These issues are further exacerbated by the adverse properties of mmWave propagation, including severe free-space attenuation [13] and vulnerability to blockage [14]. Finally, the capability of modern antenna arrays to utilize multiple beams at the same time with potentially varying half-power beamwidth (HPBW) adds further degrees of freedom to multicast group formation and scheduling, significantly complicating their design. However, when multiple beams are available, the width of numerous beams to be swept simultaneously

N. Chukhno, O. Chukhno, A. Molinaro, and G. Araniti are with Mediterranea University of Reggio Calabria, Reggio Calabria, Italy and CNIT, Italy. Email: nadezda.chukhno@unirc.it, olga.chukhno@unirc.it, araniti@unirc.it, antonella.molinaro@unirc.it

O. Chukhno, D. Moltchanov, and Y. Koucheryavy are with Tampere University, Tampere, Finland. Email: dmitri.moltchanov@tuni.fi, evgeny.kucheryavy@tuni.fi

N. Chukhno is also with Universitat Jaume I, Spain.

A. Molinaro is also with Université Paris-Saclay, Gif-sur-Yvette, France.

Yu. Gaidamaka and K. Samouylov are with Peoples' Friendship University of Russia (RUDN University), Moscow, Russia. Email: gaydamaka-yuv@rudn.ru, samuylov-ke@rudn.ru

Yu. Gaidamaka, K. Samouylov are also with Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Russia.

The authors gratefully acknowledge funding from European Union's Horizon 2020 Research and Innovation programme under the Marie Skłodowska Curie grant agreement No. 813278 *(A-WEAR: A network for dynamic wearable applications with privacy constraints).*

This paper has been supported by the RUDN University Strategic Academic Leadership Program (recipients Yu.G., K.S., Sections 1, and 2). The reported study was funded by RFBR, project number 19-07-00933, and 20-07-01064.

has to be properly selected, under the total transmission power constraint per antenna. This means that compared to single-beam systems, power has to be split among beams in a sophisticated manner.

The question of efficient multicasting in wireless systems has been addressed recently, see Section 2 for review. Particularly, optimal solutions for single-beam antenna design have been proposed so far in [15]. Furthermore, there are a number of heuristic solutions for single-beam antennas [16], [17]. While several heuristics for multi-beam NR antenna designs have also been proposed [18]–[20], no globally optimal solution is available. Without a globally optimal solution, it is impossible to fully benchmark existing solutions and develop enhancements.

This paper fills the above-mentioned gap by presenting a globally optimal solution for multi-beam antenna operation by explicitly considering mmWave specifics, including directional multi-beam antennas, signal propagation, and blockage. The optimization problem is first reduced to the special case of multi-period variable cost and size bin packing problem (BPP) having well-known numerical solution algorithms, such that one may not place any constraints on the number of the beams and their HPBWs. To account for multi-beam specifics, we select the optimization criterion to be the ratio of the amount of occupied resources to the overall resources in the system. We then proceed to formulate heuristic algorithms capable of approaching the globally optimal solution. We also benchmark heuristics proposed so far in the literature against the developed optimal solution.

The main contributions of our study are:

- the optimal solution for multi-beam mmWave BS operation minimizing the amount of resources required to serve UEs based on multi-period variable cost and size bin packing problem;
- the heuristic algorithms characterized by polynomial complexity and allowing to achieve close approximations of the optimal solution;
- assessment of the maximal deployment density of NR BSs required to satisfy a given intensity of multicast UEs.

The rest of the paper is organized as follows. Related work is covered in Section 2. Section 3 details the system model utilized in our work. The optimization problem is formulated in Section 4, where we also introduce heuristic algorithms. Illustrative results and algorithms' performance comparison are delivered in Section 5. Conclusions are drawn in the last section.

2 Background and Related Work

In this section, we review related work. We start addressing the current state of multicast support in 3GPP NR systems. Then, we proceed by outlining the solutions proposed so far for single-beam antennas. We conclude this section by exposing the gaps related to advanced multi-beam antenna design.

2.1 NR Multicasting

Multicasting in NR systems is expected to entirely reuse the physical layer of unicast NR to increase the possibility of accelerated commercial application of multicast communications [4]. The 3GPP has defined two modes of Multimedia Broadcast/Multicast Service (MBMS) operation: the broadcast and multicast modes. The multicast mode enables unidirectional point-to-multipoint (PMP) transmission of multimedia data from a single source to a group of users in a multicast area. In multicast mode, the network defines a relevant multicast zone and can selectively transmit data to those cells in the area that contain members of the multicast group [21].

A multicast service may involve one or more successive multicast sessions. Such service might, for example, consist of a single ongoing session (e.g., a multimedia stream) or may involve several intermittent multicast sessions over an extended period (e.g., messages). Unlike the broadcast mode, the multicast service can only be received by users subscribed to the specific service and joined the multicast group associated with this service. The subscription may be managed by the Public Land Mobile Network (PLMN) operator, the user, or a third party.

In the two recent 3GPP NR releases, Rel-15 and Rel-16, no support for broadcast/multicast NR functionality is provided. The architectural enhancements to the 5G system to support multicast functionality are presented in 3GPP Rel. 17 [4]. The high-level architecture for the 5G Multicast-Broadcast Service (MBS) considers only NR as radio access technology (RAT) in the new-generation radio access network (NG-RAN). The physical layer is limited to the current Rel-15 numerologies, physical channels (PDCCH/PDSCH), and waveforms. As general guidelines, the overall impact of multicast support in NR should be kept limited, and the UE complexity should be minimized (for instance, one should avoid exposure to the device's hardware) to facilitate implementation and deployment functions. Moreover, flexible allocation of resources between unicast and broadcast/multicast services should be possible.

The sequence to establish and deliver an MBS session is the following [4]: (i) optional delivery of 5G MBS service information from the service layer to the 5G Core network (CN); (ii) UEs participate in receiving the MBS flow by requesting to join an MBS session; (iii) establishment of MBS flow transport; this step may happen before step (ii) for individual UEs joining an ongoing MBS session; (iv) MBS data delivery to UEs; (v) UEs stop receiving MBS flow; (vi) MBS transport is released.

From the viewpoint of 5G CN, two delivery methods are defined [4], as illustrated in Fig. 1. In both

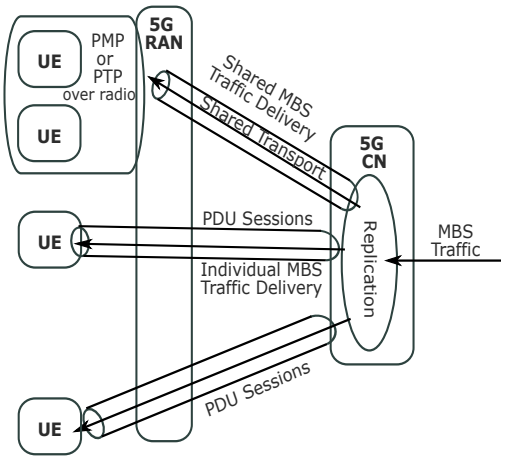


Fig. 1. Illustration of the delivery methods [4].

cases, 5G CN receives a single copy of MBS data packets, and then the operation of the two methods diverges. According to the first so-called 5GC Individual MBS traffic delivery method, 5G CN delivers separate copies of those packets to individual UEs via per-UE protocol data unit (PDU) sessions. The second, 5GC Shared MBS traffic delivery method, which is the focus of our study, assumes that 5G CN delivers the MBS data packet to a RAN node, which then multicasts it to one or multiple UEs. If the 5GC Individual MBS traffic delivery method is supported, the same received copy of MBS data packets by the 5G CN may be delivered via both 5GC Individual MBS traffic delivery method for some UE(s) and 5GC Shared MBS traffic delivery method for other UEs.

From the viewpoint of RAN, the following two delivery methods are available for the transmission of MBS packets over the radio access interface, in the case of shared MBS traffic delivery [4]. According to the point-to-point (PTP) delivery method, a RAN node delivers separate copies of the MBS data packet over the radio to individual UEs. In the case of the PMP delivery method, a RAN node delivers a single copy of MBS data packets over the radio to a set of UEs. Note that a RAN node may use a combination of PTP/PMP to deliver an MBS packet to UEs.

2.2 Related Studies

2.2.1 Single-Beam Antennas

The problem of multicast group formation and associated optimal resource utilization in wireless systems with directional antennas has received considerable attention so far. In [16], a heuristic group-aware multicast scheme (GAMS) aimed at system throughput maximization is proposed for IEEE 802.11ad networks. Specifically, multicast beamforming is performed during an association beamforming training interval in 802.11ad beacon. First, devices are classified into different multicast groups by combining only those UEs whose distances are smaller than a reference value. Second, the beamwidth is obtained for the multicast group by utilizing the law of cosine with respect to the coordinates of two edge UEs. When the beamwidth is found, the optimal data rate for multicast transmission is determined using single lobe antennas. To this aim, the farthest UE is found, and then the optimal data rate according to the MCS table satisfying the power constraint is determined.

An alternative algorithm for multicast grouping is presented in [17], wherein the beamwidth is adaptively determined based on the users' locations and the requested data rates in view of maximizing the sum rate of devices. This approach assumes an exhaustive search. The simulation results show that the multicast grouping scheme presented in [17] can improve the overall throughput by 28% to 79% compared with the conventional multicast schemes.

In [15] and more recently in [22], a multicast transmission strategy for mmWave in NR is proposed that aims to find an optimal trade-off between serving many users simultaneously, thus reducing the BS's resource consumption (channel usage time) and achieving high SNR by sweeping narrow beams. Unlike the aforementioned studies, the unreliable channel nature is explicitly accounted for, and the number of packets transmitted within the beam is optimized. In [15], the authors investigate optimal and suboptimal multicast schemes for mmWave communications with single lobe antenna patterns. Particularly, the problem has been solved using a Markov Decision Process. Because of the super-exponential complexity of the optimal solution, the authors propose a practical hierarchical optimization strategy.

2.2.2 Multi-Beam Antennas

Compared to single-beam antenna configurations, the problem of multicast group formation and associated optimal resource allocation for multi-beam antennas has received much less attention so far. In [18], a trade-off between multicasting and beamforming is investigated by designing greedy algorithms with performance guarantees that generate and schedule multi-lobe antenna patterns. To this end, switched beamforming antennas are utilized, where a set of pre-determined beams cover the entire azimuth of 360°. The authors consider both continuous (Shannon capacity) and discrete rate functions under two power allocation models, where the power is either equally split (EQP) or asymmetrically split (ASP) between the lobes. Both optimal and heuristic solutions are designed for the continuous rate function, while for the discrete rate case, the greedy solutions (e.g., GRASP2) are provided. The objective in [18] is to minimize the total time required for data dissemination to the multicast users, assuming 100% guarantee of packet delivery for all the users. In the continuous rate case, the greedy solution (GREP) provides near-optimal performance, almost coinciding with the optimal one.

A similar multicast system with switched beamforming antennas is considered in [19]. For the EQP model, the authors provide a low-complexity, dynamic-programming-based optimal solution (the algorithm's complexity is $O(B^2)$, in comparison with the $O(B^7)$ complexity of the optimal solution in [18], where B is the total number of nonoverlapping single-lobe beams) for both continuous and discrete rate functions. Under the ASP model, there exist no optimal nor approximate solutions. In the case of discrete rate function, the multicast-beamforming problem studied in [18] can be converted to a generalized version of the bin-packing problem. This allows applying generalized-bin-packing algorithms to obtain asymptotic polynomial-time approximation schemes. In [19], an asymptotic approximation solution has been developed for discrete rate functions. The solution enables drastic improvement over GRASP2 in [18], which handles ASP for the discrete rate case.

Differently from [18], [19], a joint user scheduling and adaptive beamformer design problem for multi-lobe antenna pattern to minimize the time of the data dissemination to the multicast users is addressed in [20]. The problem is stated to be non-convex and NP-hard for both discrete and continuous versions. Thus, obtaining an optimal beamformer with general channel vectors is not feasible, even for a small number of users. For this purpose, the authors propose efficient algorithms implemented in an adaptive beamforming system for multicasting (ADAM) and suitable for a practical system design.

Summarizing the related work, we observe that for multi-beam antenna systems, there are no optimal solutions simultaneously providing multicast group formation and resource allocation for the practical case of discrete rate function with adaptive beamforming. As a result, the performance of the heuristic solutions proposed for such systems cannot be reliably benchmarked, providing a numerical assessment of how close those solutions are to the optimal one.

3 System Model

In this section, we introduce our system model by specifying deployment, traffic, resource models, and radio part parameters. We then present our optimization criterion. The notation utilized in the paper is provided in Table 1.

3.1 Deployment, Traffic, and Resources

We assume a cellular deployment of NR BSs with the intersite distance of D meters and consider a randomly chosen sector (referred to as a "cell") of an NR BS having three sectoral antennas, each serving an area of 120°, as illustrated in Fig. 2. The height of UEs, NR BS, and blockers are assumed to be constant and given by h_U, h_A, and h_B, respectively. We consider a single multicast session to be provided to K UEs.

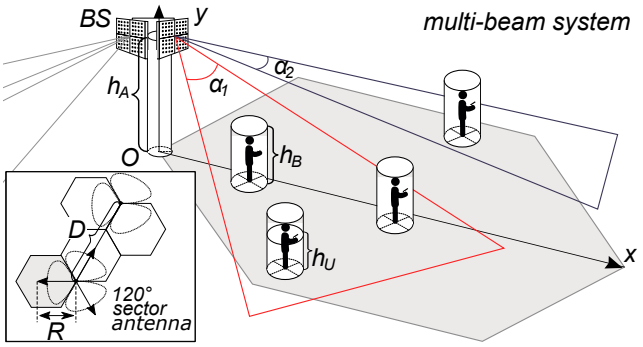


Fig. 2. Illustration of the deployment scenario.

The geometric locations of UEs are assumed to be uniformly distributed in the cell area. The bitrate of the multicast session is assumed to be C Mbps. Depending on UE locations, the amount of resources needed to serve UEs might be different and can be computed using NR MCS [24].

The bandwidth available for a sector antenna is assumed to be W MHz. Following the 3GPP NR

TABLE 1
Notation and parameters used in this work.

Fixed parameters with default values		
Parameter	**Definition**	**Value**
f_c	Carrier frequency, GHz	28 GHz
W	Available bandwidth, MHz	50 MHz
h_A	Height of NR BS, m	10 m
h_U	Height of UEs, m	1.5 m
h_B	Height of blockers, m	1.7 m
μ	5G NR numerology	3
M	Number of time slots in 1 ms subframe	8
L	Number of beams in the system	1,3,5
$P_{\max}$	Total available power, W	33 dBm
G_A, G_U	Antenna array gains at NR BS and UE ends, dBi	var/5.57 dBi
N_0	Power spectral density of noise, dBm/Hz	-174 dBm/Hz
N_A, N_U	Number of antenna elements at NR BS and UE	var/4 el
M_I	Interference margin	3 dB [23]
K	Number of multicast users	2-30
C	Bitrate of multicast session, Mbps	25 Mbps
w_{PRB}	Size of PRB, MHz	1.44 MHz
Δ	Subcarrier spacing, MHz	0.12 MHz
S_{th}	SINR threshold, dB	-9.47 dB
R_b	Number of available PRBs	32
R	Service (cell) area radius, m	250 m
Intermediate parameters		
Parameter	**Definition**	
$L(y)$	Path loss in linear scale	
$L_{dB}(y)$	Path loss in decibel scale	
X_A, Y_A	Coordinates of NR BS	
X_U, Y_U	Coordinates of UEs	
D	BSs intersite distance, m	
y	Three-dimensional distance between UE and NR BS, m	
y_{2D}	Two-dimensional distance between UE and NR BS, m	
$\theta^{\pm}_{3db}$	Upper and lower 3-dB points of antenna array, $^\circ$	
θ_m	Location of array maximum, $^\circ$	
β	Antenna array orientation, $^\circ$	
A_i, ζ	Propagation coefficients	
α	HPBW of a linear antenna array, rad	
$p_B(y)$	Distance-dependent blockage probability	
$S(y)$	Signal-to-interference-plus-noise ratio, SINR, dB	
P_A	Transmit power, W	
s_j	Spectral efficiency of the worst user in group G_j, bit/s/Hz	
Q	Number of carriers in a time slot	
c_j	Channel gain-to-noise ratio for beam j	
h_j	Channel gain for beam j	
σ_j	Standard deviation of the noise for beam j	

TABLE 2
5G NR numerologies [24].

μ	$\Delta = 2^{\mu} \cdot 15$ [kHz]	Bandwidth per RB [kHz]	TTI [ms]	Time slots, M
0	15	180	1	1
1	30	360	0.5	2
2	60	720	0.25	4
3	120	1440	0.125	8
4	240	2880	0.0625	16

standard [24], the resources are divided in time and frequency following the orthogonal frequency division multiple access (OFDMA) scheme. NR utilizes the scalable numerology that determines the subcarrier spacing, the number of slots in a subframe, the slot duration, and the cyclic prefix, see Table 2. We consider mmWave band with carrier frequency of $f_c = 28$ GHz and corresponding NR numerology $\mu = 3$ with the physical resource block (PRB) size of $w_{\text{PRB}} = 1.44$ MHz. The duration of a subframe is 1 ms, and it consists of exactly M time slots [24]. The number of carriers in each time slot is $Q = \lceil W/1.44 \rceil$. For the considered numerology, the number of subcarriers in each time slot is 12, and the number of time slots, M is set to 8 [24].

3.2 Propagation and Blockage Models

The SINR at the receiver located at the distance of y from the NR BS along the propagation path is

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

where P_A is the NR BS transmit power, G_A and G_U are the antenna array gains at the NR BS and the UE ends, respectively, N_0 is the power spectral density of noise, W is the operating bandwidth, $L(y)$ is the linear path loss. We capture the interference from the adjacent NR BSs via interference margin M_I in (1). For a given NR BS deployment density, one may estimate it by employing stochastic geometry-based models [25], [26].

Following [27], the path loss measured in dB is

$$L_{dB}(y) = 32.4 + 21 \log_{10} y + 20 \log_{10} f_c, \quad (2)$$

where f_c is the carrier frequency in GHz and y is the three-dimensional (3D) distance between the NR BS and the UE. By concentrating on the averaged traffic load and channel conditions, we omit the consideration of small-scale fading. Nevertheless, the framework provided in what follows allows utilizing more complex models to capture propagation conditions. For example, the small-scale fading can be added to the model by assuming certain fading phenomena, such as Rayleigh, Rician, Nakagami-m, or Weibull phenomena [28]. Those fading channels include multipath scattering effects, time dispersion, and Doppler shifts that arise from relative motion between the transmitter and receiver. Note that the introduction of an additional random variable to the considered propagation model, i.e., $P_R = FAy^{\gamma}$, where F follows the desired distribution, will affect the results quantitatively while preserving the same qualitative trend.

We assume that blockers might temporarily block the line-of-sight (LoS) path between the UE and the NR BS. Depending on the current link state (LoS non-blocked or blocked) and the distance between the NR BS and the UE, the session employs an appropriate MCS to maintain reliable data transmission. The attenuation due to the human body blockage is assumed to be 15 dB [29].

The path loss in the form of (2) can be represented in the linear scale by utilizing the model in the form of Ay^{ζ}, where A and ζ are the propagation coefficients. Introducing the coefficients (A_1, ζ) and (A_2, ζ) that correspond to LoS non-blocked and blocked conditions, we have

$$A_1 = 10^{2 \log_{10} f_c + 3.24}, A_2 = 10^{2 \log_{10} f_c + 4.74}, \zeta = 2.1. \quad (3)$$

We note that the considered model can be extended to a model with building blockages and corresponding LoS/nLoS states. To this purpose, one may introduce the coefficients (A_1, ζ_1), (A_1, ζ_2), (A_2, ζ_1), and (A_2, ζ_2) that correspond to LoS non-blocked, nLoS non-blocked, LoS blocked, and nLoS blocked conditions, respectively with $\zeta_1 = 2.1$, $\zeta_2 = 3.19$.

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

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

where $p_B(y)$ is the blockage probability at the 3D distance y [14], which is calculated as

$$y = \sqrt{(X_A - X_U)^2 + (Y_A - Y_U)^2 + (h_A - h_U)^2}, \quad (5)$$

where (X_A, Y_A, h_A) and (X_U, Y_U, h_U) are the coordinates of the NR BS and the multicast user, respectively.

3.3 Antenna Model

We consider planar antenna arrays at both NR BS and UEs. Following [30], [31], we assume 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, see Fig. 2. 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 [32] as

$$\alpha = 2|\theta_m - \theta_{3db}|, \quad (6)$$

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 given by $\theta_m = \arccos(-\beta/\pi)$, where β is the phase excitation difference affecting the physical orientation of the array. We assume $\theta_m = \pi/2$ for $\beta = 0$.

The gain over the HPBW can be found as in [32]:

$$G = \frac{1}{\theta_{3db}^{+} - \theta_{3db}^{-}} \int_{\theta_{3db}^{-}}^{\theta_{3db}^{+}} \frac{\sin(N\pi\cos(\theta)/2)}{\sin(\pi\cos(\theta)/2)} d\theta, \quad (7)$$

where the upper and the lower 3-dB points are

$$\theta_{3db}^{\pm} = \arccos[-\beta \pm 2.782/(N\pi)], \quad (8)$$

and N is the number of antenna elements.

Generally, it is cost-efficient to build a transceiver having fewer digital transceivers than total antennas. Analog beamforming [33] is a method to reduce the number of transceivers. Here, multiple active antennas are linked to each transceiver, and a network of analog phase shifters controls the signal phase on each antenna. In analog beamforming, each transceiver creates one beam directed at one user. The number of transceivers can be configured to be significantly less than the number of antennas when the number of simultaneously served users is rather small. Digital beamforming may be used to enable multiple data stream precoding on top of analog beamforming to improve performance [34].

The advantages of digital beamforming include (i) improved dynamic range, (ii) controlling of multiple beams, and (iii) better and faster control of amplitude and phase. Meanwhile, hybrid analog and digital beamforming is a promising candidate for large-scale mmWave multiple-input multiple-output (MIMO) systems because of its ability to significantly reduce the hardware complexity of the conventional fully-digital beamforming schemes while being capable of approaching the performance of fully-digital methods.

We consider that more than one beam can be simultaneously generated at NR BS. Each of these beams might be steered in a different direction, i.e., we assume hybrid analog-digital or digital beamforming techniques to enable multibeaming (multiple beams can be used independently at the same time). Recall that to enable this functionality, multiple radio frequency (RF) chains, proportional to antenna elements in use, need to be provided. We also assume that the HPBW of these beams depends on the number of the involved antenna elements to form a beam and is lower bounded by (6). Note that there are multiple antenna arrays connected to a separate RF chain, and the HPBW of each beam corresponds to the number of elements in the corresponding array. The HPBW can be approximated by using $\alpha = 102/N$ [35][1]. We denote the maximum number of beams by L and the total power available by $P_{\max}$. HPBWs and gains of beams computed according to $102/N$ and (7) are provided in Table 3.

1. The table with comparison of direct HPBW calculation according to (6) and its approximation can be found in [36].

TABLE 3
Parameters induced by 5G NR BS antenna arrays.

Array	HPBW,° [35]	Gain, dBi	Gain, linear
64x4	1.59	17.59	57.51
32x4	3.18	14.58	28.76
16x4	6.37	11.57	14.38
8x4	12.75	8.57	7.20
4x4	25.5	5.57	3.61
2x4	51	2.643	1.84
1x4	102	2.58	1

3.4 Optimization Criterion

Contrary to other studies that mainly examine the problem of multicast rate maximization, to optimize the mmWave NR resource utilization with multi-beam directional antennas, we consider the ratio of occupied resources to the overall set of available resources, ρ, as the optimization criterion. The rationale beyond this choice is that in the context of network design, resource consumption is one of the most crucial aspects for future systems [37]. We emphasize that in the case of a single RF chain, when the transmission can be performed over one beam at a time, the problem can be formulated as minimization of the amount of PRBs utilized. However, PRBs minimization might not provide actual resource minimization for a system with multiple antennas since the power has to be split among the beams. In this case, the increase in the number of beams adds new resources to the system. However, these resources might not be fully utilized due to maximum emitted power constraints.

As a result of the optimization, the following two metrics must be simultaneously determined: (i) ρ – the ratio of the occupied to available resources, (ii) L_{opt} – the optimal number of beams in the multi-beam system. By solving the optimization problem, we also determine the intersite distance D and, thus, η – the minimum NR BS deployment density required for multicast service provisioning.

We note that in real deployments, both multicast and unicast sessions may coexist in the system. In this paper, we omit unicast connections in the optimization problem and also limit our attention to a single multicast session. The rationale for choosing exactly one multicast session for this work is twofold. First, we would like to study the accuracy of the proposed heuristic algorithms without "external disturbances" that are always present in the system serving the mixture of traffic types and multiple competing multicast sessions. Secondly, to model the system with both unicast and multicast traffic, one needs to take additional assumptions as multicast traffic is known to occupy resources of the system more aggressively and eventually, under high multicast load conditions may fully occupy the system resources as we have shown in [8]. Nevertheless, the approach proposed in this paper can be extended to capture multiple

multicast sessions and the presence of unicast sessions as well.

For the extension of the model with the unicast traffic, one may need to introduce priorities between unicast and multicast traffic, such as unicast maximization, equal sharing, and equal competition, or as proposed in [38]. Note that the chosen prioritization scheme heavily depends on the operator. Assuming that multicast traffic is prioritized, the task reduces to optimizing multiple multicast sessions (e.g., according to the designed framework) and then allocating the remaining resources for unicast sessions according to some algorithms. Alternatively, if unicast is prioritized, the situation is reversed. Note that in both cases, the solution may not exist due to the lack of resources.

4 Optimization Framework

In this section, we first introduce our optimization framework. We start by commenting on the class of the considered problem. Then, we formalize the problem of optimal resource allocation of UEs in a multi-beam environment as a bin packing formalism [39]. Finally, we introduce our heuristic algorithms.

4.1 Preliminaries

We formalize the multi-beam operation optimization problem as a special class of BPP, one of the most studied combinatorial optimization problems. Generally, in BPP, a set of items of various sizes has to be either packed into a minimum number of identical bins, filled in the most time-efficient way, or packed so that the items are distributed evenly. A new variant of BPP, the variable-sized bin packing problem, which aims to minimize the cost of assigning the items to the particular bins, is considered in [39]. More precisely, the authors present a BPP setting wherein the cost of assigning an item to a bin is explicitly accounted for and may, or may not, depend solely on the item's volume. In our work, we consider multicast users' groups as the items, whereas directional beams represent the bins. We aim to minimize the cost of assigning users to multicast groups covered by the directional beams.

General BPPs, wherein a given set of items with various sizes has to be packed into the fewest number of unit capacity bins, belong to the NP-hard problem (NP-complete for the decision version). Accordingly, we propose efficient heuristic algorithms to solve the problem at hand in Subsection 4.3.

We also note that the generic formalism in [39] aims to minimize the cost of selecting the bins and the costs of assigning the items specific bins, which may depend on other criteria than the volumes of the items. These provide additional possibilities to account for other different system properties, but we leave them out of the scope of this paper.

4.2 Formalization of Optimization Task

We consider the 5G NR BS sectoral coverage served by an antenna array system having $L \geq 1$ beams and K users, denoted by set $\mathcal{K} = \{1, ..., K\}$. We assume an OFDMA scheme, i.e., M designates the time horizon's length, the number of time slots in the time horizon (one subframe), with the index $t \in \mathcal{T}$, $\mathcal{T} = \{1, \ldots, M\}$ of each time slot. We consider K users that utilize antenna arrays featuring multiple elements forming directional radiation patterns. The maximum number of PRBs available in the system is MLR_b, where M is the maximum number of time slots in the time horizon, L is the number of beams per antenna at each time slot t, and R_b is the available number of resource blocks in the system for the beam at time slot t. The potential maximum number of groups served within the time horizon is restricted by ML.

As one may deduce, there are $2^K - 1$ options to assign K users to multicast groups [15], i.e., $\mathcal{K}_j$ represents the set of users forming group j, $j \in \mathcal{J}, \mathcal{J} = \{1, \ldots, 2^K - 1\}$, and $|\mathcal{K}_j|$ is the number of users in group j. For example, for $K = 3$ users, these feasible options are $\mathcal{K}_1$=$\{1\}$, $\mathcal{K}_2$=$\{2\}$, $\mathcal{K}_3$=$\{3\}$, $\mathcal{K}_4$=$\{1, 2\}$, $\mathcal{K}_5$=$\{1, 3\}$, $\mathcal{K}_6$=$\{2, 3\}$, $\mathcal{K}_7$=$\{1, 2, 3\}$. We can combine these groups to form the so-called "suits", i.e., subsets $\mathcal{G}_k$ in such a manner that each $\mathcal{G}_k$ covers all the users without their repetition, $k = 1, 2, ..., |\Omega|$, where Ω is the set of all such suits. Therefore, suits $\mathcal{G}_k$ satisfy the following conditions:

$$\bigcup_{j \in \mathcal{G}_k} \mathcal{K}_j = \mathcal{K}, \tag{9}$$

$$\mathcal{K}_{j_1} \bigcap \mathcal{K}_{j_2} = \emptyset, j_1 \neq j_2, \ \forall j_1, j_2 \in \mathcal{G}_k, \tag{10}$$

meaning that each multicast user has to be served individually in a separate group. Thus, for $K = 3$, set Ω includes suit $\mathcal{G}_1 = \{1, 2, 3\}$, covering K groups with one user in each, i.e., $\mathcal{G}_1 \sim \mathcal{K}_1 \bigcup \mathcal{K}_2 \bigcup \mathcal{K}_3$, which corresponds to the unicast transmissions. Another extreme is to serve a user by only one group. This corresponds to the suit $\mathcal{G}_5 = \{7\}$, which is included in set Ω and containts all the users, i.e., $\mathcal{G}_5 \sim \mathcal{K}_7$. We emphasize that the directionality of the beam, which covers each multicast group, is already included in the definition of $\mathcal{K}_j$. Namely, the set of users in the group j defines distance L_j from the BS to the farthest user and HPBW α_j.

For $L = 1$, all the groups included in suit $\mathcal{G}_k$ are served consistently by one beam. For $L > 1$, we define subsuits $\mathcal{G}_k^l$ as the subset of groups from $\mathcal{G}_k$, which is planned for beam l by the scheduler, $\mathcal{G}_k^l \subseteq \mathcal{G}_k$, $l = 1, 2, ..., L$. Therefore, suits $\mathcal{G}_k^l$ satisfy the following conditions:

$$\mathcal{G}_k = \bigcup_{l=1}^{L} \mathcal{G}_k^l, \tag{11}$$

$$\mathcal{G}_k^{l_1} \bigcap \mathcal{G}_k^{l_2} = \emptyset, l_1 \neq l_2, \ \forall l_1, l_2 \in \{1, 2, ..., L\}. \tag{12}$$

To model the optimization problem of serving users by a suit of groups (served by directional beams), we take a binary indicator, $g_j^t \in \{0,1\}$, to denote the group assignment decision variable. Let $g_j^t = 1$ if group j is served at time slot t, and $g_j^t = 0$ otherwise. Then, we have a vector-indicator, $\mathbf{g}^t = (g_1^t, \ldots, g_{|\mathcal{J}|}^t)$, of groups that are served at time slot t.

We assume the following constraint on the maximum number of groups to be served (or beams to be swept) in the system at each time slot t:

$$\sum_{j \in \mathcal{G}_k} g_j^t \leq L, \forall t \in \mathcal{T}, \tag{13}$$

meaning that at time slot t at most L beams can be simultaneously swept (or groups that can be served).

The model does not restrict the scheduler's beam assignment, however, a suit service time should not exceed the time horizon:

$$\sum_{j \in \mathcal{G}_k^l} \sum_{t \in \mathcal{T}} g_j^t \leq M, \forall l = 1, ..., L, \forall k = 1, ..., |\Omega|. \tag{14}$$

Let P_j be the transmit power of beam that serves group j, i.e., $P_j \leq P_{\max}, \forall j \in \mathcal{J}$, whereas α_j and L_j are, respectively, the width and length of the beam determined by the number of N_j antenna elements used to form the radiation pattern of an antenna, $\alpha_j = f(N_j)$. We assume that $\alpha_j = 102/N_j$ [35], G_A, G_U are in linear scale, and then, the required P_j for each beam to serve the user located at distance L_j from the BS is

$$P_j = \frac{A_1 A_2 S_{th}(N_0 W + M_I)}{G_{\mathrm{A}} G_{\mathrm{U}} L_j^{\zeta} \left[A_2(1 - p_B(L_j)) + A_1 p_B(L_j)\right]}, \tag{15}$$

where S_{th} is the SINR threshold corresponding to a chosen NR MCS, whereas G_A depends on α_j.

Further, in our optimization problem, we have to ensure the following constraint to be held on the transmit power budget per antenna that serves group j:

$$\sum_{j \in \mathcal{G}_k} g_j^t P_j \leq P_{\max}, \forall t \in \mathcal{T}. \tag{16}$$

The SINR at UE can be written as (4). We assume that the session requires a constant bit rate C. Technically, to determine the amount of resources required from NR BS to serve a session with bit rate C, we have to know the channel quality indicator (CQI) and MCS values as well as SINR to spectral efficiency mapping. As these parameters are usually vendor-specific, in our study, we use MCS mappings from [40].

Then, cost a_j is represented in terms of the number of PRBs for the assigned beam for group j, such as $a_j = f(P_j, N_j, C)$, where C is the required session bit rate and P_j is the transmit power. That is,

$$a_j = \frac{C}{s_j w_{\mathrm{PRB}}}, \tag{17}$$

where s_j is a spectral efficiency in bit/s/Hz of the farthest user in the group j, w_{PRB} is a PRB size.

Note that the scheduler's time slot assignment is reflected in vector $\mathbf{g}_j = (g_j^1, \ldots, g_j^M)$ with

$$\sum_{t \in \mathcal{T}} g_j^t = \left\lceil \frac{a_j}{R_b} \right\rceil, j \in \mathcal{J}. \tag{18}$$

We assume that the scheduler assigns a beam to the group such that the following holds true

$$a_j \leq M R_b, j \in \mathcal{J}. \tag{19}$$

Finally, in constraints (14) and (19), the following condition for the maximum available resources in the system should be satisfied:

$$\sum_{j \in \mathcal{G}_k} a_j \leq M L R_b, j \in \mathcal{J}, k = 1, ..., |\Omega|. \tag{20}$$

The goal of the model is to determine the optimal grouping of multicast users, which minimizes the total cost of service in terms of the ratio of occupied PRBs to the total available number of PRBs for the entire time horizon. We now proceed with specifying the objective functions for two cases, $L \geq 1$ and $L = 1$.

4.2.1 Multi-Beam Antennas Optimization

In the case $L \geq 1$, we have to minimize the ratio of occupied to available resources. Thus, the optimization problem takes the following form:

$$\min_{k \in 1, \ldots, |\Omega|} \sum_{j \in \mathcal{G}_k} \frac{a_j}{M L R_b}, \tag{21}$$

$$\text{s.t. } (9), (10), (11), (12), (13), (14), (16), (19), (20).$$

4.2.2 Single-Beam Antennas Optimization

In the case $L = 1$, all the transmit power at BS is allocated to a single beam, $P_j = P_A$, and we can utilize the optimization problem defined above or use the conventional resource minimization task [17], i.e.,

$$\min_{k \in 1, \ldots, |\Omega|} \sum_{j \in \mathcal{G}_k} a_j, \tag{22}$$

$$\text{s.t. } (9), (10), (13), (14), (16), (19), (20).$$

The pseudo-code in Algorithm 1 describes the globally optimal solution according to (21) for $L \geq 1$ and (22) for $L = 1$. The algorithm employs our analytical framework to obtain optimal multicast group formation and resource allocation in Subsection 4.2.

4.3 Proposed Heuristic Solutions

Algorithm 1 is characterized by exponential complexity. To provide a practical algorithm with reduced computational requirements, we now proceed with proposing a heuristic algorithm for the case $L \geq 1$.

The proposed algorithm is divided into the following two stages: 1) grouping users into subgroups (see Subsection 4.3.1) and 2) beam assignment and power allocation (Subsection 4.3.2). The latter stage is also logically divided into the following steps: (i) selection

Algorithm 1: Optimal Solution, $L \geq 1$

1 **Input:** $(X_U(i), Y_U(i), h_U), i \in \mathcal{K}$
2 **Output:** Optimally global solution for multicast grouping
3 Create $2^K - 1$ multicast groups of users
4 **for** *each group* $\mathcal{K}_j$ **do**
5 find the farthest user i and the distance from BS to this user: $y \leftarrow \max_{i \in \mathcal{K}_j} y_i$ as (5);
6 find HPBW needed to cover the group $\mathcal{K}_j$ $\alpha_j = \arccos\left(\frac{(X_U(i)X_U(i')+Y_U(i)Y_U(i')+h_U^2}{y(i)y(i')}\right)$ ▷ *α_j is chosen according to the angle between two edge users i and i'*
7 find P_j from (15) using $L_j = y$; ▷ *$P_j = P_A$ is fixed for $L = 1$*
8 find the cost a_j from (17);
9 **end**
10 Solve the problem by using (21) with exhaustive search for $L \geq 1$ or (22) for $L = 1$.

of the multicast groups, which have to be served at a time slot simultaneously, (ii) water-filling stage for detecting the maximum power allocation that can be assigned to all of the beams simultaneously, and (iii) the subsequent refinement of the allocations for selected beams. We emphasize that starting from the second stage, we consider multi-beam transmissions, which implies that the power-budged constraints (16) per antenna have to be satisfied. In other words, for the single-beam antennas, $L = 1$, only the first stage is required, whereas in the case of $L \geq 1$ additional steps have to be performed.

4.3.1 Subgroups Formation

At this stage, we create subgroups to serve all users in the system during a time horizon. This process can be carried out in two ways as described below.

Option 1.1. To facilitate beam assignment, we extend the incremental multicast grouping algorithm for directional mmWave networks originally proposed in [17] to the case of $L > 1$. Note that for $L = 1$, we execute the method presented in [17] with the only modification on the optimization function. More specifically, we determine the number and width of the beams required to optimize the multicast transmission performance in terms of resource utilization. The pseudo-code is presented in Algorithm 2. The output of the algorithm contains the number of multicast groups, n, required to serve set $\mathcal{K}$, $1 \leq n \leq |\mathcal{J}|$; the set of multicast groups, $\mathcal{S}_1^M, ..., \mathcal{S}_n^M$, that covers all UEs from $\mathcal{K}$; and required power, $P_1^M, ..., P_n^M$.

Let us denote the set of users in the multicast session as $\mathcal{A}$. Initially, we set $\mathcal{A}$ to the set of all multicast users in the system $\mathcal{K}$ (line 3). We also introduce the 3D distance-vector $\mathbf{y} = (y_1, y_2, ..., y_i, ..., y_K)$, where each element represents the distance between the NR BS antenna and user i as per (5), where i is the index of the user, as well as vector $\mathbf{\Phi} = (\phi_1, ..., \phi_K)$ counting reference angles in the azimuth plane (lines 4-5). Line 7 sets the number of utilized resources to 0. The algorithm iteratively partitions users of set $\mathcal{A}$ into multiple subgroups, as indicated in line 9. Specifically, line 10 sets the minimization function to infinity. Here, the minimization function is assumed to represent the occupied per user resources for each multicast subgroup. The algorithm starts with choosing the farthest user from set $\mathcal{A}$ with distance y and its reference angle ϕ_y in the azimuth plane (lines 12-13).

Then, we utilize adaptive beamforming, and one beam pattern can be selected to transmit with a chosen MCS depending on the user's location. Line 15 collects all users covered by beam with width α steered toward the device of reference angle ϕ_y (corresponding to θ_m, see Section 3.3) with distance y in the multicast subgroup $\mathcal{S}_\alpha$. Note that the transmit power for each beam with width α is calculated according to (15) substituting L_j with y for $L \geq 1$. In the case of $L = 1$, the transmit power equals to the maximum available power P_A. Recall that for $L > 1$ differently from [17], we consider the minimization ratio of occupied to available resources as the objective function (line 18). Here, s_α is a spectral efficiency for a beam with width α and corresponds to s_j in (17). Thereby, the algorithm selects the best α for the chosen user in line 12 and deletes all the served users from the list (line 29). When all users are served, the algorithm stops.

Option 1.2. Another approach for group formation is based on the optimization function and is as follows. First, the algorithm selects the farthest user i, identifies the group $\mathcal{K}_j$ from $\mathcal{J} = \{1, ..., 2^K - 1\}$, such that $i \in \mathcal{K}_j$ to serve at the smallest value of $a_j/|K_j|$. The rationale is that by choosing the farthest user from the multicast group, the algorithm can capture more users when sweeping the beam. Further, to provide the solution of less complexity while preserving the intention to minimize the ratio of occupied to available resources, we select the beam with the smallest value of utilized resources per user. Then, we delete served users from the list and repeat the process for the remaining users. We emphasize that the groups containing the served users are also deleted. By doing this, we significantly reduce the complexity, see Section 4.4, while keeping comparable performance with the optimal solution, as later discussed in Section 5.

4.3.2 Beam assignment and Power Allocation

The pseudo-code of step 2 is provided in Algorithm 3. In what follows, we elaborate on the rationale and details of the algorithm.

Let $\mathcal{S}^M$ denote the set of multicast groups being selected at the first step of the proposed heuristic. The algorithm's goal is to determine the groups, which will be served simultaneously, and the corresponding transmit power to minimize the ratio of occupied to

Algorithm 2: Modified Incremental Multicast Grouping [17], $L \geq 1$

1 **Input:** $(X_U(i), Y_U(i), h_U), i \in \mathcal{K}$
2 **Output:** n; $\mathcal{S}_1^M, ..., \mathcal{S}_n^M$; $P_1^M, ..., P_n^M$;
3 $\mathcal{A} \leftarrow \mathcal{K}$, $\mathcal{K} = \{1, \ldots, K\}$;
4 $\mathbf{y} = (y_1, ..., y_K)$ as (5);
5 $\mathbf{\Phi} = (\phi_1, ..., \phi_K)$; ▷ *reference angles*
6 $n \leftarrow 0$; ▷ *subgroups counter*
7 $a_{\text{sum}} \leftarrow 0$; ▷ *occupied resources collector*
8 $\mathcal{S}_n^M \leftarrow \emptyset$;
9 **while** $\mathcal{A} \neq \emptyset$ *or* $a_{sum} < MLR_b$ *or* $n < ML$ **do**
10 $\quad \text{MIN}_Q \leftarrow \infty$;
11 $\quad n \leftarrow n + 1$;
12 $\quad y \leftarrow \max_{i \in \mathcal{A}} y_i$;
13 $\quad \phi_y \leftarrow \phi(y)$;
14 $\quad$ **for** $\alpha \in \Omega_\alpha = \{\alpha_{\min}, ..., \alpha_{\max}\}$ **do**
15 $\quad\quad \mathcal{S}_\alpha = \{i \in \mathcal{A} : \phi_y - \alpha/2 \leq \phi_i \leq \phi_y + \alpha/2\}$;
16 $\quad\quad$ calculate P_α from (15);
17 $\quad\quad$ **if** $P_\alpha \leq P_{\max}$ **then**
18 $\quad\quad\quad Q_\alpha = \frac{C}{s_\alpha w_{\text{PRB}} |\mathcal{S}_\alpha|}$;
19 $\quad\quad\quad$ **if** $MIN_Q > Q_\alpha$ **then**
20 $\quad\quad\quad\quad \text{MIN}_Q \leftarrow Q_\alpha$;
21 $\quad\quad\quad\quad \mathcal{S}_n^M \leftarrow \mathcal{S}_\alpha$;
22 $\quad\quad\quad\quad P_n^M \leftarrow P_\alpha$;
23 $\quad\quad\quad\quad a_n \leftarrow \frac{C}{s_\alpha w_{\text{PRB}}}$;
24 $\quad\quad\quad$ **end**
25 $\quad\quad$ **else**
26 $\quad\quad\quad$ go to line 29;
27 $\quad\quad$ **end**
28 $\quad$ **end**
29 $\quad \mathcal{A} \leftarrow \mathcal{A} \setminus \mathcal{S}_n^M$;
30 $\quad a_{\text{sum}} \leftarrow a_{\text{sum}} + a_n$;
31 **end**
32 **return** n, $\mathcal{S}_1^M, ..., \mathcal{S}_n^M$, $P_1^M, ..., P_n^M$.

available resources. Thus, the algorithm works until all groups are deleted from $\mathcal{S}^M$ (lines 5-22). We also use $\mathcal{D}^{(m)}$ to denote a set of groups to be served at a time slot m. The algorithm selects the worst (in the sense of the needed power) group from $\mathcal{S}^M$ and adds this group to set $\mathcal{D}^{(m)}$ (lines 7-9). If the power budget constraint allows us to add more groups to set $\mathcal{D}^{(m)}$, the algorithm selects the best group and adds it to set $\mathcal{D}^{(m)}$ (lines 10-19). The number of groups in $\mathcal{D}^{(m)}$ should be less or equal to L. When set $\mathcal{D}^{(m)}$ is determined, the power water-filling algorithm chooses the power such that the utilized resources are minimized (line 20).

Option 2.1. Traditional power water-filling. We now introduce $c_j = |h_j|/\sigma_j^2$ as a channel gain-to-noise ratio (GNR), where h_j is a channel gain, and σ_j is a standard deviation of the noise. In a traditional water-filling algorithm, the channel with high c_j receives more power, which leads to a higher system capacity. However, this approach would eventually

Algorithm 3: Heuristic Step 2, $L > 1$

1 **Input:** $\mathcal{S}_1^M, ..., \mathcal{S}_n^M$; $P_1^M, ..., P_n^M$;
2 **Output:** m, $\mathcal{D}^{(m)}$, $P_j^{*(k)}$, $j = 1, ...n$, $k = 1, ...m$;
3 $\mathcal{S}^M \leftarrow \{\mathcal{S}_1^M, ..., \mathcal{S}_n^M\}$;
4 $m \leftarrow 0$; ▷ *time slot counter*
5 **while** $\mathcal{S}^M \neq \emptyset$ **do**
6 $\quad m \leftarrow m + 1$;
7 $\quad k_{\max} \leftarrow \arg\max_{j \in \mathcal{S}^M} P_j$;
8 $\quad P_{\text{sum}} \leftarrow P_{k_{\max}}^M$;
9 $\quad \mathcal{D}^{(m)} \leftarrow \mathcal{S}_{k_{\max}}^M$;
10 $\quad$ **if** $\mathcal{S}^M \setminus \mathcal{D}^{(m)} \neq 0$ **then**
11 $\quad\quad$ **for** $j = 2 : L$ **do**
12 $\quad\quad\quad k_{\min} \leftarrow \arg\min_{j \in \mathcal{S}^M \setminus \mathcal{D}^{(m)}} P_j$;
13 $\quad\quad\quad$ **if** $P_{sum} + P_{k_{\min}}^M \leq P_{\max}$ **then**
14 $\quad\quad\quad\quad \mathcal{D}^{(m)} \leftarrow \mathcal{D}^{(m)} \cup \mathcal{S}_{k_{\min}}^M$;
15 $\quad\quad\quad$ **else**
16 $\quad\quad\quad\quad$ go to line 20;
17 $\quad\quad\quad$ **end**
18 $\quad\quad$ **end**
19 $\quad$ **end**
20 $\quad$ Perform water-filling for $\mathcal{D}^{(m)}$ and obtain $P_j^{*(m)}$ from (23)-(25);
21 $\quad \mathcal{S}^M \leftarrow \mathcal{S}^M \setminus \mathcal{D}^{(m)}$;
22 **end**
23 **return:** m, $\mathcal{D}^{(k)}$, $P_j^{*(k)}$, $j = 1, ...n$, $k = 1, ...m$.

lead to equal power distribution. Note that GNR is related to the SINR as $S_j = P_j^M c_j$, $S_j = \min_{i \in \mathcal{D}^{(m)}} S(y_i)$, $j = 1, ..., |\mathcal{D}^{(k)}|$, $k = 1, ..., m$.

The power allocations of the water-filling approach are the result of the following optimization task for the optimal power $P_j^{*(k)}$ for group j at time slot k:

$$\left(P_1^{*(k)}, ..., P_{|\mathcal{D}^{(k)}|}^{*(k)}\right) \leftarrow \max_{\left(P_1, ..., P_{|\mathcal{D}^{(k)}|}\right)} \sum_{j=1}^{|\mathcal{D}^{(k)}|} \log(1 + P_j c_j),$$

$$\text{s.t. } P_j \geq 0, \forall j = 1, \ldots, |\mathcal{D}^{(k)}|, \sum_{j=1}^{|\mathcal{D}^{(k)}|} P_j = P_{\max},$$

$$\forall k = 1, ..., m, \tag{23}$$

where $|\mathcal{D}^{(k)}|$ is the number of multicast groups that have to be served simultaneously at time slot k, $|\mathcal{D}^{(k)}| \leq L$. Note that the first constraint implies that the power allocation is non-negative, while the second constraint limits the power budget of the system. The sought optimal transmit power $P_j^{*(k)}$ is thus

$$P_j^{*(k)} = (1/\xi^* - 1/c_j)^+, \tag{24}$$

where $1/\xi^*$ is the maximum power that can be allocated for each multicast group, $x^+ = \max(x, 0)$.

The problem in (23) is convex in nature. Since the maximization of concave function (23) is equivalent

to minimization of a convex function, we have

$$\xi^* \leftarrow \min_{\xi} \sum_{j=1}^{|\mathcal{D}^{(k)}|} \log\left(1 + P_j^{*(k)} c_j\right) - \xi\left(\sum_{j=1}^{|\mathcal{D}^{(k)}|} P_j^{*(k)} - P_{\max}\right), \forall k = 1, ..., m, \quad \text{s.t. (24).} \tag{25}$$

Option 2.2. Resource-based filling. Alternatively, we implement water-filing based on the resource information. According to this option, the additional power is allocated to those groups, resulting in the largest decrease in the amount of utilized resources.

4.4 Complexity Analysis

The complexity of Algorithm 1 (optimal solution) is exponential since branch-and-bound or branch-and-cut even with relaxations are performed using the exhaustive search.

The computational complexity of the heuristic solution in Algorithm 2 is given by $O(K|\Omega_\alpha|)$, where K is the complexity due to the "while" cycle over all K multicast users in the worst case of the unicast transmission (lines 9-31). This means each user will be placed in a separate group. For the second component, which is inside the "while" cycle, $|\Omega_\alpha|$ is the complexity due to the possible beam selection from the set Ω_α (lines 14-28). As a result, in the worst case, the number of operations is in $O(K|\Omega_\alpha|)$.

Finally, the computational complexity of Algorithm 3 is $O(KL)$, where K is the complexity due to the "while" cycle over all n groups (with max K) in the worst case of the single group transmissions (lines 5-22). For the second component, which is inside the "while" cycle, $L-1$ is the complexity due to the possible selection of simultaneous groups (lines 11-18). As a result, in the worst case, the number of operations is $O(K(L-1))$. The traditional water-filling algorithm has $O(2(|\mathcal{D}^{(m)}|-1))$ complexity [41], where $|\mathcal{D}^{(m)}|$ is the number of multicast groups that have to be served simultaneously ($|\mathcal{D}^{(m)}| \leq L$), thus, the number of operations is in $O(L-1)$ for the water-filling. Here, the water-walling is also performed inside the "while" cycle, therefore, the complexity of Algorithm 3 is $O(K[(L-1)+(L-1)]) = O(KL)$.

5 Numerical Results

In this section, we evaluate the performance of the proposed multi-beam antenna optimization strategies. We first study the special case with a single beam and assess the accuracy of the algorithms. Then, we investigate optimal multicast group formation and resource allocation for a system with multiple antennas and evaluate the accuracy of the proposed heuristics. Here, we also assess the optimal usage of resources and the optimal number of antenna beams. Finally, we determine the maximal deployment density of NR BSs required to satisfy a given density of multicast UEs. We comment on the practical use of the proposed framework in the last part of this section and also provide results for antenna radiation patterns, propagation and interference models different from those defined in Section 3. The default system modeling parameters are summarized in Table 1.

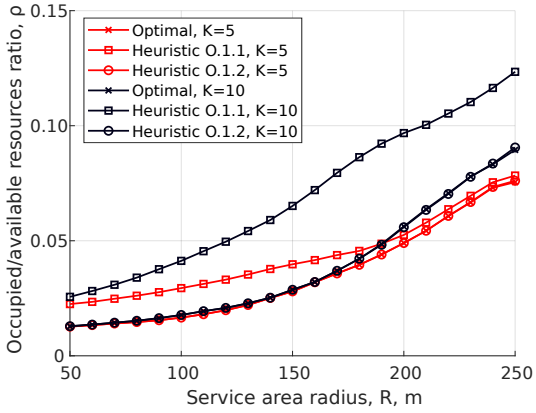


Fig. 3. Ratio of occupied to available resources as a function of the cell radius, $K = 5, 10$ users, $C = 25$ Mbps, $W = 50$ MHz.

5.1 Single-Beam Antennas Design

We start with addressing the case of single-beam antenna systems. A comparison of the ratio of occupied to available resources for $L = 1$ obtained with the developed optimization model and proposed heuristics, (O.1.1) [17] and (O.1.2), see Section 4.3, is demonstrated in Fig. 3 as a function of the cell radius, R, for the session data rates of $C = 25$ Mbps. As one may observe, there is almost a perfect match between the optimal solution and the proposed heuristic (O.1.2). At the same time, the heuristic (O.1.1) leads to a noticeably higher ratio of occupied to available resources, ρ, compared to both the heuristic (O.1.2) and the optimal solution. The underlying reason is the group formation algorithm. Specifically, the (O.1.1) algorithm does not utilize an exhaustive search resulting in much lower complexity. However, the resulting number of groups is usually higher compared to the exhaustive search algorithm employed in (O.1.2). Particularly, the algorithm selects the farthest user and then, for this user, sweeps the beam. As this user can be located at any place in the cell area, it might not be possible to cover all the users with a single beam. Inversely, until R reaches approximately 230 m, the optimal solution forms a single multicast group for the case of $L = 1$.

Analyzing the effect of the cell radius in Fig. 3, one may learn that for both the considered numbers

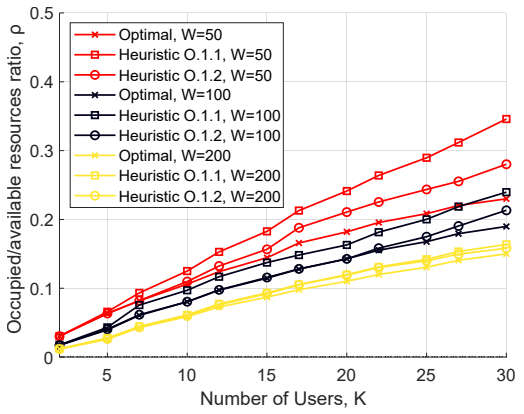


Fig. 4. Ratio of occupied to available resources as a function of the number of users, $R = 250$ m, $C = 25$ Mbps, $W = 50, 100, 200$ MHz.

of users (i.e., $K = 5$ and $K = 10$), an increase in the cell radius leads to faster growth of the ratio of occupied to available resources for all solutions for $K = 5$ users as compared to the case of $K = 10$ users. We also emphasize that the gap between (O.1.1) and the optimal solution is smaller for the lower values of K. Specifically, for $K = 10$ and all the considered cell radii, the difference between optimal solution and (O.1.1) heuristic is around 100%, while for $K = 5$, it gradually converges to zero as R increases. The rationale is that both optimal and heuristic solutions start to select more groups, progressively shifting to unicast transmissions due to the large distances.

We now proceed with assessment of the effect of the number of users, K, shown in Fig. 4 for cell radius of $R = 250$ m, requested rate of $C = 25$ Mbps, and three bandwidths, $W = 50$, $W = 100$, and $W = 200$ MHz. Note that for values of K higher than 20, we utilize quadratic extrapolation to construct the curves for the optimal solution. Analyzing the presented data, one may notice that the increase in the number of users leads to a rise in the ρ ratio for all the considered solutions. Indeed, higher values of K theoretically lead to either a higher number of groups or a higher number of users in the group (which may worsen the multicast group channel condition), thus increasing the ratio of utilized to available resources. Further, with the increase in K, the gap between the optimal and heuristic (O.1.2) solutions becomes larger. The impact of the increase in the available bandwidth W is also evident from Fig. 4. One may learn that for larger bandwidth of $W = 200$ MHz, the gap between the optimal and heuristic solutions is lower compared to $W = 100$ and $W = 50$. The rationale is that the data transmission is much faster with a larger bandwidth. This, in turn, leads to a lower ratio of occupied to available resources. Therefore, this is an inherently quantitative effect as the difference between smaller values of ρ for a larger bandwidth is lower compared to the difference for a smaller bandwidth. The gap between optimal solution and heuristics in terms of percentage is represented in Table 4.

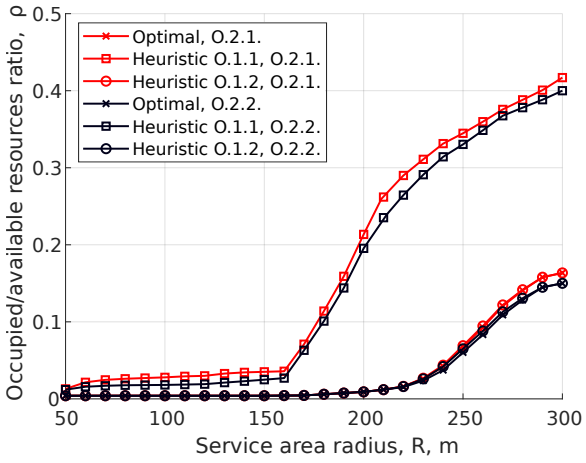
5.2 Multi-beam Antennas Design

Having studied the performance of the single-beam systems, we are in the position to proceed with the performance results of the multi-beam systems. We start with Fig. 5 presenting the ratio of occupied to available resources, ρ, for the maximum number of beams $L = 3$ and $L = 5$ as a function of the cell area radius R. From these illustrations, we observe that the curves for $L = 3$, Fig. 5(a), grow much slower with the increase in the cell radius than for $L = 5$, Fig. 5(b). It is important to highlight that at smaller values of R (e.g., approximately 50-100 m), heuristic (O.1.2) and optimal solutions combine users into a single group. This explains the fact that the curves for $L = 5$ first show better performance and then demonstrate higher ρ values for all schemes. We also note that the reason behind the gap between the optimal solution and (O.1.2) for (O.2.1) and (O.2.2) heuristic options for $L = 5$ lies in the selected number of beams per time slot. More precisely, at R of approximately 150-250 m optimal solution utilizes one beam and several time slots, whereas heuristic solutions serve users with more than one beam within one time slot. Hence, we may deduce that at large distances, such as 150-240 m, it is crucial to utilize one beam at a time to minimize ρ. Note that all the considered strategies utilize unicast mode to serve multicast users starting from around $R = 250$ m.

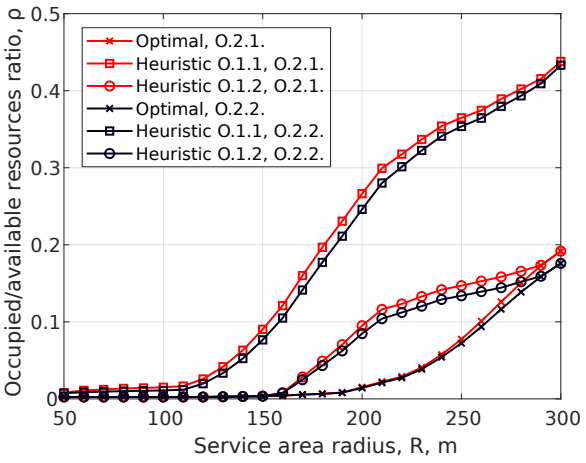
Analyzing the presented data further, one may also observe no significant difference between the types of power water-filling schemes, i.e., options (O.2.1) and (O.2.2), with the latter slightly outperforming the former approach. This modest superiority is intuitive

TABLE 4
The gap between optimal solution and heuristics in percentage

%/N	2	5	7	10	12	15	17	20	22	25	27	30
O.1.1., W=50	0.1	3.8	5.9	37.4	11.8	24.8	31.2	28.6	39.9	36	37.7	50.4
O.1.2., W=50	0	0.8	0	1	0	11	14	10.6	26	16	10.1	22
O.1.1., W=100	3.2	10.8	0.6	37.9	31.2	8.7	14.2	15	14.4	17	22	26.2
O.1.2., W=100	0	1.7	3.4	1.2	0.7	0	1.6	3.6	0	0	6.2	12.4
O.1.1., W=200	5.2	0	10.8	1.4	3.8	5.4	10.7	11.6	6.2	9.5	8.5	9.2
O.1.2., W=200	0	0	1.3	0.8	0.1	3.4	12.4	8.1	8.8	6.9	6.9	5



(a) $L = 3$.



(b) $L = 5$.

Fig. 5. Ratio of occupied to available resources as a function of cell radius, $K = 10$, $C = 25$ Mbps, $W = 50$ MHz.

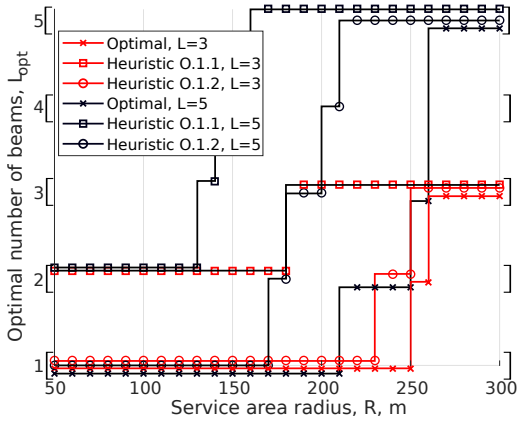


Fig. 6. Optimal number of beams in the multi-beam system as a function of the cell radius, $K = 10$ users, $C = 25$ Mbps, $W = 50$ MHz.

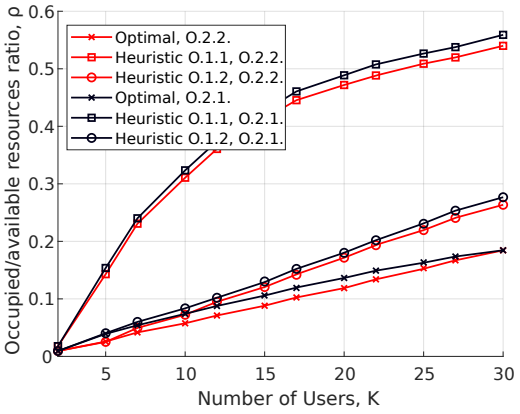


Fig. 7. The ratio of occupied to available resources as a function of number of users, $R = 250$ m, $C = 25$ Mbps, $W = 50$ MHz, $L = 3$.

and stems from the fact that water-filling (O.2.2) is based on the resource information feature. Similarly to $L = 1$, the heuristic option with exhaustive search (O.1.2) provides the best approximation of the optimal solution. However, as the maximum number of beams, L, increases, even this approximation starts to deviate from the optimal solution.

The abovementioned conclusions on the utilized number of beams are further complemented by Fig. 6, which demonstrates the optimal number of beams, L_{opt}, as a function of the cell area radius. One may observe that the optimal solution selects only one beam per time slot until R reaches 230 m and 250 m for $L = 5$ and $L = 3$ beams. Further, as one may learn from the curves for $L = 3$, the optimal solution chooses one beam and several time slots when R is in the range of 240-250 m, whereas the proposed heuristics (O.1.2) and (O.1.1) sweep two and three beams per time slot, respectively. Analyzing both Fig. 5 and Fig. 7, we can conclude that for the practical ranges of

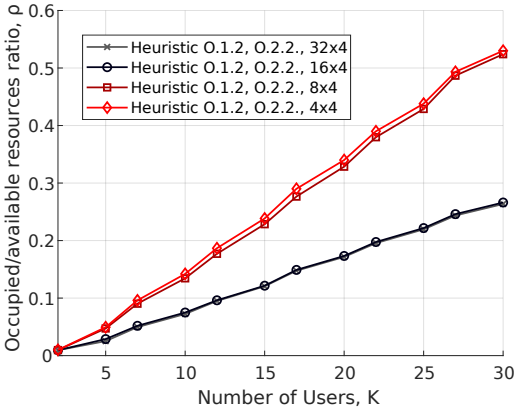


Fig. 8. The ratio of occupied to available resources as a function of number of users for different antenna arrays, $R = 250$ m, $C = 25$ Mbps, $W = 50$ MHz, $L = 3$.

cell size and considered number of users, the optimal solution always utilizes no more than 2-3 beams.

Similarly to the single-beam system, we now evalu-

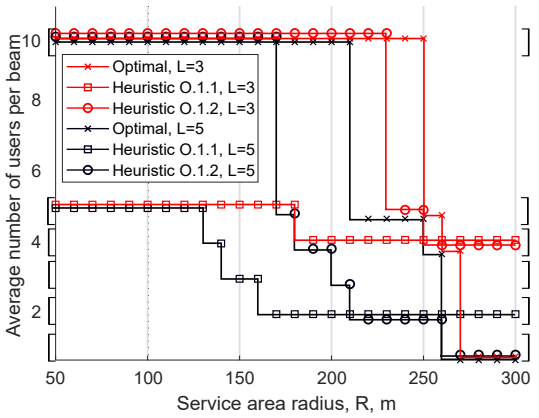


Fig. 9. Average number of users per beam as a function of the cell radius, $K = 10$ users, $C = 25$ Mbps, $W = 50$ MHz.

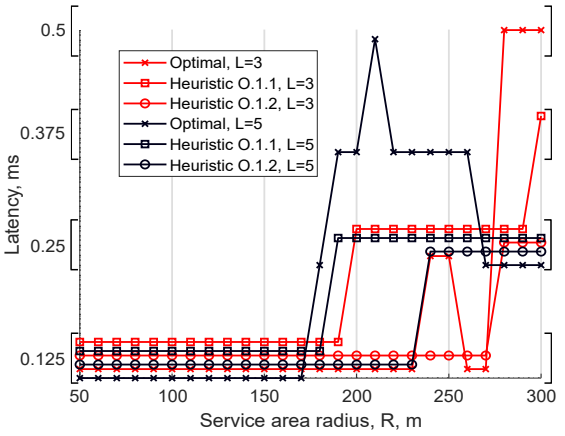


Fig. 10. Latency as a function of the cell radius, $K = 10$ users, $C = 25$ Mbps, $W = 50$ MHz.

ate the impact of the number of users on the optimal resource allocation. To this aim, Fig. 7 offers the ratio of occupied to available resources, ρ, for $L = 3$ as a function of the number of users in the system. As one may notice, the observations from Fig. 4 related to the rise of the gap between the optimal and (O.1.2) heuristic solutions with the increase in the number of users are confirmed by Fig. 7. Moreover, one may deduce by comparing the results of Fig. 4 and Fig. 7 that the heuristic solution (O.1.1) with both water-filling strategies for $L = 3$ works less efficiently than for the case of $L = 1$.

Going further, we assess the effect of the antenna array size on the optimal multicast grouping and resource allocation. To this end, Fig. 8 quantifies the ratio of occupied to available resources as a function of a number of users for different antenna arrays and the best-identified heuristic algorithm (O.1.2). Note that the smaller the number of antenna elements, the greater the occupied to available resource ratio, ρ. This is explained by the antenna directionality, which increases with the number of antenna elements forming the radiation pattern of the transmit antenna. Note that the reduction in the antenna array size does not affect the system performance for lower cell radius values R as the BS transmits using one wide beam.

We proceed with Fig. 9, which displays the average number of users served by a beam per time slot. The rationale for considering this metric is to assess the number of transmissions exploited to serve multiple users for various radii. The presented results confirm the statement derived from Fig. 5 that starting from 250 m almost all the schemes use the unicast mode for $L = 5$ beams. Hence, Fig. 9 provides an insight into the efficiency of the multicast transmissions in mmWave networks. More precisely, it reflects situations, where the system utilizes a lower resource ratio than that required by the unicast service, where users are serviced by individual beams (one user per beam). One may observe that the system with $L = 3$ beams

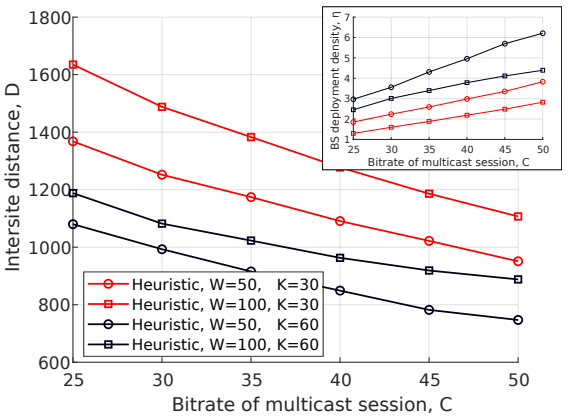


Fig. 11. NR BS intersite distance and deployment density as a function of session data rate, $K = 30, 60$, $W = 50, 100$ MHz, $L = 3$.

works better in terms of serving more users within a beam, which can be explained by the fact that, in general, the increase in the number of beams leads to a decrease in the number of users per beam.

Even though in this paper, we mainly concentrate on resource utilization, now we consider a different important metric for the system performance evaluation, which is the latency, as demonstrated in Fig. 10. Consistently with the results presented above, one may deduce that a single multicast group provides the best performance in terms of latency and utilized resources. This performance is achieved due to the absence of sequential service over multiple beams. By recalling the results presented in Fig. 6 and Fig. 9, we also may conclude that for the heuristic O.1.1 at radii distances of 50-150 m all the considered schemes occupy exactly one time slot, even if the number of used beams is more than one. The other considered strategies exploit one beam only at these distances.

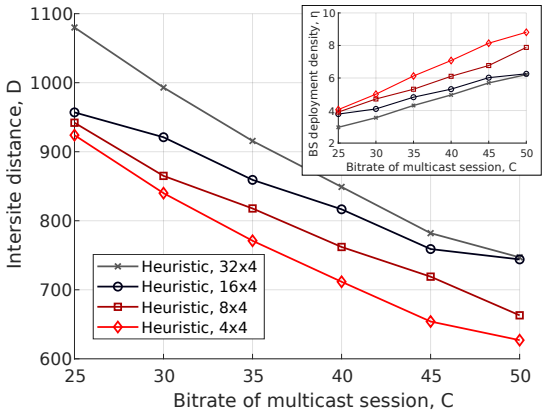


Fig. 12. NR BS intersite distance and deployment density as a function of session data rate for different antenna arrays, $K = 60$, $W = 50$ MHz, $L = 3$.

5.3 Deployment Density Assessment

We finally analyze the minimum NR BS deployment density required for multicast service delivery. To this end, in Fig. 11 we demonstrate the maximum intersite distance between NR BSs and the associated NR BS deployment density as a function of session bitrate for $K = 30, 60$, $W = 50, 100$ MHz, two heuristic solutions, (O.1.1) and (O.2.2) and the maximum number of $L = 3$ beams. Recall that the NR BSs intersite distance in the case of tri-sector antenna deployment is calculated as $D = 3R$ [42]. Fig. 11 allows us to obtain insights on the ideal extent of network densification for different values of available bandwidth W. One may learn that the NR BSs deployment density grows linearly with the number of users, while the system with a larger bandwidth guarantees lower deployment density.

Finally, we proceed with Fig. 12, where the maximum NR BS intersite distance and the NR BS deployment density are illustrated for different antenna array sizes. We note that a large intersite distance corresponds to the cells of a bigger size. In more detail, multicast users have to be served in unicast fashion at larger distances as only narrow beams can reach those users. However, the antennas with the lower number of elements fail to reach the farthest users, which is confirmed by Fig. 12.

5.4 Notes on the Practical Use of Algorithms

The algorithms presented in the paper need to be further adapted for use in practical scenarios. First, recall that to illustrate the behavior of the optimal solution and benchmark the heuristic algorithms, we omitted unicast connections in the optimization problem and limited our attention to a single multicast session. To extend the model with the unicast traffic, one has to specify additional service specifics such as priorities between unicast and multicast traffic, such as unicast maximization, equal sharing, and equal competition. These are heavily operator-dependent but can still be further incorporated into the model as briefly discussed at the end of Section 3.

Further, the critical parameter of the model is the choice of the optimization time horizon. Currently, it is chosen to coincide with the scheduling interval in NR, which is known to be 1 ms, and the model assumes that all the traffic demands need to be served in this interval. However, in practical systems, the traffic load may vary in time, and schedulers operating at the packet level may induce complex behavior such as delaying some packets for the next scheduling interval. These specifics need to be accounted for when choosing the time horizon for the proposed optimization algorithms. More specifically, it has to be chosen such that the average traffic demand remains nearly constant in time.

There are additional specifics of the models that need to be carefully aligned with realistic conditions. Here, the critical point is that the model assumes conical beam patterns parameterized with an angle of α coinciding with the HPBW of the beam. In practice, especially for multi-beam operation, radiation patterns are characterized by a more complex structure that may lead to better or worse performance of algorithms depending on user locations or beam directions. Specifically, performance degradation may happen when beams overlap in space, and some users may not have sufficient SNR even though the model states so. Thus, the proposed framework needs to be supplemented with practical algorithms allowing users to fallback to unicast service in these conditions.

Furthermore, the proposed framework allows the utilization of more complex sub-models than those considered in Section 3. To illustrate it and highlight the effect of different environmental and system parameters, Fig. 13 shows the considered metric of interest as a function of the distance when the impact of fast fading (we incorporated Rayleigh fading), more realistic antenna radiation patterns, and explicitly cal-

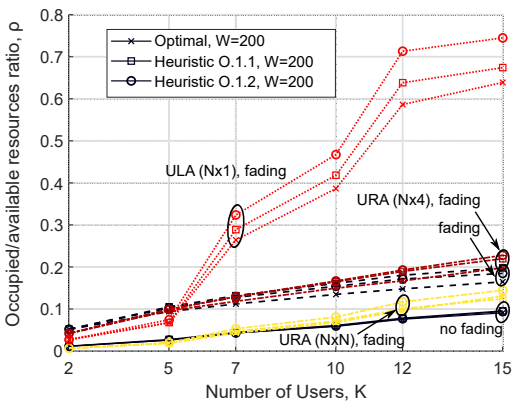


Fig. 13. Ratio of occupied to available resources as a function of the number of users, $R = 250$ m, $C = 25$ Mbps, 200 MHz.

culated interference are included. The antenna radiation patterns are now constructed in MATLAB using Sensor Array Analyzer (uniform rectangular arrays (URAs), Nx4 and NxN, and uniform linear arrays (ULAs), Nx1, where $N = [4, 8, 16, 32]$) that generally reflect the recommended patterns by 3GPP in TR 36.931), while the interference has been calculated by using the model from [30]. By analyzing these results, one may deduce that adding fading to the propagation model and also capturing interference and more realistic antenna radiation pattern leads to the increased requirements in terms of the ratio of occupied to available resources. The systems with added fading and interference for both the approximation and real radiation patterns (i.e., Nx4 URAs) show almost matching results in the metric of interest (see, "URA (Nx4), fading" and "fading" curves in Fig. 13). In contrast, the system with symmetrical URAs (i.e., NxN antenna elements) results in lower values of ρ due to higher antenna gains. Finally, the linear array is characterized by the highest ratio of the utilized to available resources. The rationale is smaller gains as compared to considered planar arrays due to the smaller total number of antenna elements (that is, Nx1).

Summarizing the presented results in Fig. 13, in general, the qualitative trends remain the same while the results slightly deviate quantitatively when fading, interference, and realistic arrays radiation patterns are taken into account. Thus, when applying the proposed framework, one has to account for specifics of the deployment and type of utilized equipment that is entirely feasible within the proposed performance optimization framework.

6 Conclusions

The capability of modern antenna arrays to utilize multiple beams simultaneously with potentially varying half-power beamwidth and asymmetric power allocation makes the problem of efficient multicast transmission in mmWave NR systems an extremely complex one. In this paper, we solve this problem by developing an optimal multicast grouping and resource allocation solution. The approach is based on a variable-sized bin packing problem and is thus NP-hard. We have developed several heuristics with different complexities and approximation accuracies to provide practical algorithms with reduced computational requirements.

In our numerical results, we utilize the developed optimal approach for benchmarking heuristic solutions. We show that a widely used group formation algorithm originally proposed in [16], [17] may drastically overestimate the amount of resources. The proposed exhaustive search group formation is nearly optimal but computationally intensive for large values of the number of users. The difference between the optimal and heuristic solutions increases with the number of users and the maximum number of supported beams by the antenna array and decreases with the amount of available bandwidth. The type of power allocation among the identified number of beams does not drastically affect the performance of the heuristic algorithms. Finally, for practical ranges of cell sizes and ranges of the number of users (10-50), the optimal amount of beams is always in the range of 2-3. For small cell radii, a single beam is almost always utilized, while unicast service is only feasible for higher ones. This makes the development of heuristic algorithms easier and levels down the requirements for practical antenna array implementations.

References

[1] Cisco VNI Forecast, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017–2022," 2019.

[2] 3GPP, "Release 15 Description; Summary of Rel-15 Work Items (Release 15)," 3GPP TR 21.915 V15.0.0, Oct 2019.

[3] ——, "Release 16 Description; Summary of Rel-16 Work Items (Release 16)," 3GPP TR 21.916 V0.6.0, Sept 2020.

[4] ——, "Study on Architectural Enhancements for 5G Multicast-Broadcast Services (Release 17)," TR 23.757 V1.2.0, November 2020.

[5] ——, "NG-RAN; Xn Application Protocol (XnAP)," 3GPP TR 38.423 (Draft), June 2018.

[6] Y. Gaidamaka and K. Samouylov, "Analytical Model of Multicast Network and Single Link Performance Analysis," in *Proc. of the 6-th International Conference on Telecommunications, Zagreb, Croatia*, 2001, pp. 169–175.

[7] O. Vikhrova, S. Pizzi, A. Iera, A. Molinaro, K. Samuylov, and G. Araniti, "Performance Analysis of Paging Strategies and Data Delivery Approaches for Supporting Group-Oriented IoT Traffic in 5G Networks," in *2019 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. IEEE, 2019, pp. 1–5.

[8] 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 Transactions on Wireless Communications*, vol. 19, no. 5, pp. 3421–3434, 2020.

[9] N. Chukhno, O. Chukhno, S. Pizzi, A. Molinaro, A. Iera, and G. Araniti, "Efficient Management of Multicast Traffic in Directional mmWave Networks," *IEEE Transactions on Broadcasting*, 2021.

[10] S. Ahmadi, *5G NR: Architecture, Technology, Implementation, and Operation of 3GPP New Radio Standards*. Academic Press, 2019.

[11] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "A Tutorial on Beam Management for 3GPP NR at mmWave Frequencies," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 173–196, 2018.

[12] E. Garro, M. Fuentes, J. Carcel, H. Chen, D. Mi, F. Tesema, J. Gimenez, and D. Gomez-Barquero, "5G Mixed Mode: NR Multicast-Broadcast Services," *IEEE Transactions on broadcasting*, vol. 66, no. 2, pp. 390–403, 2020.

[13] G. R. MacCartney, J. Zhang, S. Nie, and T. S. Rappaport, "Path Loss Models for 5G Millimeter Wave Propagation Channels in Urban Microcells," in *2013 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2013, pp. 3948–3953.

[14] 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 *2016 IEEE International Conference on Communications (ICC)*. IEEE, 2016, pp. 1–7.

[15] A. Biason and M. Zorzi, "Multicast Transmissions in Directional mmWave Communications," in *European Wireless 2017; 23th European Wireless Conference*. VDE, 2017, pp. 1–7.

[16] H. Park and C.-H. Kang, "A Group-aware Multicast Scheme in 60GHz WLANs," *TIIS*, vol. 5, no. 5, pp. 1028–1048, 2011.

[17] 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.

[18] 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*, 2009, pp. 205–216.

[19] H. Zhang, Y. Jiang, K. Sundaresan, S. Rangarajan, and B. Zhao, "Wireless Multicast Scheduling with Switched Beamforming Antennas," *IEEE/ACM Transactions on Networking*, vol. 20, no. 5, pp. 1595–1607, 2012.

[20] E. Aryafar, M. A. Khojastepour, K. Sundaresan, S. Rangarajan, and E. Knightly, "ADAM: An Adaptive Beamforming System for Multicasting in Wireless LANs," *IEEE/ACM Transactions on Networking*, vol. 21, no. 5, pp. 1595–1608, 2013.

[21] 3GPP, "Multimedia Broadcast/Multicast Service (MBMS); Stage 1 (Release 16)," 3GPP TS 22.146 V16.0.0, Jul 2020.

[22] 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.

[23] V. Begishev, D. Moltchanov, E. Sopin, A. Samuylov, S. Andreev, Y. Koucheryavy, and K. Samouylov, "Quantifying the Impact of Guard Capacity on Session Continuity in 3GPP New Radio Systems," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 12, pp. 12 345–12 359, 2019.

[24] 3GPP, "NR; Physical Channels and Modulation (Release 15)," 3GPP TR 38.211, Dec 2017.

[25] R. Kovalchukov, D. Moltchanov, A. Pyattaev, and A. Ometov, "Evaluating Multi-Connectivity in 5G NR Systems with Mixture of Unicast and Multicast Traffic," in *International Conference on Wired/Wireless Internet Communication*. Springer, 2019, pp. 118–128.

[26] 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. 10 064–10 069, 2018.

[27] 3GPP, "Technical Specification Group Radio Access Network; Study on Channel Model for Frequency Spectrum above 6 GHz (Release 14)," 3GPP TR 38.900 V14.2.0, Tech. Rep., December 2016.

[28] M. K. Simon and M.-S. Alouini, *Digital Communication Over Fading Channels*. John Wiley & Sons, 2005, vol. 95.

[29] G. R. MacCartney, T. S. Rappaport, and S. Rangan, "Rapid Fading Due to Human Blockage in Pedestrian Crowds at 5G Millimeter-Wave Frequencies," in *GLOBECOM 2017-2017 IEEE Global Communications Conference*. IEEE, 2017, pp. 1–7.

[30] 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.

[31] S. Singh, R. Mudumbai, and U. Madhow, "Interference Analysis for Highly Directional 60-GHz Mesh Networks: The Case for Rethinking Medium Access Control," *IEEE/ACM Transactions on networking*, vol. 19, no. 5, pp. 1513–1527, 2011.

[32] A. B. Constantine, *Antenna Theory: Analysis and Design*. Wiley-Interscience, 2005.

[33] W. Roh, J.-Y. Seol, J. Park, B. Lee, J. Lee, Y. Kim, J. Cho, K. Cheun, and F. Aryanfar, "Millimeter-wave Beamforming as an Enabling Technology for 5G Cellular Communications: Theoretical Feasibility and Prototype Results," *IEEE communications magazine*, vol. 52, no. 2, pp. 106–113, 2014.

[34] F. Sohrabi and W. Yu, "Hybrid Analog and Digital Beamforming for mmWave OFDM Large-Scale Antenna Arrays," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 7, pp. 1432–1443, 2017.

[35] C. A. Balanis, *Antenna Theory: Analysis and Design*. John wiley & sons, 2015.

[36] M. Gerasimenko, D. Moltchanov, M. Gapeyenko, S. Andreev, and Y. Koucheryavy, "Capacity of multiconnectivity mmwave systems with dynamic blockage and directional antennas," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 4, pp. 3534–3549, 2019.

[37] 5G AI, "European Vision for the 6G Network Ecosystem ," The 5G Infrastructure Association, Tech. Rep., June 2021.

[38] S. Pizzi, M. Condoluci, G. Araniti, A. Molinaro, A. Iera, and G.-M. Muntean, "A Unified Approach for Efficient Delivery of Unicast and Multicast Wireless Video Services," *IEEE Transactions on Wireless Communications*, vol. 15, no. 12, pp. 8063–8076, 2016.

[39] T. G. Crainic, F. D. Fomeni, and W. Rei, *The Multi-Period Variable Cost and Size Bin Packing Problem with Assignment Cost: Efficient Constructive Heuristics*. CIRRELT, 2019.

[40] J. Fan, Q. Yin, G. Y. Li, B. Peng, and X. Zhu, "MCS Selection for Throughput Improvement in Downlink LTE Systems," in *2011 Proceedings of 20th international conference on computer communications and networks (ICCCN)*. IEEE, 2011, pp. 1–5.

[41] Q. Qi, A. Minturn, and Y. Yang, "An Efficient Water-Filling Algorithm for Power Allocation in OFDM-based Cognitive Radio Systems," in *2012 International Conference on Systems and Informatics (ICSAI2012)*. IEEE, 2012, pp. 2069–2073.

[42] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) system scenarios (Release 15)," 3GPP 36.942 V15.0.0, Oct 2018.

Nadezhda Chukhno is an Early Stage Researcher at A-WEAR and Doctoral Researcher at Mediterranea University of Reggio Calabria, Italy and Jaume I University, Spain. She graduated from RUDN University, Russia, and received her B.Sc. in Business Informatics (2017) and M.Sc. in Fundamental Informatics and Information technologies (2019). Her current research activity mainly focuses on wireless communications, 5G+ networks, multicasting, D2D, and wearable technologies.

Olga Chukhno is an Early Stage Researcher within H2020 MCSA ITN/EJD A-WEAR project and a PhD student at Mediterranea University of Reggio Calabria, Italy and Tampere University, Finland. She received M.Sc. (2019) in Fundamental Informatics and Information Technologies and B.Sc. (2017) in Business Informatics from RUDN University, Russia. Her current research interests include wireless communications, social networking, edge computing, and wearable applications.

Dmitri Moltchanov received the M.Sc. and Cand.Sc. degrees from the St. Petersburg State University of Telecommunications, Russia, in 2000 and 2003, respectively, and the Ph.D. degree from the Tampere University of Technology in 2006. Currently he is University Lecturer in with the Laboratory of Electronics and Communications Engineering, Tampere University, Finland. He has (co-)authored over 150 publications on wireless communications, heterogeneous networking, IoT applications, applied queuing theory. In his career he has taught more than 50 full courses on wireless and wired networking technologies, P2P/IoT systems, network modeling, queuing theory, etc. His current research interests include research and development of 5G/5G+ systems, ultra-reliable low-latency service, industrial IoT applications, mission-critical V2V/V2X systems and blockchain technologies.

Antonella Molinaro graduated in Computer Engineering (1991) at the University of Calabria, received a Master degree in Information Technology from CEFRIEL/Polytechnic of Milano (1992), and a Ph.D. degree in Multimedia Technologies and Communications Systems (1996). She is currently an associate professor of telecommunications at the University Mediterranea of Reggio Calabria, Italy. Her research activity mainly focuses on wireless and mobile networking, vehicular networks, and future Internet.

Yuliya Gaidamaka received a PhD in 2001 and a Full Doctor of Sciences degree in 2017 in Mathematics from the Peoples' Friendship University of Russia (RUDN University). Since 2001, she has been an associate professor and currently a professor in the university's Applied Probability and Informatics Department. She is the author of more than 200 scientific and conference papers, co-author of three monographs on multiplicative solutions of finite Markov chains, matrix and analytical methods for performance analysis of wireless heterogeneous networks. Her current research focuses on performance analysis of 5G+ networks, queuing theory, mathematical modeling of communication networks including admission control, radio resource management using artificial intelligence.

Konstantin Samouylov received his Ph.D. degree from the Moscow State University and Doctor of Sciences degree from the Moscow Technical University of Communications and Informatics. During 1985–1996, he held several positions at the Faculty of Sciences of the Peoples' Friendship University of Russia where he became the head of the Telecommunication Systems Department in 1996. From 2014, he became the head of the Department of Applied Informatics and Probability Theory of RUDN University (previously named PFUR). During last two decades, Konstantin Samouylov has been conducting research projects for the Helsinki and Lappeenranta Universities of Technology, Moscow Central Science Research Telecommunication Institute, several Institutes of Russian Academy of Sciences and a number of Russian network operators. His current research interests are performance analysis of 4G networks (LTE, WiMAX), teletraffic of triple play networks, signalling network (SIP) planning, and cloud computing. He has written more than 150 scientific and technical papers and three books.

Yevgeni Koucheryavy received the Ph.D. degree from the Tampere University of Technology (TUT), Finland. He is currently a Professor at the Laboratory of Electronics and Communications Engineering, TUT. He is the author of numerous publications in the field of advanced wired and wireless networking and communications. His current research interests include various aspects in heterogeneous wireless communication networks and systems, the Internet of Things and its standardization, and nanocommunications. He is Associate Technical Editor of the IEEE Communications Magazine and an Editor of the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS.

Giuseppe Araniti (Senior Member, IEEE) received the Laurea degree and the Ph.D. degree in electronic engineering from the University Mediterranea of Reggio Calabria, Italy, in 2000 and 2004, respectively. He is currently an Assistant Professor of telecommunications with the University Mediterranea of Reggio Calabria. His major area of research is on 5G/6G networks and it includes personal communications, enhanced wireless and satellite systems, traffic and radio resource management, multicast and broadcast services, device-to-device (D2D), and machine-type communications (M2M/MTC).