Title

Priority-Aware Resource Allocation for 5G mmWave Multicast Broadcast Services

Permalink
https://escholarship.org/uc/item/3fk6n9bs

Journal
IEEE Transactions on Broadcasting, 69(1)

ISSN
0018-9316

Authors
Su, Pan-Yang
Lin, Kuang-Hsun
Li, Yi-Yun
et al.

Publication Date
2023-03-01

DOI
10.1109/tbc.2022.3221696

Priority-Aware Resource Allocation for 5G mmWave Multicast Broadcast Services

Pan-Yang Su*, Kuang-Hsun Lin†, Yi-Yun Li†, Hung-Yu Wei†
*Electrical Engineering and Computer Sciences, University of California, Berkeley
†Graduate Institute of Communication Engineering, National Taiwan University, Taiwan

***Abstract*—5G Multicast Broadcast Services (MBS) are viewed as a promising 5G New Radio (NR) application, as standardization begins in 3GPP Release 17. With MBS, one next generation Node B (gNB) delivers data to multiple user equipments (UE) simultaneously, thus improving spectrum efficiency. Millimeter wave (mmWave) beamforming further enhances system performance by focusing signals in a dedicated direction. However, despite the advantages, we identify three issues of multicast with beamforming techniques. First, link directionality causes the gNB to transmit data over the beams sequentially, resulting in a combinatorial resource allocation problem. Confronting this beam scheduling issue, we develop an optimization algorithm that obtains an optimal solution in polynomial time. Second, UEs may falsely report their valuations over beam resources to gain more utility. Under this scenario, the gNB cannot allocate the resources to those in need due to the lack of accurate UE information. Therefore, we propose a Vickrey–Clarke–Groves (VCG) auction-based mechanism to incentivize the UEs to reveal their valuations over resources truthfully. This mechanism guarantees solution efficiency and maximizes social welfare. Third, as 3GPP standards allow for different priorities for different multicast flows, and video content providers distinguish between ordinary and premium UEs, we take UE priority into account. In this regard, we extend the valuation-based mechanism to a multi-priority one. Finally, the mathematical analysis validates some desirable properties of the proposed scheme, such as incentive-compatibility. Simulation results also justify our superior performance in the 5G MBS system compared with other resource allocation schemes.**

***Index Terms*—Multicast Broadcast Services (MBS), millimeter wave (mmWave), game theory, Vickrey–Clarke–Groves (VCG) auction, resource allocation, incentive mechanism.**

I. Introduction

Millimeter wave (mmWave) communication is a promising technology in alleviating the scarcity of spectrum resources as global data traffic increases exponentially. Ranging from tens to hundreds of GHz, mmWave bands are not utilized by conventional cellular systems since the transmissions over high frequencies suffer from high path loss and are sensitive to blockages. However, with the developed beamforming technology, base stations (BS) can provide directional beams by tuning the transmitted signal of massive antenna arrays to compensate for the path loss. In addition, such oriented beams cause less interference. Therefore, the mmWave communication becomes a cornerstone of 5G systems with beamforming compensating its weak propagation characteristics, and it has found widespread applications in the wireless networks [1]–[4].

On the other hand, Multicast and Broadcast Services (MBS) also bring significant benefits to spectrum utilization. In increasingly popular applications, such as multimedia entertainment, HD live streaming, virtual reality gaming, etc., common data are requested by multiple user equipments (UE) simultaneously. Therefore, multicast is critical to the improvement of spectral efficiency by transmitting the same contents to multiple UEs in a single transmission. As mobile video data account for considerable spectrum usage [5], the utilization of abundant mmWave spectrum that helps multicast transmission appears necessary. With beamforming and substantial available bandwidth, mmWave gNB can further enhance the transmission rate of MBS communications.

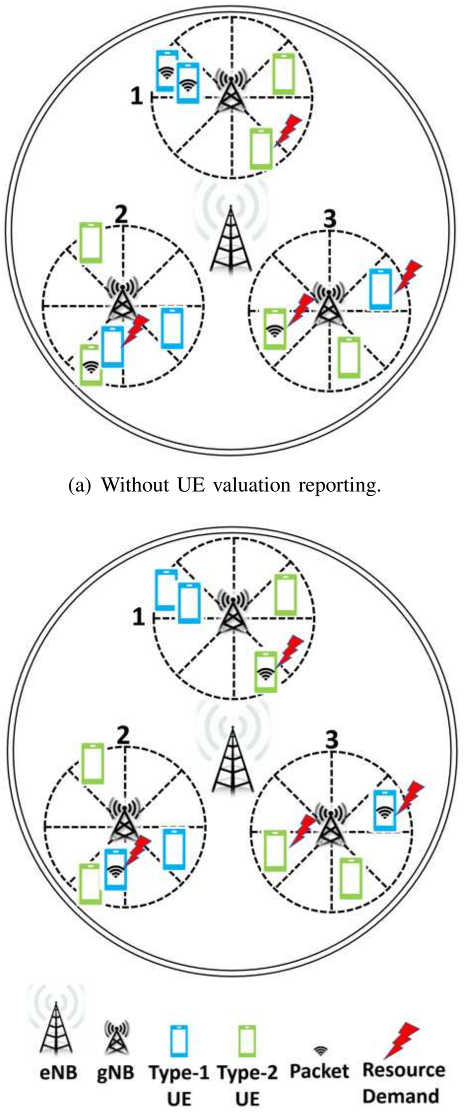
In the early version of LTE, 3GPP standardized the Multimedia Broadcast Multicast Service (MBMS) to enable the evolved Node B (eNB) to distribute multimedia contents via broadcasting or multicasting. The UEs with multimedia services, such as mobile TV and live video streaming, often request the same data from the content providers. Therefore, the MBMS can extensively enhance the system capacity. Each UE within the coverage can receive the requested data via a broadcast channel at once. Although 3GPP kept improving the MBMS techniques in the following releases of LTE, it was not widely deployed by the network service providers. To enable the next generation Node B (gNB) to support more types of services and satisfy the corresponding requirements for the Quality of Service (QoS), 3GPP decided to standardize MBS for the newly developed 5G NR systems, specifically, in the Release 17 of the 3GPP standard [6], [7]. In the newly added protocol layer, Service Data Adaptation Protocol (SDAP), the traffic data for different MBS services would be mapped to different MBS bearers so that the lower layer protocols could handle the priority better. Several topics, including more flexible network structure, deployment, resource usage, and field trial, are further studied to meet the requirements for multicast and broadcast applications in the future [8], [9].

In this paper, we examine three issues of multicast services, such as video streaming services or smart factories' AR auxiliary systems [10], [11], under mmWave communication with beamforming techniques. First, the combinatorial nature of beam scheduling makes it difficult to develop an efficient optimization algorithm [12]. In particular, multicasting with beamforming transmits signals in a dedicated direction in every timeslot. Hence, the serving sequence of different UE groups in a period becomes an essential issue. To cope with this beam scheduling problem, we devise an optimization

Corresponding Author: Hung-Yu Wei, email: hywei@ntu.edu.tw algorithm that obtains an optimal solution in polynomial time. Second, an incentive mechanism is required to aid the gNB in allocating the resources suitably. Without an incentive mechanism, UEs may falsely report their valuation functions, and the gNB does not know the accurate system information. The lack of correct information may lead to some UEs obtaining a surplus of resources while those in need are poorly served, as shown in Fig. 1(a). In this regard, we propose a mechanism based on a modified Vickrey–Clarke–Groves (VCG) auction to motivate UEs to report their valuation functions truthfully. Therefore, with the truthful pricing rule of the VCG mechanism, UEs' private valuations over resources can be obtained correctly. This incentive mechanism guarantees the efficiency of resource allocation outcomes, as shown in Fig. 1(b). Also, the mechanism maximizes social welfare, a commonly used notion to evaluate system performance by aggregating all the UEs' valuations. Third, UEs may have different priorities. In 3GPP standards, priority is considered in NR MBS with multiple flows having different QoS requirements. Also, in video streaming service, priority is a common consideration, such as the premium membership of Youtube or Netflix. Thus, UE priority is an essential issue in 5G MBS, but it cannot be characterized by valuation functions. In this regard, we further enhance the optimization algorithm and incentive mechanism to consider UE priority.

The main contributions of this paper are summarized as follows.

1) To the best of our knowledge, we are the first to propose a resource allocation mechanism in a 5G mmWave MBS system that considers the UEs' valuations over beam resources.
2) We propose an incentive-compatible mechanism that reaches a social-welfare-maximization solution in which we optimize the efficiency of resource allocation according to the UEs' valuations. Moreover, the proposed mechanism is individually rational and (weakly) budget-balanced.
3) We take into account the premium membership in 5G MBS as many multimedia content providers offer[1]. Our scheme guarantees utility superiority of high-priority UEs and also suggests a reasonable membership charge.



(a) Without UE valuation reporting.

(b) With UE valuation reporting.

Fig. 1. An example for the proposed system model. When two types of MBS services are in the system, a gNB could serve at most one MBS service for the UEs on one beam within one timeslot. The gNB should decide its target MBS group to serve in every timeslot. (a) Without UE valuation reporting, the gNB cannot cater to the UEs in need of resources. (b) With UE valuation reporting, the gNB can better serve the UEs in need of resources.

II. Related Work

Early research on directional multicast scheduling problem includes [14]–[16]. However, mmWave was not discussed, and high-frequency characteristic was not considered then. As mmWave systems become a key enabling technology for 5G cellular networks, more complex beamforming issues under mmWave systems are investigated. In [17], [18], the authors argued that the multi-lobes beam model proposed in [15], [16] required many radio frequency (RF) chains and was thus expensive and not compatible with chipsets with a single RF chain. In contrast, the single-lobe model was compatible with both analog and digital beamforming systems. Also, Bai and Heath justified the feasibility of the sectored antenna model in terms of practical resolution of antenna, radiation pattern, boresight direction, and side-lobe effects under mmWave directional multicast service [19]. Thus, in our paper, we also address the mmWave multicast resource allocation problem with a beamforming antenna model as shown in Fig. 1.

So far, some work proposed strategies for multicast under mmWave systems [12], [17], [18], [20]–[29]. Park *et al.* proposed incremental multicast grouping scheme where adaptive beam widths were determined according to the UE distribution to maximize the sum of data rates [20]. The solution in [17] was a beam-grouping algorithm approximating the minimum group data multicast time. Specifically, beam training for an access point (AP) to obtain per-client per-beam received signal strength indication (RSSI) measurements for the multicast group members was performed before grouping. In [21], a

[1] In our preliminary conference paper [13], we only considered homogeneous UEs.

transmission strategy considering both packet transmission deadline and packet loss due to low signal-to-noise ratio (SNR) was proposed. The trade-off between serving many UEs simultaneously and providing high SNR was studied in [12] as the authors tried to optimize serving beam width and UE groups. The retransmission mechanism was also considered. In [18], reflection properties of mmWave were modeled, and a minimum-delay approach to the corresponding mmWave multicast problem was proposed. Chukhno *et al.* proposed a radio resource management for an AP to determine the number and the width required to serve multicast UEs [22]. Apart from the other work, the proposed efficient resource allocation solution considered the energy efficiency aspect in addition to throughput enhancement. In [29], Chukhno *et al.* devised a machine learning-based approach to tackle the computational intensive resource allocation problem in multicast.

Recently, the synergy of multicast and Non-Orthogonal Multiple Access (NOMA) was also explored [30], [31]. In [30], subgrouping techniques and Time Division Multiple Access (TDMA) were utilized to improve the resource allocation outcome of a NOMA multicast system, while a joint power allocation and subgrouping scheme was developed in [31]. Moreover, multicast can be integrated with other communication systems to boost system performance. In [23]–[26], different scenarios of combining unicast and multicast are discussed. Some work also considers caching in a multicast system [27], [28]. However, they all investigated solutions that emphasize throughput or energy-efficient aspects.

On the other hand, some work examined the applications of MBS on video streaming service [32]–[34]. Zhang *et al.* proposed a multicast scheme with NOMA and scalable video coding (SVC) to improve the overall Quality of Experience (QoE) [32]. Similarly, aiming to improve QoE, Li *et al.* formulated the multicast adaptation problem as a minimum dominating set problem and developed a mechanism with QoE guarantee [33]. In addition to video quality, Guo *et al.* also took power consumption into account and solved the optimal transmission time, power allocation, and encoding rate to maximize video quality and minimize power consumption [34].

Unlike previous work on mmWave multicast, we consider resource utilization efficiency based on UEs' perspectives and propose a social-welfare-maximization approach to the mmWave directional multicast resource allocation problem. Social welfare is a measure of the overall system performance by aggregating all the UEs' valuations of the obtained resources. Maximizing social welfare is thus a desirable system property, but a devised algorithm can only achieve this goal with accurate knowledge of UE information. Without a mechanism to collect UEs' private valuations over beam resources, the optimal solution of the optimization algorithm proposed in the previous work may not maximize social welfare in reality. To tackle this issue, we devise a modified VCG mechanism to collect UEs' valuation over the allocated beam resources. Therefore, the service provider can obtain UEs' private valuations correctly, thereby guaranteeing the resource efficiency of the optimization algorithms. Moreover, we take prioritized services into account to cope with the premium membership of many multimedia content providers such as Netflix and Youtube offer. To this end, we extend the mechanism of our preliminary conference version [13], which assumes that all UEs are homogeneous, and provide utility guarantees for high-priority UEs.

The rest of the paper is organized as follows. In Section III, we describe the MBS system model. In Section IV, we formulate the multi-priority MBS resource allocation problem and propose a mechanism as the solution. Then, some preliminary results of a simplified model without priority are presented in Section V. After that, we provide theoretical analysis of the general multi-priority problem in Section VI. In Section VII, we conduct extensive simulations to demonstrate the superiority of the proposed mechanism. Finally, we draw the conclusion in Section VIII.

III. MBS System Model

We consider a non-standalone (NSA) 5G MBS system with one eNB and multiple gNBs, as illustrated in Fig. 1. The eNB provides a broad coverage area with a moderate data rate for UEs. The gNBs with massive Multiple-Input Multiple-Output (MIMO) antennas can beamform and deliver high-throughput and low-latency services for specific applications, such as video streaming. There are r UEs in the system. Under 5G MBS, gNBs utilize beamforming to transmit data, and different data flows are transmitted in different timeslots. In this regard, we divide the UEs into N UE groups. A UE group is a set of UEs receiving the same data flow and in the same beam direction. We denote the UE group i by S_i, and the UEs in UE group i by $A_i = \{a_{i,1}, a_{i,2}, ..., a_{i,m_i}\}$, where m_i denotes the number of UEs in S_i. We also define the UE profile $\mathbf{a} = \{a_{i,j} | i = 1, 2, ..., N, j = 1, 2, ..., m_i\}$.

In this paper, we examine the beam scheduling problem of a gNB. We consider a system with the duration of a radio frame equal to 10 ms, and there are k timeslots in each frame. From [35], we know k could be 10, 20, 40, . . . , 640 under different numerology settings, where 10, 20, and 40 are for sub-6GHz bands and the others are for mmWave bands. We mainly consider the numerologies for mmWave bands. In each timeslot, the gNB transmits data to one UE group. T_i represents the number of timeslots allocated to S_i in a frame, and $\mathbf{T}$ is the resource allocation profile, i.e., $\mathbf{T} = \{T_1, T_2, ..., T_N\}$.

To illustrate, Fig. 1 is an example with one eNB and three gNBs. The coverage area of each gNB is divided into eight beam directions. There are two kinds of data flows, so there are two types of UEs. The UEs that receive the same data flow and are located in the same beam direction are clustered into a UE group. Thus, the numbers of UE groups N in gNB 1, gNB 2, and gNB 3 are 3, 4, and 3, respectively. gNBs transmit data to different UE groups in different timeslots.

In the following, we describe the propagation model and give some definitions to characterize UEs' experienced QoS. Also, we define a UE's priority. Table I summarizes the notations, and the details will be explained in Section IV. Note that we leverage $\mathbb{Z}_n$ to denote $\{1, 2, ..., n\}$ when n is a positive integer.

A. Propagation Model

We follow the propagation model in 3GPP TR 36.776 and TR 38.901 [36], [37] and use the urban macro (UMa) low power/low tower (LPLT) network architecture. We consider a network setting with a gNB. The channel bandwidth B is 100 MHz, the carrier frequency f_c is 28 GHz, and the noise power is n_0. The gNB transmit power is P, antenna gain is g_B, and antenna height is h_B. The UE antenna gain is g_U and antenna height is h_U. There are two types of UEs: UEs with a line of sight (LOS) and UEs without a line of sight (NLOS). The probability of a UE having LOS Pr_{LOS} is given by the following equation, where d_U is the distance between the UE and the gNB.

$$Pr_{LOS}(d_U) = \begin{cases} 1 & \text{if } d_U \leq 18, \\ \frac{18}{d_U} + (1 - \frac{18}{d_U})e^{-\frac{d_U}{63}} & \text{if } 18 < d_U. \end{cases} \quad (1)$$

The pathloss from the gNB to a UE with LOS PL_{LOS} (dB) is given by the following equation. Note that $d_{3D} = \sqrt{(h_B - h_U)^2 + d_U^2}$, $d'_{BP} = \frac{4(h_B-1)(h_U-1)f_c}{c}$ (c: speed of light), and $A = (d'_{BP})^2 + (h_B - h_U)^2$.

$$\begin{aligned} & PL_{LOS}(d_U) \\ &= \begin{cases} 56.9 + 22\log_{10}(d_{3D}) & \text{if } d_U \leq d'_{BP}, \\ \\ 56.9 + 40\log_{10}(d_{3D}) - 9\log_{10}(A) & \text{if } d'_{BP} < d_U. \end{cases} \end{aligned} \quad (2)$$

The pathloss from the gNB to a UE with NLOS PL_{NLOS} (dB) is given by the following equation, where PL_{LOS} is given by (2).

$$\begin{aligned} & PL_{NLOS}(d_U) \\ &= \max(PL_{LOS}, 42.48 + 39.08\log_{10}(d_{3D}) - 0.6(h_U - 1.5)). \end{aligned} \quad (3)$$

Also, we consider a Gaussian shadow fading SF (dB) with 0 mean and σ_{SF} standard deviation.

B. UEs' Valuation

To transmit data under the unreliable channels in a wireless network, we adopt the generalized erasure coding to encode the packets [38], [39]. With the generalized erasure coding, the order of the packets does not matter. Therefore, we can define a UE's valuation function as follows. Note that a UE's valuation of the channel is the private information of the UE and is unknown to the BS.

Definition 1 (UE's valuation function)**.** *A UE $a_{i,j}$'s valuation function $v_{i,j} : \mathbb{N} \rightarrow \{0\} \bigcup \mathbb{R}^+$ is the valuation of the UE when getting t timeslots. We assume that $v_{i,j}(t)$ is concavely increasing and $v_{i,j}(0) = 0$.*

Since there are k timeslots in a frame, we denote a UE $a_{i,j}$'s valuation function $v_{i,j}$ as a k-tuple $V_{i,j} = (v_{i,j}(1), v_{i,j}(2), ..., v_{i,j}(k))$.

Definition 2 (Valuation profile)**.** *The valuation profile of all the UEs is $\boldsymbol{V} = \{V_{i,j} | i = 1, 2, ..., N, j = 1, 2, ..., m_i\}$.*

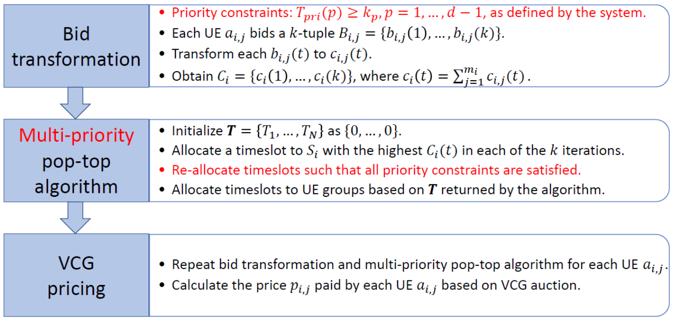

Bid transformation
• Priority constraints: $T_{pri}(p) \geq k_p, p = 1, \dots, d-1$, as defined by the system.
• Each UE $a_{i,j}$ bids a k-tuple $B_{i,j} = \{b_{i,j}(1), \dots, b_{i,j}(k)\}$.
• Transform each $b_{i,j}(t)$ to $c_{i,j}(t)$.
• Obtain $C_i = \{c_i(1), \dots, c_i(k)\}$, where $c_i(t) = \sum_{j=1}^{m_i} c_{i,j}(t)$.
Multi-priority pop-top algorithm
• Initialize $\boldsymbol{T} = \{T_1, \dots, T_N\}$ as $\{0, \dots, 0\}$.
• Allocate a timeslot to S_i with the highest $C_i(t)$ in each of the k iterations.
• Re-allocate timeslots such that all priority constraints are satisfied.
• Allocate timeslots to UE groups based on $\boldsymbol{T}$ returned by the algorithm.
VCG pricing
• Repeat bid transformation and multi-priority pop-top algorithm for each UE $a_{i,j}$.
• Calculate the price $p_{i,j}$ paid by each UE $a_{i,j}$ based on VCG auction.

Fig. 2. Flow chart of the proposed resource allocation mechanism. When no priority is concerned, the mechanism only performs the steps in black color. When the system distinguishes between different priorities, the mechanism performs all the steps.

C. UEs' Utility

A UE will get higher utility when it obtains more timeslots since getting more timeslots means being served for a longer period of time. The relationship between the number of acquired timeslots and a UE's utility is prescribed by the valuation function defined previously. However, UEs will also need to pay a corresponding price to acquire the resources, which serves as the cost of the UE. In this regard, after the resource allocation, the gNB will charge a price $p_{i,j}$ for the UE $a_{i,j}$ according to the modified VCG auction described later. We define $a_{i,j}$'s utility below.

Definition 3 (UE's utility)**.** *A UE $a_{i,j}$'s utility of getting t timeslots and the bid profile being $\boldsymbol{B}$ is $u_{i,j}(t, \boldsymbol{B}) = v_{i,j}(t) - p_{i,j}(\boldsymbol{B})$, where $\boldsymbol{B}$ will be defined later.*

In the following, we drop the value in the parenthesis to simplify notation, e.g., using $u_{i,j}$ instead of $u_{i,j}(t, \mathbf{B})$ or $p_{i,j}$ instead of $p_{i,j}(\mathbf{B})$.

D. UEs' Priority

Definition 4 (UE's priority)**.** *We denote $pri(a_{i,j})$ as a UE $a_{i,j}$'s priority. If $a_{i,j}$ is a p-priority UE, $pri(a_{i,j}) = p$.*

With a slight abuse of notations, since a UE group consists of UEs with the same priority, $pri(S_i)$ is used to denote the priority of S_i.

Definition 5 (p-priority timeslots)**.** *The total timeslots allocated to UE groups with priority at least p is $T_{pri}(p) = \sum_{pri(S_i) \geq p} T_i$.*

IV. Multi-Priority Resource Allocation Problem and Mechanism Design

A. Multi-Priority Resource Allocation Problem

We consider a 5G MBS system with different UE groups having different priorities. This happens when a content provider wants to distinguish between different services or ordinary UEs and premium UEs. This section considers a network with d priorities. We assume that the packets are transmitted over a quasi-static Rayleigh fading channel. Hence, the UEs' valuations are the same during each frame and independent between different frames, and it is sufficient to consider one frame.

TABLE I
NOTATIONS

Notation	Definition
N	Number of UE groups
r	Number of UEs
d	Number of priorities
k	Number of timeslots per frame
S_i	UE group i
$pri(S_i)$	S_i's priority
m_i	Number of UEs in S_i
$a_{i,j}$	UE j in S_i
$pri(a_{i,j})$	$a_{i,j}$'s priority
A_i	Set of UEs in S_i $A_i = \{a_{i,1}, a_{i,2}, ..., a_{i,m_i}\}$
a	UE profile $\mathbf{a} = \{a_{i,j} \mid i = 1, 2, ..., N, j = 1, 2, ..., m_i\}$
T_i	Number of timeslots allocated to S_i
$T_{pri}(p)$	p-priority timeslots $T_{pri}(p) = \sum_{pri(S_i) \geq p} T_i$
k_p	Minimum number of p-priority timeslots $T_{pri}(p) \geq k_p$
T	Resource allocation profile $\mathbf{T} = \{T_1, T_2, ..., T_N\}$
$v_{i,j}(t)$	$a_{i,j}$'s valuation for t timeslots
$V_{i,j}$	$a_{i,j}$'s valuation function $V_{i,j} = (v_{i,j}(1), v_{i,j}(2), ..., v_{i,j}(k))$
$v_i(t)$	S_i's valuation for t timeslots $v_i(t) = \sum_{j=1}^{m_i} v_{i,j}(t)$
V_i	S_i's valuation function $V_i = (v_i(1), v_i(2), ..., v_i(k))$
V	Valuation profile $\mathbf{V} = \{V_{i,j} \mid i = 1, 2, ..., N, j = 1, 2, ..., m_i\}$
$b_{i,j}(t)$	$a_{i,j}$'s bid for t timeslots
$B_{i,j}$	$a_{i,j}$'s bid function $B_{i,j} = (b_{i,j}(1), b_{i,j}(2), ..., b_{i,j}(k))$
$b_i(t)$	S_i's bid for t timeslots $b_i(t) = \sum_{j=1}^{m_i} b_{i,j}(t)$
B_i	S_i's bid function $B_i = (b_i(1), b_i(2), ..., b_i(k))$
B	Bid profile $\mathbf{B} = \{B_{i,j} \mid i = 1, 2, ..., N, j = 1, 2, ..., m_i\}$
$c_{i,j}(t)$	$a_{i,j}$'s marginal transformed bid for t timeslots $c_{i,j}(t) = \begin{cases} \underset{1 \leq p \leq t}{min} (b_{i,j}(p) - b_{i,j}(p-1)), t = 1, 2, ..., k \\ 0, t = 0 \end{cases}$
$C_{i,j}$	$a_{i,j}$'s marginal transformed bid function $C_{i,j} = (c_{i,j}(1), c_{i,j}(2), ..., c_{i,j}(k))$
$c_i(t)$	S_i's marginal transformed bid for t timeslots $c_i(t) = \sum_{j=1}^{m_i} c_{i,j}(t)$
C_i	S_i's marginal transformed bid function $C_i = \{c_i(1), c_i(2), ..., c_i(k)\}$
C	Marginal transformed bid profile $\mathbf{C} = \{c_i(t) \mid i = 1, 2, ..., N, t = 1, 2, ..., k\}$
ϕ	Social welfare $\phi = \sum_{i=1}^{N} \sum_{j=1}^{m_i} v_{i,j}(T_i)$
$W(\mathbf{B})$	Total transformed bid when the bid profile is **B** $W(\mathbf{B}) = \sum_{i=1}^{N} \sum_{j=1}^{m_i} \sum_{t=1}^{T_i(\mathbf{B})} c_{i,j}(t)$
$p_{i,j}(\mathbf{B})$	$a_{i,j}$'s payment when the bid profile is **B** $p_{i,j}(\mathbf{B}) = W_{\mathbf{a}-a_{i,j}}(\mathbf{B} - B_{i,j}) - W_{\mathbf{a}-a_{i,j}}(\mathbf{B})$
$u_{i,j}(t, \mathbf{B})$	$a_{i,j}$'s utility of getting t timeslots and the bid profile being **B** $u_{i,j}(t, \mathbf{B}) = v_{i,j}(t) - p_{i,j}(\mathbf{B})$
n_0	Noise power
P	gNB transmit power
g_B	gNB antenna gain
h_B	gNB antenna height
g_U	UE antenna gain
h_U	UE antenna height
B	Channel bandwidth (100 MHz)
f_c	Carrier frequency (28 GHz)
σ_{SF}	Standard deviation of Gaussian shadow fading

Under the MBS system, the resources are the k timeslots allocated to the UEs by the gNB, and we formulate the resource allocation problem as an auction described below. At the beginning of each frame, each UE bids a k-tuple $B_{i,j}$, which is the UE's bid function. Note that we consider a typical 5G NSA architecture, which consists of eNBs and gNBs. Therefore, UEs can transmit control messages, e.g., bid functions, via the LTE control plane. Bid functions in the same UE group S_i are collected as a total bid B_i, which is also a k-tuple $\{b_i(1), b_i(2), ..., b_i(k)\}$, where $i = 1, 2, ..., N$. Also, we define **B** as the bid profile, i.e., $\mathbf{B} = \{B_{i,j} | i = 1, 2, ..., N, t = 1, 2, ..., m_i\}$.

To characterize UE priority, we develop different service guarantees for different UE group priorities. The k timeslots are divided into d subperiods with priorities ranging from 0 to $d-1$. The p-priority subperiod contains k_p timeslots, where $0 \leq p \leq d-1$. Also, we have $k = \sum_{i=0}^{d-1} k_i$. The timeslots of the p-priority subperiod can only be allocated to the UE groups with q-priority UEs, where $q \geq p$. The gNB decides how to allocate the k timeslots to the N UE groups based on the N total bids and the priority constraints, and our system model can be formulated as the following social-welfare-maximization optimization problem (P1). Note that $v_{i,j}(t)$ is UEs' private information, but the gNB only obtains $b_{i,j}(t)$. Therefore, we need to devise an incentive mechanism to motivate UEs to report truthfully, i.e., $b_{i,j}(t) = v_{i,j}(t)$.

$$(P1) : \max_{\mathbf{T}} \sum_{i=1}^{N} \sum_{j=1}^{m_i} v_{i,j}(T_i). \tag{4}$$

$$s.t. \sum_{i=1}^{N} T_i = k, \tag{5}$$

$$T_{pri}(p) \geq k_p, \forall p \in \mathbb{Z}_{d-1}. \tag{6}$$

Note that there may be no valid solution to (P1). We first find the highest priority among all UE groups, denoted as p_{max}. Then, we check whether there exists $q > p_{max}$ such that $k_q > 0$. If so, then no valid solution exists. For example, when $d = 2$, we cannot find a valid solution if the coverage area of the gNB only contains 0-priority UEs and $k_1 > 0$. In this case, no allocation profile satisfies the priority constraint because there is no 1-priority UE group. To tackle this issue, the gNB allocates the timeslots originally belonging to the 1-priority UE groups to the 0-priority ones by setting $k_1 = 0$ and $k_0 = k$. In the following, we assume that a valid solution exists to simplify the arguments.

Thus, we can see this auction as a d-tier Stackelberg game with high-priority UEs as leaders and low-priority UEs as followers. Usually, a Stackelberg game is solved by backward induction, which involves finding the optimal solution of the d-tier game given any outcome of the $(d-1)$-tier game and then solving the optimal solution of the d-tier game. However, solving a Stackelberg game by backward induction is time-consuming due to the need to enumerate all the possible d-tier solutions given the outcomes of the $(d-1)$-tier game.

To this end, we propose the following mechanism, which consists of three stages: bid transformation, multi-priority pop-top algorithm, and VCG pricing. First, the gNB pre-processes each UE's bid. Then, the multi-priority pop-top algorithm is proposed to solve a reformulated optimization problem of (P1). The optimization problem is identical to (P1) when UEs report their valuation functions honestly. In this regard, we utilize VCG pricing to motivate each UE to bid truthfully. Fig. 2 is the flow chart of the proposed mechanism, and the details are described below.

B. Bid Transformation

Each UE $a_{i,j}$ bids a k-tuple bid function: $B_{i,j} = \{b_{i,j}(1), b_{i,j}(2), ..., b_{i,j}(k)\}$. Then, each $B_{i,j}$ is transformed to the marginal transformed bid function: $C_{i,j} = \{c_{i,j}(1), c_{i,j}(2), ..., c_{i,j}(k)\}$, where $c_{i,j}(t)$ is the marginal transformed bid of $b_{i,j}(t)$, and the formula is given below. Note that we have assumed that $b_{i,j}(0) = 0$.

$$c_{i,j}(t) = \begin{cases} \underset{1\le p\le t}{min}(b_{i,j}(p) - b_{i,j}(p-1)), & t = 1, 2, ..., k \\ 0, \quad t = 0. \end{cases} \tag{7}$$

After the bid transformation of each UE in S_i, the total bid $C_i = \{c_i(1), c_i(2), ..., c_i(k)\}$ is obtained by accumulating all the transformed bids of the UEs in S_i. That is, $c_i(t) = \sum_{j=1}^{m_i} c_{i,j}(t)$. Thus, C_i is the marginal transformed bid function of S_i. We also define the marginal transformed bid profile $\mathbf{C} = \{c_i(t) | i = 1, 2, ..., N, t = 1, 2, ..., k\}$.

C. Multi-Priority Pop-Top Algorithm

After bid transformation, the gNB allocates the resources to the UEs according to the multi-priority pop-top algorithm described in Algorithm 1. This algorithm solves the following optimization problem (P2), which becomes (P1) when UEs truthfully report their valuation functions, as prescribed by Lemma 1 given in Appendix B. We will give a formal proof later. Note that we label the constraints in (10) by (10-1) to (10-($d-1$)), and we call them priority constraints.

$$(P2) : \max_{\mathbf{T}} \sum_{i=1}^{N} \sum_{j=1}^{m_i} \sum_{t=0}^{T_i} c_{i,j}(t). \tag{8}$$

$$s.t. \sum_{i=1}^{N} T_i = k, \tag{9}$$

$$T_{pri}(p) \geq k_p, \forall p \in \mathbb{Z}_{d-1}. \tag{10}$$

Design 1 (Multi-priority pop-top algorithm)**.** *The multi-priority pop-top algorithm is specified in Algorithm 1. The algorithm first initializes T_1 to T_N. Then, it allocates a timeslot to the UE group with the highest marginal transformed bid in each iteration. When there is a tie, we can allocate the timeslot to UE groups with higher priority or break the tie arbitrarily. After k iterations, the resource allocation terminates. Then, the algorithm will check whether $T_{pri}(p)$ is smaller than k_p for $p = 0, 1, ..., d-1$. If so, it re-allocates the resources to the high-priority UE groups until $T_{pri}(p) = k_p$. This process will be repeated d times until all the priority constraints are satisfied.*

Algorithm 1 Multi-Priority Pop-Top Algorithm

```
Input: $k, k_0, ..., k_{d-1}, N, C_1, C_2, ..., C_N$
Output: $T_1, T_2, ..., T_N$
  Initialization :
  $Stack\ candidate[0], candidate[1], ..., candidate[d-1].$
  $Stack\ value[0], value[1], ..., value[d-1].$
  for $i = 1$ to $N$ do
      $T_i = 0.$
  end for
  Resource Allocation :
  for $i = 1$ to $k$ do
      $s = \arg\max_j c_j(T_j + 1).$
      $T_s = T_s + 1.$
      $candidate[pri(S_s)].push(s).$
      $value[pri(S_s)].push(c_s(T_s)).$
  end for
  Resource Re-Allocation :
  for $p = 1$ to $d-1$ do
      while $\sum_{j=p}^{d-1} candidate[j].size() < k_p$ do
          $s = \arg\max_{j, pri(S_j) \geq p} c_j(T_j + 1).$
          $T_s = T_s + 1.$
          $candidate[pri(S_s)].push(s).$
          $value[pri(S_s)].push(c_s(T_s)).$
          $int\ tmp = \arg\min_{0 \leq j \leq p-1} value[j].top().$
          $int\ tmpid = candidate[tmp].top()$
          $T_{tmpid} = T_{tmpid} - 1.$
          $candidate[tmp-1].pop().$
          $value[tmp-1].pop().$
      end while
  end for
  return $T_1, T_2, ..., T_N$
```

Note that when there is a tie, we can break it arbitrarily. However, in the following, we assume that there is no tie to simplify the arguments, although the arguments can be easily extended to encompass the situations when there are ties.

The multi-priority pop-top algorithm will satisfy the constraint one by one. In the resource allocation phase, Algorithm 1 satisfies the constraint $\sum_{i=1}^{N} T_i = k$. Then, at the p-th iteration of the resource re-allocation phase, it satisfies the constraint $T_{pri}(p) \geq k_p$ while maintaining the previously satisfied constraints. In re-allocating resources, Algorithm 1 always allocates the timeslot of the low-priority UE group with the lowest transformed marginal bid to the high-priority UE group with the highest one. In this way, the optimal solution is guaranteed. We will prove this property later.

In Algorithm 1, $candidate[p]$ stores the allocated UE groups with priority p, and $value[p]$ stores the marginal transformed bid of the allocated UE groups with priority p. There are d levels of priority, so there are $2d$ stacks to store the information. We leverage these data structures to reduce the need to traverse the whole bid profile and reduce running time.

D. VCG Pricing

After the resource allocation, each UE $a_{i,j}$ is charged a price $p_{i,j}$ according to the VCG auction, which is a technique commonly used in mechanism design. In the VCG auction, each player pays an amount equal to the utility loss of all the other players due to its presence. Using this payment rule, the auctioneer, which is the gNB in our system model, incentivizes the players to reveal their true valuations. In the following, we give a more detailed description of VCG pricing.

First, we define the total transformed bid W as follows.

Definition 6 (Total transformed bid)**.** *The total transformed bid of all the UEs when the bid profile is* $\boldsymbol{B}$ *is* $W(\boldsymbol{B}) = \sum_{i=1}^{N} \sum_{j=1}^{m_i} \sum_{t=1}^{T_i(\boldsymbol{B})} c_{i,j}(t)$.

Note that we will use a subscript to denote the region of summation, e.g.,$W_{\mathbf{a}}(\mathbf{B}) = W(\mathbf{B})$. Also, we have the following notation, where the resource allocation profile $\mathbf{T} = \{T_s | s \in \mathbb{Z}_N\}$ is based on the bid profile $\mathbf{B} = \{B_{s,t} | s \in \mathbb{Z}_N, t \in \mathbb{Z}_{m_i}\}$.

$$W_{\mathbf{a}-a_{i,j}}(\mathbf{B}) = \sum_{\substack{s\in\mathbb{Z}_N, t\in\mathbb{Z}_{m_i} \\ (s,t)\neq(i,j)}}\sum \sum_{t=0}^{T_i} c_{s,t}(t). \tag{11}$$

On the other hand, when $a_{i,j}$ is not in the system, we have a different total transformed bid as described below, where the resource allocation profile $\mathbf{T}' = \{T'_s | s \in \mathbb{Z}_N\}$ is based on the bid profile $\mathbf{B} - B_{i,j} = \{B_{s,t} | s \in \mathbb{Z}_N, t \in \mathbb{Z}_{m_i}, (i,j) \neq (s,t)\}$.

$$W_{\mathbf{a}-a_{i,j}}(\mathbf{B} - B_{i,j}) = \sum_{\substack{s\in\mathbb{Z}_N, t\in\mathbb{Z}_{m_i} \\ (s,t)\neq(i,j)}}\sum \sum_{t=0}^{T'_i} c_{s,t}(t). \tag{12}$$

With the above notations, we formally define the VCG pricing rule.

Design 2 (VCG Pricing)**.** *A UE* $a_{i,j}$ *will be charged* $p_{i,j}$ *as follows.*

$$p_{i,j} = W_{\boldsymbol{a}-a_{i,j}}(\boldsymbol{B} - B_{i,j}) - W_{\boldsymbol{a}-a_{i,j}}(\boldsymbol{B}). \tag{13}$$

V. Preliminary Results

In this section, we demonstrate some properties of the proposed mechanism by considering a simplified scenario where there is no priority constraints, i.e., $d = 1$. This section serves as the preliminary results of Section VI.

A. Resource Allocation Problem

Since there are no priority constraints, the gNB decides how to allocate the k timeslots to the N UE groups based on the N total bids, and our system model can be formulated as the following social-welfare-maximization optimization problem (P3).

$$(P3) : \max_{\mathbf{T}} \sum_{i=1}^{N} \sum_{j=1}^{m_i} v_{i,j}(T_i). \tag{14}$$

$$s.t. \sum_{i=1}^{N} T_i = k. \tag{15}$$

Thus, the mechanism is the same as the multi-priority one with some simplification to Algorithm 1. Fig. 2 in black color is the flow chart of the proposed mechanism.

B. Pop-Top Algorithm

When there are no priority constraints, the multi-priority pop-top algorithm reduces to the pop-top algorithm described in Algorithm 2 This algorithm aims to solve the following optimization problem (P4), which becomes (P3) when UEs truthfully report their valuation functions.

$$(P4) : \max_{\mathbf{T}} \sum_{i=1}^{N} \sum_{j=1}^{m_i} \sum_{t=0}^{T_i} c_{i,j}(t). \tag{16}$$

$$s.t. \sum_{i=1}^{N} T_i = k. \tag{17}$$

Algorithm 2 Pop-Top Algorithm

```
Input: $k$, $N$, $C_1, C_2, ..., C_N$
Output: $T_1, T_2, ..., T_N$
    Initialization :
for $i = 1$ to $N$ do
    $T_i = 0$.
end for
    Resource Allocation :
for $i = 1$ to $k$ do
    $s = \arg\max_j c_j(T_j + 1)$.
    $T_s = T_s + 1$.
end for
return $T_1, T_2, ..., T_N$
```

Design 3 (Pop-top algorithm)**.** *The pop-top algorithm is specified in Algorithm 2. The algorithm first initializes* T_1 *to* T_N*. Then, it allocates a timeslot to the UE group with the highest marginal transformed bid in each iteration. After* k *iterations, the allocation terminates, and each UE group* S_i *will be allocated* T_i *timeslots.*

Some properties of the pop-top algorithm are given in Appendix A.

C. Incentive-Compatibility

In this subsection, we prove that the proposed mechanism is incentive-compatible. To maximize its utility, an UE may falsely report its valuation function, thus reducing social welfare. However, under the proposed incentive-compatible mechanism, each UE can get maximum utility when it truthfully reports the valuation function, i.e., truthful bidding is each UE's dominant strategy.

Theorem 1. *The proposed mechanism is incentive-compatible. That is, truthful bidding is the dominant strategy.*

Proof. This is proved in Appendix C. □

D. Social Welfare Maximization

In this subsection, we prove that the proposed mechanism maximizes social welfare, which is the summation of all UEs' valuations as defined below.

Definition 7 (Social welfare)**.** *Social welfare ϕ is defined as follows.*

$$\phi = \sum_{i=1}^{N} \sum_{j=1}^{m_i} v_{i,j}(T_i). \tag{18}$$

Theorem 2. *The proposed mechanism maximizes social welfare ϕ.*

Proof. This is proved in Appendix D. □

E. Individual Rationality

In this subsection, we prove that the proposed mechanism is individually rational. A mechanism achieves individual rationality when no player gets negative utility by participating. Thus, all players are willing to join the auction.

Theorem 3. *The proposed mechanism is individually rational. That is, if a player $a_{i,j}$ is truthful, then $u_{i,j} \geq 0$.*

Proof. This is proved in Appendix E. □

F. Budget Balance

After proving the willingness of UEs to participate previously, we show that the gNB is also willing to be the auctioneer because it can earn money. In particular, we demonstrate that the proposed mechanism is (weakly) budget-balanced, i.e., the total payment of the players is non-negative.

Theorem 4. *The proposed mechanism is (weakly) budget-balanced. That is, $\sum_{i=1}^{N}\sum_{j=1}^{m_i} p_{i,j} \geq 0$.*

Proof. This is proved in Appendix F. □

G. Polynomial-Time Complexity

Finally, we analyze the time complexity of the proposed mechanism and show that it runs in polynomial time in this subsection. Different from [13], we give a tighter bound for the mechanism by removing unnecessary operations in VCG pricing.

Theorem 5. *The proposed mechanism runs in $O(Nrk)$, which is polynomial-time.*

Proof. This is proved in Appendix G. □

VI. Incentive-Compatibility and Other Desirable Properties of Multi-Priority Model

A. Incentive-Compatibility

In this subsection, we will prove that the multi-priority pop-top algorithm maximizes the total transformed bid and is incentive-compatible.

First, we consider $d = 2$, where we only have one priority constraint. There are two possible situations for this optimization problem: The solution is the same without (10) or the solution is different without (10). Here we borrow the terminology from convex optimization [40] and discuss the two situations.

Definition 8 (Inactive priority constraint)**.** *We say the priority constraint in (P2) is inactive when the solution is the same with or without the priority constraint.*

Proposition 1. *The multi-priority pop-top algorithm maximizes the total transformed bid if the priority constraint is inactive.*

Proof. When the priority constraint is inactive, Algorithm 1 is the same as Algorithm 2. Hence, it maximizes the total transformed bid by Proposition 10. □

Definition 9 (Active priority constraint)**.** *We say the priority constraint in (P2) is active when the solution is different with or without the priority constraint.*

Proposition 2. *If the priority constraint is active, $T_{pri}(1) = k_1$.*

Proof. We will prove this by contradiction, and we suppose that the priority constraint is active and $T_{pri}(1) > k_1$. We denote $T'_{pri}(1)$ as the total timeslots allocated to the high-priority UE groups without the priority constraint. Since the priority constraint is active, $T'_{pri}(1) \leq k_1$, for otherwise, the priority constraint is inactive.

Then, we can construct a different allocation profile by allocating a timeslot of the high-priority UE group with the lowest transformed marginal bid to the low-priority UE group with the highest transformed marginal bid. The new allocation profile still satisfies the priority constraint and has a higher total transformed bid than the original one. This contradicts the claim that the original allocation profile is optimal. Thus, $T_{pri}(1) = k_1$. □

Note that we cannot use the relationship between $T_{pri}(1)$ and k_1 to determine whether the constraint is inactive. When $T_{pri}(1) = k_1$, the constraint may be inactive or active, so our definitions utilize the optimal solutions with or without the constraint.

Proposition 3. *The multi-priority pop-top algorithm maximizes the total transformed bid if $d = 2$ and the priority constraint is active.*

Proof. When the priority constraint is active, Algorithm 1 enters the resource re-allocation phase, re-allocating some resources from low-priority UE groups to high-priority ones.

Since the the priority constraint is active, $T_{pri}(1) = k_1$. Therefore, we can allocate k_1 timeslots to the high-priority UE groups and k_0 timeslots to the low priority UE groups according to Algorithm 2. This allocation mechanism will give the optimal solution.

In the resource re-allocation phase, Algorithm 1 allocates the timeslots for low-priority UE groups after the k_0-th largest marginal transformed bid to the high-priority UE groups. This is equivalent to using Algorithm 2 to allocate k_0 timeslots to low-priority UE groups and k_1 timeslots to high-priority UE groups. Thus, the solution is optimal since Algorithm 2 gives the optimal solutions for both low-priority and high-priority UE groups by Proposition 10. □

Proposition 4. *The multi-priority pop-top algorithm maximizes the total transformed bid if $d = 2$.*

Proof. When the priority constraint is inactive, the multi-priority pop-top algorithm maximizes the total transformed bid according to Proposition 1. Conversely, when the priority constraint is active, it maximizes the total transformed bid according to Proposition 3.

Since the priority constraint is either inactive or active, the above results prove the proposition. □

Proposition 5. *The multi-priority pop-top algorithm maximizes the total transformed bid.*

Proof. We will prove this by mathematical induction.

Claim: The resource allocation profile $\mathbf{T}$ after the p-th iteration of the resource re-allocation phase in Algorithm 1 is the optimal solution of (P2) with priority constraints (10-1) to (10-p).

Base Case: For $p = 1$, the claim holds by Proposition 4.

Inductive Step: Assume that the claim holds for $p = t$, and we will prove that the claim holds for $p = t + 1$.

If the resource allocation profile already satisfies priority constraint (10-p), the claim holds for $p = t + 1$.

If not, then $T_{pri}(p) = k_p$ after the resource re-allocation phase, which can be proved in a similar way as that in Algorithm 1. We denote the total timeslots allocated to the UE groups with priorities higher than p before and after the iteration as $T'_{pri}(p)$ and $T_{pri}(p)$, respectively.

We consider the UE groups with priorities at least p. Since the original resource allocation profile is optimal, the first $T'_{pri}(p)$ timeslots are the same, and the rest $T_{pri}(p) - T'_{pri}(p)$ timeslots are allocated to the UE groups with the highest marginal transformed bids as Algorithm 1 does.

Then, we consider the UE groups with priorities lower than p. Since the original resource allocation profile is optimal, the first $k - T_{pri}(p)$ timeslots are the same, and the rest $T'_{pri}(p) - T_{pri}(p)$ timeslots are removed from the UE groups with the lowest marginal transformed bids as Algorithm 1 does.

Therefore, the p-th iteration of the resource re-allocation phase preserves optimality, thus completing the inductive step. □

Theorem 6. *The proposed multi-priority mechanism is incentive-compatible.*

Proof. Each UE $a_{\alpha,\beta}$ will choose the strategy that will optimize its utility as follows. We denote the number of timeslots allocated to S_i with or without $a_{\alpha,\beta}$ by T_i and T'_i, respectively.

$$\begin{aligned}
\max_{B_{\alpha,\beta}} u_{\alpha,\beta} &= \max_{B_{\alpha,\beta}} v_{i,j}(T_\alpha) - p_{\alpha,\beta} \\
&= \max_{B_{\alpha,\beta}} [v_{\alpha,\beta}(T_\alpha) + \sum\sum_{\substack{i\in\mathbb{Z}_N, j\in\mathbb{Z}_{m_i} \\ (i,j)\neq(\alpha,\beta)}} \sum_{t=0}^{T_i} c_{i,j}(t) \\
&\quad - \sum\sum_{\substack{i\in\mathbb{Z}_N, j\in\mathbb{Z}_{m_i} \\ (i,j)\neq(\alpha,\beta)}} \sum_{t=0}^{T'_i} c_{i,j}(t)] \qquad (19) \\
&= \max_{B_{\alpha,\beta}} [v_{\alpha,\beta}(T_\alpha) + \sum\sum_{\substack{i\in\mathbb{Z}_N, j\in\mathbb{Z}_{m_i} \\ (i,j)\neq(\alpha,\beta)}} \sum_{t=0}^{T_i} c_{i,j}(t)] \\
&\quad - \sum\sum_{\substack{i\in\mathbb{Z}_N, j\in\mathbb{Z}_{m_i} \\ (i,j)\neq(\alpha,\beta)}} \sum_{t=0}^{T'_i} c_{i,j}(t).
\end{aligned}$$

Since $\sum\sum_{\substack{i\in\mathbb{Z}_N, j\in\mathbb{Z}_{m_i} \\ (i,j)\neq(\alpha,\beta)}} \sum_{t=0}^{T'_i} c_{i,j}(t)$ does not depend on $B_{i,j}$, we do not need to include them in the optimization problem. Then, the optimization problem becomes the following.

$$\max_{B_{\alpha,\beta}} [v_{\alpha,\beta}(T_i) + \sum\sum_{\substack{i\in\mathbb{Z}_N, j\in\mathbb{Z}_{m_i} \\ (i,j)\neq(\alpha,\beta)}} \sum_{t=0}^{T_i} c_{i,j}(t)]. \qquad (20)$$

Also, the multi-priority pop-top algorithm solves the following optimization problem.

$$\max_{B_{\alpha,\beta}} [\sum_{t=0}^{T_i} c_{\alpha,\beta}(t) + \sum\sum_{\substack{i\in\mathbb{Z}_N, j\in\mathbb{Z}_{m_i} \\ (i,j)\neq(\alpha,\beta)}} \sum_{t=0}^{T_i} c_{i,j}(t)]. \qquad (21)$$

Thus, if $a_{\alpha,\beta}$ bids truthfully, (21) and (20) are the same. Hence, truthful bidding is the player's dominant strategy. □

B. Social Maximization

Theorem 7. *The proposed multi-priority mechanism maximizes social welfare.*

Proof. By Theorem 6, the proposed mechanism is incentive-compatible. Therefore, social welfare is equal to the summation of all winners' marginal transformed bids, which is the total transformed bid. Also, by Proposition 5, the multi-priority pop-top algorithm maximizes the total transformed bid. Therefore, the proposed mechanism maximizes social welfare. □

C. Individual Rationality

Theorem 8. *The proposed multi-priority mechanism is individually rational. That is, if a player $a_{i,j}$ is truthful, then $u_{i,j} \geq 0$.*

Proof. If $a_{i,j}$ bids truthfully, $u_{i,j}$ can be written as follows.

$$u_{i,j} = W_{\mathbf{a}}(\mathbf{B}) - W_{\mathbf{a}-a_{i,j}}(\mathbf{B} - B_{i,j}). \qquad (22)$$

Since $\mathbf{B} - B_{i,j} \subset \mathbf{B}$, and the multi-priority pop-top algorithm will maximize the total transformed bid by Proposition 5. Therefore, $W_{\mathbf{a}}(\mathbf{B}) \geq W_{\mathbf{a}-a_{i,j}}(\mathbf{B} - B_{i,j})$, and $u_{i,j} \geq 0$. □

D. Budget Balance

Theorem 9. *The proposed multi-priority mechanism is (weakly) budget-balanced. That is, $\sum_{i=1}^{N}\sum_{j=1}^{m_i} p_{i,j} \geq 0$.*

Proof. The price $p_{i,j}$ paid by $a_{i,j}$ is as follows.

$$p_{i,j} = W_{\mathbf{a}-a_{i,j}}(\mathbf{B} - B_{i,j}) - W_{\mathbf{a}-a_{i,j}}(\mathbf{B}). \tag{23}$$

We denote the resource allocation profile produced by $\mathbf{B} - B_{i,j}$ as $\mathbf{T}'$, and the resource allocation profile produced by $\mathbf{B}$ as $\mathbf{T}$. Also, combining Proposition 5 and Theorem 6, the proposed mechanism maximizes social welfare. If $W_{\mathbf{a}-a_{i,j}}(\mathbf{B} - B_{i,j}) < W_{\mathbf{a}-a_{i,j}}(\mathbf{B})$, $\mathbf{T}$ will give higher social welfare than $\mathbf{T}'$ when the bid profile is $\mathbf{B} - B_{i,j}$, contradicting with the above claim. Therefore, $W_{\mathbf{a}-a_{i,j}}(\mathbf{B} - B_{i,j}) \geq W_{\mathbf{a}-a_{i,j}}(\mathbf{B})$. Hence, $p_{i,j} = W_{\mathbf{a}-a_{i,j}}(\mathbf{B} - B_{i,j}) - W_{\mathbf{a}-a_{i,j}}(\mathbf{B}) \geq 0$, meaning that $\sum_{i=1}^{N}\sum_{j=1}^{m_i} p_{i,j} \geq 0$. □

E. Polynomial-Time Complexity

Theorem 10. *The proposed multi-priority mechanism runs in polynomial time.*

Proof. We denote the time complexity of bid transformation, multi-priority pop-top algorithm, and VCG pricing by γ_1, γ_2, and γ_3, respectively.

As in Theorem 5, $\gamma_1 = O(rk)$. As for the multi-priority pop-top algorithm, we analyze the time complexity of the three for loops. The first for loop (lines 3-5) takes $O(N)$, and the second for loop (lines 6-11) takes $O(Nk)$, as in Theorem 5. As for the third for loop (lines 12-24), we use aggregate analysis, one of the amortized analysis techniques. The total number of while loop is upper bounded by k since at most k timeslots are re-allocated. Inside each iteration, line 14 and line 18 take $O(N)$ due to the need to traverse all UE groups. Thus, the third for loop takes $O(Nk)$, and $\gamma_2 = O(N) + O(Nk) + O(Nk) = O(Nk)$. The VCG pricing will run the bid transformation and the multi-priority pop-top algorithm for every winning UE. By the same technique in Theorem 5, $\gamma_3 = O(Nrk)$. Thus, $\gamma_1 + \gamma_2 + \gamma_3 = O(Nrk)$, which is polynomial-time. □

F. High-Priority Superiority

In this subsection, we prove that the expected utility of a high-priority UE is at least that of a low-priority UE. We consider $d = 2$, but the arguments can be generalized to any d. To simplify the proof, we have the following three assumptions.

1) The UEs' location distributions, i.e., which UE group each UE belongs to, follow independent and identically distributed (i.i.d.) distributions, no matter the UE priority.
2) The UEs' valuation functions follow i.i.d. distributions, no matter the UE priority.
3) The total number of UEs r is large.

Theorem 11. *The expected value of a high-priority UE's utility is higher than or equal to the expected value of a low-priority UE's utility under the proposed multi-priority mechanism with the assumptions specified above.*

Proof. We denote the expected value of a high-priority UE's utility by $\mathbb{E}_{pri(a_{i,j})=1}[u_{i,j}]$ and the expected value of a low-priority UE's utility by $\mathbb{E}_{pri(a_{i,j})=0}[u_{i,j}]$.

First, we consider the situation when $k_1 = 0$. Since there is no difference between a high-priority UE and a low-priority one without the priority constraint, we have the following.

$$\mathbb{E}_{pri(a_{i,j})=1}[u_{i,j}] = \mathbb{E}_{pri(a_{i,j})=0}[u_{i,j}]. \tag{24}$$

Then, we consider $k_1 > 0$, and we will prove that $\mathbb{E}'_{pri(a_{i,j})=1}[u_{i,j}] \geq \mathbb{E}'_{pri(a_{i,j})=0}[u_{i,j}]$.

We can express $u_{i,j}$ as $u_{i,j} = v_{i,j} - p_{i,j}$, where $p_{i,j}$ is the social welfare lose it has incurred on other UEs when participating.

Note that when the number of UEs is high, it is unlikely that a single UE's decision will change the allocation profile significantly. Moreover, since the UEs' valuation functions follow i.i.d. distributions, we obtain the following relationship.

$$\frac{p_{i,j}}{v_{i,j}} = O(\frac{1}{r}). \tag{25}$$

Therefore, when r is large, we can ignore $p_{i,j}$ and focus on $v_{i,j}$. Since $v_{i,j}$ depends on the timeslots a high-priority UE $a_{i,j}$ gets, we consider the following two situations.

First, we consider the situation when the priority constraint is inactive. Since the priority constraint is inactive, the solution is the same with or without the priority constraint. Thus, Algorithm 1 reduces to Algorithm 2. Therefore, we arrive at the following.

$$v'_{i,j} = v_{i,j}. \tag{26}$$

On the other hand, when the priority constraint is active, Algorithm 1 will allocate some timeslots from low-priority UE groups to high-priority ones. Thus, the utility of a high-priority UE will increase after each iteration of the resource re-allocation phase, and we have the following.

$$v'_{i,j} \geq v_{i,j}. \tag{27}$$

Combining (26) and (27) and taking the expectation, we have the following.

$$\mathbb{E}'_{pri(a_{i,j})=1}[v_{i,j}] \geq \mathbb{E}_{pri(a_{i,j})=1}[v_{i,j}]. \tag{28}$$

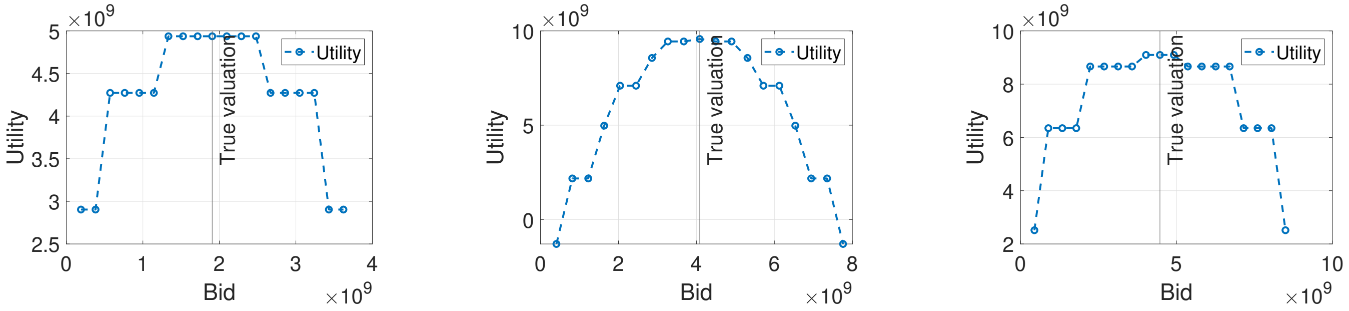
Moreover, we have (29), for otherwise, $\phi' > \phi$, contradicting Theorem 2.

$$\mathbb{E}'_{pri(a_{i,j})=0}[u_{i,j}] \leq \mathbb{E}_{pri(a_{i,j})=0}[u_{i,j}]. \tag{29}$$

Therefore, based on (24), (28), and (29), we can compare the expected utility of a high-priority UE and a low-priority one as follows, thus completing the proof.

$$\begin{aligned} \mathbb{E}'_{pri(a_{i,j})=1}[u_{i,j}] &\geq \mathbb{E}_{pri(a_{i,j})=1}[u_{i,j}] \\ &= \mathbb{E}_{pri(a_{i,j})=0}[u_{i,j}] \\ &\geq \mathbb{E}'_{pri(a_{i,j})=0}[u_{i,j}]. \end{aligned} \tag{30}$$

□



(a) No priorities.

(b) Two priorities.

(c) Six priorities.

Fig. 3. Incentive-compatibility: The UE's utility with respect to different bids in systems with different numbers of priorities. The UE's true valuation is plotted in each subfigure with a vertical black line. When the UE's bid equals its true valuation, the UE can get maximum utility. This means UEs are incentivized to report their valuations truthfully, validating the incentive-compatibility of the proposed mechanism. (a) No priorities. (b) Two priorities. (c) Six priorities.

TABLE II
PARAMETERS FOR SIMULATION

Notation	Meaning	Default value
N	Number of UE groups	32
r	Number of UEs	1000
d	Number of priorities	$\{1, 2, ..., 6\}$
μ	Numerology setting	3
	Subcarrier spacing	120 kHz
	Beam number	8
	Beam width	15°
k	Number of timeslots per frame	80
n_0	Noise power	-100 dBm
P	gNB transmit power	30 dBm
g_B	gNB antenna gain	15 dB
h_B	gNB antenna height	15 m
g_U	UE antenna gain	-7.35 dB
h_U	UE antenna height	1.5 m
B	Channel bandwidth	100 MHz
f_c	Carrier frequency	28 GHz
σ_{SF}	Standard deviation of Gaussian shadow fading	LOS: 4, NLOS: 6

VII. PERFORMANCE EVALUATION

A. Evaluation Methodology

In this section, we simulate the proposed resource allocation mechanism to demonstrate its properties numerically. We consider a network setting with a gNB, which has a coverage radius of 500 m. All the UEs are uniformly and randomly distributed within the coverage of the gNB, and the minimum distance between a UE and the gNB is set to be 10 m. We follow the propagation model in Section III-A, and the parameters for simulation are specified in Table II. We use the numerology set $\mu = 3$ in our evaluation [35], where the subcarrier spacing is 120 kHz. The beam number is 8 per sector, and the gNB uses three sectors to serve its coverage area. The width of each beam is 15°.

The valuation function of each UE $a_{i,j}$ is given below, where $SNR_{i,j}$ is $\frac{P \cdot g_B \cdot g_U \cdot PL(d_{i,j}) \cdot SF}{n_0}$, and $\sum_{p=1}^{t} U_c[0,1]^{(p)}$ captures the concavity of the valuation function.

$$v_{i,j}(t) = B \log_2(1 + SNR_{i,j}) \sum_{p=1}^{t} U_c[0,1]^{(p)}. \tag{31}$$

Note that we use the following notations.

- $U_c[a,b]$ denotes the continuous uniform distribution over $[a,b]$.
- $U_c[a,b]^{(p)}$ denotes the p-th largest element of k $U_c[a,b]$s.

If not specified explicitly, the simulation result of each experiment is averaged over 100 simulations to avoid the effect of randomness.

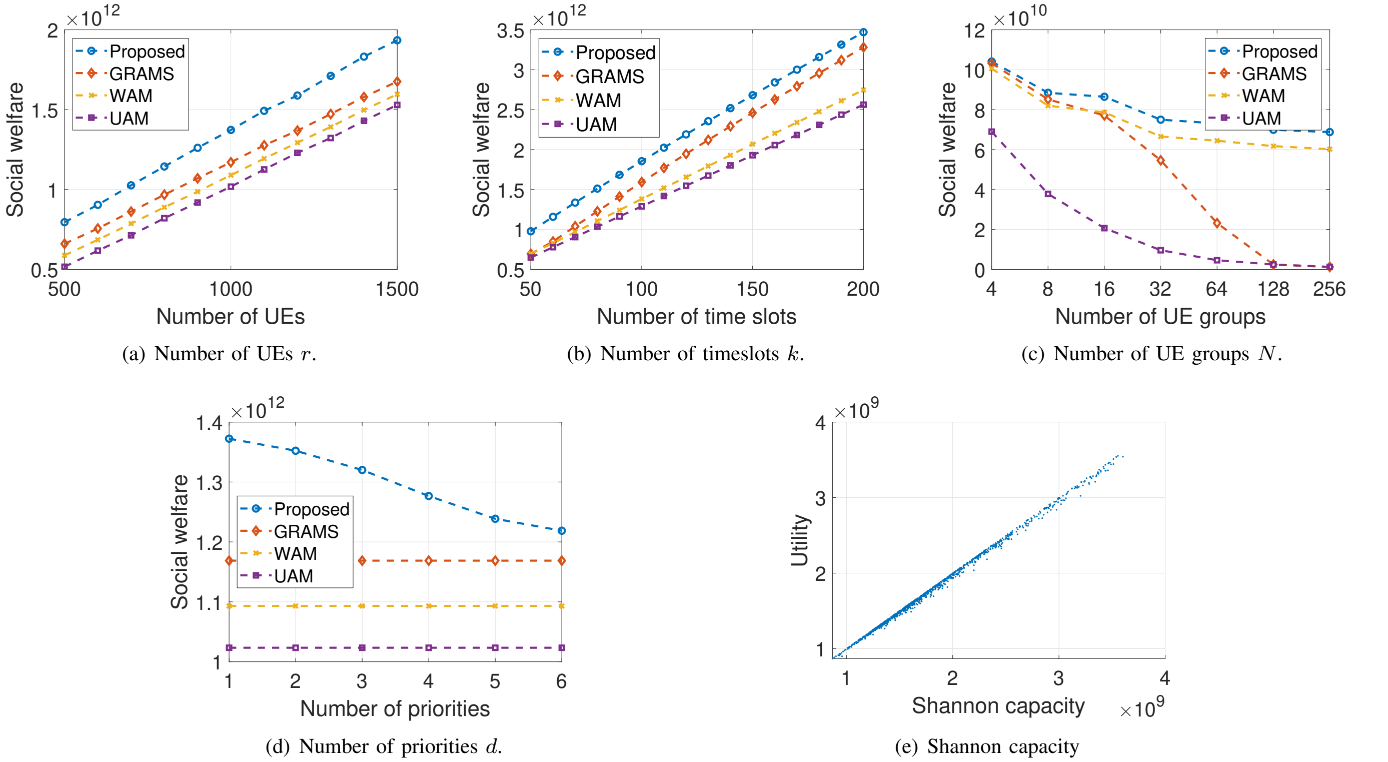
For the purpose of showing the advantages of our proposed mechanism, three following schemes are used for comparison.

- **GRAph-based Multicast Scheduling (GRAMS)**: The graph-based multicast scheduling mechanism is from [18]. This mechanism first allocates one timeslot to each UE group to ensure basic needs. For the remaining resources, it allocates the timeslots to maximize social welfare.
- **Weighted Allocation Mechanism (WAM)**: The weighted allocation mechanism takes valuation into account when distributing resources [41]. It allocates timeslots to each UE group proportional to the sum of the first elements of the valuation functions, i.e., $\sum_{j=1}^{m_i} c_{i,j}(1)$, in that direction.
- **Uniform Allocation Mechanism (UAM)**: The uniform allocation mechanism allocates resources from a fairness standpoint [42]. It is a per-group uniform allocation mechanism, allocating timeslots to all the UE groups in a round-robin fashion.

B. Incentive-Compatibility

In this subsection, we demonstrate incentive-compatibility of the proposed mechanism. Fig. 3(a), Fig. 3(b), and Fig. 3(c) are a UE's utility with respect to different bids when there are no priorities, two priorities, and six priorities, respectively. The number of UE groups N is 8, the priorities of different UE groups are randomly chosen, and other parameters are in Table II. We consider the bid function $B_{1,1}$ of a UE $a_{1,1}$ in UE group S_1 with different bids for the fifth timeslot $b_{1,1}(5)$, while other values are the same, i.e., $b_{1,1}(t) = v_{1,1}(t)$, $t \neq 5$.

The UE will get maximum utility if it bids truthfully, as indicated by the vertical black line in each subfigure. When the UE under-bids, it may get fewer timeslots, thus getting lower utility, as indicated by the region to the left of the vertical



(a) Number of UEs r.

(b) Number of timeslots k.

(c) Number of UE groups N.

(d) Number of priorities d.

(e) Shannon capacity

Fig. 4. Social welfare comparison with different parameters. (a) Different numbers of UEs. (b) Different numbers of timeslots. (c) Different numbers of UE groups. (d) Different numbers of priorities. (e) Different Shannon capacities.

black line. When the UE over-bids, it may get more timeslots. While the valuation of more timeslots is higher, it has to pay more money under the VCG pricing, thus getting lower utility, either. Thus, the UEs are motivated to report their valuation functions truthfully, as prescribed by Theorem 6.

C. Social Welfare Maximization

In this subsection, we present social welfare maximization of the proposed mechanism. In Fig. 4(a), Fig. 4(b), Fig. 4(c), and Fig. 4(d), we compare our proposed mechanism with GRAMS, WAM, and UAM under different scenarios.

Fig. 4(a) demonstrates social welfare comparison with different numbers of UEs r. With the growth of r, social welfare increases since there are more UEs. Also, the proposed mechanism achieves higher social welfare than the other three baselines. As demonstrated in Fig. 4(a), when the number of UEs increases, social welfare increases nearly linearly. Because we assume that each UE's valuation function follows the same distribution, which is reasonable as there are a lot of UEs, social welfare increases when more UEs are involved.

Fig. 4(b) demonstrates social welfare comparison with different numbers of total timeslots k. With the increase of k, social welfare increases since UEs can get more timeslots and more utility in each frame. It can be seen that the proposed mechanism achieves higher social welfare than the other allocation schemes because the proposed mechanism solves the optimization problem and always attains a social-welfare-maximization solution. Moreover, note that Fig. 4(a) and Fig. 4(b) follow similar trends in that both r and k are positively related to social welfare.

Fig. 4(c) explores social welfare comparison with different numbers of UE groups N. When N increases and other parameters remain the same, each UE group contains fewer MUs. Therefore, fewer MUs can get packets in each timeslot, thus reducing social welfare. With the increase of N, social welfare decreases rapidly initially but slows down when N is high. UAM achieves the lowest social welfare no matter the number of UE groups due to its inability to capture the UE distribution. On the other hand, with the increase of N, the proposed allocation mechanism achieves the highest social welfare because it yields an allocation profile by solving an optimization problem aiming at maximizing social welfare. While GRAMS and WAM can also capture the UE distribution, their allocation criteria are still heuristic. Therefore, compared to the proposed mechanism, which directly solves the optimization problem, GRAMS and WAM get lower social welfare.

Fig. 4(d) presents social welfare comparison with different numbers of priorities d. We set the relative number of UE groups with different priorities as an exponential relationship, i.e., the number of UE groups with i-priority is twice that with $(i+1)$-priority. On the other hand, the timeslot numbers of different subperiods are the same, i.e., $k_0 = k_1 = ... = k_{d-1}$. With the increase of d, more resources are concentrated on high-priority UEs while social welfare decreases slightly due to more constraints on the resource allocation optimization problem. Therefore, there is a trade-off between social welfare

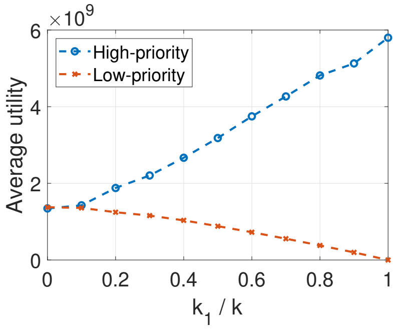


Fig. 5. Utility comparison of high-priority UEs and low-priority UEs with different high-priority percentages k_1/k.

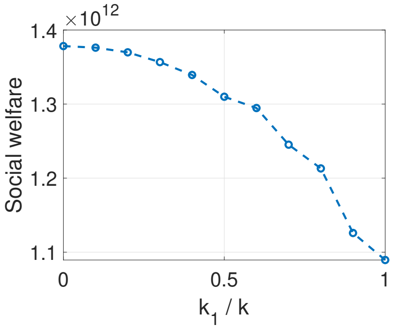


Fig. 6. Social welfare of the proposed mechanism with different k_1/k.

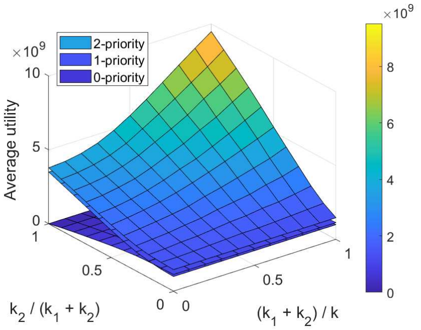


Fig. 7. Utility comparison of 2-priority UEs, 1-priority UEs, and 0-priority UEs with different k_1 and k_2.

and high-priority superiority. For other allocation schemes, social welfare remains unchanged because they do not consider priority when allocating resources. Conversely, the proposed mechanism considers UE priority. Under priority constraints, a small number of high-priority UEs acquire a higher number of timeslots. However, due to the non-increasing property of each UE's marginal valuation function, the high-priority UEs contribute less to social welfare when they have already obtained abundant beam resources, resulting in a decrease in social welfare when d increases. However, although social welfare declines with the rise of d, it still outperforms the other baselines because the proposed optimization algorithm suitably re-allocates the timeslots in a way that maximizes social welfare while satisfying priority constraints.

Finally, UEs' utility under different channel qualities is examined in Fig. 4(e). A UE's utility is approximately linear to its Shannon capacity. Therefore, when a UE has a higher Shannon capacity, it has higher utility. As shown in Fig. 4(e), the utility of each UE suitably reflects the channel quality.

D. Priority Comparison

In this subsection, we examine the impact of the percentage between the high-priority subperiod timeslot number k_1 and the low-priority subperiod timeslot number k_0 on the utility of UEs with different priorities. We consider a system with 32 UE groups and two priorities. The high-priority UEs account for about 20% of all the UEs. This percentage is similar to the Premium subscribers' percentage of Youtube [43]. Thus, we set 20% of the UE groups as high-priority consisting of high-priority UEs.

In Fig. 5, we compare the utility of UEs under different high-priority percentages k_1/k. As can be seen in Fig. 5, high-priority UEs can get higher utility under the proposed mechanism. Moreover, as k_1/k increases, the average utility of high-priority UEs increases. Since the percentage of high-priority UEs is 20%, the difference in utility between high-priority UEs and low-priority ones is not significant when k_1/k is less than 0.2. After that, high-priority UEs will get considerably higher utility than low-priority UEs do. The difference in utility can be set as the membership charge of the UEs. On the other hand, the social welfare of the proposed mechanism remains nearly the same when k_1/k is less than 0.2 and decreases slightly after that, as can be seen from Fig. 6. Combining Fig. 5 and Fig. 6, there is a trade-off between high-priority superiority and social welfare maximization. Moreover, k_1/k can be set as a number around the high-priority UE percentage. As such, the system ensures the superiority of the high-priority UEs while guaranteeing social welfare.

Furthermore, we explore the utility of different UEs in a three-priority system in Fig. 7. When high-priority timeslots increase, both 2-priority UEs and 1-priority UEs get higher utility. Moreover, when the timeslots allocated to 2-priority UEs increase and those allocated to 1-priority UEs decrease, the utility of 2-priority UEs goes up while that of the 1-priority UEs goes down. However, no matter the parameters, high-priority UEs always achieve a higher utility.

E. Latency Comparison

Besides utility comparison, we analyze UE latency and link reliability under a 3-priority system in Fig. 8 and Fig. 9. We assume each UE has a packet to receive, and the packet size is denoted as δ. The timeslot duration is set as $td = \frac{1}{8}$ ms. The throughput of each UE $a_{i,j}$ is $R_{i,j} = td \times B \log_2(1+SNR_{i,j})$. To capture the link reliability using the above parameters, we set a UE's valuation function as the following.

$$v_{i,j}(t) = \min(1, \log_2(1 + \frac{R_{i,j}t}{\delta})). \tag{32}$$

When the UE receives no timeslot, its valuation is 0. When the UE receives the entire packet successfully, meaning that $R_{i,j}t \geq \delta$, its valuation is 1. The valuation between obtaining no resources and a successful transmission is characterized by a concave log function that captures the decreasing marginal utility.

The gNB allocates the timeslots to different UE groups with priority consideration. When considering priority, a UE group with a higher priority gets beam resources earlier than another one with a lower priority. When there is a tie, we allocate the resources to the UE group with more UEs. The priority consideration dictates the serving sequence in each round, and the transmission is in a round-robin fashion between different rounds. So, the gNB will first serve each UE group once before serving any of them the second time, and this process

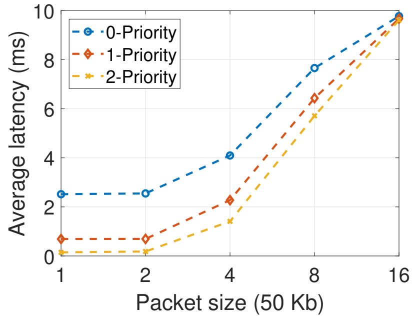


Fig. 8. Average latency of 2-priority UEs, 1-priority UEs, and 0-priority UEs.

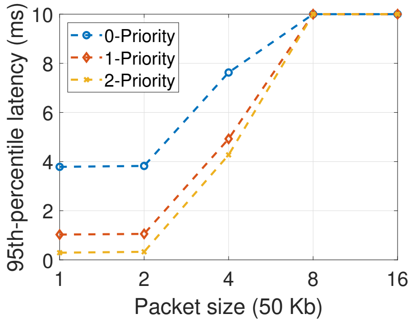
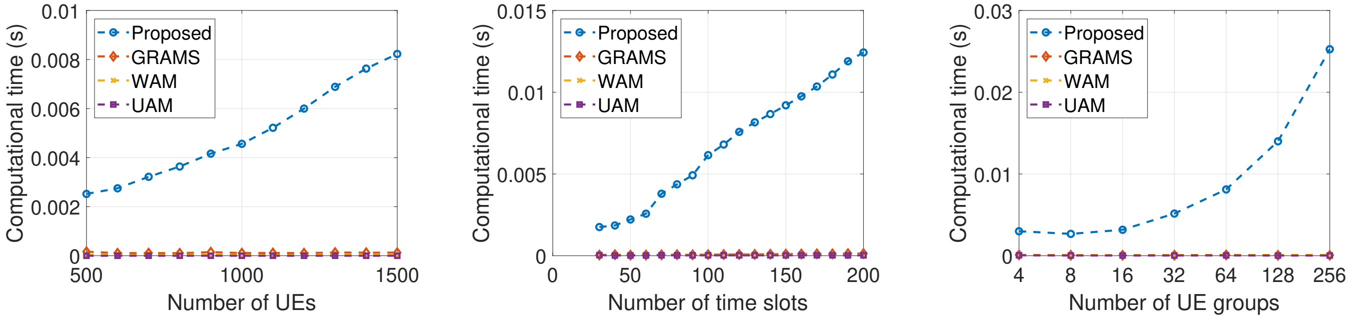


Fig. 9. 95th-percentile latency of 2-priority UEs, 1-priority UEs, and 0-priority UEs.



(a) Number of UEs r.

(b) Number of timeslots k.

(c) Number of UE groups N.

Fig. 10. Computational efficiency comparison with different parameters. (a) Different numbers of UEs. (b) Different numbers of timeslots. (c) Different numbers of UE groups.

continues. The latency is defined as the time at which the transmission of the packet is complete. If the transmission is not finished within the whole frame, the latency is set as $k \times td = 10$ ms.

Fig. 8 demonstrates the latency of three kinds of UEs. When the packet size is small, the gNB can finish the transmission in a timeslot. So, the latency is determined by the first time a UE receives the packet. When the packet size increases, a single timeslot is not sufficient to transmit a packet, so the gNB needs to utilize more timeslots for transmission. However, beam scheduling is a round-robin fashion. So, the gap between two serving times is huge, leading to increased latency. Finally, when the packet size is too large to transmit with a few timeslots, the gNB cannot transmit successfully, and all the UEs have a latency of 10 ms.

Fig. 9 explores the 95-th percentile latency of three kinds of UEs. The 95-th percentile latency indicates a threshold under which 95% of the UEs can finish the transmission, and this criterion is indicative of the link reliability. Fig. 9 follows the same trend as that in Fig. 8, but the 95-th percentile in Fig. 9 is always greater than the average latency in Fig. 8. When the packet size is small, the gNB can guarantee successful transmission within a short time. Thus, the reliability of links is higher in terms of the latency guarantee. However, the 95-th percentile latency reaches the upper limit of 10 ms even when the average latency is lower. This phenomenon means that while some UEs may have finished the transmission earlier due to higher throughput, the system only guarantees a moderate delay for the UEs with poor SNR. When the packet size is larger than 400 Kb, all 95-th percentile latency is limited to 10 ms, the length of a radio frame, because of the resource allocation cycle in our setting. If a UE's latency reaches the upper limit, it is regarded as an outage UE since it cannot successfully receive its packet in time.

As for the comparison between different UE priorities, Fig. 8 and Fig. 9 demonstrate that UEs with a higher priority have lower latency because they can be served first. Combining this result with that of Fig. 7, we observe that a higher priority has both utility and latency advantages. Also, the links are reliable when the packet size is small. When the packet size is large, the latency hugely depends on the throughput of each UE, and the links are not reliable.

F. Computational Efficiency

Finally, we demonstrate the computational efficiency of the proposed mechanism by some simulations. Fig. 10(a), Fig. 10(b), and Fig. 10(c) show the computational efficiency of the proposed mechanism with respect to different numbers of UEs r, timeslots k, and UE groups N, respectively. The results show that the computational time of the proposed mechanism is linear to r, k, and N, as indicated by Theorem 5. As for the benchmark schemes, the computational time stays almost the same. The reason for this phenomenon is that these schemes do not consider the whole UE valuations, so they do not traverse the UE groups to achieve a resource allocation outcome.

VIII. Conclusion

In this paper, we considered a resource allocation framework for a 5G mmWave MBS system with beamforming techniques. The goal was to devise a low-complexity and resource-efficient mechanism that allocates the resources to different UE groups. Unlike the previous work, we treated the UEs' valuations as their private information and formulated the optimization problem as an auction. Moreover, as content providers provide premium membership, we took UE priority into account and extended the system model to a multi-priority one.

With the modified VCG-based mechanism, UEs are incentivized to report their true valuations of the beam resources. In this way, the optimization algorithms correctly reflect the situations of the UEs, thus guaranteeing the social-welfare-maximization outcomes in the actual system. On the other hand, the proposed mechanism can provide higher utility and lower latency for high-priority UEs. Besides, our mechanism is individually rational, (weakly) budget-balanced, and computationally efficient. Finally, simulation results validated the superiority of the proposed mechanism under different scenarios and provided insight into the premium membership system.

Appendix A
Properties of the Pop-Top Algorithm

The pop-top algorithm can also be formulated as the sorting algorithm described in Design 4. These two algorithms are identical with different purposes. The pop-top algorithm has lower time complexity, so we leverage it in the resource allocation mechanism. Conversely, the sorting algorithm has a higher time complexity, but its sorting property is useful for the proofs in the following sections.

Definition 10 (Sorted marginal transformed bid profile)**.** *$\boldsymbol{D} = (\boldsymbol{D}(1), \boldsymbol{D}(2), ..., \boldsymbol{D}(Nk))$ is the sorted marginal transformed bid profile, where $\boldsymbol{D}(i) = \boldsymbol{C}^{(i)}$, and $\boldsymbol{C}^{(i)}$ is the ith largest element in $\boldsymbol{C}$.*

We denote the ith element of $\boldsymbol{D}$ by $\boldsymbol{D}(i)$ and $\{\boldsymbol{D}(i), \boldsymbol{D}(i+1), ..., \boldsymbol{D}(j)\}$ by $\boldsymbol{D}(i:j)$.

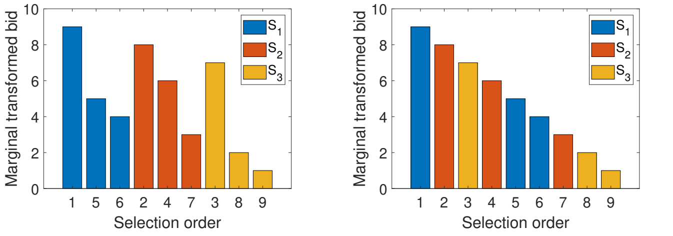
Design 4. *(Sorting algorithm) Sort $\boldsymbol{C}$ to get $\boldsymbol{D}$. A UE group S_i is allocated T_i timeslots, where $T_i = |C_i \cap \boldsymbol{D}(1:k)|$.*

Now, we prove the equivalence of the pop-top algorithm and the sorting algorithm.

Proposition 6. *The resource allocation profile $\boldsymbol{T}$ returned by the pop-top algorithm is equal to the resource allocation profile $\tilde{\boldsymbol{T}}$ returned by the sorting algorithm if the bid profile $\boldsymbol{B}$ is the same.*

Proof. This is proved in Appendix B □

Fig. 11 demonstrates Proposition 6. There are three UE groups, labeled by S_1, S_2, and S_3. The x-axis shows the selection order of different algorithms. Fig. 11(a) presents the allocation procedure of the pop-top algorithm. The pop-top algorithm allocates the first timeslot to S_1 because it has the largest marginal transformed bid. Then, the pop-top algorithm allocates the second timeslot to S_2 because it has the largest marginal transformed bid among the unselected bids. This process goes on until all the timeslots are fully allocated.



(a) Pop-top algorithm.

(b) Sorting algorithm.

Fig. 11. A simple example to demonstrate Proposition 6. (a) Pop-top algorithm. (b) Sorting algorithm.

Fig. 11(b) presents the allocation procedure of the sorting algorithm. The sorting algorithm first sorts all the marginal transformed bids. Then, it allocates the timeslots according to the order. Obviously, both the pop-top algorithm and the sorting algorithm return the same resource allocation profile, as demonstrated by the selection order in Fig. 11(a) and Fig. 11(b).

Appendix B
Proof of Proposition 6

First, we prove two properties of the marginal transformed bid function.

Lemma 1. *If $a_{i,j}$ bids truthfully ($B_{i,j} = V_{i,j}$), then $\sum_{t=0}^{T} c_{i,j}(t) = v_{i,j}(T)$.*

Proof. If $B_{i,j} = V_{i,j}$, then the bid function $B_{i,j}$ is concavely increasing. Hence, we have the following relationship for $t = 1, 2, ..., k$.

$$\begin{aligned} c_{i,j}(t) &= \underset{1\leq p\leq t}{min}\,(b_{i,j}(p) - b_{i,j}(p-1)) \\ &= b_{i,j}(t) - b_{i,j}(t-1). \end{aligned} \tag{33}$$

Summing up $c_{i,j}(t)$, and using $b_{i,j}(0) = 0$ and $c_{i,j}(0) = 0$, we arrive at the final result.

$$\begin{aligned} \sum_{t=0}^{T} c_{i,j}(t) &= b_{i,j}(T) - b_{i,j}(0) + c_{i,j}(0) \\ &= b_{i,j}(T) \\ &= v_{i,j}(T). \end{aligned} \tag{34}$$

□

Lemma 2. *The marginal transformed bid function C_i is non-increasing. That is, $c_i(t) \leq c_i(t-1)$, $t = 2, 3, ..., k$.*

Proof. First, we prove $c_{i,j}(t) \leq c_{i,j}(t-1)$ for $j = 1, 2, ..., m_i$.

$$\begin{aligned} c_{i,j}(t) &= \underset{1\leq p\leq t}{min}\,(b_{i,j}(p) - b_{i,j}(p-1)) \\ &\leq \underset{1\leq p\leq t-1}{min}\,(b_{i,j}(p) - b_{i,j}(p-1)) \\ &= c_{i,j}(t-1). \end{aligned} \tag{35}$$

Then, we prove the lemma.

$$c_i(t) = \sum_{j=1}^{m_i} c_{i,j}(t) \leq \sum_{j=1}^{m_i} c_{i,j}(t-1) = c_i(t-1). \tag{36}$$

□

Now, we prove the proposition.

Proof. We will prove this proposition by mathematical induction.

Claim: The resource allocation profile $\mathbf{T}$ returned by the pop-top algorithm is equal to the resource allocation profile $\tilde{\mathbf{T}}$ returned by the sorting algorithm.

Base Case: For $k = 1$, we are going to show that $\mathbf{T} = \tilde{\mathbf{T}}$. Assume that $T_i = 1$, and $T_p = 0$, $p \neq i$. Since the pop-top algorithm will give the only timeslot to the UE group with the largest marginal transformed bid at $t = 1$, we have $c_i(1) \geq c_p(1), p \neq i$. Since C_i is non-increasing by Lemma 2, we have $c_i(1) \geq c_p(q)$, $p = 1, 2, ..., N$, $q = 1, 2, ..., k$.

$c_i(1)$ is the largest element of $\mathbf{C}$, meaning that $c_i(1) = \mathbf{D}(1)$. Hence, the sorting algorithm will also give the only timeslot to S_i. Therefore, $\mathbf{T} = \tilde{\mathbf{T}}$ for $k = 1$, completing the base case.

Inductive Step: Assume that the claim holds for $k = t$, and we will prove that the claim holds for $k = t + 1$.

Assume that the $(t + 1)$th timeslot is allocated to S_i by Algorithm 2. Since Algorithm 2 gives the $(t + 1)$th timeslot to the UE group with the largest marginal transformed bid, we have $c_i(T_i(t) + 1) \geq c_p(T_p(t) + 1)$, $p \neq i$, where we denote $T_j(p)$ as the value of T_j after the pth iteration of Algorithm 2.

Since C_i is non-increasing, $c_i(T_i(t) + 1) \geq c_p(q)$, $p = 1, 2, ..., N$, $q = T_p(t)+1, ..., k$. Thus, $c_i(T_i(t)+1)$ is the largest element of $\mathbf{D}(t + 1 : Nk)$, meaning $c_i(T_i(t) + 1) = \mathbf{D}(t + 1)$. Hence, the sorting algorithm gives the $(t+1)$th timeslot to S_i. Thus, $\mathbf{T} = \tilde{\mathbf{T}}$ for $k = t+1$, completing the inductive step. □

Appendix C
Proof of Theorem 1

In the following, we analyze a UE's different bidding strategies and show that a UE cannot get higher utility by untruthfully reporting its valuation function. We consider a UE $a_{\alpha,\beta}$ in S_α to untruthfully report $V_{\alpha,\beta}$, and we denote the received number of timeslots and the utility of of $a_{\alpha,\beta}$ as $T^{'}_\alpha$ and $u^{'}_{\alpha,\beta}$, respectively. Also, T_α and $u_{\alpha,\beta}$ denote the received number of timeslots and the utility of of $a_{\alpha,\beta}$ when it truthfully reports $V_{\alpha,\beta}$. We will prove that $u^{'}_{\alpha,\beta} \leq u_{\alpha,\beta}$ no matter its reported valuation function.

Lemma 3. *If $T^{'}_\alpha = T_\alpha$, then $T^{'}_j = T_j, j \in \mathbb{Z}_N \setminus \alpha$.*

Proof. Since $T_\alpha = T^{'}_\alpha$ and $C_j = C^{'}_j, j \in \mathbb{Z}_N \setminus \alpha$, $\mathbf{D}(1 : k) \setminus C_\alpha = \mathbf{D}'(1 : k) \setminus C'_\alpha$. By the sorting algorithm, $T_j = |C_j \cap \mathbf{D}(1 : k)| = |C'_j \cap \mathbf{D}'(1 : k)| = T^{'}_j, j \in \mathbb{Z}_N \setminus \alpha$. □

Proposition 7. *If $T^{'}_\alpha = T_\alpha$, then $u^{'}_{\alpha,\beta} = u_{\alpha,\beta}$.*

Proof. The utility of a UE $a_{\alpha,\beta}$ when getting T timeslots is given by the following equation.

$$\begin{aligned} u_{\alpha,\beta}(T) &= v_{\alpha,\beta}(T) - p_{\alpha,\beta} \\ &= v_{\alpha,\beta}(T) - (W_{\mathbf{a}-a_{\alpha,\beta}}(\mathbf{B} - B_{\alpha,\beta}) - W_{\mathbf{a}-a_{\alpha,\beta}}(\mathbf{B})). \end{aligned} \tag{37}$$

$W_{\mathbf{a}-a_{\alpha,\beta}}(\mathbf{B} - B_{\alpha,\beta}) - W_{\mathbf{a}-a_{\alpha,\beta}}(\mathbf{B})$ is the price calculated by the VCG auction. $W_{\mathbf{a}-a_{\alpha,\beta}}(\mathbf{B} - B_{\alpha,\beta})$ is independent of UE $a_{\alpha,\beta}$. $W_{\mathbf{a}-a_{\alpha,\beta}}(\mathbf{B})$ is the same for the the UEs in S_α if T_α is the same. Also, if T_α does not change, the UEs not in S_α will get the same number of timeslots as before by Lemma 3. Hence, $W_{\mathbf{a}-a_{\alpha,\beta}}(\mathbf{B})$ is the same for the UEs not in S_α.

Moreover, $v_{\alpha,\beta}$ depends on T and $T^{'}_\alpha = T_\alpha$. Thus, all the terms in $u_{\alpha,\beta}$ won't change, meaning that $u^{'}_{\alpha,\beta} = u_{\alpha,\beta}$. □

Proposition 8. *If $T^{'}_\alpha > T_\alpha$, then $u^{'}_{\alpha,\beta} \leq u_{\alpha,\beta}$.*

Proof. The change of utility is given below.

$$\begin{aligned} & u^{'}_{\alpha,\beta} - u_{\alpha,\beta} \\ & = [v_{\alpha,\beta}(T^{'}_i) - (W_{\mathbf{a}-a_{\alpha,\beta}}(\mathbf{B}^{'} - B^{'}_{\alpha,\beta}) - W_{\mathbf{a}-a_{\alpha,\beta}}(\mathbf{B}^{'})))] \\ & - [v_{\alpha,\beta}(T_i) - (W_{\mathbf{a}-a_{\alpha,\beta}}(\mathbf{B} - B_{\alpha,\beta}) - W_{\mathbf{a}-a_{\alpha,\beta}}(\mathbf{B}))] \\ & = v_{\alpha,\beta}(T^{'}_i) - v_{\alpha,\beta}(T_i) + W_{\mathbf{a}-a_{\alpha,\beta}}(\mathbf{B}^{'}) - W_{\mathbf{a}-a_{\alpha,\beta}}(\mathbf{B}). \end{aligned} \tag{38}$$

Since S_α gets more timeslots than before, the other UE groups must get fewer timeslots than before or the same number of timeslots as before. $\mathbf{E} = \{e | e \in \mathbf{D}(1 : k), e \notin \mathbf{D}'(1 : k)\}$ denotes the set of such bids. Note that we have used Proposition 6.

Also, we have $e \geq c_i(t)$, $e \in \mathbf{E}$, $t = T_\alpha + 1, T_\alpha + 2, ..., T^{'}_\alpha$, for otherwise $e \in \mathbf{D}'(1 : k)$, contradicting the definition of $\mathbf{E}$.

Thus, we can write (38) as $\sum_{t=T_\alpha+1}^{T^{'}_\alpha} c_\alpha(t) - \sum_{e\in\mathbf{E}} e \leq 0$, so $u^{'}_{\alpha,\beta} \leq u_{\alpha,\beta}$. □

Proposition 9. *If $T^{'}_\alpha < T_\alpha$, then $u^{'}_{\alpha,\beta} \leq u_{\alpha,\beta}$.*

Proof. The change of utility is given below.

$$\begin{aligned} & u^{'}_{\alpha,\beta} - u_{\alpha,\beta} \\ & = [v_{\alpha,\beta}(T^{'}_i) - (W_{\mathbf{a}-a_{\alpha,\beta}}(\mathbf{B}^{'} - B^{'}_{\alpha,\beta}) - W_{\mathbf{a}-a_{\alpha,\beta}}(\mathbf{B}^{'})))] \\ & - [v_{\alpha,\beta}(T_i) - (W_{\mathbf{a}-a_{\alpha,\beta}}(\mathbf{B} - B_{\alpha,\beta}) - W_{\mathbf{a}-a_{\alpha,\beta}}(\mathbf{B}))] \\ & = v_{\alpha,\beta}(T^{'}_i) - v_{\alpha,\beta}(T_i) + W_{\mathbf{a}-a_{\alpha,\beta}}(\mathbf{B}^{'}) - W_{\mathbf{a}-a_{\alpha,\beta}}(\mathbf{B}). \end{aligned} \tag{39}$$

Since S_α gets fewer timeslots than before, the other UE groups must get more timeslots than before or the same number of timeslots as before. $\mathbf{E} = \{e | e \notin \mathbf{D}(1 : k), e \in \mathbf{D}'(1 : k)\}$ denotes the set of such bids. Note that we have used Proposition 6.

Also, we have $e \leq c_\alpha(t)$, $e \in \mathbf{E}$, $t = T^{'}_\alpha + 1, T^{'}_\alpha + 2, ..., T_\alpha$, for otherwise $e \in \mathbf{D}(1 : k)$, contradicting the definition of $\mathbf{E}$.

We can write (39) as $\sum_{e\in\mathbf{E}} e - \sum_{t=T^{'}_\alpha+1}^{T_\alpha} c_\alpha(t) \leq 0$.

Thus, $u^{'}_{\alpha,\beta} \leq u_{\alpha,\beta}$. □

Then, we have the proof of Theorem 1.

Proof. Theorem 1 is the direct result of Proposition 7, Proposition 8, and Proposition 9. □

Appendix D
Proof of Theorem 2

Proposition 10. *The pop-top algorithm maximizes the total transformed bid W.*

Proof. By the sorting algorithm, the pop-top algorithm allocates the timeslots to the k largest marginal transformed bids, thus maximizing the total transformed bid W. □

Then, we have the proof of Theorem 2.

Proof. By Theorem 1, the proposed mechanism is incentive-compatible. Therefore, social welfare is equal to the summation of all winners' marginal transformed bids, which is the total transformed bid. Also, by Proposition 10, the pop-top algorithm maximizes the total transformed bid. Therefore, the proposed mechanism maximizes social welfare ϕ. □

Appendix E
Proof of Theorem 3

Proof. If $a_{i,j}$ bids truthfully, $u_{i,j}$ can be written as follows.

$$u_{i,j} = W_{\mathbf{a}}(\mathbf{B}) - W_{\mathbf{a}-a_{i,j}}(\mathbf{B} - B_{i,j}). \tag{40}$$

We denote the sorted marginal transformed bid profile produced by $\mathbf{B} - B_{i,j}$ as $\mathbf{D}'$, and the sorted marginal transformed bid profile produced by $\mathbf{B}$ as $\mathbf{D}$. Since $\mathbf{B} - B_{i,j} \subset \mathbf{B}$, we have $u_{i,j} = \sum_{t=1}^{k} \mathbf{D}(t) - \sum_{t=1}^{k} \mathbf{D}'(t) \geq 0$. □

Appendix F
Proof of Theorem 4

Proof. The price $p_{i,j}$ paid by $a_{i,j}$ is as follows.

$$p_{i,j} = W_{\mathbf{a}-a_{i,j}}(\mathbf{B} - B_{i,j}) - W_{\mathbf{a}-a_{i,j}}(\mathbf{B}). \tag{41}$$

We denote the resource allocation profile produced by $\mathbf{B} - B_{i,j}$ as $\mathbf{T}'$, and the resource allocation profile produced by $\mathbf{B}$ as $\mathbf{T}$. If $W_{\mathbf{a}-a_{i,j}}(\mathbf{B} - B_{i,j}) < W_{\mathbf{a}-a_{i,j}}(\mathbf{B})$, $\mathbf{T}$ will give higher social welfare than $\mathbf{T}'$ when the bid profile is $\mathbf{B} - B_{i,j}$, contradicting with Theorem 2 that the proposed mechanism maximizes social welfare. Therefore, $W_{\mathbf{a}-a_{i,j}}(\mathbf{B} - B_{i,j}) \geq W_{\mathbf{a}-a_{i,j}}(\mathbf{B})$. Hence, $p_{i,j} = W_{\mathbf{a}-a_{i,j}}(\mathbf{B} - B_{i,j}) - W_{\mathbf{a}-a_{i,j}}(\mathbf{B}) \geq 0$, meaning that $\sum_{i=1}^{N} \sum_{j=1}^{m_i} p_{i,j} \geq 0$. □

Appendix G
Proof of Theorem 5

Proof. We denote the time complexity of bid transformation, pop-top algorithm, and VCG pricing by γ_1, γ_2, and γ_3, respectively.

Since the number of UEs is r and each UE will bid a k-tuple, $\gamma_1 = O(rk)$. For the pop-top algorithm, the second for loop has k iterations, and each iteration takes $O(N)$ to find the minimum marginal transformed bid by traversing all the UE groups. Thus, $\gamma_2 = O(Nk)$. The naive analysis will show that since the VCG pricing needs to run the bid transformation and the pop-top algorithm for every winning UE, $\gamma_3 = r(\gamma_1 + \gamma_2) = O(Nrk + r^2k)$. Thus, $\gamma_1 + \gamma_2 + \gamma_3 = O(Nrk + r^2k)$. However, some operations are redundant in the VCG pricing, so we can bound the time complexity more tightly.

Note that while the VCG pricing needs to go through the bid transformation and the pop-top algorithm for every winning UE, the bid transformation and the pop-top algorithm do not need to be performed repeatedly for every winning UE.

For the bid transformation of VCG pricing for a UE $a_{i,j}$, we only need to consider the bid transformation of S_i without $a_{i,j}$. Thus, the running time is not $O(rk)$ but $O(k)$ since only one UE's bid will change, and the total running time on the bid transformation is $O(rk)$.

Similarly, for the pop-top algorithm of VCG pricing for a UE $a_{i,j}$, we only need to re-allocate T_i timeslots since the other timeslots will not change. Hence, the running time is not $O(Nk)$ but $O(NT_i)$, and the total VCG pricing running time on the pop-top algorithm will be $O(Nk \times \max_i m_i)$ since $\sum_{i=1}^{N} T_i = k$. While the worst-case time complexity of $O(Nk \max_i m_i)$ is $O(Nrk)$, the time complexity will become $O(Nk \times \frac{r}{N}) = O(rk)$ if each UE group has the same number of UEs. Thus, $\gamma_3 = O(rk + Nrk) = O(Nrk)$, and the total time complexity $\gamma_1 + \gamma_2 + \gamma_3 = O(rk + Nk + Nrk) = O(Nrk)$, which is polynomial-time. □

References

[1] H. Shokri-Ghadikolaei, C. Fischione, G. Fodor, P. Popovski, and M. Zorzi, "Millimeter wave cellular networks: A MAC layer perspective," *IEEE Transactions on Communications*, vol. 63, no. 10, pp. 3437–3458, 2015, doi: 10.1109/TCOMM.2015.2456093.

[2] G. Sanfilippo, O. Galinina, S. Andreev, S. Pizzi, and G. Araniti, "A concise review of 5G new radio capabilities for directional access at mmWave frequencies," in *Internet of Things, Smart Spaces, and Next Generation Networks and Systems*, O. Galinina, S. Andreev, S. Balandin, and Y. Koucheryavy, Eds. Cham: Springer International Publishing, 2018, pp. 340–354, doi: 10.1007/978-3-030-01168-0_65.

[3] D. Wu, J. Wang, Y. Cai, and M. Guizani, "Millimeter-wave multimedia communications: challenges, methodology, and applications," *IEEE Communications Magazine*, vol. 53, no. 1, pp. 232–238, 2015, doi: 10.1109/MCOM.2015.7010539.

[4] A. V. Lopez, A. Chervyakov, G. Chance, S. Verma, and Y. Tang, "Opportunities and challenges of mmWave NR," *IEEE Wireless Communications*, vol. 26, no. 2, pp. 4–6, 2019, doi: 10.1109/MWC.2019.8700132.

[5] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2017–2022," Cisco Visual Networking Index (VNI) Forecast, White paper C11-738429-01, 2019. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

[6] 3GPP, "Architectural enhancements for 5G multicast-broadcast services," 3rd Generation Partnership Project (3GPP), Technical specification (TS) 23.247, Sep. 2021, version 17.0.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3854

[7] V. K. Shrivastava, S. Baek, and Y. Baek, "5G evolution for multicast and broadcast services in 3GPP release 17," *TechRxiv*, Jul. 2021, doi: 10.36227/techrxiv.14985354.v1.

[8] H.-H. Liu and H.-Y. Wei, "Towards NR MBMS: A flexible partitioning method for SFN areas," *IEEE Transactions on Broadcasting*, vol. 66, no. 2, pp. 416–427, 2020, doi: 10.1109/TBC.2020.2983847.

[9] L. Richter and U. H. Reimers, "A 5G new radio-based terrestrial broadcast mode: System design and field trial," *IEEE Transactions on Broadcasting*, vol. 68, no. 2, pp. 475–486, 2022.

[10] J. Montalban, P. Scopelliti, M. Fadda, E. Iradier, C. Desogus, P. Angueira, M. Murroni, and G. Araniti, "Multimedia Multicast Services in 5G Networks: Subgrouping and Non-Orthogonal Multiple Access Techniques," *IEEE Communications Magazine*, vol. 56, no. 3, pp. 91–95, 2018, doi: 10.1109/MCOM.2018.1700660.

[11] M. M. Mabkhot, A. M. Al-Ahmari, B. Salah, and H. Alkhalefah, "Requirements of the smart factory system: A survey and perspective," *Machines*, vol. 6, no. 2, 2018, doi: 10.3390/machines6020023.

[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, doi: 10.1109/MCOM.2019.1700679.

[13] P.-Y. Su, Y.-Y. Li, and H.-Y. Wei, "Strategy-proof beam-aware multicast resource allocation mechanism," in *2021 30th Wireless and Optical Communications Conference (WOCC)*, 2021, pp. 75–79, doi: 10.1109/WOCC53213.2021.9602884.

[14] S. Sen, J. Xiong, R. Ghosh, and R. R. Choudhury, "Link layer multicasting with smart antennas: No client left behind," in *2008 IEEE International Conference on Network Protocols*, 2008, pp. 53–62, doi: 10.1109/ICNP.2008.4697024.

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

[16] 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, doi: 10.1109/TNET.2012.2191977.

[17] 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, doi: 10.1109/TNET.2017.2717901.

[18] I.-S. Cho and S. J. Baek, “Optimal multicast scheduling for millimeter wave networks leveraging directionality and reflections,” in *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, 2021, pp. 1–10, doi: 10.1109/INFOCOM42981.2021.9488427.

[19] T. Bai and R. W. Heath, “Coverage and rate analysis for millimeter-wave cellular networks,” *IEEE Transactions on Wireless Communications*, vol. 14, no. 2, pp. 1100–1114, 2015, doi: 10.1109/TWC.2014.2364267.

[20] H. Park, S. Park, T. Song, and S. Pack, “An incremental multicast grouping scheme for mmWave networks with directional antennas,” *IEEE Communications Letters*, vol. 17, no. 3, pp. 616–619, 2013, doi: 10.1109/LCOMM.2013.011513.122519.

[21] A. Biason and M. Zorzi, “Multicast transmissions in directional mmWave communications,” in *European Wireless 2017; 23th European Wireless Conference*, 2017, pp. 1–7.

[22] 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, doi: 10.1109/TBC.2021.3061979.

[23] G. Araniti, P. Scopelliti, G.-M. Muntean, and A. Iera, “A hybrid unicast-multicast network selection for video deliveries in dense heterogeneous network environments,” *IEEE Transactions on Broadcasting*, vol. 65, no. 1, pp. 83–93, 2019, doi: 10.1109/TBC.2018.2822873.

[24] A. Samuylov, D. Moltchanov, R. Kovalchukov, R. Pirmagomedov, Y. Gaidamaka, S. Andreev, Y. Koucheryavy, and K. Samouylov, “Characterizing resource allocation trade-offs in 5G NR serving multicast and unicast traffic,” *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, pp. 3421–3434, 2020, doi: 10.1109/TWC.2020.2973375.

[25] L. Yang, J. Chen, Q. Ni, J. Shi, and X. Xue, “NOMA-enabled cooperative unicast–multicast: Design and outage analysis,” *IEEE Transactions on Wireless Communications*, vol. 16, no. 12, pp. 7870–7889, 2017, doi: 10.1109/TWC.2017.2754261.

[26] Y. Mao, B. Clerckx, and V. O. K. Li, “Rate-splitting for multi-antenna non-orthogonal unicast and multicast transmission: Spectral and energy efficiency analysis,” *IEEE Transactions on Communications*, vol. 67, no. 12, pp. 8754–8770, 2019, doi: 10.1109/TCOMM.2019.2943168.

[27] W. Huang, Y. Huang, S. He, and L. Yang, “Cloud and edge multicast beamforming for cache-enabled ultra-dense networks,” *IEEE Transactions on Vehicular Technology*, vol. 69, no. 3, pp. 3481–3485, 2020, doi: 10.1109/TVT.2020.2968466.

[28] X. Pei, H. Yu, Y. Chen, M. Wen, and G. Chen, “Hybrid multicast/unicast design in NOMA-based vehicular caching system,” *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 16 304–16 308, 2020, doi: 10.1109/TVT.2020.3041260.

[29] N. Chukhno, O. Chukhno, D. Moltchanov, A. Gaydamaka, A. Samuylov, A. Molinaro, Y. Koucheryavy, A. Iera, and G. Araniti, “The use of machine learning techniques for optimal multicasting in 5G NR systems,” *IEEE Transactions on Broadcasting*, pp. 1–14, 2022, doi: 10.1109/TBC.2022.3206595.

[30] E. Iradier, M. Fadda, M. Murroni, P. Scopelliti, G. Araniti, and J. Montalban, “Nonorthogonal multiple access and subgrouping for improved resource allocation in multicast 5G NR,” *IEEE Open Journal of the Communications Society*, vol. 3, pp. 543–556, 2022, doi: 10.1109/OJCOMS.2022.3161312.

[31] M. N. Dani, D. K. C. So, J. Tang, and Z. Ding, “Resource allocation for layered multicast video streaming in NOMA systems,” *IEEE Transactions on Vehicular Technology*, pp. 1–15, 2022, doi: 10.1109/TVT.2022.3193122.

[32] M. Zhang, H. Lu, F. Wu, and C. W. Chen, “NOMA-based scalable video multicast in mobile networks with statistical channels,” *IEEE Transactions on Mobile Computing*, vol. 20, no. 6, pp. 2238–2253, 2021, doi: 10.1109/TMC.2020.2977639.

[33] Z. Li, Q. Wang, and H. Zou, “QoE-aware video multicast mechanism in fiber-wireless access networks,” *IEEE Access*, vol. 7, pp. 123 098–123 106, 2019, doi: 10.1109/ACCESS.2019.2938422.

[34] C. Guo, Y. Cui, and Z. Liu, “Optimal multicast of tiled 360 VR video,” *IEEE Wireless Communications Letters*, vol. 8, no. 1, pp. 145–148, 2019, doi: 10.1109/LWC.2018.2864151.

[35] 3GPP, “NR; physical channels and modulation,” 3rd Generation Partnership Project (3GPP), Technical specification (TS) 38.211, Sep. 2022, version 17.3.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3213

[36] ——, “Evolved universal terrestrial radio access (E-UTRA); study on LTE-based 5G terrestrial broadcast,” 3rd Generation Partnership Project (3GPP), Technical report (TR) 36.776, Mar. 2019, version 16.0.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3500

[37] ——, “Study on channel model for frequencies from 0.5 to 100 GHz,” 3rd Generation Partnership Project (3GPP), Technical report (TR) 38.901, Feb. 2022, version 16.1.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3173

[38] J. Blömer, M. Kalfane, R. Karp, M. Karpinski, M. Luby, and D. Zuckerman, “An XOR-based erasure-resilient coding scheme,” 1995.

[39] W. Lin, D. Chiu, and Y. Lee, “Erasure code replication revisited,” in *Proceedings. Fourth International Conference on Peer-to-Peer Computing, 2004. Proceedings.*, 2004, pp. 90–97, doi: 10.1109/PTP.2004.1334935.

[40] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004, doi: 10.1017/CBO9780511804441.

[41] H. Zheng, H. Li, S. Hou, and Z. Song, “Joint resource allocation with weighted max-min fairness for NOMA-enabled V2X communications,” *IEEE Access*, vol. 6, pp. 65 449–65 462, 2018, doi: 10.1109/ACCESS.2018.2877199.

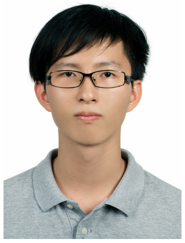
[42] D. T. Ngo, C. Tellambura, and H. H. Nguyen, “Efficient resource allocation for OFDMA multicast systems with spectrum-sharing control,” *IEEE Transactions on Vehicular Technology*, vol. 58, no. 9, pp. 4878–4889, 2009, doi: 10.1109/10.1109/TVT.2009.2027331.

[43] L. Ceci, “Global YouTube premium subscribers 2024,” Sep. 2021. [Online]. Available: https://www.statista.com/statistics/1261865/youtube-premium-subscribers/

Pan-Yang Su (Member, IEEE) received the B.S. degree in Electrical Engineering from National Taiwan University, Taipei, Taiwan, in 2022. He is currently pursuing the Ph.D. degree with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, USA.

In 2021, he received the Outstanding Performance Award for research in wireless multicast from National Taiwan University, and his research about quantum game theory was sponsored by the Ministry of Science and Technology. He is currently a Phi Tau Phi Scholastic Honor Society Member.

He is broadly interested in game theory and mechanism design, especially their intersection with algorithms, optimization, and quantum information theory. His research spans wireless communication, federated learning, dynamic tolling, nonlocal games, etc.



Kuang-Hsun Lin (Member, IEEE) received the B.S. degree in Electrical Engineering from National Taiwan University, Taipei, Taiwan, in 2015. He received the Ph.D. in Communication Engineering from National Taiwan University, GICE, Taipei in 2022. Since 2015, he has been working with Wireless Mobile Networking Laboratory led by Professor Hung-Yu Wei. He held summer internships at Mediatek in the summer of 2015 and 2018. His research interests include wireless mobile networks, devices/network power saving mechanisms, mobility management, and other MAC protocol design. He is working as a graduate research assistant for Taiwan's MOST 6G program.

Yi-Yun Li received the B.S. degree in Electrical Engineering from National Taiwan University, Taipei, Taiwan, in 2020, where he is currently pursuing the Ph.D. degree with the Graduate Institute of Communication Engineering. He has been working with Wireless Mobile Networking Laboratory led by Professor Hung-Yu Wei.

Hung-Yu Wei (Senior Member, IEEE) received the B.S. degree in electrical engineering from National Taiwan University in 1999 and the M.S. and Ph.D. degrees in electrical engineering from Columbia University in 2001 and 2005, respectively.

He is a Professor with the Department of Electrical Engineering and Graduate Institute of Communications Engineering, National Taiwan University, where he currently serves as an Associate Chair with the Department of Electrical Engineering. He was a summer intern with Telcordia Applied Research in 2000 and 2001. He was with NEC Labs America from 2003 to 2005. In July 2005, he joined the Department of Electrical Engineering, National Taiwan University. His research interests include next-generation wireless broadband networks, IoT, fog/edge computing, cross-layer design for wireless multimedia, and game-theoretic models for communications networks. He received NTU Excellent Teaching Award in 2008 and 2018. He also received "Recruiting Outstanding Young Scholar Award" from the Foundation for the Advancement of Outstanding Scholarship in 2006, the K. T. Li Young Researcher Award from ACM Taipei/Taiwan Chapter and The Institute of Information and Computing Machinery in 2012, the Excellent Young Engineer Award from the Chinese Institute of Electrical Engineering in 2014, the Wu Ta You Memorial Award from MOST in 2015, and the Outstanding Research Award from MOST in 2020. He has been actively participating in NGMN, IEEE 802.16, 3GPP, IEEE P1934, and IEEE P1935 standardization. He serves as the Vice-Chair of IEEE P1934 Working Group to standardize fog computing and networking architecture. He serves as the Secretary for IEEE ComSoC Fog/Edge Industry Community. He is an Associate Editor for IEEE SYSTEM JOURNAL and *IEEE Internet of Things Magazine* and was an Associate Editor for IEEE INTERNET OF THINGS JOURNAL. He is an IEEE-certified Wireless Communications Professional. He was the Chair of the IEEE VTS Taipei Chapter from 2016 to 2017. He serves as a Program Co-Coordinator for Taiwan's MOST 6G program. He is currently the Chair of the IEEE P1935 working group for edge/fog management and orchestration standard.