# Non-Orthogonal Multiple Access for Cooperative Multicast Millimeter Wave Wireless Networks

Zhengquan Zhang, *Student Member, IEEE*, Zheng Ma, *Member, IEEE*, Yue Xiao, Ming Xiao, *Senior Member, IEEE*, George K. Karagiannidis, *Fellow, IEEE*, and Pingzhi Fan, *Fellow, IEEE*

***Abstract*— Millimeter wave (mmWave) wireless networks can operate in single-cell point-to-multipoint mode to provide local multicast services efficiently. In this paper, the performance of multicast mmWave wireless networks is studied, through stochastic geometry. Then, the use of power domain non-orthogonal multiple access (NOMA) for enhancing mmWave multicasting is also investigated. Furthermore, we study multicasting in two-tier mmWave heterogeneous networks, and propose a novel cooperative NOMA multicast scheme. Analytical expressions for the signal-to-interference-plus-noise ratio coverage probability, the average number of served users, and the sum multicast rate are derived, in order to assess the performance of these schemes. Finally, we discuss the maximum sum multicast rates, by formulating them as optimization problems, and also develop efficient golden section search algorithms to solve them. The offered solutions reveal the impact of data transmission rate and power allocation on the sum multicast rate. Both analytical and numerical results demonstrate that NOMA can significantly improve the mmWave multicasting, while the proposed cooperative NOMA mmWave multicast scheme can further improve the NOMA mmWave multicasting.**

***Index Terms*— Non-orthogonal multiple access, multicast transmission, millimeter wave communications, stochastic geometry.**

## I. Introduction

MULTICASTING enables the same content to be transmitted to all users or a specific group of users on the identical radio resources as point-to-multipoint (PTM) transmission, which is one spectrum-efficient mechanism for multimedia communications [1]. Multicasting has been adopted by the fourth-generation (4G) networks to efficently provide multimedia broadcast/multicast services (MBMS) in the form of multi-cell PTM transmission [2], [3] and single-cell PTM transmission [2], [4], and will be an important enabling technology to satisfy the 1000-fold data growth in the fifth-generation (5G) wireless networks and beyond [5]–[9].

### A. Related Literature

Multicasting was studied in detail in wireless networks [10], heterogeneous networks (HetNets) [11], and device-to-device (D2D) communications [12]. In 4G networks, multi-cell PTM transmission mainly provides services, as mobile TV, within relative static and wide coverage, while single-cell PTM transmission can be used to achieve specific mission critical communications (MCC)[1] [13] and group communications [14], [15]. In order to promote the development of 4G broadcast/multicast, fourteen business cases have been identified in [16]. Furthermore, with the evolution towards 5G networks, some works on multicasting for 5G and beyond have been done and several multicast applications were identified in [7]–[9] and [15], which can be categorized into *human-oriented* and *machine-oriented* [7]. The next generation mobile networks (NGMN) 5G initiative identified four broadcast/multicast-like services, which require 200 Mbps downlink data rate and <200 ms end-to-end latency [5]. However, conventional orthogonal multicasting operated in low frequency band cannot satisfy these requirements for 5G multicast services, as it suffers from low data rate transmission.[2] To provide high quality-of-experience (QoE) and data rate for 5G multicast services, millimeter-wave (mmWave) communications and non-orthogonal multiple access (NOMA) will be two important enabling technologies.

MmWave communications [17]–[21] utilizes the 30-300 GHz frequency band with rich spectrum resources to achieve multi-gigabit transmissions, which is one of the most promising technologies for 5G and beyond.

Manuscript received December 13, 2016; revised April 19, 2017; accepted April 28, 2017. Date of publication June 1, 2017; date of current version July 15, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 61571373, in part by the Key International Cooperation Project of Sichuan Province under Grant 2017HH0002, in part by the NSFC China-Swedish Project under Grant 61611012997, in part by the National Science and Technology Major Project under Grant 2016ZX03001018, and in part by the 111 Project under Grant 111-2-14, in part by the National Natural Science Foundation of China under Grant 61371105, and in part by the EU Marie Curie Project, "QUICK", No. 612652. The work of Z. Zhang was supported by the KTH-CSC Programme. *(Corresponding author: Zhengquan Zhang.)*

Z. Zhang, Z. Ma, Y. Xiao, and P. Fan are with the Key Lab of Information Coding and Transmission, Southwest Jiaotong University, Chengdu 610031, China (e-mail: zhang.zhengquan@hotmail.com; zma@home.swjtu.edu.cn; alice_xiaoyue@hotmail.com; pzfan@home.swjtu.edu.cn).

M. Xiao is with the Department of Information Science and Engineering, School of Electrical Engineering, KTH Royal Institute of Technology, SE-100 44, Stockholm, Sweden (e-mail: mingx@kth.se).

G. K. Karagiannidis is with the Key Lab of Information Coding and Transmission, Southwest Jiaotong University, 610031 Chengdu, China, and also with the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, 54 124 Thessaloniki, Greece (e-mail: geokarag@auth.gr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSAC.2017.2710918

[1]According to terrestrial trunked radio (TETRA) and critical communications association (TCCA), reliability, availability, stability and security are vital to ensure continuous availability of functions critical for society. The typical applications are law enforcement and emergency services for public safety and disaster relief, and also for general commercial applications (e.g., utility companies and railways).

[2]In order to ensure that all users can successfully decode the media content, the data transmission rate for conventional orthogonal multicasting relies on weak users with bad channel conditions.

MmWave channel measurement campaigns [18], [22] were conducted to understand mmWave propagation characteristics and facilitate channel modeling, which demonstrate that mmWave frequency band can be allocated to cellular communications, by employing highly directional antenna arrays. In [23], medium access control (MAC) layer techniques for mmWave cellular networks were investigated. The signal-to-interference-plus-noise ratio (SINR) coverage and rate performance of mmWave cellular networks were studied by using stochastic geometry in [24], and a comprehensive overview of mathematical models and analytical techniques for mmWave cellular networks, was provided in [25]. Random beamforming for mmWave NOMA networks was studied to improve the performance of unicast users, and the sum rate and outage probability were analysed in [26]. Recently, multicast mmWave networks, which can provide high data rate and low-latency multicast services, have also attracted some attention. The 5G-MiEdge project was launched by European Union (EU)-Japan to study the combination of mmWave access/backhauling and mobile edge computing (MEC), in order to provide mission critical low-latency applications and enhanced mobile broadband (eMBB) services [27]. In [28], an incremental multicast grouping scheme for mmWave networks with adaptive beamwidth was studied, to improve the throughput. In [29], two-stage transmit and receive beamforming for 60 GHz multicast mmWave networks, was studied. In [30], a joint optimization of relay selection and power allocation for cooperative multicast mmWave networks was studied, to minimize the outage probability. Finally, the optimization of beam sequence and multicast rate for sequential multicast transmission in directional mmWave networks was studied in [31].

On the other hand, NOMA can achieve excellent spectrum efficiency, by performing superposition transmission in power domain [32]. This has been adopted by 4G long term evolution advanced (LTE-A) networks to enhance downlink unicast transmission, named as multiuser superposition transmission (MUST) [33], and is being further studied for 5G systems in 3GPP. Several works have been devoted to the performance analysis and optimization of NOMA, including sum rate [34], [35], outage probability [34]–[37], and energy efficiency [38]. Recently, the application of NOMA for multicast enhancements has also attracted some attention. NOMA for enhancing MBMS transmission by two-layer superposition transmission to increase data transmission rate and spectral efficiency, was studied in [39]. Choi [40] studied minimum power multicast beamforming in NOMA systems for multiresolution broadcast. Furthermore, Ding *et al.* [41] studied the superposition transmission of multicast and unicast streams by NOMA to improve spectrum efficiency. Finally, in [42], a cooperative multicast for cognitive non-orthogonal multiple access scheme was studied, to improve the outage probability of primary users.

### *B. Motivation and Contribution*

From the aforementioned literature, it is concluded that although some works studied multicasting in mmWave networks, there still lacks not only tractable models for analysing the performance of multicast mmWave networks, but also the research on NOMA for multicast mmWave enhancements. More specifically, in [12], system modeling and performance analysis of multicast D2D transmissions by using stochastic geometry were provided, but the authors did not consider mmWave networks and NOMA. Bai and Heath [24] and Andrews *et al.* [25] studied system modeling and performance analysis of mmWave cellular networks by using stochastic geometry. However, they also did not study multicasting and NOMA. Beamforming for mmWave NOMA networks was studied in [26], which just considers unicasting rather than multicasting. Although multicast mmWave transmissions were studied in [28]–[31], they did not study tractable modes for performance analysis by using stochastic geometry, as well as NOMA. NOMA for multicast enhancements was studied in [39]–[42], but they did not consider mmWave wireless networks. Motivated by the above analysis, in this paper, we study multicast mmWave networks and the application of NOMA for multicast mmWave enhancements, and then propose cooperative NOMA multicast scheme for two-tier mmWave HetNets, followed by developing a tractable model for performance analysis, which are the main novelty and difference from existing works. More specifically, we first present system modeling and performance analysis of multicast mmWave wireless networks, by using stochastic geometry. Then, we further study the use of NOMA for enhancing multicast mmWave networks. Furthermore, we study multicast transmission in a two-tier mmWave HetNet and propose cooperative NOMA multicast scheme. Based on the tractable model, analytical expressions for the SINR coverage probability, average number of served users, and sum multicast rate are also derived. Finally, network optimization on multicast transmission rate and NOMA power allocation is also discussed.

From a practical point of view, multicast mmWave wireless networks, as the integration of mmWave communications with multi-gigabit transmittability and multicasting having high spectrum efficiency, can provide small cell access for a large amount of multicast groups to achieve high-speed and low-latency data transmission, which is very attractive to provide local multicast services [5] rather than regional and national multicast services. For human-oriented multicast applications, multicast mmWave networks are envisioned to efficiently provide services for the use cases, as mobile video and public safety etc. Multicast mmWave networks can provide high-speed data transmission for group-based video services for the typical scenarios, including video conferences, sporting events, concerts, operas, as well as emerging augmented reality (AR) multicast application [7], [8], especially for tourist or commercial services. Public safety [7], [16] is another important multicast application for festivals/events/fairs in densely populated local areas, as stadiums, squares, shopping malls, airports, and stations. In this case, multicast mmWave networks can deliver very time-sensitive information (such as text, pictures, or videos) to a myriad of users. The 5G-MiEdge project is devoted to developing 5G mmWave architecture together with MEC, in order to enable eMBB services and

mission critical low-latency applications at the 2020 Tokyo Olympic Games. With the application and popularization of machine-type communications (MTC), the emerging machine-oriented multicast applications have attracted an increasing attention. The typical use cases for machine-oriented multicast applications include smart environments, intelligent transport, and software/firmware update etc. Multicast mmWave networks can provide low-latency and low-energy group-based multicast services for smart environments [7], including smart homes/offices/shops/industrial plants etc., to reduce costs, improve the quality of life, and optimize industrial processes. Intelligent transport systems can also benefit from multicast mmWave networks, which can achieve high data rate and low-latency traffic data exchange to optimize fleet management [7] and autonomous vehicles [20], [21] etc., such that traffic efficiency and traffic safety can be significantly improved. Software/firmware update [7], [16] for smart devices, as sensors and smart phones, is another important machine-oriented multicast case. In this case, multicast mmWave networks can simultaneously support a large amount of multicast groups to distribute large files with high data rate. Therefore, multicast mmWave networks have a brightly practical perspective for providing local multicast services efficiently, which are worthy to be intensively studied. First, in order to characterize the average system performance and obtain tractable performance models for facilitating network design and optimization, we study system modeling and performance analysis of multicast mmWave networks by using stochastic geometry. Then, in order to overcome the loss of data transmission rate caused by orthogonal transmission in multicast mmWave networks and increase spectrum efficiency, we study the application of NOMA to multicast mmWave networks. This enables scalable multimedia to be delivered by superposition transmission in power domain, such that users can decode appropriate multicast data layers according to their channel conditions. Furthermore, in order to achieve excellent multicast performance in mmWave HetNets, we study multicast transmission in a two-tier mmWave HetNet and propose cooperative NOMA multicasting scheme. Finally, in order to obtain optimal system performance, we also use optimization theory to study optimal network configurations, based on the tractable performance model.

The contribution of this paper can be summarized as follows.

- We study multicast mmWave wireless networks and present the system modeling and performance analysis by using stochastic geometry. Based on this tractable model, analytical expressions for the SINR coverage probability, average number of served users, and sum multicast rate are derived. Besides, we also investigate the effects of data transmission rate on the sum multicast rate with/without QoS constraints, by formulating them as optimization problems, and develop golden section search (GSS)-based algorithms.
- We further study the use of NOMA for enhancing mmWave multicasting. Without loss of generality, a two-layer NOMA for multicast mmWave wireless networks is presented, which enables the secondary layer (SL) with low transmit power to be superposed to the original primary layer (PL) in power domain. This superposition transmission improves the spectrum efficiency of multicasting, because with NOMA, users can decode their desired data layers according to their channel conditions. To be specific, weak users can decode the primary layer to obtain the basic QoS, while strong users can further decode the secondary layer to obtain the enhanced data to improve the QoS. Furthermore, we present the system modeling and performance analysis by using stochastic geometry, and derive analytical expressions for the SINR coverage probability, average number of served users, and sum multicast rate. Besides, we also discuss the maximum sum multicast rates for NOMA multicast and study the impacts of data transmission rate and power allocation on that by formulating them as optimization problems. In order to solve these optimization problems, we also develop GSS-based algorithms.
- We also study multicasting in a two-tier mmWave HetNet consisting of one low frequency MBS tier and one mmWave small cell tier, and propose cooperative NOMA multicast to further improve the NOMA multicast performance. This scheme enables the MBS tier to cooperatively transmit the primary layer with low data rate. Therefore, the proposed scheme can increase the success probability that users can decode the primary and secondary layers, as a result it can improve the NOMA multicast performance. We also present system modeling and performance analysis of the proposed scheme, by using stochastic geometry.

### C. Paper Outline

The rest of the paper is organized as follows. Section II describes the system model and explains the main concept of NOMA multicast. A detailed system modeling and performance analysis of mmWave multicast, NOMA for mmWave multicast enhancement and cooperative NOMA multicast in mmWave HetNets, are presented in Section III, IV, and V, respectively. Analytical results, Monte Carlo simulations and discussion are presented in Section VI, followed by the conclusions in Section VII.

## II. System Model

Fig. 1 illustrates a two-tier HetNet, which consists of one MBS tier with low frequency bands and one mmWave small cell tier. According to [44] and [45], independently homogeneous Poisson point process (PPP) can be used to model the locations of MBSs and mmWave small cells, denoted as, $\Phi_M$, with density, $\lambda_M$, and, $\Phi_S$, with density, $\lambda_S$, respectively. Similarly, the locations of users can be also represented by some PPP, $\Phi_U$, with density, $\lambda_U$. The maximum transmit power of MBSs and mmWave small cells are assumed to be, $P_M$, and, $P_S$, respectively.

### A. NOMA for Multicast Communications

Multicast transmission can be enhanced through power domain division multiplexing (PDM)-based NOMA, as multi-rate multicasting [43]. Without loss of generality, two-layer

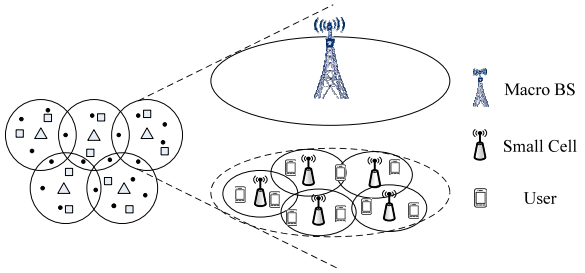



Fig. 1. System model of a two-tier HetNet.

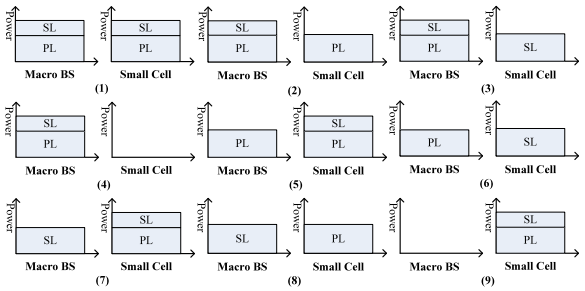



Fig. 2. NOMA for multicast communications in a two-tier HetNet.

multiplexing in power domain is considered, where the primary layer has higher priority and more power is allocated, while the rest of the power is allocated to the secondary layer. Multi-rate multicasting is very efficient for scalable multimedia transmission, where the data are first coded to the base data and enhanced data with different QoS requirements through source layered coding, and then they are delivered by superposition transmission in power domain. In this case, the primary layer carries data with basic QoS requirement, while the data with enhanced QoS requirement are transmitted in the secondary layer. As a result, users can decode the corresponding data sub-layers through successive interference cancellation (SIC) [32], [34], according to their channel conditions. More specifically, the users with weak channel conditions can decode the primary layer, in order to obtain the basic QoS, while the users with strong channel conditions can obtain better QoS, through decoding both the primary and secondary layers. Therefore, NOMA multicast can fully utilize the channel non-similarity between users to increase multicast data transmission rate and improve the spectrum efficiency. Furthermore, Fig. 2 illustrates NOMA schemes for multicast communications in two-tier HetNets. In this paper, we mainly focus on cooperative NOMA multicast scheme for a two-tier mmWave HetNet, that can achieve more reasonable performance tradeoffs.

### *B. Channel Modeling*

*1) Path Loss:* For macro BSs, the power-law path loss model with path loss exponent, $\alpha_M > 2$, is used to characterize the signal attenuation [45]. Therefore, the average power received at the user is, $P_{\mathrm{Rx}}(d) = P_M d^{-\alpha_M}$. However, there are obvious line-of-sight (LOS) and non-LOS (NLOS) links in mmWave communications, which require different path loss exponents. Given that a link has distance, $d$, the path-loss model can be given by [25]

$$L_S(d) = \begin{cases} C_{S,L} d^{-\alpha_{S,L}}, & \text{wp } p_L(d), \\ C_{S,N} d^{-\alpha_{S,N}}, & \text{wp } p_N(d), \end{cases} \tag{1}$$

where $\alpha_{S,L}$ and $\alpha_{S,N}$ are the path loss exponents for mmWave LOS and NLOS links respectively, $C_{S,L}$ and $C_{S,N}$ are the corresponding intercepts of path loss formulas for mmWave LOS and NLOS links, which are the function of reference distance and wavelength (or carrier frequency) and are equal to $10^{-2\log_{10}(4\pi/\lambda_c)}$ for the same close-in reference distance $d_{\mathrm{ref}} = 1$ m [22], $p_L(d)$ is the probability that a link having length $d$ is LOS, and $p_N(d) = 1 - p_L(d)$ is the probability of the NLOS one.

*2) Small Scale Fading:* The links between users and macro BSs are assumed to be subjected to Rayleigh fading. Let $h_{M,i}$ be the channel coefficient of the link between the $i$-th macro BS and the user. Then, $H_{M,i} = |h_{M,i}|^2$ follows exponential distribution with mean one, i.e., $H_{M,i} \sim \exp(1)$ [45]. According to [24] and [25], it is assumed that each mmWave link follows independent Nakagami-$m$ fading. Besides, different Nakagami-$m$ fading parameters, $N_{S,L}$ and $N_{S,N}$, are applied to the mmWave LOS and NLOS links, respectively. Let $h_{S,i}$ is the channel coefficient of the link between the $i$-th mmWave small cell and the user. Then, $H_{S,i} = |h_{S,i}|^2$, follows a normalized Gamma distribution. Finally, similar to [24], [25], shadowing is ignored.

*3) Directivity Gain:* Beamforming [19] can enable antenna arrays to form directional beams, which is one key technology to overcome the high path loss in mmWave communications. The deployment of antenna arrays[3] at both BSs and user equipments (UEs) is assumed. Similar to [23]–[25], sectored antenna model is used to approximate the beamforming pattern for tractable analysis. In this case, the directivity gain that can be obtained is given by

$$G(\phi) = \begin{cases} G_M, & |\phi| \le \theta, \\ G_m, & |\phi| > \theta, \end{cases} \tag{2}$$

where $G_M$ and $G_m$ are the main and side lobe gains, respectively, $\phi$ is some angle, and $\theta$ is the beamwidth of the main lobe. Furthermore, according to [23], $G_M$ and $G_m$ can be equal to $\frac{2\pi-(2\pi-\theta)\epsilon}{\theta}$ and $\epsilon$, respectively. At the mmWave small cell side, $G_{S,M}$, $G_{S,m}$, and $\theta_S$ denote the main lobe gain, side lobe gain, and beamwidth, respectively, while their corresponding symbols are $G_{U,M}$, $G_{U,m}$, and $\theta_U$ at the user side. Therefore, the total directivity gain of the communication link between the user and its serving mmWave small cell is, $G_{S,0} = G_S(\phi_S)G_U(\phi_U)$, where $\phi_S$ and $\phi_U$ are the angle of departure (AoD) and the angle of arrival (AoA) of the signal, respectively. According to [24] and [25], the directivity gain of the $i$-th interference mmWave link is assumed to be a discrete random variable (RV), whose probability distribution is $G_{S,i} = a_k$ with probability $b_k, k \in \{1, 2, 3, 4\}$, where $a_k$ and $b_k$ are constants defined in Table I. Note that for macro BSs, we consider omnidirectional antennas, i.e., $\theta = 2\pi$, as a result there is no directivity gain.

[3]According to 3GPP TR 38.913, up to 256 Tx and Rx BS antenna elements and 32 Tx and Rx UE antenna elements are assumed.

TABLE I
PROBABILITY MASS FUNCTION OF $G_{S,i}(i \geq 1)$

| $k$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $a_k$ | $G_{U,M}G_{S,M}$ | $G_{U,M}G_{S,m}$ | $G_{U,m}G_{S,M}$ | $G_{U,m}G_{S,m}$ |
| $b_k$ | $\frac{\theta_U\theta_S}{(2\pi)^2}$ | $\frac{\theta_U(2\pi-\theta_S)}{(2\pi)^2}$ | $\frac{(2\pi-\theta_U)\theta_S}{(2\pi)^2}$ | $\frac{(2\pi-\theta_U)(2\pi-\theta_S)}{(2\pi)^2}$ |

## III. PERFORMANCE ANALYSIS OF mmWAVE MULTICAST

In this section, the system modeling and performance analysis of multicast mmWave wireless networks by using stochastic geometry, are presented. Analytical expressions for the SINR coverage probability, average number of served users and sum multicast rate are derived. We also discuss the maximum sum multicast rates with/without QoS constraints, by formulating them as optimization problems and GSS-based algorithms are developed to solve them.

### A. Received SINR

The user with a random distance, $d_0$, receives the signal from its serving mmWave small cell, and interference signals from other mmWave small cells as well. Thus, the sum signal received at the user can be expressed as

$$y_S^1 = h_{S,0}\sqrt{G_{S,0}P_S L_S(d_0)}x_0 + \underbrace{\sum_{\mathrm{X_i}\in\Phi_S\setminus\mathrm{B}_0} h_{S,i}\sqrt{G_{S,i}P_S L_S(d_i)}x_i}_{I_S} + n_S, \quad (3)$$

where $\mathrm{X}_i$ is the location of the $i$-th interference mmWave small cell, $d_i = \|\mathrm{X}_i\|_2$ is the distance between the $i$-th interference mmWave small cell and the user, and $n_S$ is the thermal noise. Let $\sigma_S^2$ be the normalized thermal noise power by the transmit power, $P_S$.

Therefore, the SINR received at the user can be written as

$$\mathrm{SINR_S^1} = \frac{\mathrm{H_{S,0}G_{S,0}L_S(d_0)}}{\underbrace{\sum_{\mathrm{X_i}\in\Phi_S\setminus\mathrm{B}_0} H_{S,i}G_{S,i}L_S(d_i)}_{\mathrm{I_S}} + \sigma_S^2}. \quad (4)$$

### B. Downlink SINR Coverage Probability

The SINR coverage probability [24], [25] characterizes the average quality of network coverage, and is defined as the probability that the received SINR is beyond a given threshold, $T$

$$P_c(T) = \mathbb{P}(\mathrm{SINR} > \mathrm{T}), \quad (5)$$

where SINR relies on the specific cases and will be given explicitly later.

*Proposition 1:* For fixed SINR thresholds, $T$, the downlink SINR coverage probability can be approximately expressed as [24], [25]

$$\begin{aligned} P_{c,S}^1(T) \approx\ & A_L\sum_{n=1}^{N_L}(-1)^{n+1}\binom{N_L}{n} \\ & \times\int_0^\infty e^{-\left(\frac{n\eta_L x^{\alpha_{S,L}}T\sigma_S^2}{C_L G_{S,0}}+Q_n(T,x)+V_n(T,x)\right)}\hat{f}_L(x)dx \\ & + A_N\sum_{n=1}^{N_N}(-1)^{n+1}\binom{N_N}{n} \\ & \times\int_0^\infty e^{-\left(\frac{n\eta_N x^{\alpha_{S,N}}T\sigma_S^2}{C_N G_{S,0}}+W_n(T,x)+Z_n(T,x)\right)}\hat{f}_N(x)dx, \end{aligned} \quad (6)$$

where $s \in \{L, N\}$, $A_s$ is the probability that the user is associated with one mmWave small cell, $\eta_s = N_s(N_s!)^{-1/N_s}$ and $N_s$ are the parameters of Nakagami-$m$ small-scale fading, $\hat{f}_s(.)$ is the conditional probability density function (PDF) of the distance to the nearest mmWave small cell, $Q_n(.)$ and $V_n(.)$ are the corresponding terms related with LOS and NLOS interference links for the serving LOS mmWave small cell, $W_n(.)$ and $Z_n(.)$ are the corresponding terms related with LOS and NLOS interference links for the serving NLOS one, which are defined in detail in works [24], [25].

*Proof:* Substituting (4) into (5), we can get

$$P_{c,S}^1(T) = \mathbb{P}(\mathrm{SINR_S^1} > \mathrm{T}) = \mathbb{P}\left(\mathrm{H_{S,0}} > \frac{\mathrm{T}(\sigma_S^2+\mathrm{I_S})}{\mathrm{G_{S,0}L_S(d_0)}}\right). \quad (7)$$

According to [24] and [25], the SINR downlink coverage probability of mmWave small cells can be obtained as in (6). ■

### C. Average Number of Served Users

We consider the multicast cluster, $B_o(0,\ R)$, with radius, $R = (\pi\lambda_S)^{-1/2}$, because its performance can imply the spatially averaged performance over all multicast cells. For an arbitrary user, $y \in \Phi_{U,B_o}$, let $E^1(y) = \{\log(1+\mathrm{SINR_{S,y}^1}) \geq \mathrm{R_T}\}$, which represents the event that the user is in the coverage of multicasting with data rate, $R_T$. In this case, its SINR form can be written as, $E^1(y) = \{\mathrm{SINR_{S,y}^1} \geq 2^{\mathrm{R_T}} - 1\}$. Therefore, the average number of served users by the multicast cluster, $B_o$, with data rate, $R_T$, is given by [12]

$$\mathbb{E}^o[N^1] \triangleq \mathbb{E}^o\left[\sum_{y\in\Phi_{U,B_o}}\mathbb{I}(E^1(y))\right], \quad (8)$$

where $\mathbb{I}(.)$ is the indicator function.

*Lemma 1:* Given fixed multicast data rate, $R_T$, then the average number of served users by the mmWave multicast cluster, can be expressed as

$$\mathbb{E}^o[N^1] = \lambda_U P_{c,S}^1(T)\theta_S(2\pi\lambda_S)^{-1}, \quad (9)$$

where, $T = 2^{R_T} - 1$ is the SINR threshold.

*Proof:* See Appendix A. ■

Note that for omnidirectional antennas, i.e., $\theta_S = 2\pi$, the average number of served users can be rewritten as

$$\mathbb{E}^o[N^1] = \lambda_U P_{c,S}^1(T)/\lambda_S. \quad (10)$$

### D. Sum Multicast Rate

The sum multicast rate is defined as the mean of the sum rate of all users, who are in the multicast area and decode the media with data rate, $R_T$, successfully. It can be expressed as

$$\bar{R}^1_{\text{sum}} = R_T \mathbb{E}^o[N^1]. \tag{11}$$

*Theorem 1:* Given fixed multicast data rate, $R_T$, then the sum multicast rate for the mmWave multicast cluster is

$$\bar{R}^1_{\text{sum}} = R_T \lambda_U P^1_{c,S}(T) \theta_S (2\pi \lambda_S)^{-1}, \tag{12}$$

where, $T = 2^{R_T} - 1$, is the SINR threshold.

*Proof:* Combining (6), (9), and (11), the sum multicast rate can be obtained as in (12) and the proof is completed. ■

Note that with omnidirectional antennas, i.e., $\theta_S = 2\pi$, the sum multicast rate for the multicast cluster can be written as

$$\bar{R}^1_{\text{sum}} = R_T \lambda_U P^1_{c,S}(T) / \lambda_S. \tag{13}$$

Next, we investigate the maximization of the sum multicast rate under different constraints.

*1) Fixing Beamwidth and Optimize Multicast Data Transmission Rate Without QoS Constraints:* Given the beamwidth, then the sum multicast rate varies with different multicast data transmission rate, $R_T$. On the one hand, with larger $R_T$, more multicast data can be transmitted per time-frequency resource. However, larger $R_T$ leads to a decrease of the average number of served users, because less users can decode the multicast data successfully. Therefore, in order to achieve the maximum sum multicast rate, the optimization problem can be formulated as

$$\begin{aligned} &\max_{R_T} \ \bar{R}^1_{\text{sum}} \\ &\text{s. t. } R_T > 0. \end{aligned} \tag{P1}$$

problem P1 and to obtain the optimal multicast data rate, $R^*_T$. Note that in order to limit the search space, we arbitrarily give an upper limit, based on the simulation results.

**Algorithm 1** Multicast Data Transmission Rate Optimization Without QoS Constraints

1: Initialization: $R_{T,a} \leftarrow 0.1$, $R_{T,b} \leftarrow 12$, $\epsilon \leftarrow 0.1$;
2: Construct object function $f(R_T)$ according to (6) and (12);
3: Call GOLDENSECTIONSEARCHALG $f$, $R_T$, $R_{T,a}$, $R_{T,b}$, $\epsilon$ shown in **Subalgorithm 1**.

*2) Fixing Beamwidth and Optimize Multicast Data Rate With QoS Constraints:* In general, it is required to not only maximize the sum multicast rate, but also ensure that at least, $0 < N_T \leq \lambda_U R^2 \theta_S / 2$, users can be served by the mmWave multicast area, with the minimum data rate, $R_{T,m}$. This is equivalent to the fact that the SINR coverage probability with threshold, $T = 2^{R_T} - 1$, should be no smaller than, $\eta_T = \frac{2N_T}{\lambda_U R^2 \theta_s}$. Consequently, the optimization problem can be formulated as

$$\begin{aligned} &\max_{R_T} \ \bar{R}^1_{\text{sum}} \\ &\text{s. t. } R_T \geq R_{T,m}, \quad P^1_{c,S}(T) \geq \eta_T. \end{aligned} \tag{P2}$$

**Subalgorithm 1** Golden Section Search

**Input:**
1: $f$: Objective function;
2: $var$: Function variable;
3: $a$: The left limit of the search interval;
4: $b$: The right limit of the search interval;
5: $\epsilon$: Predefined precision;
**Output:**
6: optimal point $var^*$ and corresponding objective function value $f^*$;
7: **function** GOLDENSECTIONSEARCHALG($f, var, a, b, \epsilon$)
8: Obtain two golden section points: $var^{(1)} \leftarrow b - 0.618(b-a)$ and $var^{(2)} \leftarrow a + 0.618(b-a)$;
9: Compute objective function values: $f_1 \leftarrow f(var^{(1)})$ and $f_2 \leftarrow f(var^{(2)})$;
10: **while** $b - a > \epsilon$ **do**
11: **if** $f_1 > f_2$ **then**
12: Update $b$, $var^{(1)}$, $var^{(2)}$, $f_1$, and $f_2$: $b \leftarrow var^{(2)}$, $var^{(2)} \leftarrow var^{(1)}$, $f_2 \leftarrow f_1$,
13: $var^{(1)} \leftarrow b - 0.618(b-a)$, and $f_1 = f(var^{(1)})$;
14: **else**
15: Update $a$, $var^{(1)}$, $var^{(2)}$, $f_1$, and $f_2$: $a \leftarrow var^{(1)}$, $var^{(1)} \leftarrow var^{(2)}$, $f_1 \leftarrow f_2$,
16: $var^{(2)} \leftarrow a + 0.618(b-a)$, and $f_2 = f(var^{(2)})$;
17: **end if**
18: **end while**
19: Compute the optimal point $var^*$ and its corresponding objective function value $f^*$:
20: $var^* \leftarrow (a+b)/2$ and $f^* \leftarrow f(var^*)$;
21: **return** $var^*$, $f^*$.
22: **end function**

Furthermore, the constraint, $P^1_{c,S}(T) \geq \eta_T$, can be converted into, $R_T \leq f^{-1}(\eta_T)$, due to non-increasing function, $P^1_{c,S}(T)$. As a result, the original problem P2 can be transformed into the problem as

$$\begin{aligned} &\max_{R_T} \ \bar{R}^1_{\text{sum}} \\ &\text{s. t. } R_{T,m} \leq R_T \leq f^{-1}(\eta_T). \end{aligned} \tag{P2A}$$

Similarly, **Algorithm 2** based on the GSS method is also developed to solve the optimization problem P2A and obtain the optimal multicast data rate, $R^*_T$, under QoS constraint.

**Algorithm 2** Multicast Data Transmission Rate Optimization With QoS Constraints

1: Initialization: $R_{T,a} \leftarrow R_{T,m}$ and $\epsilon \leftarrow 0.1$;
2: Construct objective function $f(R_T)$ according to (6) and (12);
3: Obtain the right limit of the search interval: $R_{T,b} \leftarrow f^{-1}(\eta_T)$;
4: Call GOLDENSECTIONSEARCHALG $f$, $R_T$, $R_{T,a}$, $R_{T,b}$, $\epsilon$ shown in **Subalgorithm 1**.

## IV. Performance Analysis of NOMA for mmWave Multicast Enhancements

In this section, NOMA for enhancing the performance of multicast mmWave wireless networks is studied, and the system modeling and performance analysis of NOMA multicast by using stochastic geometry are presented. Analytical expressions for the SINR coverage probability, average number of served users and sum multicast rate are also derived. The impacts of data transmission rate and power allocation on the sum multicast rates for NOMA multicast are also investigated by formulating them as optimization problems, and GSS-based algorithms are developed to solve them.

### A. Received SINR

With NOMA, the mmWave small cell transmits a superposed signal to all users within its coverage as

$$x = \alpha_p^{1/2} x_{PL} + (1-\alpha_p)^{1/2} x_{SL}, \tag{14}$$

where $0 < \alpha_p < 1$ is the power allocation factor (PAF), $x_{PL}$ and $x_{SL}$ are the transmit signals of the primary and secondary layers, respectively. The signal received at the user with random distance, $d_0$, from its serving and interference mmWave small cells can be expressed as

$$y_S^2 = h_{S,0}\sqrt{G_{S,0}P_S L_S(d_0)}x_0 + \underbrace{\sum_{\mathrm{X_i}\in\Phi_S\setminus \mathrm{B}_0} h_{S,i}\sqrt{G_{S,i}P_S L_S(d_i)}x_i}_{I_S} + n_S. \tag{15}$$

After receiving the superposed signal, the user first decodes the primary layer, and then cancel it from the received signal before decoding the secondary layer. Therefore, substituting (14) into (15), the SINRs of detecting the primary and secondary layers can be written, respectively, as

$$\mathrm{SINR}_{\mathrm{S,PL}}^2 = \frac{\alpha_p H_S G_{S,0} L_S(d_0)}{(1-\alpha_p)H_S G_{S,0} L_S(d_0) + \underbrace{\sum_{\mathrm{X_i}\in\Phi_S\setminus \mathrm{B}_0} H_{S,i} G_{S,i} L_S(d_i)}_{I_S} + \sigma_S^2}, \tag{16}$$

and

$$\mathrm{SINR}_{\mathrm{S,SL}}^2 = \frac{(1-\alpha_\mathrm{p})\mathrm{H_S G_{S,0} L_S(d_0)}}{\underbrace{\sum_{\mathrm{X_i}\in\Phi_S\setminus \mathrm{B}_0} H_{S,i} G_{S,i} L_S(d_i)}_{\mathrm{I_S}} + \sigma_\mathrm{S}^2}. \tag{17}$$

Note that $\sigma_S^2$ is the thermal noise power, normalized by transmit power, $P_S$.

### B. Downlink SINR Coverage Probability

*Proposition 2:* For fixed SINR thresholds for the primary and secondary layers, $T_{PL}$ and $T_{SL}$, the downlink SINR coverage probability of the primary layer can be expressed as in (18), as shown at the top of the next page, while the downlink SINR coverage probability of both primary and secondary layers can be expressed as

$$P_{c,S,PSL}^2(T_{PL}, T_{SL}, \alpha_p) = \begin{cases} P_{c,S,PL}^2(T_{PL}, \alpha_p), \alpha_p \le \dfrac{T_{PL}(1+T_{SL})}{T_{SL}+T_{PL}(1+T_{SL})}, \\ P_{c,S,SL}^2(T_{SL}, \alpha_p), \alpha_p > \dfrac{T_{PL}(1+T_{SL})}{T_{SL}+T_{PL}(1+T_{SL})}, \end{cases} \tag{19}$$

where $P_{c,S,PL}^2(T_{PL}, \alpha_p)$ is expressed as in (18), and $P_{c,S,SL}^2(T_{SL}, \alpha_p)$ is approximated as

$$\begin{aligned} P_{c,S,SL}^2(T_{SL}, \alpha_p) &\approx A_L \sum_{n=1}^{N_L} (-1)^{n+1} \binom{N_L}{n} \\ &\times \int_0^\infty e^{-\left(\frac{n\eta_L x^{\alpha_{S,L}} T_{SL}\sigma_S^2}{C_L(1-\alpha_p)G_S} + Q_n(T_{SL},x) + V_n(T_{SL},x)\right)} \hat{f}_L(x)dx \\ &+ A_N \sum_{n=1}^{N_N} (-1)^{n+1} \binom{N_N}{n} \\ &\times \int_0^\infty e^{-\left(\frac{n\eta_N x^{\alpha_{S,N}} T_{SL}\sigma_S^2}{C_N(1-\alpha_p)G_S} + W_n(T_{SL},x) + Z_n(T_{SL},x)\right)} \hat{f}_N(x)dx. \end{aligned} \tag{20}$$

*Proof:* The downlink SINR coverage probability that users can decode the primary layer relative to the SINR threshold $T_{PL}$ can be written as

$$P_{c,PL}^2(T_{PL}) = \mathbb{E}_R[\mathbb{P}[\mathrm{SINR}_{\mathrm{PL}}^2 > \mathrm{T_{PL}} \mid \mathrm{R} = \mathrm{r}]]. \tag{21}$$

Considering the maximum SINR for detecting the primary layer, $\lim_{H_s\to\infty} \mathrm{SINR}_{\mathrm{PL}}^2 = \frac{\alpha_\mathrm{p}}{1-\alpha_\mathrm{p}}$, the integral domain is

$$\begin{aligned} D &= \left\{(H_S) \mid \mathrm{SINR}_{\mathrm{PL}}^2 > \mathrm{T_{PL}}\right\} \\ &= \left\{(H_S) \mid H_S > \frac{T_{PL}(I_S+\sigma_S^2)}{(\alpha_p-(1-\alpha_p)T_{PL})G_{S,0}L_S(r_0)} \mid \right. \\ &\qquad \left. T_{PL} < \frac{\alpha_p}{1-\alpha_p}\right\}. \end{aligned} \tag{22}$$

Therefore, the SINR coverage probability can be expressed as

$$P_{c,S,PL}^2(T_{PL}, \alpha_p) = \mathbb{E}_R\left[\mathbb{P}\left[H_S > \frac{T_{PL}(I_S+\sigma_S^2)}{(\alpha_p-(1-\alpha_p)T_{PL})G_{S,0}L_S(r_0)} \mid R = r\right]\right]. \tag{23}$$

According to [24] and [25] and after some manipulations, the downlink SINR coverage probability of the primary layer of the multicast can be obtained as in (18).

$$P^2_{c,S,PL}(T_{PL},\alpha_p) \approx \begin{cases} A_L \sum_{n=1}^{N_L}(-1)^{n+1}\binom{N_L}{n}\int_0^{\infty} e^{-\left(\frac{n\eta_L x^{\alpha_{S,L}} T_{PL}\sigma_S^2}{C_L(\alpha_p-(1-\alpha_p)T_{PL})G_S}+Q_n(T_{PL},x)+V_n(T_{PL},x)\right)}\hat{f}_L(x)dx \\ +A_N \sum_{n=1}^{N_N}(-1)^{n+1}\binom{N_N}{n}\int_0^{\infty} e^{-\left(\frac{n\eta_N x^{\alpha_{S,N}} T_{PL}\sigma_S^2}{C_N(\alpha_p-(1-\alpha_p)T_{PL})G_S}+W_n(T_{PL},x)+Z_n(T_{PL},x)\right)}\hat{f}_N(x)dx, & T_{PL} < \frac{\alpha_p}{1-\alpha_p}, \\ 0, & T_{PL} \geq \frac{\alpha_p}{1-\alpha_p}. \end{cases} \quad (18)$$

The downlink SINR coverage probability by both the primary and secondary layers relative to the SINR thresholds $T_{PL}$ and $T_{SL}$ can be written as

$$P^2_{c,PSL}(T_{PL},T_{SL},\alpha_p) = \mathbb{E}_R[\mathbb{P}[\{\mathrm{SINR}^2_{\mathrm{PL}} > \mathrm{T_{PL}} \cap \mathrm{SINR}^2_{\mathrm{SL}} > \mathrm{T_{SL}}\} \mid \mathrm{R=r}]]. \quad (24)$$

Its integral domain is

$$D = \left\{(H_S) \mid \{\mathrm{SINR}^2_{\mathrm{PL}} > \mathrm{T_{PL}} \cap \mathrm{SINR}^2_{\mathrm{SL}} > \mathrm{T_{SL}}\}\right\} = \left\{(H_S) \mid \left\{H_S > \frac{T_{PL}(I_S+\sigma_S^2)}{(\alpha_p-(1-\alpha_p)T_{PL})G_{S,0}L_S(r_0)} \cap H_S > \frac{T_{SL}(I_S+\sigma_S^2)}{(1-\alpha_p)G_{S,0}L_S(r_0)}\right\}\right\}. \quad (25)$$

Let $\frac{T_{PL}(I_S+\sigma_S^2)}{(\alpha_p-(1-\alpha_p)T_{PL})G_{S,0}L_S(r_0)} = \frac{T_{SL}(I_S+\sigma_S^2)}{(1-\alpha_p)G_{S,0}L_S(r_0)}$, we can obtain $\alpha_p = \frac{T_{PL}(1+T_{SL})}{T_{SL}+T_{PL}(1+T_{SL})}$. As a result, the integral domain can be divided into two sub-domains as

$$D1 = \left\{(H_S) \mid H_S > \frac{T_{PL}(I_S+\sigma_S^2)}{(\alpha_p-(1-\alpha_p)T_{PL})G_{S,0}L_S(r_0)} \mid \alpha_p \leq \frac{T_{PL}(1+T_{SL})}{T_{SL}+T_{PL}(1+T_{SL})}\right\}, \quad (26)$$

and

$$D2 = \left\{(H_S) \mid H_S > \frac{T_{SL}(I_S+\sigma_S^2)}{(1-\alpha_p)G_{S,0}L_S(r_0)} \mid \alpha_p > \frac{T_{PL}(1+T_{SL})}{T_{SL}+T_{PL}(1+T_{SL})}\right\}. \quad (27)$$

Thus, the downlink SINR coverage probability $P^2_{c,PSL}(T_{PL},T_{SL},\alpha_p)$ can be expressed as

$$P^2_{c,S,PSL}(T_{PL},T_{SL},\alpha_p) = \begin{cases} \mathbb{E}_R\left[\mathbb{P}\left[H_S > \frac{T_{PL}(I_S+\sigma_S^2)}{(\alpha_p-(1-\alpha_p)T_{PL})G_{S,0}L_S(r_0)} \mid R=r\right]\right], & \alpha_p \leq \frac{T_{PL}(1+T_{SL})}{T_{SL}+T_{PL}(1+T_{SL})}, \\ \mathbb{E}_R\left[\mathbb{P}\left[H_S > \frac{T_{SL}(I_S+\sigma_S^2)}{(1-\alpha_p)G_{S,0}L_S(r_0)} \mid R=r\right]\right], & \alpha_p > \frac{T_{PL}(1+T_{SL})}{T_{SL}+T_{PL}(1+T_{SL})}. \end{cases} \quad (28)$$

Let

$$P^2_{c,S,SL}(T_{SL},\alpha_p) = \mathbb{E}_R\left[\mathbb{P}\left[H_S > \frac{T_{SL}(I_S+\sigma_S^2)}{(1-\alpha_p)G_{S,0}L_S(r_0)} \mid R=r\right]\right]. \quad (29)$$

Thus, combining (23) and (29), the SINR coverage probability $P^2_{c,PSL}(T_{PL},T_{SL},\alpha_p)$ can finally be written as in (19). According to [24] and [25] and after some manipulations, the SINR coverage probability $P^2_{c,S,SL}(T_{SL},\alpha_p)$ can be solved as in (20). The proof is completed. ■

### C. Average Number of Served Users

*Lemma 2:* For fixed data rates for the primary and secondary layers, $R_{PL}$ and $R_{SL}$, the average number of served users by the primary layer, can be expressed as

$$\mathbb{E}^o[N^2_{PL}] = \lambda_U P^2_{c,S,PL}(T_{PL},\alpha_p)\theta_S(2\pi\lambda_S)^{-1}, \quad (30)$$

while the average number of served users by both primary and secondary layers, can be expressed as

$$\mathbb{E}^o[N^2_{PSL}] = \lambda_U P^2_{c,S,PSL}(T_{PL},T_{SL},\alpha_p)\theta_S(2\pi\lambda_S)^{-1}, \quad (31)$$

where, $T_{PL} = 2^{R_{PL}} - 1$ and $T_{SL} = 2^{R_{SL}} - 1$ are the corresponding SINR thresholds for decoding the primary and secondary layers, and $\alpha_p$ is the power allocation factor.

*Proof:* For the mmWave multicast cluster, $B_o$, the average number of served users by the primary layer can be expressed as

$$\mathbb{E}^o[N_{PL}^2] \triangleq \mathbb{E}^o\left[\sum_{y\in\Phi_{U,B_o}} \mathbb{I}(E_{PL}^2(y))\right], \tag{32}$$

where $E_{PL}^2(y) = \{\text{SINR}_{\text{S,PL}}^2 \geq 2^{\text{R}_{\text{PL}}} - 1\}$. Then, the average number of served users, who can decode the data contained in both the primary and secondary layers, can be written as

$$\mathbb{E}^o[N_{PSL}^2] \triangleq \mathbb{E}^o\left[\sum_{y\in\Phi_{U,B_o}} \mathbb{I}(E_{PSL}^2(y))\right], \tag{33}$$

where $E_{PSL}^2(y) = \{\{\text{SINR}_{\text{S,PL}}^2 \geq 2^{\text{R}_{\text{PL}}} - 1\} \cap \{\text{SINR}_{\text{S,SL}}^2 \geq 2^{\text{R}_{\text{SL}}} - 1\}\}$.

Since $\mathbb{E}^o[N_{PL}^2]$ and $\mathbb{E}^o[N_{PSL}^2]$ have similar form as $\mathbb{E}^o[N^1]$, their derivations are similar to Lemma 1. Now, substituting $P_{c,S,PL}^2(T_{PL}, \alpha_p)$ for $P_{c,S}^1(T)$ in (9), the average number of served users by the primary layer can be expressed as in (30). After replacing $P_{c,S}^1(T)$ in (9) with $P_{c,S,PSL}^2(T_{PL}, T_{SL}, \alpha_p)$, the average number of served users by both the primary and secondary layers can be obtained as in (31). This completes the proof. ∎

### D. Sum Multicast Rate

The sum multicast rate for NOMA multicast is defined as the mean of the sum rate of all users in coverage of the multicast cluster, who successfully decode the primary layer with data rate, $R_{PL}$, or both the primary and secondary layers with data rate, $R_{PL} + R_{SL}$. This is given by

$$\begin{aligned}\bar{R}_{\text{sum}}^2 &= (\mathbb{E}^o[N_{PL}^2] - \mathbb{E}^o[N_{PSL}^2])R_{PL} + \mathbb{E}^o[N_{PSL}^2](R_{PL}+R_{SL}) \\ &= \mathbb{E}^o[N_{PL}^2]R_{PL} + \mathbb{E}^o[N_{PSL}^2]R_{SL}.\end{aligned} \tag{34}$$

Note that $\mathbb{E}^o[N_{PL}^2] - \mathbb{E}^o[N_{PSL}^2]$ is the average number of served users by the mmWave multicast cluster, who can only decode the primary layer.

*Theorem 2:* For fixed data rates for the primary and secondary layers, $R_{PL}$ and $R_{SL}$, the sum rate for the mmWave multicast cluster can be expressed as

$$\bar{R}_{\text{sum}}^2 = \begin{cases} \dfrac{(R_{PL}+R_{SL})\lambda_U P_{c,S,PL}^2(T_{PL},\alpha_p)\theta_S}{2\pi\lambda_S}, \\ \qquad \alpha_p \leq \dfrac{T_{PL}(1+T_{SL})}{T_{SL}+T_{PL}(1+T_{SL})} \text{ and } \text{T}_{\text{PL}} < \dfrac{\alpha_{\text{p}}}{1-\alpha_{\text{p}}}, \\ \dfrac{R_{PL}P_{c,S,PL}^2(T_{PL},\alpha_p)\lambda_U\theta_S}{2\pi\lambda_S} \\ +\dfrac{R_{SL}P_{c,S,SL}^2(T_{PL},T_{SL},\alpha_p)\lambda_U\theta_S}{2\pi\lambda_S}, \\ \qquad \alpha_p > \dfrac{T_{PL}(1+T_{SL})}{T_{SL}+T_{PL}(1+T_{SL})} \text{ and } \text{T}_{\text{PL}} < \dfrac{\alpha_{\text{p}}}{1-\alpha_{\text{p}}}, \\ 0, \qquad T_{PL} \geq \dfrac{\alpha_p}{1-\alpha_p}, \end{cases} \tag{35}$$

where, $T_{PL} = 2^{R_{PL}} - 1$ and $T_{SL} = 2^{R_{SL}} - 1$, are the SINR thresholds for the primary and secondary layers.

*Proof:* Combining (18), (20), (30), (31), and (34), the sum multicast rate can be obtained as in (35) and the proof is completed. ∎

Next, we discuss the maximization of the sum multicast rate for NOMA multicast.

*1) Optimizing Power Allocation for Fixed Data Transmission Rate for the Primary and Secondary Layers:* We first optimize the power allocation, in order to maximize the sum multicast rate, assuming fixed values for the primary and secondary layer data rates, $R_{PL}$ and $R_{SL}$. An appropriate power allocation factor, $\alpha_p$, should be chosen, as the increase of the power allocated to the primary layer can enable more users to decode the primary layer, but less users can decode the secondary layer. Therefore, the optimization problem can be formulated as

$$\begin{aligned} \max_{\alpha_p} \quad & \bar{R}_{\text{sum}}^2 \\ \text{s. t} \quad & 0 < \alpha_p < 1. \end{aligned} \tag{P3}$$

Similarly, **Algorithm 3** based on the GSS method is developed to solve the optimization problem P3.

**Algorithm 3** Power Allocation Optimization of NOMA Multicast

1: Initialization: $\alpha_{p,a} \leftarrow 0.01$, $\alpha_{p,b} \leftarrow 1$, $\epsilon \leftarrow 0.01$;
2: Construct objective function $f(\alpha_p)$ according to (18), (20) and (35);
3: Call GOLDENSECTIONSEARCHALG $f, \alpha_p, \alpha_{p,a}, \alpha_{p,b}, \epsilon$ shown in **Subalgorithm 1**.

*2) Optimizing the Secondary Layer's Data Transmission Rate for Fixed Power Allocation and the Primary layer's Data Transmission Rate:* In order to guarantee the basic QoS of the multicast service, we first fix the power allocation and data transmission rate for the primary layer, then choose an appropriate data transmission rate for the secondary layer to achieve the maximum sum multicast rate. Therefore, the optimization problem can be formulated as

$$\begin{aligned} \max_{R_{SL}} \quad & \bar{R}_{\text{sum}}^2 \\ \text{s. t} \quad & R_{SL} > 0. \end{aligned} \tag{P4}$$

Similarly, **Algorithm 4** based on the GSS method is developed to solve the optimization problem P4, and obtain the optimal data transmission rate for the secondary layer of NOMA multicast. As above, in order to limit the search space, an arbitrary large value is given, based on the simulation results.

**Algorithm 4** Data Transmission Rate Optimization for the Secondary Layer of NOMA Multicast

1: Initialization: $R_{SL,a} \leftarrow 0.1$, $R_{SL,b} \leftarrow \log_2(1+10^4)$, $\epsilon \leftarrow 0.1$;
2: Construct objective function $f(R_{SL})$ according to (18), (20) and (35);
3: Call GOLDENSECTIONSEARCHALG $f, R_{SL}, R_{SL,a}, R_{SL,b}, \epsilon$ shown in **Subalgorithm 1**.

## V. Performance Analysis of Cooperative NOMA Multicast in mmWave HetNets

In this section, we further study multicasting in a two-tier mmWave HetNet consisting of one low frequency MBS tier and one mmWave small cell tier, which is the typical deployment for mmWave wireless networks, and propose cooperative NOMA multicast scheme to further improve the NOMA multicast performance. The system modeling and performance analysis of cooperative NOMA multicast by using stochastic geometry, are provided. Analytical expressions for the SINR coverage probability, average number of served users and sum multicast rate are also derived.

### A. Received SINR

Cooperative NOMA multicast enables the MBS tier to cooperatively transmit the primary layer with low data rate. With MBS cooperative multicast, the users who are failed to decode the primary layer of the multicast data from the mmWave small cell tier, will try to decode it from the macro BS tier. If the primary layer is decoded, they cancel the primary layer from the superposed signal and further decode the secondary layer. Thus, this increase the success probability that decodes the primary and secondary layers, such that the NOMA multicast performance can be improved.

The signal received at the user with random distance, $d_0$, from its serving and interference MBSs can be expressed as

$$y_M^3 = h_{M,0}\sqrt{P_M}d_0^{-\alpha_M/2}x_{0,PL} + \underbrace{\sum_{\mathrm{X_i}\in\Phi_M\setminus \mathrm{B_0}} h_{M,i}\sqrt{P_M}d_i^{-\alpha_M/2}x_{i,PL}}_{I_M} + n_M, \quad (36)$$

where, $n_M$ is the thermal noise with power $\sigma_M^2$. Therefore, the SINR of decoding the data can be written as

$$\mathrm{SINR}_{\mathrm{PL}}^3 = \frac{\mathrm{H_{M,0}P_M d_0^{-\alpha_M}}}{\underbrace{\sum_{\mathrm{X_i}\in\Phi_M\setminus \mathrm{B_0}} H_{M,i}P_M d_i^{-\alpha_M}}_{\mathrm{I_M}} + \sigma_{\mathrm{M}}^2}. \quad (37)$$

The SINR of decoding the primary layer, $\mathrm{SINR}_{\mathrm{S,PL}}^3$, from the mmWave small cell tier, can be written as in (16), while the SINR of the secondary layer, $\mathrm{SINR}_{\mathrm{S,SL}}^3$, can be expressed as in (17).

### B. Downlink SINR Coverage Probability

*Proposition 3:* For fixed SINR thresholds for the primary and secondary layers, $T_{PL}$ and $T_{SL}$, the downlink SINR coverage probabilities of cooperative NOMA multicast can be expressed, respectively, as

$$P_{c,PL}^3(T_{PL}) = P_{c,M,PL}^3(T_{PL}) + P_{c,S,PL}^2(T_{PL},\alpha_p) - P_{c,M,PL}^3(T_{PL})P_{c,S,PL}^2(T_{PL},\alpha_p), \quad (38)$$

and

$$P_{c,PSL}^3(T_{PL},T_{SL},\alpha_p) = P_{c,S,PSL}^2(T_{PL},T_{SL},\alpha_p)(1-P_{c,M,PL}^3(T_{PL})) + P_{c,S,SL}^2(T_{SL},\alpha_p)P_{c,M,PL}^3(T_{PL}), \quad (39)$$

where

$$P_{c,M,PL}^3(T_{PL}) = \pi\lambda_M\int_0^\infty e^{-\pi\lambda_M(1+\rho(T_{PL},\alpha_M))x - T_{PL}(P_M/\sigma_M^2)^{-1}x^{\alpha_M/2}}dx, \quad (40)$$

and

$$\rho(T_{PL},\alpha_M) = T_{PL}^{2/\alpha_M}\int_{T_{PL}^{-2/\alpha_M}}^\infty (1+t^{\alpha_M/2})^{-1}dt. \quad (41)$$

*Proof:* See Appendix B. ∎

### C. Average Number of Served Users

*Lemma 3:* For fixed data rates for the primary and secondary layers, $R_{PL}$ and $R_{SL}$, the average number of served users by the primary layer can be expressed as

$$\mathbb{E}^o[N_{PL}^3] = \lambda_U P_{c,PL}^3(T_{PL})\theta_S(2\pi\lambda_S)^{-1}, \quad (42)$$

while the average number of served users by the secondary layer can be expressed as

$$\mathbb{E}^o[N_{PSL}^3] = \lambda_U P_{c,PSL}^3(T_{PL},T_{SL},\alpha_p)\theta_S(2\pi\lambda_S)^{-1}, \quad (43)$$

where $T_{PL} = 2^{R_{PL}} - 1$ and $T_{SL} = 2^{R_{SL}} - 1$.

*Proof:* For the mmWave multicast cluster, $B_o$, the average number of served users by the primary and secondary layers can be expressed, respectively, as

$$\mathbb{E}^o[N_{PL}^3] \triangleq \mathbb{E}^o\left[\sum_{y\in\Phi_{U,B_o}}\mathbb{I}(E_{PL}^3(y))\right], \quad (44)$$

and

$$\mathbb{E}^o[N_{PSL}^3] \triangleq \mathbb{E}^o\left[\sum_{y\in\Phi_{U,B_o}}\mathbb{I}(E_{PSL}^3(y))\right], \quad (45)$$

where $E_{PL}^3(y) = \{\mathrm{SINR}_{\mathrm{M,PL}}^3 \geq 2^{\mathrm{R_{PL}}}-1 \cup \mathrm{SINR}_{\mathrm{S,PL}}^3 \geq 2^{\mathrm{R_{PL}}}-1\}$, $E_{PSL}^3(y) = \{\{\mathrm{SINR}_{\mathrm{M,PL}}^3 \geq 2^{\mathrm{R_{PL}}}-1 \cup \mathrm{SINR}_{\mathrm{S,PL}}^3 \geq 2^{\mathrm{R_{PL}}}-1\} \cap \{\mathrm{SINR}_{\mathrm{S,SL}}^3 \geq 2^{\mathrm{R_{SL}}}-1\}\}$.

Since $\mathbb{E}^o[N_{PL}^3]$ and $\mathbb{E}^o[N_{PSL}^3]$ have similar form as $\mathbb{E}^o[N^1]$, their derivations are similar to Lemma 1. Therefore, substituting $P_{c,PL}^3(T_{PL},\alpha_p)$ for $P_{c,S}^1(T)$ in (9), the average number of served users by the primary layer can be expressed as in (42). Furthermore, replacing $P_{c,S}^1(T)$ in (9) with $P_{c,PSL}^3(T_{PL},T_{SL},\alpha_p)$, the average number of served users by both the primary and secondary layers can be obtained as in (43) and the proof is completed. ∎

### D. Sum Multicast Rate

*Theorem 3:* For fixed data rates for the primary and secondary layers, $R_{PL}$ and $R_{SL}$, the sum multicast rate for MBS cooperative NOMA multicast can be expressed as

$$\bar{R}_{\mathrm{sum}}^3 = (R_{PL}P_{c,PL}^3(T_{PL}) + R_{SL}P_{c,PSL}^3(T_{PL},T_{SL},\alpha_p))\lambda_U\theta_S(2\pi\lambda_S)^{-1}, \quad (46)$$

where $T_{PL} = 2^{R_{PL}} - 1$ and $T_{SL} = 2^{R_{SL}} - 1$ are the SINR thresholds for the primary and secondary layers.

TABLE II
SIMULATION PARAMETERS

| Parameter | Value |
|---|---|
| Macro BS carrier frequency | 2 GHz |
| Macro BS radius | 1000 m |
| mmWave small cell carrier frequency | 28 GHz |
| Small cell radius | 200 m |
| Path loss exponent for macro BSs | 3 |
| LOS Path loss exponent for mmWave small cells | 2 |
| NLOS Path loss exponent for mmWave small cells | 4 |
| User density | 2000/km$^2$ |

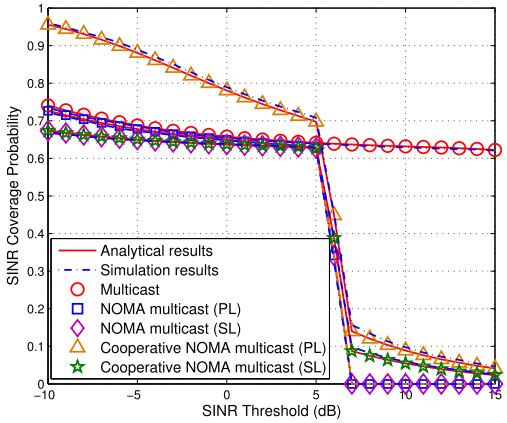



Fig. 3. SINR coverage probabilities of multicast, NOMA multicast, cooperative NOMA multicast.

*Proof:* The sum multicast rate for cooperative NOMA multicast is the mean of the sum rate for all users in coverage of the multicast cluster and is equal to

$$\begin{aligned}\bar{R}^3_{\text{sum}} &= R_{PL}(\mathbb{E}^o[N^3_{PL}]-\mathbb{E}^o[N^3_{PSL}])+(R_{PL}+R_{SL})\mathbb{E}^o[N^3_{PSL}] \\ &= R_{PL}\mathbb{E}^o[N^3_{PL}]+R_{SL}\mathbb{E}^o[N^3_{PSL}]. \end{aligned} \quad (47)$$

Combining (42), (43), and (47), the sum multicast rate can be obtained as in (46) and the proof is completed. ■

## VI. NUMERICAL RESULTS AND DISCUSSION

In this section, numerical results and Monte Carlo simulations for the SINR coverage probability, average number of served users, and sum multicast rate, are presented. From these figures, an exact match is evident between simulations and analytical results. The simulation parameters are illustrated in Table II.

Fig. 3 depicts the corresponding SINR coverage probabilities of multicast, NOMA multicast, and cooperative NOMA multicast with power allocation factor, $\alpha_p = 0.8$, according to Propositions 1, 2, and 3. It can be observed that NOMA multicast can provide similar coverage layer as the conventional one, and also achieve an extra coverage layer. However, when the SINR threshold $T$ is larger than the maximum SINR, $\frac{\alpha_p}{1-\alpha_p}$, that users can detect the primary layer, the SINR coverage probabilities of the primary and secondary layers are equal to zero. This is because if users are failed to decode the primary layer, they do not further decode the NOMA secondary layer through SIC. We also observe that cooperative NOMA multicast can further improve the SINR coverage probability of the primary and secondary layers. More specifically, compared with multicast, cooperative NOMA multicast can achieve superior coverage in the low SINR threshold region. Compared with NOMA multicast, cooperative NOMA multicast can achieve better primary coverage, especially in the low and high SINR threshold regions, while it provide better secondary coverage, especially in the high SINR threshold region. These are because cooperative NOMA multicast enables MBS tier to transmit a copy of the primary layer without power split as well, as a result, users can receive two copies of the primary layer, which increases the success probability of decoding the primary layer. This also increases the success probability of decoding the secondary layer, as users can try to further decode the secondary layer when they are failed to decode the primary layer from mmWave small cells but succeed to decode the primary layer from MBSs.

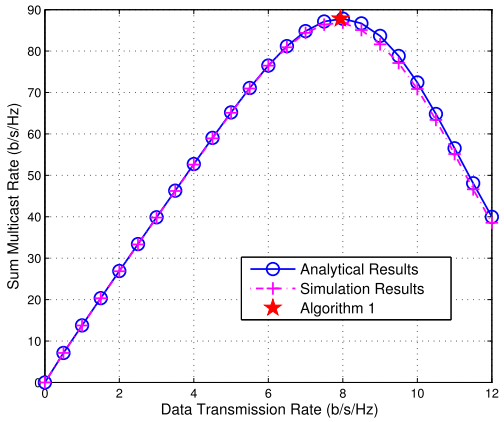



Fig. 4. Sum multicast rates for multicasting with different data transmission rates.

Fig. 4 shows the sum multicast rates for multicast mmWave cellular networks varied with different data transmission rates without QoS constraints, according to Theorem 1. With the increase of multicast data transmission rate, the sum multicast rate also increases up to the peak. Then, it decreases as the multicast data transmission rate continues to increase. That is, the sum multicast rate is a unimodal function on the data transmission rate. It is also demonstrated that **Algorithm 1** can search the optimal point, denoted as pentagram. Comparing the problems P1 and P2A, the only difference is data transmission rate region, thus the optimal point of the problem P2A under QoS constraints obtained by **Algorithm 2** still falls on the performance curve, shown in Fig. 4. Considering QoS constraints, data transmission rate for multicasting is generally limited in the low region, thus the $X$-coordinate of the optimal point with QoS constraints is lower than that of the optimal point without QoS constraints. Due to its monotonicity in the low data transmission rate region to the medium region, the $X$-coordinate of the optimal point for the problem P2A under QoS constraints is the upper limit of the constraint function.

Given {0.4, 1} b/s/Hz for the primary layer and {2, 4, 6} b/s/Hz for the secondary layer, the sum rate for NOMA multicast with different power allocations is plotted in Fig. 5. It is shown that for each pair of the primary and secondary layer rates, as the power allocation factor grows, the sum multicast rate first experiences a sharp rise to the peak, then it falls slowly in the medium power allocation factor region. Finally, it declines rapidly to the lowest point in the

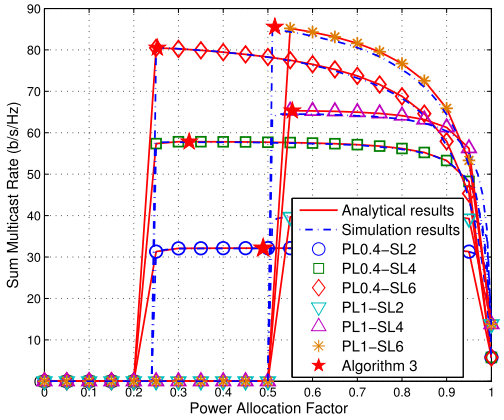



Fig. 5. Sum multicast rates for NOMA multicast with different power allocations.

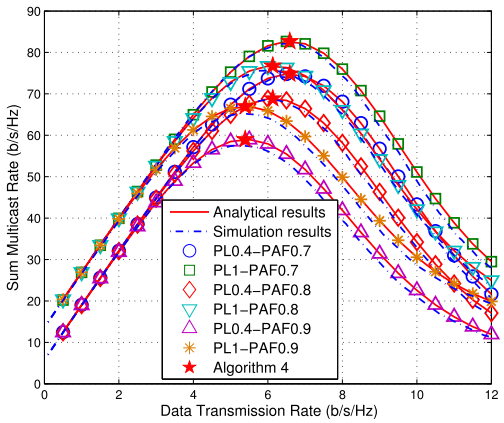
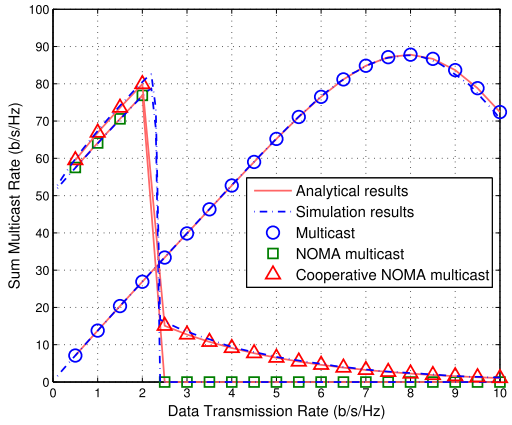
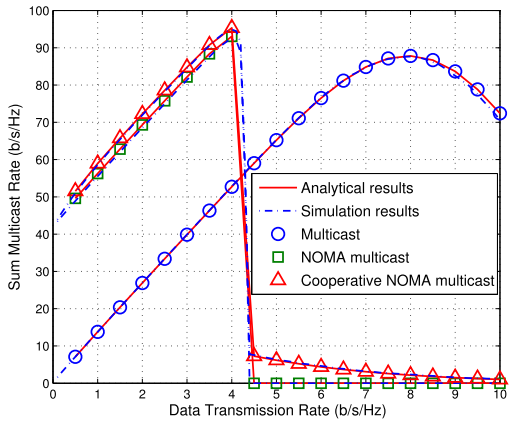



Fig. 6. Sum multicast rates for NOMA multicast with different data transmission rates for the secondary layer.




(a)




(b)

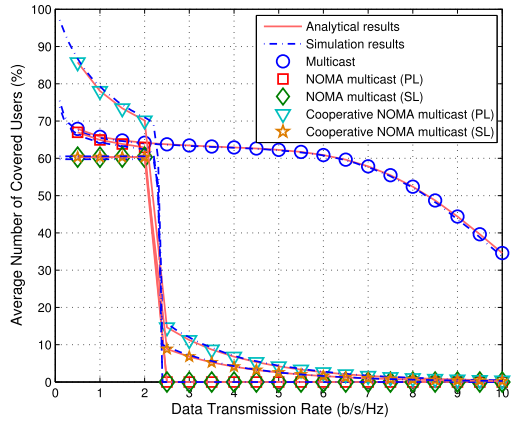
Fig. 7. Sum multicast rates for multicast, NOMA multicast, and cooperative NOMA multicast: a) $\alpha_p = 0.8$; b) $\alpha_p = 0.95$.

high power allocation factor region. Furthermore, with fixed secondary layer rate, the sum multicast rate for the primary layer rate 1 b/s/Hz is higher than that of the primary layer rate 0.4 b/s/Hz, which requires more power to be allocated to the primary layer. The sum multicast rate for the secondary layer rate 6 b/s/Hz is higher than that of the secondary layer rates 4 and 2 b/s/Hz, when the primary layer rate is fixed. The results also show that **Algorithm 3** can obtain the optimal points (pentagram).

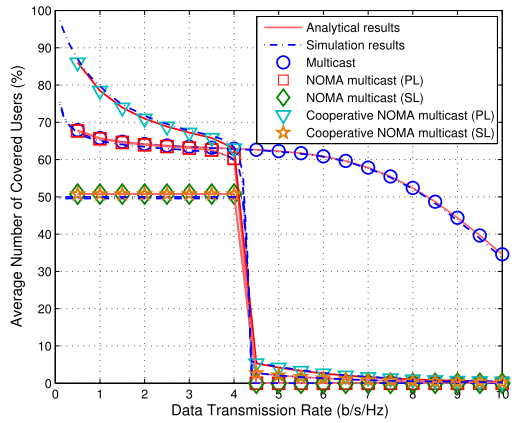
Given {0.4, 1} b/s/Hz for the primary layer and power allocation factors {0.7, 0.8, 0.9}, the sum rates for NOMA multicast with different data rates for the secondary layer are compared in Fig. 6. The results show that the sum multicast rate is a unimodal function on the data transmission rate for the secondary layer. Furthermore, more power is allocated to the secondary layer, higher data rate can be transmitted. Also, it is evident that **Algorithm 4** can obtain the optimal points (pentagram).

Fixing the secondary layer data rate, $R_{SL} = 4$ b/s/Hz, and the power allocation factors, $\alpha_p = \{0.8,\ 0.95\}$, the sum rates for multicast, NOMA multicast, and cooperative NOMA multicast are plotted in Fig. 7, according to Theorems 1, 2, and 3, respectively. The results show that NOMA multicast can achieve a significant gain, compared with conventional multicast in the low data transmission rate region. This is because NOMA multicast can fully utilize the channel conditions of strong users. However, the maximum SINR for detecting the primary layer is $\frac{\alpha_p}{1-\alpha_p}$, thus, NOMA multicast cannot work when data transmission rate exceeds $\log_2(1+\frac{\alpha_p}{1-\alpha_p})$. The results also show that cooperative NOMA multicast can achieve higher sum multicast rate than NOMA multicast, especially in the medium data transmission rate region. This is because in cooperative NOMA multicast scheme, the MBS transmits a replica of the primary layer as well, which overcomes the maximum SINR limit, $\frac{\alpha_p}{1-\alpha_p}$, of decoding the primary layer caused by NOMA transmission. As a result, this increases the success probability of decoding the primary and secondary layers. Comparing Fig. 7(a) and Fig. 7(b), more power is allocated to the primary layer, higher data transmission rate can be provided, yet lower sum multicast rate is achieved. This means that NOMA multicast gradually degrades to conventional multicast, with the increase of power allocated to the primary layer.

In Fig. 8, the percentages of served users in multicast, NOMA multicast, and cooperative NOMA multicast with fixed secondary layer rate $R_{SL} = 4$ b/s/Hz and power allocation factor $\alpha_p = \{0.8,\ 0.95\}$ are plotted, according to Lemmas 1, 2, and 3, respectively. The results show that in the low data transmission rate region, NOMA multicast offers similar coverage as conventional one, as well as additional secondary




(a)




(b)

Fig. 8. Percentage of served users in multicast, NOMA multicast, and cooperative NOMA multicast: a) $\alpha_p = 0.8$; b) $\alpha_p = 0.95$.

layer coverage with high data rate. Furthermore, cooperative NOMA multicast performs better primary and secondary layer coverage than NOMA multicast. Comparing Fig. 8(a) and Fig. 8(b), as more power is allocated to the primary layer, the coverage gain gradually declines. These results shown in Fig. 8 can used to explain the observed results in Fig. 7.

## VII. Conclusions

In this work, we studied multicasting in mmWave wireless networks and presented the system modeling and performance analysis by using stochastic geometry. Analytical expressions for the SINR coverage probability, average number of served users and sum multicast rate were also derived. We further studied the use of NOMA for mmWave multicasting to improve the multicast performance, which superposes a secondary layer with low transmit power to the original primary layer in power domain. Furthermore, we studied multicasting in a two-tier mmWave HetNet consisting of one low frequency MBS tier and one mmWave small cell tier, and proposed cooperative NOMA multicast scheme to further improve the NOMA multicast performance. This scheme enables the primary layer with low data rate to be transmitted cooperatively with the help of the MBS tier, such that more users can decode the primary and secondary layers through SIC. According to the numerical results verified by Monte Carlo simulations, we can conclude: (1) NOMA multicast can significantly increase the sum multicast rate, compared with multicast; (2) the proposed cooperative NOMA multicast can further improve the performance of NOMA multicast. In a future work, it would be interesting to further explore cooperative schemes for improving the performance of NOMA multicast in mmWave HetNets.

## Appendix A
## Proof of Lemma 1

Since users follow a homogeneous PPP with mean, $\lambda_U$, the mean of the mmWave multicast cluster with radius, $R$, and beamwidth, $\theta_S$, is $\lambda_U R^2\theta_S/2$. Therefore, for fixed multicast data rate, $R_T$, the average number of served users by the multicast area, shown in (8), can be rewritten as

$$\begin{aligned}
&\mathbb{E}^o[N^1] \\
&= \mathbb{E}_{\mathbb{R}}\left[\sum_{m=0}^{\infty}\frac{(\lambda_U R^2\theta/2)^m}{m!}e^{-\lambda_U R^2\theta_S/2}\right. \\
&\quad\left.\times\left(\sum_{i=1}^{m}\mathbb{P}(\log(1+\mathrm{SINR_{y_i}})\geq \mathrm{R_T}\mid \mathrm{R}=\mathrm{r})\right)\right] \\
&= \mathbb{E}_{\mathbb{R}}\left[\sum_{m=0}^{\infty}\frac{(\lambda_U R^2\theta_S/2)^m}{m!}e^{-\lambda_U R^2\theta_S/2}\right. \\
&\quad\left.\times\left(\sum_{i=1}^{m}\mathbb{P}(\mathrm{SINR_{y_i}}\geq 2^{\mathrm{R_T}}-1\mid \mathrm{R}=\mathrm{r})\right)\right] \\
&\overset{(a)}{=} \sum_{m=0}^{\infty}\frac{(\lambda_U R^2\theta_S/2)^m}{m!}e^{-\lambda_U R^2\theta_S/2}m P^1_{c,S}(T) \\
&= (\lambda_U R^2\theta_S/2)e^{-\lambda_U R^2\theta_S/2}P^1_{c,S}(T)\sum_{m=0}^{\infty}\frac{(\lambda_U R^2\theta_S/2)^m}{m!}, \qquad (48)
\end{aligned}$$

where (a) follows the change, $T = 2^{R_T} - 1$. Using $R = (\pi\lambda_S)^{-1/2}$ and applying $e^{\lambda_U R^2\theta_S/2} = \sum_{m=0}^{\infty}\frac{(\lambda_U R^2\theta_S/2)^m}{m!}$, the average number of served users can be obtained as in (9) and the proof is completed.

## Appendix B
## Proof of Theorem 3

With cooperative NOMA multicast in a two-tier mmWave HetNet, the SINR coverage probability of the primary layer can be expressed as

$$\begin{aligned}
P^3_{c,PL}(T_{PL},\alpha_p) &= \mathbb{P}\{\{\mathrm{SINR^3_{M,PL}}>\mathrm{T_{PL}}\}\cup\{\mathrm{SINR^3_{S,PL}}>\mathrm{T_{PL}}\}\} \\
&= \mathbb{P}\{\mathrm{SINR^3_{M,PL}}>\mathrm{T_{PL}}\}+\mathbb{P}\{\mathrm{SINR^3_{S,PL}}>\mathrm{T_{PL}}\} \\
&\quad-\mathbb{P}\{\mathrm{SINR^3_{M,PL}}>\mathrm{T_{PL}}\}\mathbb{P}\{\mathrm{SINR^3_{S,PL}}>\mathrm{T_{PL}}\}, \qquad (49)
\end{aligned}$$

where,

$$\begin{aligned}
&P^3_{c,M,PL}(T_{PL}) \\
&= \mathbb{P}\{\mathrm{SINR^3_{M,PL}}>\mathrm{T_{PL}}\} \\
&= \int_{r>0}\mathbb{P}\left[\frac{H_M P_M R^{-\alpha_M}}{I_M+\sigma^2_M}>T_{PL}\mid R=r\right]f_R(r)dr. \qquad (50)
\end{aligned}$$

According to [45], the SINR coverage probability can be obtained as in (40). Since $\mathrm{SINR}^3_{\mathrm{S,PL}} = \mathrm{SINR}^2_{\mathrm{S,PL}}$, the term, $\mathbb{P}\{\mathrm{SINR}^3_{\mathrm{S,PL}} > \mathrm{T_{PL}}\}$, can be obtained as in (18). Therefore, $P^3_{c,PL}(T_{PL}, \alpha_p)$ can be solved as in (38).

The SINR coverage probability of the secondary layer can be expressed as

$$\begin{aligned} &P^3_{c,PSL}(T_{PL}, T_{SL}, \alpha_p) \\ &\quad = \mathbb{P}\{\{\mathrm{SINR}^3_{\mathrm{S,SL}} > \mathrm{T_{SL}} \cap \mathrm{SINR}^3_{\mathrm{S,PL}} > \mathrm{T_{PL}}\} \\ &\quad \cup \{\mathrm{SINR}^3_{\mathrm{S,SL}} > \mathrm{T_{SL}} \cap \mathrm{SINR}^3_{\mathrm{M,PL}} > \mathrm{T_{PL}}\}\}. \end{aligned} \tag{51}$$

After some manipulations,

$$\begin{aligned} &P^3_{c,PSL}(T_{PL}, T_{SL}, \alpha_p) \\ &\quad = \mathbb{P}\{\mathrm{SINR}^3_{\mathrm{S,SL}} > \mathrm{T_{SL}} \cap \mathrm{SINR}^3_{\mathrm{S,PL}} > \mathrm{T_{PL}}\} \\ &\qquad + \mathbb{P}\{\mathrm{SINR}^3_{\mathrm{S,SL}} > \mathrm{T_{SL}}\}\mathbb{P}\{\mathrm{SINR}^3_{\mathrm{M,PL}} > \mathrm{T_{PL}}\} \\ &\qquad - \mathbb{P}\{\mathrm{SINR}^3_{\mathrm{S,SL}} > \mathrm{T_{SL}} \cap \mathrm{SINR}^3_{\mathrm{S,PL}} > \mathrm{T_{PL}}\} \\ &\qquad \times \mathbb{P}\{\mathrm{SINR}^3_{\mathrm{M,PL}} > \mathrm{T_{PL}}\}. \end{aligned} \tag{52}$$

Since $\mathrm{SINR}^3_{\mathrm{S,SL}} = \mathrm{SINR}^2_{\mathrm{S,SL}}$ and $\mathrm{SINR}^3_{\mathrm{S,PL}} = \mathrm{SINR}^2_{\mathrm{S,PL}}$, $\mathbb{P}\{\mathrm{SINR}^3_{\mathrm{S,SL}} > \mathrm{T_{SL}} \cap \mathrm{SINR}^3_{\mathrm{S,PL}} > \mathrm{T_{PL}}\}$ can be written as in (19), and $\mathbb{P}\{\mathrm{SINR}^3_{\mathrm{S,SL}} > \mathrm{T_{SL}}\}$ can be expressed as in (20). Therefore, combining (19), (20), (40), and (52), $P^3_{c,PSL}(T_{PL}, T_{SL}, \alpha_p)$ can be finally obtained as in (39). This completes the proof.

## REFERENCES


[1] M. Gruber and D. Zeller, "Multimedia broadcast multicast service: New transmission schemes and related challenges," *IEEE Commun. Mag.*, vol. 49, no. 12, pp. 176–181, Dec. 2011.

[2] *Evolved Universal Terrestrial Radio Access (E-UTRA) Evolved Universal Terrestrial Radio Access Netw. (E-UTRAN); Overall Description; Stage 2*, document TS 36.300, 3GPP, Mar. 2017.

[3] *Multimedia Broadcast/Multicast Service (MBMS); Architecture and Functional Description*, document TS 23.246, 3GPP, Dec. 2016.

[4] *Study on Single-Cell Point-to-Multipoint Transmission for E-UTRA*, document TR 36.890, 3GPP, Jun. 2015.

[5] (2015). NGMN. *5G White Paper*. [Online]. Available: http://www.ngmn.org/5g-white-paper.html

[6] K. Poularakis, G. Iosifidis, V. Sourlas, and L. Tassiulas, "Exploiting caching and multicast for 5G wireless networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2995–3007, Apr. 2016.

[7] G. Araniti, M. Condoluci, P. Scopelliti, A. Molinaro, and A. Iera, "Multicasting over emerging 5G networks: Challenges and perspectives," *IEEE Netw.*, vol. 31, no. 2, pp. 80–89, Mar./Apr. 2017.

[8] A. de la Fuente, R. P. Leal, and A. G. Armada, "New technologies and trends for next generation mobile broadcasting services," *IEEE Commun. Mag.*, vol. 54, no. 11, pp. 217–223, Nov. 2016.

[9] M. Condoluci, G. Araniti, T. Mahmoodi, and M. Dohler, "Enabling the IoT machine age with 5G: Machine-type multicast services for innovative real-time applications," *IEEE Access*, vol. 4, pp. 5555–5569, Sep. 2016.

[10] S. R. Mirghaderi, A. Bayesteh, and A. K. Khandani, "On the multicast capacity of the wireless broadcast channel," *IEEE Trans. Inf. Theory*, vol. 58, no. 5, pp. 2766–2780, May 2012.

[11] Y. Li, Q. Peng, and X. Wang, "Multicast capacity with max-min fairness for heterogeneous networks," *IEEE/ACM Trans. Netw.*, vol. 22, no. 2, pp. 622–635, Apr. 2014.

[12] X. Lin, R. Ratasuk, A. Ghosh, and J. G. Andrews, "Modeling, analysis, and optimization of multicast device-to-device transmissions," *IEEE Trans. Wireless Commun.*, vol. 13, no. 8, pp. 4346–4359, Apr. 2014.

[13] *Mission Critical Push To Talk (MCPTT) Over LTE*, document TS 22.179, 3GPP, Dec. 2016.

[14] *Group Communication System Enablers for LTE; (GCSE_LTE)*, document TS 22.468, 3GPP, Mar. 2017.

[15] J. Kim, S.-W. Choi, W.-Y. Shin, Y.-S. Song, and Y.-K. Kim, "Group communication over LTE: A radio access perspective," *IEEE Commun. Mag.*, vol. 54, no. 4, pp. 16–23, Apr. 2016.

[16] (2014). Expway. *14 LTE Broadcast Business Cases*. [Online]. Available: http://www.expway.com/page-white-paper/

[17] S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter-wave cellular wireless networks: Potentials and challenges," *Proc. IEEE*, vol. 102, no. 3, pp. 366–385, Mar. 2014.

[18] T. S. Rappaport *et al.*, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, May 2013.

[19] W. Roh *et al.*, "Millimeter-wave beamforming as an enabling technology for 5G cellular communications: Theoretical feasibility and prototype results," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 106–113, Feb. 2014.

[20] J. Choi, V. Va, N. G.-Prelcic, R. Daniels, C. R. Bhat, and R. W. Heath, "Millimeter-wave vehicular communication to support massive automotive sensing," *IEEE Commun. Mag.*, vol. 54, no. 12, pp. 160–167, Dec. 2016.

[21] L. Kong, M. K. Khan, F. Wu, G. Chen, and P. Zeng, "Millimeter-wave wireless communications for IoT-cloud supported autonomous vehicles: Overview, design, and challenges," *IEEE Commun. Mag.*, vol. 55, no. 1, pp. 62–68, Jan. 2017.

[22] T. S. Rappaport, G. R. Maccartney, M. K. Samimi, and S. Sun, "Wideband millimeter-wave propagation measurements and channel models for future wireless communication system design," *IEEE Trans. Commun.*, vol. 63, no. 9, pp. 3029–3056, Sep. 2015.

[23] H. Shokri-Ghadikolaei, C. Fischione, G. Fodor, P. Popovski, and M. Zorzi, "Millimeter wave cellular networks: A MAC layer perspective," *IEEE Trans. Commun.*, vol. 63, no. 10, pp. 3437–3458, Oct. 2015.

[24] T. Bai and R. W. Heath, Jr., "Coverage and rate analysis for millimeter-wave cellular networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 2, pp. 1100–1114, Feb. 2015.

[25] J. G. Andrews, T. Bai, M. Kulkarni, A. Alkhateeb, A. Gupta, and R. W. Heath, Jr., "Modeling and analyzing millimeter wave cellular systems," *IEEE Trans. Commun.*, vol. 65, no. 1, pp. 403–430, Jan. 2017.

[26] Z. Ding, P. Fan, and H. V. Poor, "Random beamforming in millimeter wave NOMA networks," *IEEE Access*, vol. 5, pp. 7667–7681, Jun. 2017.

[27] K. Sakaguchi *et al.* (Mar. 2017). *Use Cases and Scenario Definition, 5G-MiEdge Deliverable D1.1*. [Online]. Available: http://5g-miedge.eu/wp-content/uploads/2017/05/5G-MiEdge_D1.1_v6.3_final.pdf

[28] H. Park, S. Park, T. Song, and S. Pack, "An incremental multicast grouping scheme for mmWave networks with directional antennas," *IEEE Commun. Lett.*, vol. 17, no. 3, pp. 616–619, Mar. 2013.

[29] W. Feng, B. Gao, D. Jin, and L. Zeng, "Transmit and receive beamforming for 60-GHz physical-layer multicasting," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Sep. 2014, pp. 329–334.

[30] H. Chu, P. Xu, W. Wang, C. Yang, and W. Zhu, "BF-assisted joint relay selection and power control for cooperative multicast in mmWave networks," in *Proc. IEEE 26th Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, Apr. 2015, pp. 2255–2259.

[31] (2017). A. Biason and M. Zorzi. *Multicast Transmissions in Directional mmWave Communications*. [Online]. Available: https://arxiv.org/abs/1703.01190.

[32] L. Dai, B. Wang, Y. Yuan, S. Han, C.-L. I, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[33] *Study on Downlink Multiuser Superposition Transmission (MUST) for LTE*, document TR 36.859, 3GPP, Dec. 2015.

[34] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014.

[35] Z. Yang, Z. Ding, P. Fan, and G. K. Karagiannidis, "On the performance of non-orthogonal multiple access systems with partial channel information," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 654–667, Feb. 2016.

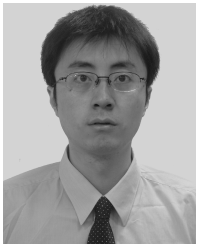
[36] P. Xu, Y. Yuan, Z. Ding, X. Dai, and R. Schober, "On the outage performance of non-orthogonal multiple access with 1-bit feedback," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 6716–6730, Oct. 2016.

[37] Z. Ding, M. Peng, and H. V. Poor, "Cooperative non-orthogonal multiple access in 5G systems," *IEEE Commun. Lett.*, vol. 19, no. 8, pp. 1462–1465, Aug. 2015.

[38] W. Han, Y. Zhang, X. Wang, M. Sheng, J. Li, and X. Ma, "Orthogonal power division multiple access: A green communication perspective," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3828–3842, Dec. 2016.

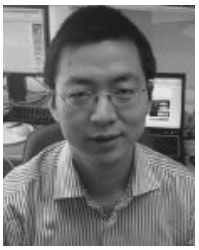
[39] *On PMCH Enhancement by MUST Technologies*, document 3GPP TSG RAN WG1 Meeting #81 R1-153235, Huawei and HiSilicon, Fukuoka, Japan, May 2015.
[40] J. Choi, "Minimum power multicast beamforming with superposition coding for multiresolution broadcast and application to NOMA systems," *IEEE Trans. Commun.*, vol. 63, no. 3, pp. 791–800, Mar. 2015.
[41] Z. Ding, Z. Zhao, M. Peng, and H. V. Poor, "On the spectral efficiency and security enhancements of NOMA assisted multicast-unicast streaming," *IEEE Trans. Commun.*, to be published, doi: 10.1109/TCOMM.2017.2696527.
[42] L. Lv, J. Chen, Q. Ni, and Z. Ding, "Design of cooperative non-orthogonal multicast cognitive multiple access for 5G systems: User scheduling and performance analysis," *IEEE Trans. Commun.*, to be published, doi: 10.1109/TCOMM.2017.2677942.
[43] A. M. C. Correia, J. C. M. Silva, N. M. B. Souto, L. A. C. Silva, A. B. Boal, and A. B. Soares, "Multi-resolution broadcast/multicast systems for MBMS," *IEEE Trans. Broadcast.*, vol. 53, no. 1, pp. 224–234, Mar. 2007.
[44] H. S. Dhillon, R. K. Ganti, F. Baccelli, and J. G. Andrews, "Modeling and analysis of K-tier downlink heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 550–560, Apr. 2012.
[45] J. G. Andrews, A. K. Gupta, and H. S. Dhillon. (Apr. 2016). "A primer on cellular network analysis using stochastic geometry." [Online]. Available: http://arxiv.org/abs/1604.03183.
[46] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 7th ed. New York, NY, USA: Academic, 2007.
[47] E. K. P. Chong and S. H. Zak, *An Introduction to Optimization*, 4th ed. Hoboken, NY, USA: Wiley, 2013.

**Zhengquan Zhang** (S'16) received the M.Sc. degree in communication and information system from Southwest Jiaotong University, Chengdu, China, in 2008. He is currently pursuing the Ph.D. degree with Southwest Jiaotong University. From 2008 to 2013, he was with ZTE Corporation as a Communication Protocol Software Engineer, where he involved in the development of 3G CDMA2000 1xEV-DO and 4G LTE. Since 2016, he has been a Guest Ph.D. Student with the KTH Royal Institute of Technology. His current research interests include multiple access, millimeter-wave communications, cooperative communications, and full-duplex communications.



**Zheng Ma** (M'07) received the B.Sc. and Ph.D. degrees in communications and information system from Southwest Jiaotong University in 2000 and 2006, respectively. He was a Visiting Scholar with the University of Leeds, U.K., in 2003. In 2003 and 2005, he was a Visiting Scholar with The Hong Kong University of Science and Technology. From 2008 to 2009, he was a Visiting Research Fellow with the Department of Communication Systems, Lancaster University, U.K. He is currently a Professor with Southwest Jiaotong University, and servers as a Deputy Dean of the School of Information Science and Technology. He has authored over 40 research papers in high quality journals and conferences. His research interests include information theory and coding, signal design and applications, FPGA/DSP implementation, and professional mobile radio. He is currently the Editor of the IEEE COMMUNICATIONS LETTERS. He is also the Vice Chairman of Information Theory Chapter in the IEEE Chengdu Section.

**Yue Xiao** received the B.Sc. degree in communication engineering from Southwest Jiaotong University, Chengdu, China, in 2014, where she is currently pursuing the Ph.D. degree. Her current research interests include multiple access, channel coding, and cooperative communications.



**Ming Xiao** (S'02–M'07–SM'12) received the bachelor's and master's degrees in engineering from the University of Electronic Science and Technology of China, Chengdu, in 1997 and 2002, respectively, the Ph.D. degree from the Chalmers University of Technology, Sweden, in 2007. From 1997 to 1999, he was a Network and Software Engineer with China Telecom. From 2000 to 2002, he also held a position in the Sichuan Communications Administration. Since 2007, he has been in communication theory, School of Electrical Engineering, Royal Institute of Technology, Sweden, where he is currently an Associate Professor in communications theory. He received the best paper Awards from the International Conference on Wireless Communications and Signal Processing in 2010 and the IEEE International Conference on Computer Communication Networks in 2011. He received Chinese Government Award for Outstanding Self-Financed Students Studying Abroad in 2007, Hans Werthen Grant from the Royal Swedish Academy of Engineering Science (IVA) in 2006, and the Ericsson Research Funding from Ericsson in 2010. Since 2012, he has been an Associate Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE COMMUNICATIONS LETTERS, and the IEEE WIRELESS COMMUNICATIONS LETTERS, and has been a Senior Editor of the IEEE COMMUNICATIONS LETTERS since 2015.

**George K. Karagiannidis** (M'96–SM'03–F'14) was born in Pithagorion, Greece. He received the University Diploma and Ph.D. degrees in electrical and computer engineering from the University of Patras, in 1987 and 1999, respectively. From 2000 to 2004, he was a Senior Researcher with the Institute for Space Applications and Remote Sensing, National Observatory of Athens, Greece. In 2004, he joined the faculty of the Aristotle University of Thessaloniki, Greece, where he is currently a Professor with the Electrical and Computer Engineering Department and the Director of the Digital Telecommunications Systems and Networks Laboratory. He is also an Honorary Professor with Southwest Jiaotong University, Chengdu, China. His research interests are in the broad area of digital communications systems and signal processing, with emphasis on wireless communications, optical wireless communications, wireless power transfer and applications, molecular and nanoscale communications, and stochastic processes in biology and wireless security. He is the Author or Co-Author of over 400 technical papers published in scientific journals and presented at international conferences. He is also the Author of the Greek Edition of a book on *Telecommunications Systems* and the Co-Author of the book *Advanced Optical Wireless Communications Systems* (Cambridge Publications, 2012). He has been a General Chair, a Technical Program Chair, and a member of the Technical Program Committees in several IEEE and non-IEEE conferences. He was an Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS, a Senior Editor of the IEEE COMMUNICATIONS LETTERS, an Editor of the *EURASIP Journal of Wireless Communications and Networks*, and several times the Guest Editor of the IEEE SELECTED AREAS IN COMMUNICATIONS. From 2012 to 2015, he was the Editor-in Chief of the IEEE COMMUNICATIONS LETTERS. He is one of the highly cited authors across all areas of Electrical Engineering, recognized as a 2015 and 2016 Thomson Reuters Highly Cited Researcher.

**Pingzhi Fan** (M'93–SM'99–F'15) received the Ph.D. degree in electronic engineering from Hull University, U.K. He is currently a Professor and the Director of the Institute of Mobile Communications, Southwest Jiaotong University, China. He is an IEEE VTS Distinguished Lecturer from 2015 to 2017. He has over 200 research papers published in the IEEE and other academic journals, and eight books (including edited) published by John Wiley and Sons Ltd, Springer, and the IEEE Press. His current research interests include high mobility wireless communications, 5G technologies, and wireless networks for big data. He is a fellow of the IET, the CIE, and the CIC. He was a recipient of the U.K. ORS Award, the NSFC Outstanding Young Scientist Award, and a Chief Scientist of the National 973 Major Project. He served as a General Chair or a TPC Chair of a number of international conferences, as well as a Guest Editor-in-Chief, a Guest Editor, or an Editorial Member of several international journals. He is the Founding Chair of the IEEE VTS BJ Chapter, the IEEE ComSoc CD Chapter, and the IEEE Chengdu Section. He also served as a Board Member of the IEEE Region 10, the IET (IEE) Council, and the IET Asia-Pacific Region.