Optimal Multicasting in Dual mmWave/μ Wave 5G NR Deployments With Multi-Beam Directional Antennas

Optimal Multicasting in Dual mmWave/μWave 5G NR Deployments with Multi-Beam Directional Antennas

Olga Chukhno, Nadezhda Chukhno, Dmitri Moltchanov, Antonella Molinaro, Anna Gaydamaka, Andrey Samouylov, Yevgeni Koucheryavy, Antonio Iera, and Giuseppe Araniti

***Abstract*—The design of multicast services in the fifth-generation (5G) New Radio (NR) deployments is hampered by the directional nature of antenna radiation patterns. This complexity is further compounded by the emergence of new deployment options, such as dual millimeter wave (mmWave) and microwave (μWave) base station (BS) deployments, as well as new antenna design solutions. In this paper, the resource allocation task for multicast services in dual mmWave/μWave deployments with multi-beam directional antennas is addressed as a multi-period variable cost and size bin packing problem. We solve this problem and characterize the globally optimal solution. To decrease complexity, we then propose and test the simulated annealing approximation and relaxation techniques, i.e., local branching and relaxation-induced neighborhood search heuristic. Our results show that for the considered system parameters, the properties of the optimal solution depend on the density of dual-mode BS deployment and BS deployment type. We observe a transition point at which the system shifts from primarily utilizing mmWave resources to exclusively using μWave BS. Furthermore, the optimal number of beams is upper limited by 3 for mmWave and by 2 for μWave BSs. The efficiency of resource utilization is also affected by the utilized numerology and technology selection priority. Finally, we show that the simulated annealing technique allows for decreasing the solution complexity at the expense of slightly overestimating the amount of resources.**

***Index Terms*—5G, New Radio, millimeter Wave, microwave, multicast, dual base stations, multi-beam antennas, optimization.**

I. Introduction

Nowadays, as the standardization of the fifth-generation (5G) New Radio (NR) technology is over, many operators have already deployed microwave (μWave) NR systems operating in $3.5-7.125$ GHz band. This band provides the so-called "coverage" layer and is expected to carry most traditional cellular mobile communication traffic. The next wave of NR rollout is anticipated to happen in the millimeter wave (mmWave) band, $24.25-52.6$ GHz, focusing on short-range and high data rate capabilities providing the so-called "capacity" layer [1]. Network operators are expected to deploy both technologies to offer ubiquitous data rate services at the air interface [2].

Both μWave and mmWave NR deployments are expected to rely on antenna arrays forming directional radiation patterns to compensate for path losses and extend the coverage range of NR base stations (BS). As a side effect, high antenna directivities utilized in these systems would efficiently suppress interference in dense deployments [3]. However, this property is also expected to drastically reduce the efficiency of multicast services as the use of highly directional antenna radiation patterns does not allow to serve all the users, which belong to the same multicast session via a single transmission [4]. The provisioning of multicast services [5], [6] is further complicated by the capabilities of modern antenna arrays supporting multiple beams simultaneously with varying half-power beamwidths (HPBW) as well as by the availability of multiple radio access technologies (RAT) that can serve a multicast group [7]. Finally, it can also be affected by the mmWave propagation, including blockage [8].

With the increased usage of high-bandwidth applications in mobile systems, efficient air interface utilization becomes vital. 5G NR has evolved since its introduction in Release 15 and continues to expand both the availability and applicability of 5G NR services [9]. Mainly, multicast functionality for NR will be introduced in Release 17, emphasizing group-scheduling mechanisms, among others [10]. However, despite this interest, to the best of our knowledge, optimal multicasting in new deployment options, such as dual mmWave and μWave BS deployments with multi-beam directional antennas, has not been investigated so far.

A. Related Work

Recently, there have been multiple attempts to address the multicast problem in directional systems for both single- and multi-beam designs. For the single-beam system, the optimal solution of the multicast problem formulated in [11] has super-exponential complexity in the number of users as it requires solving a Markov decision process with large state

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

O. Chukhno, N. Chukhno, D. Moltchanov, A. Gaydamaka, A. Samouylov, and Y. Koucheryavy are with Tampere University, Tampere, Finland. Email: {nadezda.chukhno, dmitri.moltchanov, anna.gaydamaka, andrey.samuylov, evgeny.kucheryavy}@tuni.fi

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

A. Iera is with University of Calabria, Italy and CNIT, Italy. Email: antonio.iera@dimes.unical.it

This work was supported by the European Union's Horizon 2020 Research and Innovation programme under the Marie Sklodowska-Curie grant agreement No. 813278 (A-WEAR; http://www.a-wear.eu/) and by the Academy of Finland projects "Machine learning algorithms for energy efficient and QoS aware communications in heterogeneous 6G mmWave/sub-THz networks" (ML6GThz) and "Enabling Mobile Terahertz Communication for 6G Cellular Networks" (EMERGENT).

and action spaces. Hence, a heuristic algorithm for optimal multicasting with a single lobe antenna pattern by considering delay-energy trade-off has been introduced in [11], [12]. Similarly, a heuristic resource management framework aimed at simultaneously minimizing energy and maximizing network throughput has been proposed in [13]. Heuristic approaches for multicast grouping have also been proposed in [14], [15]. We may conclude that for single-beam directional multicast systems, there are exact solutions proposed to date. However, most of the solutions are heuristic in nature. Also, there are no studies comparing the accuracy of heuristic approaches and the exact solutions.

Several studies have also focused on multi-beam antenna solutions, where, in addition to the group formation, the power budget has to be properly split among the beams. To this end, multicasting and switched beamforming trade-offs have been addressed in [16], [17]. 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. Under the more complex ASP model, neither optimal nor approximate solutions exist. Differently from [16], [17], adaptive beamforming techniques minimizing the time required to serve multicast users have been offered in [18]. The problem is stated to be non-convex NP-hard for both discrete and continuous rate versions. Thus, obtaining an optimal beamformer with general channel vectors is not feasible, even for a small number of users. In view of this, the authors propose heuristics suitable for a practical system design.

Further, a cooperative multicast scheme for mmWave systems utilizing non-orthogonal multiple access (NOMA) has been introduced in [19]. Then, analytical expressions for the signal-to-interference-plus-noise (SINR) ratio coverage probability are derived to evaluate the performance of the proposed heuristics. An analog-beam splitting heuristic approach for mmWave D2D multicast system has been proposed in [20], where the antenna elements are divided into groups composing an analog beam to serve a receiver. A heuristic solution that leverages an unsupervised machine learning algorithm for multicast grouping and exploits the D2D technology to deal with the blockages is proposed in [21]. By analyzing works focused on multi-beam multicast systems, we may deduce that those studies are plagued by a lack of optimal solutions. Our work builds on top of the study in [22], where the globally optimal solution for multicast with multi-beam antenna operation is presented. Unfortunately, the complexity of the solution does not allow scaling it for a realistic number of multicast users. Differently from [22], in this work, we consider a dual deployment of mmWave/μWave systems, propose an exact solution, extend it to the case of operators-specific priorities for RAT usage, and evaluate suggested heuristic techniques allowing to scale the system to the case of tens of users.

Integrated mmWave/μWave system deployments have been addressed in just a few recent studies. In [23], a dual-mode architecture has been investigated from the power and bandwidth allocation point of view to maximize the achieving sum rate and energy efficiency. The throughput and reliability of dual-mode multi-beam operation systems have been studied in [24]. More recently, in [25], [26], blockage mitigation, prediction, and beam management issues have been addressed. A cross-layer optimization for joint scheduling and transmit precoding has been introduced in [27]. In [28], the dual operation of mmWave/μWave system for unicast traffic has been studied. We emphasize that *there are no research studies so far proposing solutions for optimal multicasting in dual mmWave/μWave systems with multi-beam antennas*.

B. Contribution

Despite several techniques proposed for single-beam multicast systems (that can be readily adapted for single-beam in dual-mode systems), none of the authors addressed the problem of optimal multicasting in dual mmWave/μWave BS deployments with highly directional multi-beam antenna design. Both dual-mode BSs and highly directional multi-beam antennas offer improved efficiency potential and will certainly continue to be utilized in various applications. Due to the need to overcome the limitations of mmWave frequencies, improve network capacity using microwave frequencies, and optimize multicasting, research on optimal multicasting in dual mmWave/μWave BS with highly directional multi-beam antenna is imperative. Since these options are expected to be essential in future NR deployments, and only a small part of the literature analyzes them, we fill this gap in this work.

To this end, we adopt a general formalism of multi-period variable cost and size bin packing problem [29] that allows us to capture multi-beam antenna operation in both considered bands. We provide both the exact solution and approximations, including the simulated annealing approach, which is pioneered in new 5G NR deployment options with dual mmWave and μWave operations and multi-beam antenna design solutions. The optimization criterion accounts for multi-beam specifics and is selected as the fraction of utilized resources to the overall available resources, while the ultimate metric of interest is the density of dual mmWave/μWave BS deployments.

The main contributions of our study are:

- problem formalization and computation of the exact (globally optimal) solution for multicast optimization in dual-mode mmWave/μWave BS deployments with multi-beam antenna design under different RAT selection criteria (mmWave priority, μWave priority, and weighted optimization function);
- application of simulated annealing to solve the multicasting problem in 5G NR dual-mode mmWave/μWave systems in an efficient way;
- analysis of the exact optimal solution showing that for considered deployment and system parameters, there is an abrupt transition between the use of mmWave and μWave technology for mmWave RAT priority scheme while the maximum number of supported beams is 3 and 2 for mmWave and μWave RATs, respectively;
- numerical results showing that local speed-up techniques do not provide any noticeable impact on the exact solution while simulated annealing allows preserving polynomial

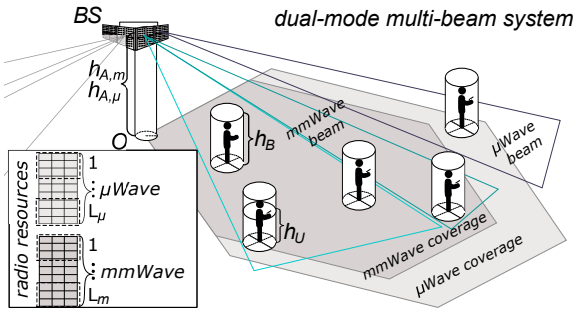


Fig. 1. Illustration of the considered dual-mode mmWave/μWave system.

time complexity for $30-60$ UEs in a multicast group at the expense of $10-40\%$ degradation in resource usage.

The rest of the paper is organized as follows. In Section II, we formulate the system model and assumptions. Further, in Section III, we develop our optimization framework and provide an exact solution algorithm. An approximate solution based on the simulated annealing approach is further developed in Section IV. Numerical results are given in Section V. Conclusions are drawn in the last section.

II. System Model and Assumptions

This section presents our system model and assumptions on deployment, traffic, propagation, blockage, and antenna models. We also introduce our optimization criterion. The notation used in the paper is provided in Table I.

A. Deployment and Traffic Models

We consider a tri-sector dual-mode co-located mmWave and μWave (sub-6 GHz) NR system deployment shown in Fig. 1 and concentrate on a randomly chosen sector, i.e., "cell". We investigate a downlink single multicast session provisioning to K user equipment (UE) devices. Locations of UEs are assumed to be uniformly distributed in the cell. The UEs, mmWave NR BS, and μWave NR BS heights are considered to be constant and given by h_U, $h_{A,m}$, and $h_{A,\mu}$, respectively.

To narrow down the considered use cases, we assume that the dual system operates in crowded open environments such as city squares, with significant traffic demands from pedestrian UEs during various happenings and festivals. This assumption affects the choice of the blockage and propagation models in what follows. Specifically, we consider the following radio part specifics: (i) small-scale blockage of the propagation paths between BS and UEs in mmWave band by small objects such as humans, (ii) large-scale blockage of the propagation paths by accounting for line-of-sight (LoS) and non-LoS (nLoS) conditions caused by buildings, (iii) BS and UE antenna directionality in both bands, (iv) multi-beam operation antennas in both bands via hybrid/digital beamforming, (v) propagation specifics resulting in principally different coverages in both bands. However, whenever addressing a specific scenario, additional impairments and propagation specifics can be added to the propagation model. Specifically, by utilizing the 3GPP LoS and nLoS model in addition to blockage and non-blockage states one can capture other types of deployments, where the LoS path can also be blocked by large-static buildings (nLoS/LoS states as defined by 3GPP). An example of such a model is provided in [30].

The BS operates in both the mmWave, 28 GHz, and μWave, 3.5 GHz, frequency bands simultaneously by utilizing separate antennas at transceivers. Each UE is also equipped with corresponding interfaces and can receive in both frequency bands. We assume the orthogonal frequency division multiple access (OFDMA) schemes at both interfaces. The available bandwidth is W_m MHz and W_μ MHz for mmWave and μWave BS, respectively. The bitrate of the multicast session is assumed to be C Mbps.

Definition. By following [31], we use the term *subgroup* to denote the subset of UEs belonging to the multicast group served by the same beam, whereas a *multicast group* contains all UEs interested in receiving a *multicast session* (i.e., data flow/content). With the term *suit* we imply a configuration of multicast subgroups that covers all UEs (i.e., a multicast group) without repetitions.

B. Blockage and Propagation Models

Compared to the mmWave band, μWave systems are much less susceptible to human body blockage effects, i.e., the induced human-body attenuation has been reported not to exceed $2-4$ dB [32]. For this reason, we neglect the blockage effect in μWave band. In mmWave band, the human body blockage attenuation is assumed to be 15 dB [33]. Here, the blockers are modeled as cylinders with height h_B and radius r_B, and their number follows a Poisson distribution with the density of λ_B per square meter. The blockage probability at the 3D distance y is determined according to [8]:

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

where λ_B is the blockers density, h_B and r_B are the blockers' height and radius, h_U is the UE height, $h_B \geq h_U$, h_A is the BS height (either mmWave or μWave BS). We denote by $h_{A,m}$ and $h_{A,\mu}$ mmWave and μWave BS heights, respectively.

To account for building blockage in city deployments, one also needs a building blockage model. The LoS probability for the 2D distance x between the mmWave BS and the UE, $p_L(x)$, can be obtained by using the 3GPP UMi street canyon model [34] as

$$p_L(x) = \begin{cases} 1, & x \leq 18\text{m}, \\ 18 + xe^{-\frac{x}{36}} - 18e^{-\frac{x}{36}}, & x > 18\text{m}. \end{cases} \quad (2)$$

To define a mmWave propagation model accounting for blockage caused by both buildings and humans, one may consider the UE in one of the four states: (LoS non-blocked), (LoS blocked), (nLoS non-blocked), (nLoS blocked). Here, nLoS state means that buildings can also block the path between the BS and the UE. Then, the associated UMi path loss measured in dB for four different states is given by [30]:

$$L_{dB}(y) = \begin{cases} 32.4 + 21\log_{10} y + 20\log_{10} f_c, & \text{LoS nBl.}, \\ 47.4 + 21\log_{10} y + 20\log_{10} f_c, & \text{LoS Bl.}, \\ 32.4 + 31.9\log_{10} y + 20\log_{10} f_c, & \text{nLoS nBl.}, \\ 47.4 + 31.9\log_{10} y + 20\log_{10} f_c, & \text{nLoS Bl.}, \end{cases} \quad (3)$$

where the first line corresponds to (4).

Following 3GPP, we adopt the 3GPP urban microcell (UMi) street canyon path loss model in [34] for mmWave and μWave frequency bands. For LoS non-blocked conditions, the path loss is provided by

$$L_{\text{dB}}(y) = 32.4 + 21\log_{10} y + 20\log_{10} f_c, \tag{4}$$

where f_c is the carrier frequency in GHz and y is the three-dimensional (3D) distance between the BS and the UE. In what follows, $f_{c,m}$ and $f_{c,\mu}$ correspond to the mmWave and μWave operating frequencies, respectively.

The path loss defined in (4) can be written in the linear scale by utilizing the generic representation Ay^{ζ}, where A, ζ are the propagation coefficients. Note that for both bands, the main difference is in carrier frequency leading to

$$A = 10^{2\log_{10} f_c + 3.24},\ \zeta = 2.1. \tag{5}$$

For the mmWave propagation impaired by both nLoS and blockage, the coefficients in (5) read as

$$\begin{aligned} A_1 &= 10^{2\log_{10} f_c + 3.24}, \zeta_1 = 2.1, \\ A_2 &= 10^{2\log_{10} f_c + 4.74},\ \zeta_2 = 3.19, \end{aligned} \tag{6}$$

where the coefficients (A_1, ζ_1), (A_1, ζ_2), (A_2, ζ_1), and (A_2, ζ_2) correspond to LoS non-blocked, nLoS non-blocked, LoS blocked, and nLoS blocked conditions, respectively. Note that for μWave technology, we differentiate between LoS and nLoS conditions only as human body blockage does not impact path loss significantly [32].

Finally, we stress that assuming multi-beam antennas, one also needs to account for sidelobe power leakage when using the same band among either mmWave beams or μWave beams. In general, this factor depends on array implementation. To this aim, we capture it with a constant of 3 dB [35].

C. Antenna Array Model

We assume planar antenna arrays with the cone radiation pattern at both BS and UEs. By following classic representation from [35], the HPBW of an array is determined as $\theta = 2|\theta_m - \theta_{3db}|$, where θ_{3db} represents the angle at which the output power drops at a level of -3 dB from its peak value, whereas θ_m corresponds to the location of the array maximum, i.e., $\theta_m = \arccos(-\beta/\pi)$, where β is the phase excitation difference. The mean gain over the HPBW is calculated as

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

where N is the number of antenna elements, whereas the upper and the lower 3-dB points are $\theta^{\pm}_{3db} = \arccos[\pm 2.782/(N\pi)]$.

We assume that up to $L_m > 1$ and $L_\mu > 1$ beams can be made simultaneously available at mmWave and μWave parts of BS, respectively. We assume hybrid analog-digital or digital beamforming techniques to enable multi-beam operation, i.e., more than one beam can be simultaneously generated at NR BS (e.g., utilizing superposition of multiple steering vectors [16]). Note that the possibility to form multiple simultaneous directional beams is subject to the fact that multiple RF chains are utilized, thus requiring hybrid and/or digital beamforming [36], [37]. The HPBW of the beams depends on the number of the antenna elements forming a beam and is upper bounded by the maximum number of antenna elements. The total power at mmWave and μWave parts of BS is upper bounded by $P_{\max,m}$ and $P_{\max,\mu}$.

In our work, we consider optimization on scales around a transmission time interval (TTI) in NR – a subframe of 1 ms (or multiple TTIs at most, depending on the practical scheduler implementation). Furthermore, beamforming in NR happens every 20 ms, i.e., the default synchronization signal block (SSB) periodicity is set to 20 ms. Additionally, the authors in [38] also demonstrated that even for highly directional terahertz (THz, 0.3-3 THz) communications systems, the link is stable for at least a few seconds, even for high rotational mobility of UEs. Thus, we can assume that at the timescale of interest, the impact of beam misalignment on the proposed multicast grouping algorithm is rather limited. On the other hand, UE arrivals and departures happen at a much larger timescale (e.g., seconds) and thus also do not impact the proposed approach.

D. Resource Allocation

In this work, we consider a resource allocation task over a finite time horizon. All UEs subscribed to the multicast session have to receive the service at the requested rate of C Mbps in this time horizon. Note that in our framework, this time horizon may or may not coincide with the NR basic units, i.e., frame/subframe in NR numerology, and, in general, may depend on the scheduler implementation, which

TABLE I
NOTATION USED IN THIS WORK.

Parameter	Definition
h_U	Height of UEs, m
$h_{A,m}, h_{A,\mu}$	Height of mmWave/μWave BS, m
W_m, W_μ	Available mmWave/μWave bandwidth, MHz
C	Session data rate, Mbps
$f_{c,m}, f_{c,\mu}$	mmWave/μWave carrier frequency, GHz
$L_{dB}(y)$	Path loss in decibel scale
y	Three-dimensional distance between UE and NR BS, m
x	Two-dimensional distance between UE and NR BS, m
h_B	Height of blocker, m
r_B	Radius of blocker, m
N	Number of planar antenna array elements
$\theta^{\pm}_{3db}$	Upper and lower 3-dB points of antenna array, rad
θ_m	Location of array maximum, rad
β	Antenna array orientation, rad
$P_{j,m}, P_{j,\mu}$	Transmit power for subgroup j, Watt
$P_{\max,m}, P_{\max,\mu}$	Total mmWave/μWave available power, Watt
G_A, G_U	Antenna array gains at NR BS and UE ends, dBi
M_m, M_μ	Number of time slots in mmWave/μWave time horizon
L_m, L_μ	Number of beams in mmWave/μWave
$R_{b,m} R_{b,\mu}$	Number of resource blocks in mmWave/μWave time slot
K	Number of multicast UEs
N_0	Power spectral density of noise, dB/Hz
A, ζ	Propagation coefficients
S_{th}	SNR threshold, dB
$p_B(y)$	Distance-dependent blockage probability
$w_{\text{PRB},m}, w_{\text{PRB},\mu}$	mmWave/μWave size of PRB, MHz
s_j	Spectral efficiency, bit/s/Hz
$S(y)$	Signal-to-noise ratio at 3D-distance y, SNR
D	BS intersite distance, m
R	Service (cell) area radius, m

is vendor-specific. In what follows, we utilize the NR subframe of duration 1 ms as the time horizon of interest. In addition, we also assume that BS knows the values of channel quality indicators of all UEs at both mmWave and μWave interfaces at the beginning of the resource allocation process.

To parameterize the OFDMA scheme, we introduce M_m and M_μ to denote the number of time slots in the time horizon (e.g., subframe) with the time slot indices t_m and t_μ. The maximum number of primary resource blocks (PRBs) available in the mmWave system is thus $M_m L_m R_{b,m}$, where $R_{b,m}$ is the available number of PRBs for the beam at a time slot in the mmWave time horizon. Observe that the potential maximum number of subgroups is limited to $M_m L_m$. Similarly, in the μWave system, the number of PRBs is limited by $M_\mu L_\mu R_{b,\mu}$, whereas the number of subgroups is bounded by $M_\mu L_\mu$.

E. Optimization Criterion

Traffic engineering is one of the most challenging topics in communication networks, which plays a key role in providing the required network services with quality of service. The term refers to applying scientific principles and strategies to operational networks to achieve optimal performance [39]. In other words, traffic is routed throughout the network to meet traffic needs while achieving certain performance objectives. Those objectives are usually congestion minimization, end-to-end delay minimization, packet loss minimization, energy consumption minimization, and resource utilization minimization. However, from a network design standpoint, resource usage is one of the most crucial perspectives for future systems [40]. Further, we note that the amount of studies addressing multi-beam multicast systems is limited.

When the transmission can be performed over one beam at a time at both technologies, the problem of optimal multicasting can be formulated as the minimization of the amount of utilized PRBs for each technology. However, for a system with multiple beams, PRB minimization might not provide actual resource minimization since the increase in the number of beams adds new resources to the system. These resources might not be fully utilized due to maximum emitted power constraints per antenna. Thus, we propose our own criterion that is a natural extension of the resource utilization often considered in single beam systems [41]. Thus, we select the ratio of occupied resources to the overall amount of available resources, ρ, as the optimization criterion to account for multi-beam operation. The motivation behind the proposed metric is that the utilized resources in one beam depend on the allocated power to this beam which is a part of the overall power that needs to be distributed across all the beams.

We consider the following RAT selection criteria: (i) mmWave priority, meaning that whenever possible mmWave capacity layer is chosen to serve UEs, and the μWave coverage layer is only employed when mmWave transmission fails to deliver the service due to propagation restrictions, (ii) μWave priority, and (iii) the weighted optimization function, where the RAT selection priority is explicitly controlled. Note that RAT selection priority depends on different factors, such as deployment area, as discussed in Section III-B2. For example, the network operator may choose to use μWave for longer distances and mmWave for short ones due to the mmWave propagation properties (severe path loss) and the fact that the data rate is higher at mmWave compared to sub-6 GHz frequency band. Further, the weight of 0.5 sets up no priority.

III. Optimization Framework

In this section, we first mathematically formalize the problem of optimal multicasting in dual mmWave/μWave deployments as a variable cost and size bin packing problem (BPP). Then, we extend it to the case of weighted priorities capturing operator-specific trade-offs in mmWave and μWave RAT usage. Finally, we propose the solution algorithm and two relaxation techniques.

A. Problem Formalization

In our task, a set of K UEs that make up a multicast group, $\mathcal{K} = \{1, ..., K\}$, is served by directional beams, implying that each beam covers a subset of UEs, the so-called multicast subgroup. There are $2^K - 1$ options to assign K UEs to multicast subgroup(s) [11], i.e., $\mathcal{K}_j$ represents the subset of UEs forming subgroup j, $j \in \mathcal{J}, \mathcal{J} = \{1, 2, \ldots, 2^K - 1\}$, and $|\mathcal{K}_j|$ is the number of UEs in subgroup j.

The goal of the model is to determine the optimal grouping of multicast UEs, 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 considering the possibility of transmission using two technologies (i.e., mmWave/μWave) while meeting all the system requirements.

We assume the time slot horizon contains M_m time slots for mmWave technology and M_μ time slots for μWave technology. Similarly, let L_m and L_μ be the numbers of beams that can be simultaneously swept during the given time horizon.

1) Suits of subgroups: We combine subgroups from set $\mathcal{J} = \{1, 2, \ldots, 2^K - 1\}$ to form so-called "suits" $\mathcal{G}_k$, $k = 1, 2, ..., |\Omega|$. The following definition summarizes the term "suit" used in this paper.

Definition 1. A "suit" is a collection of subgroup's indices $\mathcal{G}_k \in \mathcal{J}$ satisfying the following conditions:

$$\bigcup_{j \in \mathcal{G}_k} \mathcal{K}_j = \mathcal{K}, k = 1, 2, ..., |\Omega|,$$
$$\mathcal{K}_{j_1} \bigcap \mathcal{K}_{j_2} = \emptyset, j_1 \neq j_2, \ \forall j_1, j_2 \in \mathcal{G}_k, \tag{8}$$

where Ω is the set of all such suits, each covering all the UEs without repetition. Thus, each UE belongs to one and only one subgroup contained in the suit $\mathcal{G}_k$. For example, for $K = 2$ UEs, we have $\mathcal{K}_1$=$\{1\}$, $\mathcal{K}_2$=$\{2\}$, $\mathcal{K}_3$=$\{1, 2\}$ with $\mathcal{G}_1 = \{1, 2\}$, $\mathcal{G}_2 = \{3\}$. That is, $\mathcal{G}_1 \sim \mathcal{K}_1 \bigcup \mathcal{K}_2$ and $\mathcal{G}_2 \sim \mathcal{K}_3$, $\Omega = \{1, 2\}$, $|\Omega| = 2$.

Note that $\mathcal{K}_j$ preserves the directionality of the beam since all HPBWs are selected based on the UE locations. For example, when $\mathcal{K}_j$ contains one UE, a narrow beam can be utilized to serve that UE in a unicast way, whereas a wider beam can be necessary to serve a subgroup with several multicast UEs.

We define subsuits $\mathcal{G}_k^{l_m}$ and $\mathcal{G}_k^{l_\mu}$ as subsets of subgroup's indices from $\mathcal{G}_k$, which are planned for beams $l_m = 1, ..., L_m$ and $l_\mu = 1, ..., L_\mu$ by the scheduler, $\mathcal{G}_k^{l_m} \subseteq \mathcal{G}_k, \mathcal{G}_k^{l_\mu} \subseteq \mathcal{G}_k$, $k = 1, 2, ..., |\Omega|$. Therefore, we have

$$\mathcal{G}_k = \left(\bigcup_{l_m=1}^{L_m} \mathcal{G}_k^{l_m} \right) \bigcup \left(\bigcup_{l_\mu=1}^{L_\mu} \mathcal{G}_k^{l_\mu} \right), \tag{9}$$

that satisfies the constraints of serving a multicast UE via one technology and only ones, i.e.,

$$\begin{aligned} &\mathcal{G}_k^{l_{m_1}} \bigcap \mathcal{G}_k^{l_{m_2}} = \emptyset, l_{m_1} \neq l_{m_2}, \ \forall l_{m_1}, l_{m_2} \in \{1, ..., L_m\}, \\ &\mathcal{G}_k^{l_{\mu_1}} \bigcap \mathcal{G}_k^{l_{\mu_2}} = \emptyset, l_{\mu_1} \neq l_{\mu_2}, \ \forall l_{\mu_1}, l_{\mu_2} \in \{1, ..., L_\mu\}, \\ &\mathcal{G}_k^{l_m} \bigcap \mathcal{G}_k^{l_\mu} = \emptyset, \ \forall l_m \in \{1, ..., L_m\}, l_\mu \in \{1, ..., L_\mu\}. \end{aligned} \tag{10}$$

2) Decision variables: We introduce two binary indicators, $g_{j,m}^{t_m} \in \{0, 1\}$ and $g_{j,\mu}^{t_\mu} \in \{0, 1\}$, that represent decision variables of serving subgroup $j, j \in \mathcal{J}$, through mmWave/μWave beams at time slots $t_m \in \mathcal{T}_m$, $\mathcal{T}_m = \{1, \ldots, M_m\}$ and $t_\mu \in \mathcal{T}_\mu$, $\mathcal{T}_\mu = \{1, \ldots, M_\mu\}$. Let $g_{j,m}^{t_m} = 1$, if subgroup j is served in time slot t_m by a mmWave beam, $g_{j,m}^{t_m} = 0$ otherwise. Similarly, $g_{j,\mu}^{t_\mu} = 1$, if subgroup j is served in time t_μ by a μWave beam, $g_{j,\mu}^{t_\mu} = 0$ otherwise.

Therefore, we have the matrices-indicators for the two technologies, $\mathbf{G}_m$ and $\mathbf{G}_\mu$, where a row $\mathbf{g}_{j,m}$/$\mathbf{g}_{j,\mu}$ shows the time slot of serving subgroup j by mmWave/μWave technology during the time horizon $\mathcal{T}_m$/$\mathcal{T}_\mu$, whereas a column $\mathbf{g}^{t_m}$/$\mathbf{g}^{t_\mu}$ shows the subgroups that are served at time slot t_m/t_μ by mmWave/μWave beams.

3) Constraints on Beams and subgroups: First, observe that the system should comply with the constraint on the maximum number of subgroups to be served at time slots t_m and t_μ. This implies that at a time slot at most L_m and L_μ beams can be simultaneously swept, leading to

$$\begin{aligned} &\sum_{j \in \mathcal{G}_k^{l_m}} g_{j,m}^{t_m} \leq L_m, \forall t_m \in \mathcal{T}_m, \\ &\sum_{j \in \mathcal{G}_k^{l_\mu}} g_{j,\mu}^{t_\mu} \leq L_\mu, \forall t_\mu \in \mathcal{T}_\mu. \end{aligned} \tag{11}$$

Recall that the time horizon in our system is the NR subframe, which is fixed at 1 ms for all considered numerologies. Therefore, the suit service time should not exceed the subframe duration. Thus, for any $l_m = 1, ..., L_m$ and $l_\mu = 1, ..., L_\mu$, $k = 1, ..., |\Omega|$, we have the following constraints to be satisfied

$$\begin{aligned} &\sum_{j \in \mathcal{G}_k^{l_m}} \sum_{t_m \in \mathcal{T}_m} g_{j,m}^{t_m} \leq M_m, \\ &\sum_{j \in \mathcal{G}_k^{l_\mu}} \sum_{t_\mu \in \mathcal{T}_\mu} g_{j,\mu}^{t_\mu} \leq M_\mu. \end{aligned} \tag{12}$$

4) Power budget constraints: The constraints on the transmit power budget per antenna that serves subgroup $j \in \mathcal{G}_k^{l_m}$, $j \in \mathcal{G}_k^{l_\mu}$ over mmWave/μWave bands must be satisfied at any $t_m \in \mathcal{T}_m$ and $t_\mu \in \mathcal{T}_\mu$. That is, we have

$$\sum_{j \in \mathcal{G}_k^{l_m}} g_{j,m}^{t_m} P_{j,m} \leq P_{\max,m}, \forall t_m \in \mathcal{T}_m,$$

$$\sum_{j \in \mathcal{G}_k^{l_\mu}} g_{j,\mu}^{t_\mu} P_{j,\mu} \leq P_{\max,\mu}, \forall t_\mu \in \mathcal{T}_\mu, \tag{13}$$

where $P_{j,m}$ and $P_{j,\mu}$ represent the transmit power of the beam that serves subgroup j via mmWave and μWave, respectively. These powers can be determined by utilizing the path loss models defined in Section II.

5) Constraints on resource utilization: We introduce costs of the service in terms of PRBs, $a_{j,m}$ and $a_{j,\mu}$, required for serving subgroup j by a beam over mmWave and μWave bands, which take into account session bitrate, C, spectral efficiency, $s_{j,m}$/$s_{j,\mu}$, and PRB sizes of the subgroup j, w_{PRB_m} and w_{PRB_μ}, i.e.,

$$a_{j,m} = \frac{C}{s_{j,m} w_{\text{PRB}_m}}, \ a_{j,\mu} = \frac{C}{s_{j,\mu} w_{\text{PRB}_\mu}}. \tag{14}$$

We should then impose resource constraints by assigning a beam to the subgroup for all the service time. For any $j \in \mathcal{J}$ we, therefore, have

$$a_{j,m} \leq M_m R_{b,m}, \ a_{j,\mu} \leq M_\mu R_{b,\mu}, \tag{15}$$

and also the following must be satisfied for $k = 1, \ldots, |\Omega|$

$$\begin{aligned} &\sum_{j \in \mathcal{G}_k^{l_m}} a_{j,m} \leq L_m M_m R_{b,m}, \\ &\sum_{j \in \mathcal{G}_k^{l_\mu}} a_{j,\mu} \leq L_\mu M_\mu R_{b,\mu}. \end{aligned} \tag{16}$$

We emphasize that time slot assignment in the system is reflected in two vector-indicators, $\mathbf{g}_{j,m} = (g_{j,m}^1, \ldots, g_{j,m}^{M_m})$, $\mathbf{g}_{j,\mu} = (g_{j,\mu}^1, \ldots, g_{j,\mu}^{M_\mu})$. i.e., rows of matrices $\mathbf{G}_m$ and $\mathbf{G}_\mu$. The elements of these vectors give time interval duration for serving subgroup j by mmWave and μWave technologies and are written as

$$\begin{aligned} &\sum_{t_m \in \mathcal{T}_m} g_{j,m}^{t_m} = \left\lceil \frac{a_{j,m}}{R_{b,m}} \right\rceil, \\ &\sum_{t_\mu \in \mathcal{T}_\mu} g_{j,\mu}^{t_\mu} = \left\lceil \frac{a_{j,\mu}}{R_{b,\mu}} \right\rceil, j \in \mathcal{J}. \end{aligned} \tag{17}$$

B. Objective Functions

1) No Service Priorities: In our objective function, ρ, we determine the optimal grouping of multicast UEs, i.e., obtain the suit $\mathcal{G}_k$ of multicast subgroup's indices that covers all UEs without their repetition, taking into account (9). We consider minimizing the ratio of occupied PRBs to the total available number of PRBs for the time horizon. That is, the optimization problem takes the following form

$$\min_{k \in 1, \ldots, |\Omega|} \sum_{j \in \mathcal{G}_k} \left[\frac{a_{j,m}}{M_m L_m R_{b,m}} + \frac{a_{j,\mu}}{M_\mu L_\mu R_{b,\mu}} \right], \tag{18}$$

$$\text{s.t. } (8), (9), (10), (11), (12), (13), (15), (16).$$

Note that (18) can also be used to induce service priorities between RATs. In the case of mmWave priority, the system selects mmWave band to serve a set of UEs $\mathcal{K}_j$, $j \in \mathcal{J}$, if $P_{j,m} \leq P_{\max,m}$. This means that the mmWave BS is utilized up to its maximum coverage distance. Similarly, μWave priority ensures that set $\mathcal{K}_j$ is served by μWave BS, if $P_{j,\mu} \leq P_{\max,\mu}$.

2) Weighted Priority Service: Note that the problem in (18) assumes that either mmWave or μWave are assigned priority in service. Particularly, when supposing the priority is given to mmWave technology, then the solution algorithm will try to use it whenever possible, and μWave technology is only utilized when some UEs are outside the coverage of mmWave one. However, in practice, an operator may consider the choice of technology related to the available spectrum, deployment area, traffic conditions, etc. To cover these specific needs, we suggest the following weighted objective function:

$$\min_{k\in 1,\dots,|\Omega|} \sum_{j\in\mathcal{G}_k} \left[w\frac{a_{j,m}}{M_m L_m R_{b,m}} + (1-w)\frac{a_{j,\mu}}{M_\mu L_\mu R_{b,\mu}} \right], \quad (19)$$
$$\text{s.t. } (8),(9),(10),(11),(12),(13),(15),(16),$$

where w is the weight parameter.

The weight factor w in (19) can be utilized to provide weighted priority in technology selection. For example, when considering the coexistence of unicast and multicast traffic, one may want to make w proportional to the coverage distance by setting $w = \min(1, R^2/R_m^2)$, where R is the service area radius and R_m is the mmWave cell radius. The rationale is that when the geometric locations of unicast sessions are uniformly distributed in the dual-mode BS coverage area, the objective function in (19) maximizes the resources that will be available for a new session. Alternatively, the weight w can be set proportionally to the operator's utility, depending on the abovementioned factors. Since these factors are operator-dependent, we leave them outside the scope of this work.

C. More Than Two RAT Deployment

Different from the single-RAT networks, both multicast UE grouping that minimizes total service cost and mapping of these subgroups onto multiple RATs for parallel transmission in the multi-RAT networks should be determined. In our framework, the multi-RAT technologies can minimize the ratio of utilized to available resources while satisfying the service requirements. Thus, similarly to the two-RAT scheme, the scheduler aims to maximize total delivery cost in terms of the ratio of occupied PRBs to the available PRBs during the entire time horizon considering the possibility of transmission over all available technologies. For more than two RATs, one may use the formulation provided above by adding more components associated with all available technologies. Alternatively, the optimization criteria can be latency minimization, data rate maximization, etc. By combining multiple technologies, the effective service area of a multi-RAT solution will be extended to the coverage of all technologies onboard. Moreover, the reliability can be significantly improved compared to any single RAT connectivity. Note that, in general, the choice of technology depends on the application/service the UEs are involved in.

In general, when more than two RATs are considered, the optimization function takes the following form

$$\min_{k\in 1,\dots,|\Omega|} \sum_{j\in\mathcal{G}_k} \sum_{\gamma\in\Gamma} w_\gamma \frac{a_{j,\gamma}}{M_\gamma L_\gamma R_{b,\gamma}}, \quad (20)$$

where γ represents index of RAT, $\gamma \in \Gamma$, Γ is a set of RATs.

Algorithm 1: Optimal Solution

Input: $(X_U(i), Y_U(i), h_U), i \in \mathcal{K}$
Output: Optimal solution $\mathcal{G}_k^*$ for multicast grouping in form of (9)
Create $2^K - 1$ multicast subgroups of UEs
for *each subgroup* $\mathcal{K}_j$ **do**
  find the farthest UE i and the distance from BS to this UE: $y \leftarrow \max_{i\in\mathcal{K}_j} y_i$;
  find horizontal HPBW needed to cover the subgroup $\mathcal{K}_j$:
  $\theta_j = \arccos\left(\frac{(X_U(i)X_U(i')+Y_U(i)Y_U(i')+h_U^2}{y(i)y(i')}\right)$; ▷ θ_j *as the angle between two edge UEs* i *and* i'
  calculate P_j ; ▷ $P_j = P_{\max}$ *is fixed for* $L = 1$
  find the cost $a_{j,m}, a_{j,\mu}$ from (14);
end
Solve by exerting (18) with the exhaustive search.

D. Exact Solution and Relaxation Techniques

The formalized optimization problem can be classified as a special class of BPP, where items of various sizes are packed into the smallest number of unit capacity bins to minimize the cost of assigning the items to the particular bins. The pseudo-code in Algorithm 1 describes the globally optimal solution according to (18), i,e., suit $\mathcal{G}_k^*$ of multicast subgroup's indices in the form of (9). Here, in Algorithm 1 and also further in Algorithm 2, we consider the antennas having a sufficiently wide beam in the vertical direction, which allows covering all the UEs along the entire service radius of the BS. However, if the antennas have high vertical directivity, one can still use the beam with high directivity in both planes and utilize the proposed framework just by finding the HPBW in both planes.

Note that the problems (18), (19), (20) are NP-hard, while the associated complexities are exponential. Indeed, the multicast problem is formalized as a special class of BPP – allocating varying resources into multiple beams, each having finite capacity. Thus, the statement we make directly follows from a class of the problem as BPP is known to be NP-hard. The proof is provided in classic textbooks on optimization theory and operations research, e.g., [42].

For a limited number of UEs in the coverage area of BS, the direct solution of the problems in (18) and (19) can be adopted by utilizing, e.g., branch-and-cut or branch-and-bound techniques [43]. Some of these solutions allow controlling heuristic behavior with an emphasis on solution integrity instead of its optimality. In this work, we consider techniques known to provide significant improvements in the heuristic behavior of mixed-integer programming (MIP, [44]).

1) Local Branching Heuristic: Metaheuristics, which are general frameworks to build heuristics, often use combinatorial formulations. Heuristics for solving mathematical programming problems can be developed based on metaheuristic rules. One possible option, local branching (LB), is based on the idea of switching neighborhoods during the search to obtain the best possible solution [45]. LB is a technique developed based on the exact method. The difference is that the LB has the

time allotted to solve a given instance. If this time is reached before the optimal solution is found, LB stops and returns the best-known solution.

2) Relaxation Induced Neighborhood Search Heuristic: Relaxation-induced neighborhood search heuristic (RINS) is a heuristic that explores the neighborhood of a valid solution to discover an improved one [46]. The construction of a promising neighborhood is achieved by continuous relaxation of the MIP model and is formulated as another MIP (known as the subMIP). The subMIP optimization is truncated by limiting the number of nodes in the search tree.

IV. Approximate Solution

The use of the proposed algorithm at the TTI timescale places severe constraints on the execution time. Since the formulated problem is NP-hard, we propose adopting a heuristic simulated annealing solution, which is known to be efficient for BPPs [47]. We specify the initialization and implementation parts as well as parameterize the technique.

A. Simulated Annealing Approach

Stochastic heuristic methods [48] applicable to the MIP problem formulations are: local search, simulated annealing, evolutionary algorithm, simulated allocation, genetic algorithm, tabu search, etc. Out of these techniques, simulated annealing has been shown to achieve the best performance for combinatorial problems such as BPP [49], [50]. For this reason, we utilize this approach in this section to produce an approximate solution algorithm.

The idea of simulated annealing consists in randomizing the local search procedure and accepting changes that worsen the solution with some probability [51]. More specifically, it imitates the annealing of metals in thermodynamics, where metal is exposed to a very high temperature and then left to slowly cool down to form the desired shape with a defect-free structure [52]. Thus, a key concept in simulated annealing is to use an appropriate temperature cooling schedule. Several modifications of the simulated annealing algorithm differ in the distribution and the temperature reduction law, producing particular shortcomings and advantages, such as speed, complexity, and the guarantee of finding the global minimum.

In this work, we utilize the *standard simulated annealing* methodology [43], see Algorithm 2 to obtain heuristic solution $\tilde{\mathcal{G}}_k$. First, we determine problem-specific choices, including the form of the objective function $c(S)$ and the way solution S is obtained. Theoretically, the initial solution does not affect the final result. However, several experiments have shown that sometimes the initial solution obtained employing a good heuristic may result in faster convergence to the optimal solution [53], [54] due to the fact that the heuristic-based solution is more likely to be close to the optimal one. In Section V, we consider random-based simulated annealing (SA) with a random choice of the initial solution and heuristic-based simulated annealing (SA-H) with guided choice of the initial solution, as well as discuss the convergence speed of the two configurations of simulated annealing. The latter is implemented according to Algorithm 2 (Modified Incremental Multicast Grouping) provided in [22].

To achieve the global minimum, the number of steps, $MaxIt$, in the inner loop of Algorithm 2 must be larger than the number of points in the solution space, i.e., $MaxIt > |Q|$, leading to the futility of the exact approach [43].

B. Implementation

To obtain a good initial solution, we use the following heuristic from [22] proven to provide close to optimal results and capture multicast properties in directional systems. First, the farthest UE from the BS is selected. Then, by iterating through the set of predefined beamwidths θ, one is selected to provide the lowest number of utilized resources per UE. Note that all UEs covered by the selected beam are included in the multicast subgroup. The selected multicast subgroup is then deleted from the set of UEs, and the algorithm again selects the farthest UE from the remaining set of UEs. The algorithm stops when there are no UEs left.

We now describe the general logic that governs the operation of the Algorithm 2 itself. Here, the initial temperature is set in the temperature parameter, while the temperature reduction is a function of cooling α, $0 < \alpha < 1$. At each iteration k, the temperature is cooled down by α. We define the number of neighbors, $MaxIt$, to visit at each iteration. A stopping criterion can be the condition $T = 1$ or the lack of significant improvement in two consecutive executions of the objective function of the outer loop. Also, achieving a solution that does not exceed a predetermined cost may be utilized to stop the procedure. We use condition $T = 1$. The objective function $c(S)$ represents the ratio of occupied to available resources, ρ, required by solution S, where S is a set that includes all the multicast UEs only once.

Algorithm 2: Simulated Annealing Solution

Input: $(X_U(i), Y_U(i), h_U), i \in \mathcal{K}$
Output: Heuristic solution $\tilde{\mathcal{G}}_k$ for multicast grouping in form of (9)
Generate a feasible initial solution S;
Setup initial temperature $T = 10$;
Setup the cooling rate α;
while $T > 1$ *(stopping criterion* $T = 1$*)* **do**
  $k \leftarrow 0$; ▷ *number of iterations*
  while $k <$ *MaxIt* **do**
    Select a neighbor S' of S;
    $\Delta c = c(S') - c(S)$;
    if $\Delta c \leq 0$ **then**
      $S \leftarrow S'$;
    else
      $S \leftarrow S'$ if $random(0,1) < \exp(\frac{-\Delta c}{T})$;
    end
    $k \leftarrow k + 1$;
  end
  $T = T\alpha$;
end

After defining the initial solution and setting up the general execution parameters, the algorithm runs the outer "while" loop with fixed temperature (lines 6-19 of Algorithm 2). In the inner "while" loop, which executes $MaxIt$ times, the algorithm selects a random neighbor S' and performs the Metropolis test to accept the move from S to S' or not (lines 8-17). In the algorithm, the process of the random neighbor selection is as follows: (i) randomly generate set S' such that it covers all the UEs, (ii) calculate the required transmit power for S' based on the most robust SNR, (iii) perform water-filling for those multicast subgroups that can be served simultaneously at a time slot considering the power budget per antenna, $P_{\max}$, and (iv) compute $c(S') = \rho$. Note that if the cost function $\Delta c = c(S') - c(S)$ is non-positive, the move is always accepted. Otherwise, the move is accepted with probability $P = e^{-\Delta c/T}$. Once MaxIt steps are completed, the temperature decreases (line 18), and the inner loop starts again. The algorithm works until the stop criterion is met.

C. Choice of Initial Parameters

For fixed T, the acceptance probability P is an exponentially decreasing function of Δc. Hence, as Δc increases, the acceptance probability quickly becomes very small. The Metropolis test [55] allows for leaving the local minimum encountered while wandering around the solution space within the inner loop. After performing $MaxIt$ steps, the temperature is declined according to the temperature reduction function, and the inner loop is started again. For fixed Δc, the acceptance probability decreases with T, so in the consecutive execution of the inner loop, the uphill (accepted) moves are rarer.

The algorithm is relatively simple to implement, but its efficient implementation requires tinkering with parameters and figuring out ways to reduce the run-time associated with computing the solution for values in the search space. The initial temperature typically is a large number. Then the inner while-end loop is executed $MaxtIt$ times, which is another parameter of the algorithm. Choosing the proper initial temperature, cooling rate, and the number of neighbors to visit is crucial for balancing convergence speed and complexity. The optimal values of these parameters are problem-dependent and are typically determined through experimentation. As simulated annealing is a heuristic solution, in Section V, we explore the optimality and complexity of the simulated annealing algorithm when the number of neighbors to be explored, $MaxIt$, is 15, and the initial temperature is $T = 10$. We select a cooling rate $\alpha = 0.8$ that better controls the speed at which the algorithm explores the search space and reduces the temperature.

D. Computational Complexity

The SA method is stochastic in the sense that there is a random element guiding the sequence of generated solution points. By design, SA inherently operates with polynomial complexity [56]. The complexity of Algorithm 2 depends on many factors, such as initial temperature, cooling rate, the number of neighbors to visit, solution acceptance probability, and temperature reduction function, among others.

TABLE II
DEFAULT PARAMETERS FOR NUMERICAL ASSESSMENT.

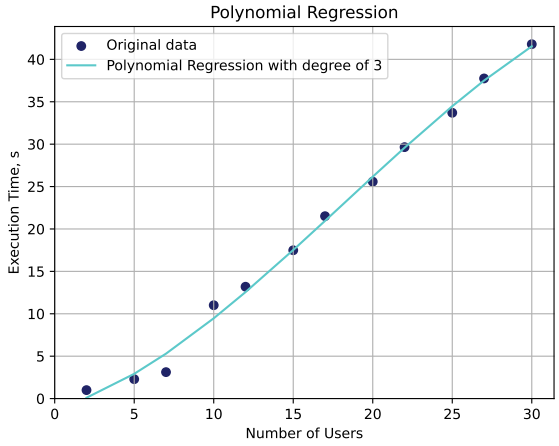
Parameter	Value
mmWave operating frequency, $f_{c,m}$	28 GHz
μWave operating frequency, $f_{c,\mu}$	3.5 GHz
mmWave bandwidth, W_m	100 MHz
μWave bandwidth, W_μ	50 MHz
mmWave PRB size, $w_{\text{PRB},m}$	1.44 MHz
μWave PRB size, $w_{\text{PRB},\mu}$	0.18 MHz
Height of mmWave/μWave BS, $h_{A,m}, h_{A,\mu}$	10 m
Height of blocker, h_B	1.7 m
Height of UE, h_U	1.5 m
SNR threshold, S_{th}	-9.47 dB
mmWave/μWave power budget, $P_{\max,m}$, $P_{\max,\mu}$	33 dBm
Power spectral density of noise, N_0	-174 dBm/Hz
Number of UE planar antenna elements, N	4 el
UE receive gain, G_U	5.57 dBi
Session data rate, C	5 Mbps
BS antenna array	32×4
BS transmit gain, G_A	14.58 dBi
Service area radius, R	400 m
Number of UEs, K	2-30
Subframe duration	1 ms
mmWave slot duration	125 μs
μWave slot duration	1 ms
5G NR numerology, mmWave, κ_m	3
5G NR numerology, μWave, κ_μ	0
Number of time slots in a subframe, mmWave, M_m	8
Number of time slots in a subframe, μWave, M_μ	1
Number of available resource blocks, mmWave, $R_{b,m}$	66
Number of available resource blocks, μWave, $R_{b,\mu}$	270
Number of beams available in the system, L_m, L_μ	5

Note that a key component that plays a crucial role in the performance of SA is the criteria under which the temperature changes, namely, the cooling rate. To provide the complexity in $O(\cdot)$ notation, let us denote the length of the cooling schedule as m. Then, the computational complexity of the SA algorithm is given by $O(m \cdot MaxIt)$, where $MaxIt$ is the complexity due to the "while" cycle over the number of neighbors to be visited before cooling the temperature (lines 8-17), and m depends on the temperature reduction function (lines 6-19). The first component is executed inside the second "while" cycle, leading to the resulted $O(m \cdot MaxIt)$, complexity. To assess further the computational complexity, in the next section, we provide Table III, which contains execution times and discusses it in detail.

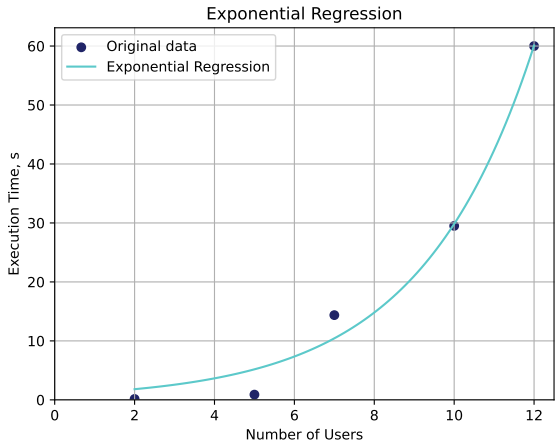
We can also construct mathematical solutions to model the trends of the execution time as a function of the number of users, K. SA has polynomial complexity. Hence, we use a polynomial regression model built in a Python environment. Based on the data from Table III for SA, the trend can be described by the following expression: $f(K) = -1.203 + 0.521K^1 + 0.067K^2 - 0.001K^3$. The optimal solution's trend is exponential, and a mathematical model can be written as $f(K) = 0.899 \cdot \exp(0.350K)$. We provide Fig. 2 to show the accuracy of the mathematical models.

V. NUMERICAL RESULTS

To assess the performance of our mmWave/μWave system with the multi-beam directional antennas and reveal engi-



(a) Polynomial regression model for SA



(b) Exponential regression model for optimal solution

Fig. 2. Execution time vs. number of users.

neering design choices, we adopt the following procedure. We initially analyze the scenario when priority is given to mmWave technology. Here, we first consider the case of the limited number of UEs, $K = 10$, and compare the results in terms of utilized/available PRBs for globally optimal solution (Algorithm 1), two relaxations applied to the optimal solution, namely, LB and RINS using ILOG CPLEX, and two versions of simulated annealing (heuristic- and random-based initial solution generations). We then proceed to emphasize the importance of NR numerology utilized at μWave BS. Further, we consider the case when priority is given to μWave BS. We then examine the weighted priority case in (19). Finally, by utilizing the simulated annealing approach, we report the optimal dual-mode BS deployment density. To run the simulations, we use a standard laptop with 8 GB of RAM and an Intel Core i5-7200U with 2 hyper-threaded cores running at a base clock of 2.50 GHz.

The default system parameters are gathered in Table II. The utilized path loss model is defined in (3) and (6), where $f_{c,m}$ and $f_{c,\mu}$ correspond to 28 GHz and 3.5 GHz for mmWave and μWave BSs, respectively. To produce the numerical results, we consider the deployment with human blockers only. The transmit power budget is fixed at 33 dBm, and the session rate C is assumed to be 5 Mbps.

By default, we assume the available bandwidths of $W_m = 100$ MHz and $W_\mu = 50$ MHz, and consider numerologies $\kappa_m = 3$ (with PRB size of 1.44 MHz) for mmWave technology and $\kappa_\mu = 0$ and $\kappa_\mu = 2$ (with PRB size of 0.18 MHz and 0.72 MHz, respectively) for μWave. For both RATs, we utilize similar antenna arrays with $\{32, 16, 8, 4, 2, 1\}$ and 4 elements forming directional beam patterns in vertical and horizontal dimensions, e.g., 32H×4V. We note that, in general, arbitrary antenna array configurations can be utilized. We note that the choice of antenna arrays is operator-specific and depends on the equipment capabilities. Typically, the array size at mmWave BSs is larger than that of μWave BSs. We also note that the antenna array size induces the upper limit on the HPBW. The actual HPBW values required to serve multicast subgroups are selected as explained in Algorithm 1 by mapping a_j to the available antenna beams.

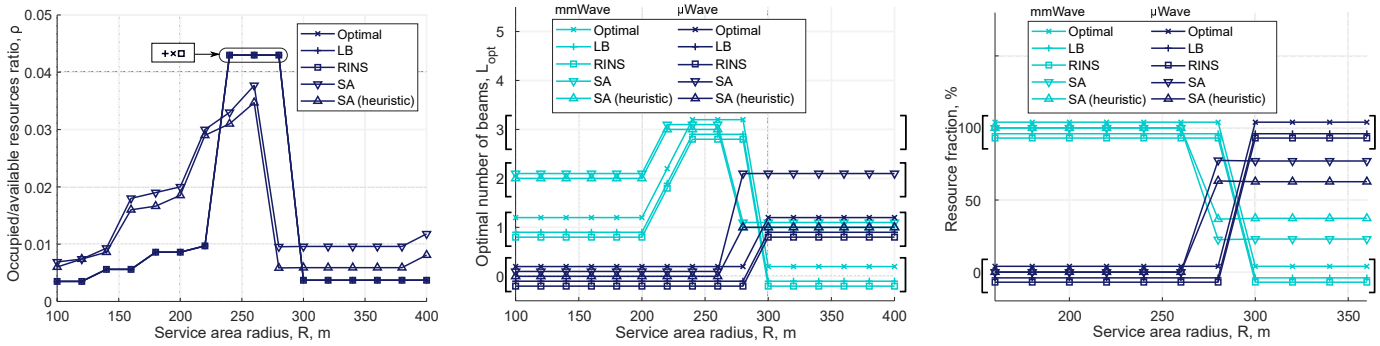
Note that for the sake of understandability in Fig. 3, Fig. 4, Fig. 6, and Fig. 7, we move apart the curves that show the same value by using $[\ ,]$ signs.

A. mmWave Priority

The results of the performance analysis, when mmWave resources are utilized whenever possible, are shown in Fig. 3 for mmWave numerology $\kappa_m = 3$, μWave numerology $\kappa_\mu = 0$, $K = 10$ UEs, $C = 5$ Mbps, $W_m = 100$ MHz, $W_\mu = 50$ MHz, $L_m = L_\mu = 5$ beams. Here, we start by analyzing the ratio of occupied to available resources, ρ, as a function of cell radius, R, illustrated in Fig. 3(a). As a general trend, one may notice that ρ grows with the increase in the service area radius until it reaches the distance at which the use of mmWave resources becomes ineffective. At this point, the system starts selecting μWave as a transmission technology. For example, in the case of the optimal solution, $R = 300$ m can be considered as a *threshold* that defines the change in the utilized transmission technology. Once this threshold is exceeded, the optimal solution always chooses the subgroup containing all K UEs for μWave transmission.

We emphasize that the relaxation techniques (LB, RINS) show a perfect match with the globally optimal solution. On the other hand, the simulated annealing algorithms demonstrate slightly worse results but with better optimality vs. complexity trade-offs than optimal solutions. By comparing the simulated annealing algorithms, we may learn that starting with a good solution (compared to the random one) at some points brings us a better value of ρ. This can be explained by the fact that heuristic-based simulated annealing can find a better solution by the time the stopping criterion is met. As our additional observation, we note that the fewer multicast subgroups are chosen, the fewer resources they demand.

Despite the simulated annealing technique leading to sub-optimal solutions as discussed above, it allows to drastically reduce the complexity of the solution as indicated in Table III providing a comparison of execution times for all the considered solutions, where SA-H and SA stand for simulated



(a) Ratio of occupied to available resources (b) Optimal number of beams (c) Utilized mmWave and μWave resources

Fig. 3. Performance metrics when mmWave resources are utilized whenever possible (mmWave RAT priority): mmWave – $\kappa_m = 3$, μWave – $\kappa_\mu = 0$.

TABLE III
ALGORITHMS' EXECUTION TIME, MINUTES.

Time/K	2	5	7	10	12	15	17	20	22	25	27	30
Optimal	0.15	0.89	14.37	29.50	60 (limited)	-	-	-	-	-	-	-
LB	0.13	0.88	14.2	28.70	60 (limited)	-	-	-	-	-	-	-
RINS	0.13	0.88	14.25	29.20	60 (limited)	-	-	-	-	-	-	-
SA-H	1 (0.091 s)	2.29 (0.060 s)	3.12 (0.053 s)	11.01 (0.066 s)	13.19 (0.050 s)	17.49 (0.059 s)	21.51 (0.039 s)	25.58 (0.091 s)	29.65 (0.060 s)	33.70 (0.056 s)	37.75 (0.040 s)	41.79 (0.044 s)
SA	1 (0.025 s)	2.29 (0.013 s)	3.12 (0.007 s)	11.01 (0.042 s)	13.19 (0.242 s)	17.49 (0.062 s)	21.51 (0.019 s)	25.58 (0.025 s)	29.65 (0.041 s)	33.70 (0.038 s)	37.75 (0.058 s)	41.79 (0.197 s)
*In brackets, the time to generate an initial solution is shown in seconds. *SA and SA-H stand for simulated annealing with random and heuristic choice of the initial solution, respectively.												

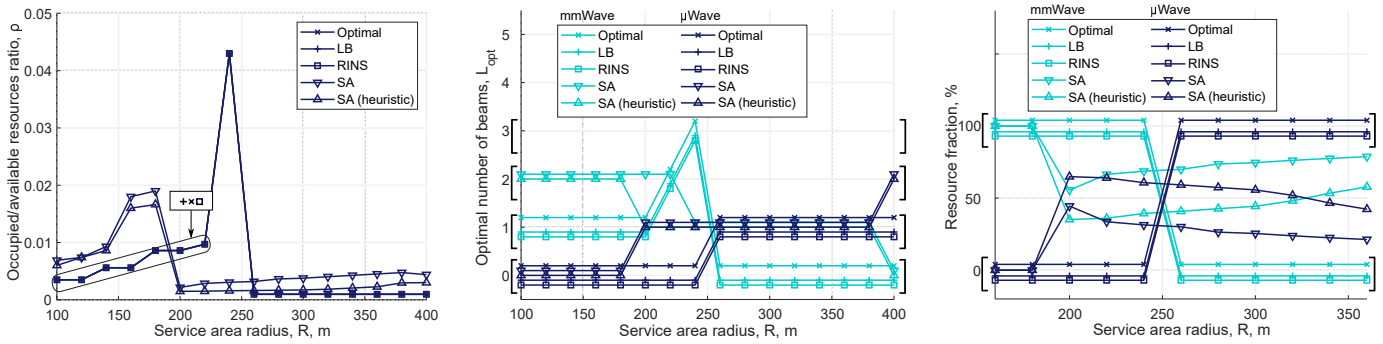
annealing with heuristic and random choice of initial solutions, respectively, see Section IV-A for details. By analyzing the presented data, one may observe that both considered relaxation techniques provide no performance improvements in solution time. Furthermore, all three exact solutions cannot solve the problem in a reasonable time when the number of UEs in a multicast group is higher than approximately 12. On the other hand, both simulated annealing solutions are characterized by linearly increasing solution time when the number of UEs grows. Note that the time to create an initial solution for both cases is short (less than a second), which does not impact the time to find a final solution (in minutes). In addition, by analyzing the results, we may deduce that the type of initial solution does not affect the computational time, due to the equal number of iterations, but impacts the final solution performance. More precisely, a good initial solution (the one in SA-H) offers slightly better final results compared to a random one (SA). The reason is that the stopping criterion is $T = 1$ that depends on the number of iterations, $MaxIt$, and cooling rate, α, which are the same for both SA and SA-H. However, the convergence speed of heuristic-based SA-H is faster, which can explain better final results.

Getting back to performance metrics, we further comment on the optimal number of beams utilized in the multi-beam dual system as a function of the cell radius illustrated in Fig. 3(b). The optimal number of mmWave beams, L_m, starts with one beam (when UEs form a single subgroup) and then increases up to 3 beams. On the contrary, up to one μWave beam can be swept at a time (and up to 2 μWave beams for random simulated annealing). As one may notice, μWave transmissions are utilized when mmWave fails to provide the service due to propagation conditions and blockage. We emphasize that μWave BS sweeps one beam as, first, it is possible to provide a multicast service to all UEs by using the wide beam (small propagation losses) and, second, it ensures the best ratio of occupied to available resources, ρ. We also note that the utilized HPBWs for μWave antennas are larger than those of mmWave technology as the former technology is employed for multicast subgroups having UEs located farther away from each other. In contrast, mmWave technology typically serves individual UEs in the unicast way or very clustered subgroups.

To complement the discussion, we show the percentage of utilized resources by each of the two technologies for all considered algorithms in Fig. 3(c). This metric shows a breakdown in the amount of utilized resources between considered technologies as a function of distance, R. Observe that, for the optimal solution, up to the distance of around 300 m, 100% of resources used to serve the multicast group are taken from the mmWave band. Once this boundary is passed, the role of mmWave and μWave technologies is reversed, and all the resources are taken from the μWave band.

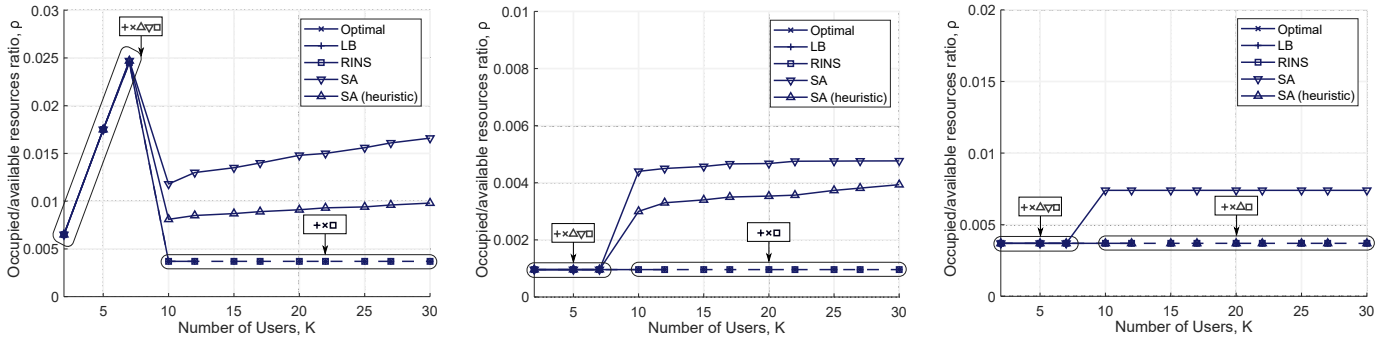
B. Effects of μWave Numerology

The set of numerologies defined for NR provides an additional degree of flexibility and adaptivity. We now proceed with assessing the system performance when utilizing numerology $\kappa_\mu = 2$ instead of $\kappa_\mu = 0$ for μWave band. In this configuration, a subframe has 4 slots with the length of 0.25 ms. By analogy with Fig. 3, in Fig. 4 we demonstrate



(a) Ratio of occupied to available resources (b) Optimal number of beams (c) Utilized mmWave and μWave resources

Fig. 4. Performance metrics when mmWave resources are utilized whenever possible (mmWave RAT priority): mmWave – $\kappa_m = 3$, μWave – $\kappa_\mu = 2$.



(a) mmWave ($\kappa_m = 3$) priority, μWave – $\kappa_\mu = 0$ (b) mmWave ($\kappa_m = 3$) priority, μWave – $\kappa_\mu = 2$ (c) μWave ($\kappa_m = 0$) priority, mmWave – $\kappa_\mu = 3$

Fig. 5. Ratio of occupied to available resources, ρ, as a function of the number of UEs.

(a) the ratio of occupied to available resources, ρ, (b) optimal number of beams in the system, and (c) percentage of utilized mmWave/μWave resources for $K = 10$ UEs, $C = 5$ Mbps, $W_m = 100$ MHz, $W_\mu = 50$ MHz, $L_m = L_\mu = 5$ beams.

As one may observe, Fig. 4 demonstrates qualitatively similar results to those illustrated for μWave numerology $\kappa_\mu = 0$ in Fig. 3. Nevertheless, there is a numerical difference between considered numerologies. More precisely, differently from $\kappa_\mu = 0$, the simulated annealing solutions start utilizing μWave band at shorter distances, i.e., at approximately $R = 200$ m. The optimal solution provides the same results when mmWave technology is selected. Here, one multicast subgroup and, respectively, one beam is selected when μWave is utilized. One may also observe that for numerology $\kappa_\mu = 2$, the gap between optimal and approximate solutions becomes less visible due to the number and size of available PRBs that the system can provide. The reason is that numerology $\kappa_\mu = 2$ provides more flexibility in terms of (i) PRBs size as the bandwidth occupied by a PRB depends upon the numerology being used (1 PRB = 12 subcarrier spacing (SCS) [kHz]) and (ii) time slots available for scheduling, $M = 2^{\kappa_\mu}$.

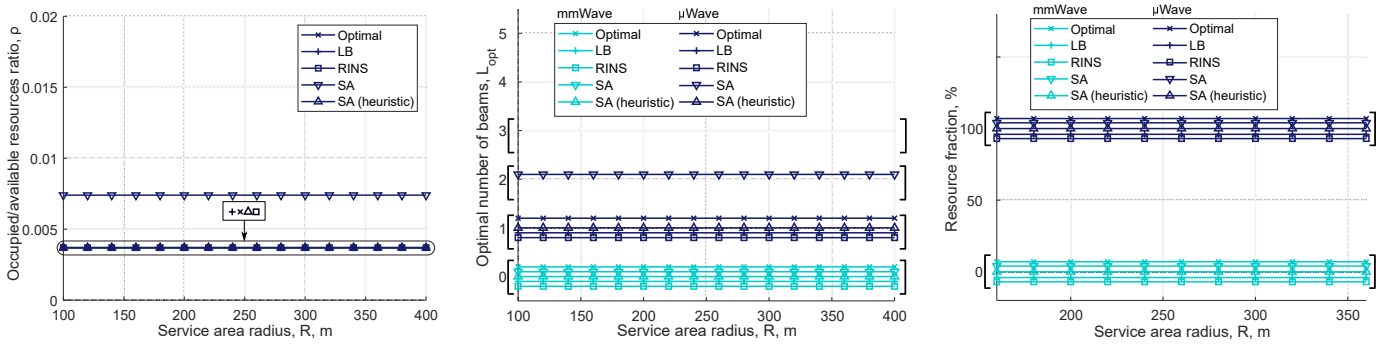
The increase in the service area of dual-mode BSs, R, makes the solution more complex in the case of the simulated annealing algorithms, which affects the system performance. It can be observed in Fig. 4(b) at $R = 400$ m, where random-based simulated annealing selects two subgroups as compared to just one subgroup in the case of the optimal solution. This can be explained by the fact that the algorithms do not converge to the optimal solution due to the increased complexity. To deal with these situations, one needs to increase the number of algorithm iterations. However, it is not reasonable to do that for smaller radii of the service area due to the complexity-accuracy trade-off. Therefore, the advantages of SA include its polynomial time complexity and the ability to adjust the quality and complexity by selecting appropriate initial parameters.

We note that, in our numerical evaluation, by way of example, we provide the effect of the selected numerology on the μWave band. One can utilize the proposed framework to change the numerology for all considered technologies.

C. Dependence on Number of Users

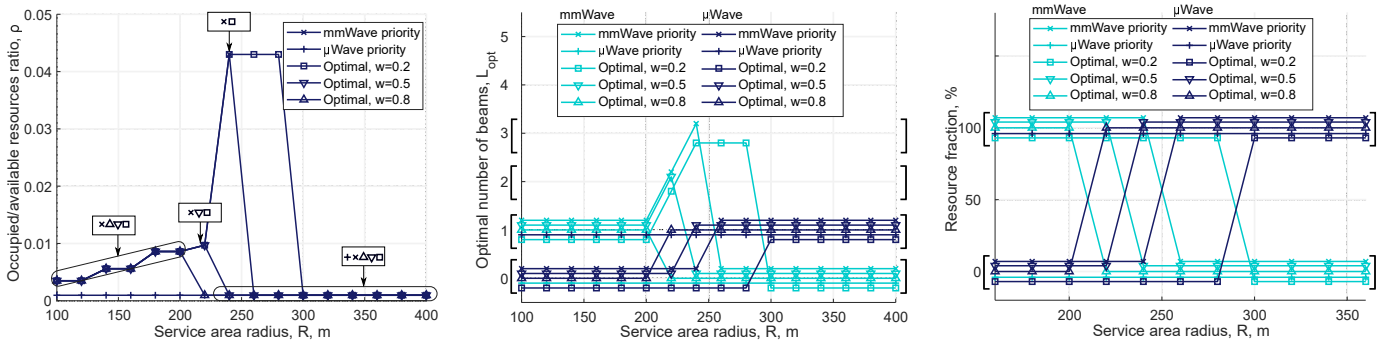
So far, we focused on the performance comparison of different solution algorithms under different parameter settings for a rather limited number of UEs in a multicast group. In Fig. 5, we investigate the behavior of the ratio of occupied to available resources as a function of the number of UEs K. Note that for the number of UEs higher than 12, we utilize quadratic extrapolation to construct the curves for the optimal solution, LB, and RINS. The rest of the parameters are $R = 400$ m, $C = 5$ Mbps, $W_m = 100$ MHz, $W_\mu = 50$ MHz, $L_m = L_\mu = 5$ beams.

We start analyzing the dependence on the number of UEs for mmWave numerology $\kappa_m = 3$ and μWave numerology $\kappa_\mu = 0$ illustrated in Fig. 5(a). First of all, observe that for a



(a) Ratio of occupied to available resources (b) Optimal number of beams (c) Utilized mmWave and μWave resources

Fig. 6. Performance metrics when μWave resources are utilized whenever possible (μWave RAT priority): mmWave – $\kappa_m = 3$, μWave – $\kappa_\mu = 0$.



(a) Ratio of occupied to available resources (b) Optimal number of beams (c) Utilized mmWave and μWave resources

Fig. 7. Performance metrics for weighted optimization function: mmWave – $\kappa_m = 3$, μWave – $\kappa_\mu = 0$.

smaller number of UEs, e.g., $K = 7$, mmWave band is utilized up to 400 m (and up to 300 m for $K = 10$). This allows us to conclude that the radius of the service area and the number of UEs affect the choice of technology. We note that for optimal solutions at $R = 400$ m, the system chooses μWave BS when the number of UEs is more than 10. Although simulated annealing solutions also mainly utilize μWave technology, the number of subgroups can be higher than in the case of the optimal solution.

By analyzing Fig. 5(b), showing the dependence of ρ on the number of UEs for mmWave numerology $\kappa_m = 3$ and μWave numerology $\kappa_\mu = 2$, one may notice a dissimilar trend to the one demonstrated in Fig. 5(a). Here, we consider a service area radius of 400 m. For this radius and numerology $\kappa_\mu = 2$, optimal solutions utilize μWave BS for any range of UEs differently from $\kappa_\mu = 0$. As one may observe, approximate solutions match the optimal solutions at lower K value, i.e., $5 - 7$ UEs. For $K \geq 10$, simulated annealing algorithms choose two subgroups (see Fig. 4(b)) and, hence, there is a gap between the optimal solution (one subgroup) and heuristics.

D. Effect of Different Objectives

Having considered the system, where mmWave resources are utilized to serve multicast UEs whenever possible, we now study the effect of different types of objective functions. Particularly, we first consider the system response to the case when μWave technology is prioritized and then address the weighted structure of the optimization function.

We start with the effect of μWave priority. The corresponding performance results are displayed in Fig. 6 for mmWave numerology $\kappa_m = 3$, μWave numerology $\kappa_\mu = 0$, $K = 10$ UEs, $C = 5$ Mbps, $W_m = 100$ MHz, $W_\mu = 50$ MHz, $L_m = L_\mu = 5$ beams. As expected, μWave priority promotes the use of μWave band while completely eliminating the use of mmWave resources, see Fig. 6(c). Furthermore, as can be deduced from Fig. 6(b), for the considered range of values for cell radius R, the service is performed by utilizing just a single beam at μWave technology. Then, as one may observe in Fig. 6(a) and Fig. 6(b), heuristic-based simulated annealing perfectly matches the optimal solutions for all the considered values of R. However, the random simulated annealing solution demonstrates a higher ρ value due to the utilization of two beams.

Getting back to Fig. 5, we are now in a position to compare mmWave and μWave priority strategies by discussing the effect of the total number of UEs requesting multicast traffic in the system as illustrated in Fig. 5(c). Addressing the choice of the solution algorithm for optimization, we note that similarly to Fig. 5(a) and Fig. 5(b), at a lower number of UEs, all schemes demonstrate the same performance. Then, starting from $K = 7$, the performance of the random-based simulated annealing solutions begins to degrade, producing a gap with the other approaches. By comparing absolute values

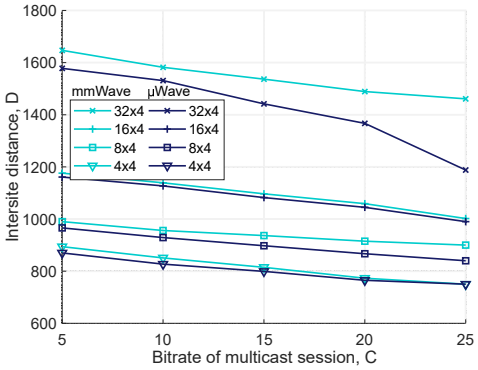


Fig. 8. Dual NR BS intersite distance.

of ρ, we notice that both mmWave and μWave priorities return quantitatively similar results. A notable exception is the mmWave priority scheme with mmWave numerology $\kappa_m = 3$ and μWave numerology $\kappa_\mu = 0$ that is characterized by a higher amount of utilized resources until approximately $7-10$ UEs.

Observe that μWave priority completely excludes mmWave resources, thereby loading μWave technology. A network operator may want to avoid it as μWave technology needs to be utilized in those areas not accessible for mmWave. On the other hand, the mmWave priority scheme exclusively utilizes mmWave resources up to a certain distance and then switches to μWave technology. In practice, an operator might have different preferences for balancing resource utilization between considered RATs. To this end, we continue by investigating the impact of the weighted optimization function on the system performance. The corresponding results are shown in Fig. 7 for mmWave numerology $\kappa_m = 3$, μWave numerology $\kappa_\mu = 2$, $K = 10$ UEs, $C = 5$ Mbps, $W_m = 100$ MHz, $W_\mu = 50$ MHz, $L_m = L_\mu = 5$ beams.

By analyzing the data presented in Fig. 7, we emphasize that the increase in w leads to the shift in the priority from mmWave to μWave. Particularly, one may learn that at lower distances R, weights $w = 0.2, 0.8, 0.5$, do not affect the performance and provide results similar to the mmWave priority scheme. This can be explained by the fact that mmWave ensures more efficient resource utilization at smaller distances. Further, note that the choice of $w = 0.5$ produces a similar effect to mmWave priority; thereby utilizing μWave band resources only when mmWave service is infeasible due to propagation and blockage conditions. Alternatively, $w = 0.2$ increases the range of mmWave technology up to 280 m (compared to 240 m in the case of mmWave priority), whereas $w = 0.8$ shortens R to 200 m, thereby allowing μWave band usage. Therefore, we can conclude that depending on the operator's preferences, weights can be properly adjusted to achieve a particular goal with respect to resource usage in dual-mode mmWave/μWave systems.

E. Dual mmWave/μWave Deployment Analysis

Finally, as an example, in Fig. 8, we present the intersite distance (ISD) between BSs (estimated based on both mmWave/μWave coverage ranges) as a function of multicast session bit rate C for different antenna arrays, mmWave priority RAT selection criteria, mmWave numerology $\kappa_m = 3$, μWave numerology $\kappa_\mu = 0$, $K = 30$, $W_m = 100$ MHz, $W_\mu = 50$ MHz, $L_m = L_\mu = 5$ beams. Recall that for tri-sector antenna deployment, the ISD corresponds to $3R$ [57]. We note that the ISD of the dual mmWave/μWave system (measured using μWave band as this technology provides larger coverage) is increased insignificantly as compared to a mmWave system. For instance, for 32x4 BS antenna array and $C = 20$ Mbps, ISDs of μWave and mmWave correspond to 1476 m and 1428 m, respectively. The reason is that μWave requires a large number of PRBs due to the utilized numerology $\kappa_\mu = 0$ with SCS of 15 kHz and one time slot available for the scheduling. It means that at 1476 m for $C = 20$ Mbps and 32x4 antenna array, the system has insufficient resources available to serve all the UEs. Note that the operator can use its own parameters in realistic deployments while exploiting the proposed methodology to calculate the optimal coverage range of dual-mode BSs.

VI. Conclusions

Inspired by the prospective 5G NR integrated mmWave/μWave deployments and advanced antenna systems designs capable of simultaneously supporting multiple directional beams, in this work, we have provided a globally optimal solution for multicast grouping. Accounting for the NP-hard nature of the problem, we have then proposed and characterized the approximate simulated annealing approach as an efficient solution methodology. The proposed approach is characterized by polynomial time complexity, potentially allowing for practical implementation.

Our numerical results illustrate that properties of the optimal solution, such as resource utilization and the type of utilized technology, heavily depend on the density of dual-mode BS deployments, RAT priority, and considered system parameters. There is a clear turning point for small dual-mode BS densities when the system switches from the regime when mmWave resources are utilized for service to the case when μWave technology is exclusively utilized. This point is dictated by the mmWave blockage and propagation conditions. The number of beams associated with optimal solution is upper limited by 3 for mmWave and by 2 for μWave technologies across all the considered densities of dual BS deployment. Moreover, in most cases, only one beam is utilized at μWave technology. Further, the utilized numerology may quantitatively affect the abovementioned conclusions, but the overall qualitative trends remain unchanged. The investigated RAT selection priorities reveal that when μWave RAT is prioritized for multicast service, mmWave resources are not utilized at all. However, by utilizing weights for mmWave and μWave resources, the operator might achieve the desired balance by fitting its needs in a particular deployment. Finally, we note that the efficiency of resource utilization for multicast service may also be affected by the number of UEs and utilized numerologies.

Concluding, we also note that the exact solution is feasible for up to 10-15 UEs in a multicast group, while relaxation techniques, such as LB and RINS heuristics, although

producing a perfect match with the exact solution, do not reduce the solution time. The approximate simulated annealing techniques decrease the complexity leading to a linear increase in the solution time with the number of UEs. However, this happens at the expense of allocating 10-40% of more resources to serve the multicast group.

REFERENCES

[1] E. Dahlman, S. Parkvall, and J. Skold, *5G NR: The Next Generation Wireless Access Technology*. Academic Press, 2020.

[2] I. Qualcomm Technologies, "Deploying 5G NR mmWave to Unleash the Full 5G Potential," tech. rep., Tech Rep, 2020.

[3] 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.

[4] 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.

[5] J. Karvo, O. Martikainen, J. Virtamo, and S. Aalto, "Blocking of Dynamic Multicast Connections," *Telecommunication Systems*, vol. 16, no. 3, pp. 467–481, 2001.

[6] G. Araniti, M. Condoluci, M. Cotronei, A. Iera, and A. Molinaro, "A Solution to the Multicast Subgroup Formation Problem in LTE Systems," *IEEE Wireless Communications Letters*, vol. 4, no. 2, pp. 149–152, 2015.

[7] N. Chukhno, O. Chukhno, G. Araniti, A. Iera, A. Molinaro, and S. Pizzi, "Challenges and Performance Evaluation of Multicast Transmission in 60 GHz mmWave," in *International Conference on Distributed Computer and Communication Networks*, pp. 3–17, Springer, 2020.

[8] 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)*, pp. 1–7, IEEE, 2016.

[9] J. Peisa, P. Persson, S. Parkvall, E. Dahlman, A. Grovlen, C. Hoymann, and D. Gerstenberger, "5G evolution: 3GPP Releases 16 & 17 Overview," *Ericsson Technol. Rev.*, vol. 9, pp. 1–5, 2020.

[10] 5G Americas, "The 5G Evolution 3GPP Releases 16-17," *5G Americas White Paper*, 2020.

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

[12] 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.

[13] 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*, vol. 67, no. 3, pp. 593–605, 2021.

[14] M. Makolkina, A. Vikulov, and A. Paramonov, "The Augmented Reality Service Provision in D2D Network," in *International Conference on Distributed Computer and Communication Networks*, pp. 281–290, Springer, 2017.

[15] S. Naribole and E. Knightly, "Scalable Multicast in Highly-Directional 60-GHz WLANs," *IEEE/ACM Transactions on Networking*, vol. 25, no. 5, pp. 2844–2857, 2017.

[16] K. Sundaresan, K. Ramachandran, and S. Rangarajan, "Optimal Beam Scheduling for Multicasting in Wireless Networks," in *Proceedings of the 15th annual international conference on Mobile computing and networking*, pp. 205–216, 2009.

[17] 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.

[18] 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.

[19] Z. Zhang, Z. Ma, Y. Xiao, M. Xiao, G. K. Karagiannidis, and P. Fan, "Non-orthogonal Multiple Access for Cooperative Multicast Millimeter Wave Wireless Networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 8, pp. 1794–1808, 2017.

[20] L. Liu, Y. Ma, N. Yi, and R. Tafazolli, "An Analogue-Beam Splitting Approach for MmWave D2D Multicast Channel," in *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pp. 1–6, IEEE, 2018.

[21] N. Chukhno, O. Chukhno, S. Pizzi, A. Molinaro, A. Iera, and G. Araniti, "Unsupervised Learning for D2D-Assisted Multicast Scheduling in mmWave Networks," in *2021 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pp. 1–6, IEEE, 2021.

[22] N. Chukhno, O. Chukhno, D. Moltchanov, A. Molinaro, Y. Gaidamaka, K. Samouylov, Y. Koucheryavy, and G. Araniti, "Optimal Multicasting in Millimeter Wave 5G NR with Multi-beam Directional Antennas," *IEEE Transactions on Mobile Computing (Early Access)*, 2021.

[23] M. Hashemi, C. E. Koksal, and N. B. Shroff, "Energy-Efficient Power and Bandwidth Allocation in an Integrated Sub-6 GHz–Millimeter Wave System," *arXiv preprint arXiv:1710.00980*, 2017.

[24] P. Zhou, X. Fang, X. Wang, and L. Yan, "Multi-beam Transmission and Dual-Band Cooperation for Control/Data Plane Decoupled WLANs," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 10, pp. 9806–9819, 2019.

[25] G. Yao, M. Hashemi, and N. B. Shroff, "Integrating Sub-6 GHz and Millimeter Wave to Combat Blockage: Delay-Optimal Scheduling," in *2019 International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOPT)*, pp. 1–8, IEEE, 2019.

[26] M. Alrabeiah and A. Alkhateeb, "Deep Learning for mmWave Beam and Blockage Prediction Using Sub-6 GHz Channels," *IEEE Transactions on Communications*, vol. 68, no. 9, pp. 5504–5518, 2020.

[27] M. Cheng, J.-B. Wang, J. Cheng, J.-Y. Wang, and M. Lin, "Joint Scheduling and Precoding for mmWave and Sub-6GHz Dual-Mode Networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 11, pp. 13098–13111, 2020.

[28] V. Begishev, E. Sopin, D. Moltchanov, R. Pirmagomedov, A. Samuylov, S. Andreev, Y. Koucheryavy, and K. Samouylov, "Performance Analysis of Multi-Band Microwave and Millimeter-Wave Operation in 5G NR Systems," *IEEE Transactions on Wireless Communications*, vol. 20, no. 6, pp. 3475–3490, 2021.

[29] 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.

[30] M. Gapeyenko, V. Petrov, D. Moltchanov, M. R. Akdeniz, S. Andreev, N. Himayat, and Y. Koucheryavy, "On the Degree of Multi-Connectivity in 5G Millimeter-wave Cellular Urban Deployments," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 2, pp. 1973–1978, 2018.

[31] 3GPP, "Multimedia Broadcast/Multicast Service (MBMS); Stage 1," 3GPP TR 22.146, March 2022.

[32] ITU-R, "Propagation Data and Prediction Methods Required for the Design of Terrestrial Line-of-Sight Systems," *ITU-R P 530-17*, November 2017.

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

[34] 3GPP, "Study on Channel Model for Frequencies from 0.5 to 100 GHz (Release 14)," 3GPP TR 38.901 V14.1.1, July 2017.

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

[36] T. E. Bogale and L. B. Le, "Beamforming for Multiuser Massive MIMO Systems: Digital versus Hybrid Analog-Digital," in *2014 IEEE Global Communications Conference*, pp. 4066–4071, IEEE, 2014.

[37] I. Ahmed, H. Khammari, A. Shahid, A. Musa, K. S. Kim, E. De Poorter, and I. Moerman, "A Survey on Hybrid Beamforming Techniques in 5G: Architecture and System Model Perspectives," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 3060–3097, 2018.

[38] N. Stepanov, D. Moltchanov, V. Begishev, A. Turlikov, and Y. Koucheryavy, "Statistical Analysis and Modeling of User Micromobility for THz Cellular Communications," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 1, pp. 725–738, 2021.

[39] Z. AlSaeed, I. Ahmad, and I. Hussain, "Multicasting in Software Defined Networks: A Comprehensive Survey," *Journal of Network and Computer Applications*, vol. 104, pp. 61–77, 2018.

[40] 5G AI, "European Vision for the 6G Network Ecosystem," tech. rep., The 5G Infrastructure Association, June 2021.

[41] B. Zhou, H. Hu, S.-Q. Huang, and H.-H. Chen, "Intracluster device-to-device relay algorithm with optimal resource utilization," *IEEE transactions on vehicular technology*, vol. 62, no. 5, pp. 2315–2326, 2013.

[42] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*. MIT press, 2022.

[43] M. Pióro and D. Medhi, *Routing, Flow, and Capacity Design in Communication and Computer Networks*. Elsevier, 2004.

[44] B. H. Korte, J. Vygen, B. Korte, and J. Vygen, *Combinatorial optimization*, vol. 1. Springer, 2011.

[45] M. Fischetti and A. Lodi, "Local Branching," *Mathematical programming*, vol. 98, no. 1, pp. 23–47, 2003.

[46] E. Danna, E. Rothberg, and C. Le Pape, "Exploring Relaxation Induced Neighborhoods to Improve MIP Solutions," *Mathematical Programming*, vol. 102, no. 1, pp. 71–90, 2005.

[47] Y. Fu and A. Banerjee, "Heuristic/Meta-Heuristic Methods for Restricted Bin Packing Problem," *Journal of Heuristics*, vol. 26, no. 5, pp. 637–662, 2020.

[48] L. Hamm, B. W. Brorsen, and M. T. Hagan, "Comparison of Stochastic Global Optimization Methods to Estimate Neural Network Weights," *Neural Processing Letters*, vol. 26, no. 3, pp. 145–158, 2007.

[49] L. Wei, Z. Zhang, D. Zhang, and S. C. Leung, "A Simulated Annealing Algorithm for the Capacitated Vehicle Routing Problem with Two-Dimensional Loading Constraints," *European Journal of Operational Research*, vol. 265, no. 3, pp. 843–859, 2018.

[50] E. Hopper and B. C. Turton, "A Review of the Application of Meta-Heuristic Algorithms to 2D Strip Packing Problems," *Artificial Intelligence Review*, vol. 16, no. 4, pp. 257–300, 2001.

[51] E. Aarts and J. Korst, *Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing*. John Wiley & Sons, Inc., 1989.

[52] M. Giesen, G. Beltramo, S. Dieluweit, J. Müller, H. Ibach, and W. Schmickler, "The Thermodynamics of Electrochemical Annealing," *Surface science*, vol. 595, no. 1-3, pp. 127–137, 2005.

[53] M. Eusuff, K. Lansey, and F. Pasha, "Shuffled Frog-Leaping Algorithm: A Memetic Meta-Heuristic for Discrete Optimization," *Engineering optimization*, vol. 38, no. 2, pp. 129–154, 2006.

[54] C. Gallo and V. Capozzi, "A Simulated Annealing Algorithm for Scheduling Problems," *Journal of Applied Mathematics and Physics*, vol. 7, no. 11, pp. 2579–2594, 2019.

[55] O. Hasançebi, S. Çarbaş, and M. P. Saka, "Improving the Performance of Simulated Annealing in Structural Optimization," *Structural and Multidisciplinary Optimization*, vol. 41, no. 2, pp. 189–203, 2010.

[56] E. Aarts, J. Korst, and W. Michiels, "Simulated Annealing," in *Search methodologies*, pp. 187–210, Springer, 2005.

[57] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) System Scenarios (Release 15)," 3GPP 36.942 V15.0.0, Oct 2018.

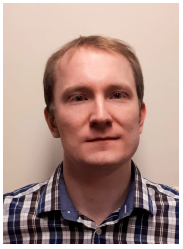
Olga Chukhno is Researcher at Mediterranea University of Reggio Calabria, Italy and Tampere University, Finland. She received M.Sc. (2019) in Fundamental Informatics and Information Technologies and B.Sc. (2017) in Business Informatics from RUDN University, Russia. She received her double Ph.D. degree from Mediterranea University of Reggio Calabria, Italy and Tampere University, Finland. Her current research interests include wireless communications and edge computing.

Nadezhda Chukhno is a Researcher at Tampere University, Finland. 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). She received her double Ph.D. degree from Mediterranea University of Reggio Calabria, Italy and Jaume I University, Spain. Her current research activity mainly focuses on wireless communications, 5G+ networks, multicasting, D2D, and ML.

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 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. 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 a Full 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.

Anna Gaydamaka received the B.Sc. (Hons) degree in Business Informatics from the Peoples' Friendship University of Russia (RUDN University) in 2018 and the M.Sc. degree in Computer Science (program Data Science and Business Informatics) from Universita di Pisa, Italy in 2021. Currently, she is pursuing a Ph.D. in Computing and Electrical Engineering at Tampere University, Finland. Her research interests include 5G and 6G wireless networks and machine learning.



Andrey Samuylov received the Ms.C. degree in applied mathematics and the Cand.Sc. degree in physics and mathematics from RUDN University, Russia, in 2012 and 2015. Currently he is pursuing a Ph.D. degree with the Unit of Electrical Engineering, Tampere University, Finland. His research interests include P2P networks performance analysis, performance evaluation of wireless networks with enabled D2D communications, and mmWave-band communications.

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.

Antonio Iera graduated in computer engineering from the University of Calabria in 1991, and received a Master's degree in IT from CEFRIEL/Politecnico di Milano in 1992 and a Ph.D. degree from the University of Calabria in 1996. From 1997 to 2019 he has been with the University Mediterranea, Italy, and currently holds the position of Full Professor of Telecommunications at the University of Calabria, Italy. His research interests include next-generation mobile and wireless systems, and the Internet of Things.

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 Associate 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).