Reliability Enhancement for Multimedia Delivery in Caching-Assisted MmWave HetNets

Fangfang Yin, An Wang, Yu Zhang, Danpu Liu, and Libiao Jin

***Abstract*—Explosive growth of bandwidth-intensive multimedia applications may impose a heavy traffic burden on wireless links of heterogeneous networks (HetNets), which also requires substantial energy consumption. Edge caching is emerging to reduce the service latency and relieve the huge burden on the backhaul links in HetNets. Moreover, millimeter wave (mmWave) spectrum with huge available bandwidth has been regarded as a promising technology to satisfy the high capacity demands in future HetNets. However, the high directionality coupled with harsh propagation environment makes mmWave communication vulnerable to blockage, which makes link establishment and maintenance challenging. In this paper, we propose a joint maximum distance separable (MDS) coded caching and transmission problem for a mmWave HetNet, aiming at minimizing the backhaul energy consumption while satisfying users' quality of service (QoS) requirements on high data rate. The formulated mixed integer non-linear programming (MINLP) problem is solved by decomposing it into two stages of subproblems. In the first stage, the convex optimization technique is applied to tackle the coded caching subproblem. Subsequently, we propose two efficient algorithms based on matching in the second stage to deal with the transmission subproblem. Simulation results demonstrate that the proposed schemes can effectively maximize the backhaul energy saving and enhance the download rate compared with the benchmark scheme.**

***Index Terms*—Coded caching, matching, mmWave, multi-beam concurrent transmissions.**

I. Introduction

With the rapid proliferation of high data rate multimedia applications, future wireless networks are required to cope with the explosive traffic surge. MmWave spectrum with huge frequency bands (e.g., 30 GHz to 300 GHz) has been regarded as a key enabling technology to overcome the bandwidth shortage at traditional microwave band (i.e., μW or sub-6 GHz), and achieve multi-Gbps data rata of future communication systems [1]. Dense deployment of mmWave small cell base stations (SBSs) in the existing sub-6 GHz macrocell base station (MBS) that forms mmWave-μW heterogeneous networks (HetNets), will boost the capacity and provide high data rates by decreasing user-base station (BS) distance and improving spectral efficiency. However, the backhauling of huge amount of data traffic has become one of the main challenges in mmWave HetNets especially when the BSs are densely deployed. Firstly, the massive implementation of traditional wired backhauling may become infeasible due to the expensive costs and hard-to-reach location of all BSs. Secondly, though wireless backhauls were proposed as the alternative approach that enables low-cost connection between BSs, the wireless backhaul solutions for the BSs have not been widely adopted due to the spectrum shortage at μW. Thirdly, the backhaul link with the limited capacity makes constraint on the content transmission, introduces extra power consumption and degrades the quality of service (QoS) for users [2]. Considering those issues, it is necessary to analyze the performance of the HetNets by taking the backhaul cost into account.

Maximum distance separable (MDS) coded caching technique at the mmWave SBSs is a good candidate to reduce the backhaul traffic as well as energy [3]. Additionally, since the requests can be served directly by a cluster of mmWave SBSs without visiting the MBS or a core network, MDS coded caching can assist content transmissions to achieve large throughput and small download delay. In general, users with high mobility can only download a small fraction of the requested video, and thus need several connections with BSs to recover the entire video. Fortunately, coded caching allows videos to be splitted into encoded packets, which are cached at different mmWave SBSs, so that a certain number of randomly encoded packets will be sufficient to recover the file. That is to say, coded caching which does not require to cache the entire content from the video library, provides high-mobility users with more flexility. With the overlay cache space of multiple SBSs, MDS coded caching can be used to provide additional robustness for multimedia delivery in mobility scenario.

In cache-enabled mmWave HetNets, the design of the MDS coded caching policy and the transmission strategy are closely coupled. On the one hand, the MDS coded caching strategy should be able to cater for the specific transmission strategy.

Manuscript received August 29, 2022 revised February 20, 2023; approved for publication by Sangheon Pack, Division 2 Editor, April 1, 2023.

This work is supported in part by National Natural Science Foundation of China: No. 61971069, and 62271065, Beijing Natural Science Foundation: No. L202003, The Open Research Project of the State Key Laboratory of Media Convergence and Communication, Communication University of China: No. SKLMCC2021KF009, and The Fundamental Research Project of Science and Technology on Complex Electronic System Simulation Laboratory: DXZT-JC-ZZ-2020-011, and the Project funded by China Postdoctoral Science Foundation No. 2021M702987.

F. Yin and L. Jin are with the State Key Laboratory of Media Convergence and Communication, School of Information and Communication, Communication University of China, Beijing 100024, China, email: {yinff, libiao}@cuc.edu.cn.

A. Wang and D. Liu are with Beijing Laboratory of Advanced Information Network, Beijing Key Laboratory of Network System Architecture and Convergence, Beijing University of Posts and Telecommunications, Beijing 100876, China, email: {BY_Wa2021, dpliu}@bupt.edu.cn.

Y. Zhang is with the State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China, with the Beijing Key Laboratory of Mobile Computing and Pervasive Device, Beijing, 100080, China, and also with University of Chinese Academy of Sciences, Beijing 100049, China, email: zhangyu@ict.ac.cn.

L. Jin is the corresponding author.

Digital Object Identifier: 10.23919/JCN.2023.000015

1229-2370/23/$10.00 © 2023 KICS

Due to the limited caching capacity, SBSs can only afford caching part of the contents. How to take full advantage of the limited cache space to assist transmissions to enhance the network throughput and reliability in mmWave systems is vital crucial. On the other hand, given a caching scenario, the download rate depends on the communication links between mmWave SBSs and users. That is to say, given the cache state, the transmission parameters of mmWave SBSs determine the download rate of users, which directly affects the backhauls. Thus, it is essential to design an effective transmission strategy that adaptive to the MDS coded caching strategy to achieve high data rate of access links. Towards this end, a solution that combines MDS coded caching with transmission strategy is urgently needed in mmWave HetNets.

Despite the wide available spectrum of mmWaves, there are still many key technical challenges that need to be addressed [4]. Since mmWave communication works at the high frequency band which yields the severe penetration attenuation, directional antennas have to be adopted for both mmWave SBSs and users to overcome the severe propagation loss and achieve high data rate. However, the single transmission link between the mmWave SBS and the user is sensitive to the blockage due to the user mobility and the motion of objects in the surroundings. Moreover, the short transmission range of mmWave communication leads to short connection durations for mobile users, which imply that high mobility users can only retrieve a small fraction of a video file from a cache enabled mmWave SBS. In order to be free of interruptions of handover or link failures in mmWave systems, multi-beam concurrent transmissions, which enable users to be served by multiple mmWave SBSs simultaneously, have shown great potentials. The novelty in multi-beam concurrent transmissions is that when blockage happens, backup links can be established to restore connectivilty, which is critical to gurantee the reliability and robustness of multimedia delivery [5].

A. Ralated Works

To tackle the interruptions of handover or link failures in mmWave systems, multi-beam transmissions have attracted much attention to improve the throughput and reliability [6]–[9]. Authors in [6] considered the multi-beam transmissions problem to maximize the uplink sum rate for heterogeneous multibeam cloud radio access network. Authors in [7] utilized multi-lobe beams to identify several channel clusters for combating blockages in mmWave systems. In [8], with μW and mmWave dualband cooperation, authors proposed a novel multi-beam transmission mechanism to improve the throughput and reliability in WLAN architecture. In addition, the authors in [9] proposed a novel non-orthogonal multiple access (NOMA)-based multi-beam transmission strategy for unmanned aerial vehicles (UAVs) over μW band. Although the works in [6]–[9] have studied the use of multi-beam transmissions, the coded caching has not been considered for the system design.

Currently, the joint optimization of coded caching and content transmission policy in cache-enabled networks has been widely studied in previous works [10]–[16]. In [10], authors considered the joint proactively cache and resource allocation strategy at the mobile edge computing server to reduce the service delay and users' energy cost. Authors in [11] jointly optimized the coded caching and multicast transmissions for satellite-UAV-vehicle networks, aims at reducing the backhaul traffic. Besides, Y. Fu *et al.* in [12] investigated the joint coded caching and resource allocation problem to reduce the system latency in terms of the content delivery latency and backhaul transmission delay for NOMA networks. To increase the content diversity and allocate the best caching node for users with high QoS requirements, authors in [13] studied the joint caching placement and coordinated multipoint (CoMP) scheme for UAV-aided cellular networks. In order to improving the quality of experince (QoE) of users, [14] proposed a resource allocation strategy based on multipath cooperative scalable video transmission for 5G HetNets. In [15], authors proposed a joint mode selection, power allocation and user pairing scheme to enhance the performance of a hybrid scheme (i.e., either NOMA or coded multicasting). [16] proposed a reliable user association and robust resource allocation algorithm to improve the energy efficiency. However, none of these works study the joint transmission and coded caching problem for mmWave systems. Of relevance to our work is [17] where authors focused on the optimization problem of uncoded cache placement, BS clustering, and multicast beamforming to minimize the total power consumption. Specifically, authors in [17] considered satellite networks as the complement to terrestrial networks (i.e., works at traditional μW band), which can serve users with broad service coverage and provide users with high-speed data services. In this paper, mmWave spectrum with huge bandwidth resources is considered as the complement to the bandwidth shortage at traditional μW band, and thus significantly enhance the capacity of terrestrial HetNets. In addition to that, authors in [17] mainly investigated the static environment without considering the network dynamics. In fact, the high mobility can significantly affect the content delivery performance since it may result in additional delays, including handovers, propagation or retransmission delays. Inspired by the contributions in the previous works [10]–[17], we would like to propose a solution that combines multi-beam concurrent transmission with the coded caching, which has a significant effect on the continuity of multimedia delivery in a highly dynamic environment.

B. Contributions

Motivated by the above concerns, we formulate an optimization problem by jointly considering MDS coded caching and multi-beam concurrent transmissions in mmWave systems. To the best of our knowledge, no existing work addresses the above problem in mmWave-μW HetNets. Our main contributions can be summarized as follows.

- Multi-beam concurrent transmissions combine with MDS code caching are introduced to provide seamless data transmissions and robustness in mmWave HetNets. We formulate the joint coded caching and multi-beam transimission problem to maximize the backhaul energy saving of the mmWave-μW HetNets, under the constraints

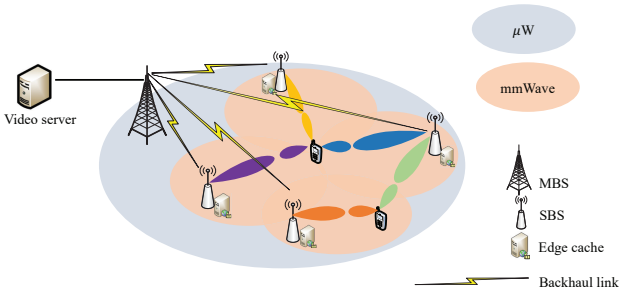


Fig. 1. System model of cache enabled mmWave-μW HetNets.

of limited storage capacities and wireless resources, as well as the users' diverse QoS requirements.

- To solve the formulated problem, we propose efficient approaches with the aid of matching theory and convex optimization technology. At first, considering the cache parameters and the storage capacity, we apply the convex optimization technology to find the optimal coded caching placement. Subsequently, based on the optimal caching placement, we propose two many-to-many matching-based multi-beam concurrent transmission algorithms for the content delivery subproblem.
- We present numerical results to illustrate the performance of the proposed algorithms by using the parameters of transmission and caching. The backhaul energy savings of the proposed algorithms are much higher than the traditional algorithms. Meanwhile, the mmWave throughput of the HetNets is enhanced by the optimization of the multi-beam concurrent transmissions.

C. Outline

The rest of this paper is organized as follows. The system model and problem formulation are described in Section II. In Section III, the original problem is divided into two subproblems, where convex optimization technique is discussed to solve the coded caching subproblem. We present two many-to-many matching-based transmission strategies in Section IV. Section V shows the simulation results, and conclusions are finally drawn in Section VI.

II. System Model and Problem Formulation

In our analysis, we firstly introduce the system model of cache enabled mmWave-μW HetNets, and the joint MDS coded caching and transmission problem is formulated.

A. System Model

As illustrated in Fig. 1, we consider a mmWave-μW HetNet consisting of a μW MBS and K mmWave SBSs, which are interconnected via backhaul links. The μW MBS has access to the video server, each mmWave SBS that equipped with a cache device can be viewed as a relay and has a limited capacity of $C_k^{\max}$ bits. Let $\mathcal{U}$ denote the set of U users with $\mathcal{U} = \{1, 2, \cdots, U\}$, $\mathcal{K}$ denote the set of K SBSs with $\mathcal{K} = \{1, 2, \cdots, K\}$. We assume that both mmWave SBSs and users are equipped with multiple antennas, then each SBS

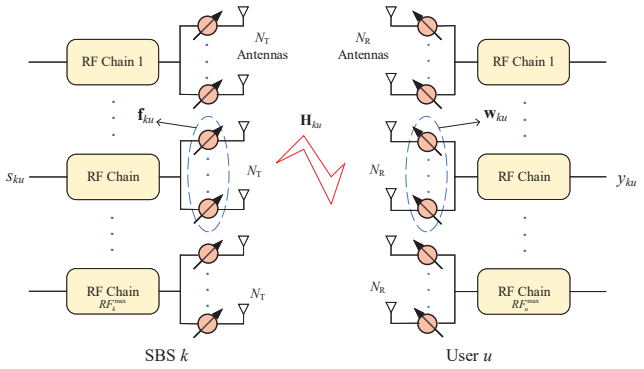


Fig. 2. Partially-connected beamforming architecture.

can establish directional beams with multiple users, and vice versa. $RF_k^{\max}$ and $RF_u^{\max}$ denote, respectively, the number of RF chains for each SBS and user. Note that we assume each mobile user can be served by multiple mmWave SBSs for high-data-rate service. As Fig. 1 shows, mmWave SBSs establish multi-beam pair links (BPLs) with each user to achieve higher data rate to gurantee the transmission reliability.

1) Communication model: We consider the partially-connected beamforming structure similar as [18], [19] in Fig. 2, where each RF chain connects with partial antennas through the phase shifter. As shown in Fig. 2, each RF chain of mmWave SBS k is connected to N_{T} antennas, and each RF chain of user u is connected to N_{R} antennas. In other words, each mmWave SBS k equipped with $RF_k^{\max} N_{\mathrm{T}}$ antennas and $RF_k^{\max}$ RF chains commnicates a user equipped with $RF_u^{\max} N_{\mathrm{R}}$ antennas.

Define the pair (z, v) as the transmitting beam of mmWave SBS z directed to user v, and set $\Omega = \{(z, v) \,|z \in \mathcal{K}_v^c, v \in \mathcal{U}\}$ as the transmitting beams of all mmWave SBSs, where $\mathcal{K}_v^c$ represents the mmWave SBS set corresponding to concurrent beams of user v. s_{zv}, $\mathbf{f}_{zv}$ and $\mathbf{w}_{zv}$ denote, respectively, the symbol that mmWave SBS z sends to v, the transmitting beamforming vector of mmWave SBS z directed to v, and the receiving beamforming vector of user v directed to z. $P_{zv} = \mathbb{E}\left(|s_{zv}|^2\right)$ represents the transmitting power for the beam of mmWave SBS z directed to the user v.

For the BPL (k, u), the receiving symbol at user u from mmWave SBS k can be written as

$$y_{ku} = \mathbf{w}_{ku}^H \mathbf{H}_{ku} \mathbf{f}_{ku} s_{ku} + \underbrace{\sum_{(z,v)\in\Omega\setminus(k,u)} \mathbf{w}_{ku}^H \mathbf{H}_{zu} \mathbf{f}_{zv} s_{zv}}_{\text{interference}} + \mathbf{w}_{ku}^H \mathbf{n}_{ku}, \tag{1}$$

where $(\cdot)^H$ represents conjugate transpose, $\mathbf{H}_{ku}$ is the $N_{\mathrm{R}} \times N_{\mathrm{T}}$ channel matrix between mmWave SBS k and user u, and $\mathbf{n}_{ku} \sim \mathcal{CN}\left(0, \sigma_{ku}^2 \mathbf{I}_{N_{\mathrm{R}}}\right)$ is the $N_{\mathrm{R}} \times 1$ Guassian noise. σ_{ku}^2 means the noise power.

Due to the limited scattering at the mmWave band, $\mathbf{H}_{ku}$ is depicted as a sum of $N_{\mathrm{cl}}+1$ scattering clusters, each of which contributes N_{ray} propagation paths, and can be expressed as [20], [21]

$$\mathbf{H}_{ku} = \sqrt{\frac{N_{\mathrm{T}} N_{\mathrm{R}}}{(N_{\mathrm{cl}}+1)N_{\mathrm{ray}}}} \sum_{i=0}^{N_{\mathrm{cl}}} \sum_{l=1}^{N_{\mathrm{ray}}} \alpha_{il} \mathbf{a}_{ku,\mathrm{R}} \left(\phi_{ku,\mathrm{R}}^{il}, \theta_{ku,\mathrm{R}}^{il}\right) \times \mathbf{a}_{ku,\mathrm{T}} \left(\phi_{ku,\mathrm{T}}^{il}, \theta_{u,\mathrm{T}}^{il}\right)^{H}, \tag{2}$$

where N_{cl} denotes the number of clusters corresponding to the non line of sight (NLOS) paths, $i = 0$ denotes the cluster corresponding to the line of sight (LOS) path, α_{il} is the complex gain, $\phi_{ku,\mathrm{T}}^{il}$ ($\theta_{ku,\mathrm{T}}^{il}$) is the azimuth (elevation) angle of departure, and $\phi_{ku,\mathrm{R}}^{il}$ ($\theta_{ku,\mathrm{R}}^{il}$) is the azimuth (elevation) angle of arrival of the lth path in the ith cluster, respectively. $\mathbf{a}_{ku,\mathrm{T}}$ and $\mathbf{a}_{ku,\mathrm{R}}$ are the array response vectors corresponding to the angles of departure and arrival. With regard to the uniform planar array (UPA), the array response vector can be expressed as

$$\mathbf{a}_{\mathrm{UPA}}(\phi,\theta) = \frac{1}{\sqrt{N_y N_z}} \Big[1, \cdots, e^{j\frac{2\pi}{\lambda} d(m \sin\phi \sin\theta + n\cos\theta)}, \cdots, e^{j\frac{2\pi}{\lambda} d((N_y-1)\sin\phi\sin\theta + (N_z-1)\cos\theta)}\Big]^{T}, \tag{3}$$

where λ is the wavelength, d is the inter-antenna distance, $0 \le m < N_y$ and $0 \le n < N_z$ are the y–axis and z-axis indices of the antenna element [22], [23].

Thus, the SINR of the BPL between mmWave SBS k and user u can be expressed as

$$SINR_{ku}(t) = \frac{P_{ku}\left|\mathbf{w}_{ku}^{H}\mathbf{H}_{ku}\mathbf{f}_{ku}\right|^2}{\sum\limits_{(z,v)\in\Omega\setminus(k,u)} P_{zv}\left|\mathbf{w}_{ku}^{H}\mathbf{H}_{zu}\mathbf{f}_{zv}\right|^2 + \sigma_{ku}^2}. \tag{4}$$

The achievable rate between SBS k and user u at time t can be defined by

$$R_{ku}(t) = W\log_2(1 + SINR_{ku}(t)). \tag{5}$$

Define $x_{ku}(t)$ as the binary beam pair selection indicator, i.e., $x_{ku}(t) = 1$ if the BPL between mmWave SBS k and user u is selected for transmission in the tth time slot, and otherwise $x_{ku}(t) = 0$. Then, by using multi-beam concurrent transmissions, the download rate of the user u can be denoted as

$$R_u^m(t) = \sum_{k\in K} x_{ku}(t) R_{ku}(t). \tag{6}$$

2) Caching model: Recall that users may have diverse requirements, e.g., videos, messages, or different sub-maps in HD maps. For the sake of generality, we consider a multimedia library $\mathcal{F} \triangleq \{F_1, F_2, \cdots, F_N\}$ consisting of N distinct files with equal size of B bits. The popularity for the file is denoted as p_j, satisfying $0 \le p_j \le 1$ and $\sum_{j=1}^{N} p_j = 1$. As adopted in most previous works [24], [25], suppose that the file popularity follows Zipf distribution with a shape parameter α that shapes the steepness of the popularity distribution. Specifically, the popularity of the file can be defined as $p_j = j^{-\alpha}/\sum_{j=1}^{N} j^{-\alpha}$.

The MDS encoded caching strategy is adopted similarly to the exsiting works [26], [27], i.e., each mmWave SBS k caches fraction q_{jk} ($0 \le q_{jk} \le 1$) encoded packets of each file F_j. Therefore, a file can be recovered when B bits are collected from any mmWave SBSs. If the requested video content is not fully retrieved from the associated mmWave SBSs, the missing part of the file content must be transported from the μW MBS via backhaul links, which also requires backhaul energy cost. In this paper, how to reap the limited storage space and RF chains of mmWave SBSs to pose less backhaul traffic on the μW MBS is the emphasis. Based on the above analysis, by associating users with the mmWave SBSs that caches the requested video content, the backhaul energy saving of the network can be expressed as

$$E_{bh,\mathrm{saving}} = e_{MBS} \sum_{k\in K} \sum_{u\in U} \sum_{j\in N} x_{ku}(t)\, p_j B q_{jk}, \tag{7}$$

where e_{MBS} (J/bit) denotes the backhaul energy consumption factor per bit transmitted, $\sum\limits_{j\in N}\sum\limits_{k\in K}\sum\limits_{u\in U} x_{ku}(t)\, p_j B q_{jk}$ represents the amount of coded data (in bits) cached at mmWave SBSs, i.e., the backhaul traffic saving of the μW MBS.

B. Problem Formulation

In this paper, our objective is to find an optimal cache fraction of files in mmWave SBSs, and multi-beam pair selection between mmWave SBSs and users, which maximizes the backhaul energy saving of the network. The joint caching and transmission problem $\mathcal{P}$ is formulated as follows:

$$\mathcal{P}: \max_{q_{jk}, x_{ku}} E_{bh,\mathrm{saving}} \tag{8}$$

$$\text{s.t.} \quad R_u^m(t) \ge R_u, \forall u \in \mathcal{U} \tag{8a}$$

$$x_{ku}(t) \in \{0,1\}, \forall u, k \tag{8b}$$

$$\sum_{k\in K} x_{ku} \le RF_u^{\max}, \forall u \in \mathcal{U} \tag{8c}$$

$$\sum_{u\in U} x_{ku} \le RF_k^{\max}, \forall k \in \mathcal{K} \tag{8d}$$

$$B\sum_{j\in N} q_{jk} \le C_k^{\max}, \forall k \in \mathcal{K} \tag{8e}$$

$$0 \le q_{jk} \le 1, \forall j, k, \tag{8f}$$

where (8a) guarantees that the QoS requirement for each user. The constraint (8b) keeps the beam pair selection indicators x_{ku} binary. Constraints (8c) and (8d) limit, respectively, the RF chains of each user and mmWave SBS. That is to say, each user can be served by maximum $RF_u^{\max}$ mmWave SBSs, and each mmWave SBS can associate with maximum $RF_k^{\max}$ users. Constraint (8e) restricts the cache capacity of the mmWave SBS k, and (8f) limits the encoded fraction of the video F_j at the cached SBS k. It is worth noting that some redundant strategies are considered to ensure the transmission reliability for high-mobility users: i) Through the MDS coded caching, high-mobility users actually have potential to be served by multiple mmWave SBSs with the accumulated cache size for high-data-rate service. ii) Applying multi-beam concurrent transmissions to mmWave systems makes it possible to reduce both service latency and service disruptions in a highly dynamic environment. In summary, both the MDS codes and enough BPLs at mmWave SBSs guarantee content downloading and retrieving, which afford the transmission reliability for high-mobility users.

III. Proposed Joint Optimization Algorithm

Given that the original problem $\mathcal{P}$ is a non-convex mixed integer non-linear programming (MINLP) problem, it is very difficult to find an optimization algorithm to make problem $\mathcal{P}$ converge to an optimal solution within polynomial time. Consequently, we decompose problem $\mathcal{P}$ into the following two sub-problems, and then solve them with the aid of convex optimization and matching theory.

A. Optimal Coded Caching Placement

In this subsection, the caching subproblem of what and through which mmWave SBS the coded video will be transmitted requires to be solved. In fact, the timescale for caching updating period (e.g., several hours) is apparently much longer than that of content transmissions. It can be observed from (4) that for any given feasible variable x_{ku}, the backhaul energy saving mainly depends on the cache placement varible Bq_{jk}. In another word, for the given multi-beam transmission strategy X^*, the problem $\mathcal{P}$ can be reformulated to determine the fraction (i.e., q_{jk}) of the video cached at mmWave SBSs. Here, X^* with element x_{ku} indicates the multi-association matric. Thus, in this case, the optimizing problem is transformed into the caching problem as follows.

$$\mathcal{P}_{\mathcal{C}}\left(\mathbf{Q} \mid \mathbf{X}^*\right): \max_{q_{jk}} E_{bh,\text{saving}} \tag{9}$$

$$\text{s.t.} \quad (8e)(8f) \tag{9a}$$

It should be noted that (8e) and (8f) in (9a) are linear. Moreover, it can be observed from (9) that the objective function of problem $\mathcal{P}_{\mathcal{C}}$ is a convex function of q_{jk} under the given X^*. Hence, problem $\mathcal{P}_{\mathcal{C}}$ is convex with respect to the optimization variables q_{jk}, which can be optimally solved by the available software packages such as CVX [28].

B. Multi-beam Concurrent Transmissions

Under the optimized caching variable q_{jk}^*, the essential problem now is how to optimize the transmission parameters, which determine how to transmit the requested videos to users in a much shorter timescale, e.g., one time frame. Thus, the multi-beam pair selection mechanism is designed, in which directional beams from multiple mmWave SBSs can serve the user concurrently. Thus, after obtaining the optimal caching variable q_{jk}^*, the original problem $\mathcal{P}$ can be equivalently transformed into

$$\mathcal{P}_{\mathcal{D}}\left(\mathbf{X}_t \mid \mathbf{Q}^*\right): \max_{x_{ku}} E_{bh,\text{saving}} \tag{10}$$

$$\text{s.t.} \quad (8a)(8b)(8c)(8d), \tag{10a}$$

where (8a)–(8d) in (10a) restrict the communication parameters for mmWave SBSs and mobile users as mentioned earlier. The formulated problem $\mathcal{P}_{\mathcal{D}}$ is also one of Karp's 21 NP-complete problems and is usually intractable [29]. It must be noticed that the formulation proposed in $\mathcal{P}_{\mathcal{D}}$ is a hard problem due to its combinatorial nature as a result of the presence of the binary variables x_{ku}. Therefore, obtaining the optimal solution using brute-force exhaustive search will be very challenging especially for a large number of users and SBSs. Moreover, as our systems are capable to support multi-beam concurrent transmissions, obtaining the optimal solution becomes more challenging. Therefore, the valid tool based on matching theory [30] can be used to solve it as described in the following section.

IV. Algorithms for Multi-beam Concurrent Transmissions

In this section, we develop the multi-beam concurrent transmission strategy based on many-to-many matching which provides polynomial time solutions for resource allocation problems [31]–[35].

Definition 1: A many-to-many matching Φ is defined by the mapping relations of two disjoint sets, i.e., SBS beams and user beams such that:

1) $\Phi(k) \subseteq \mathcal{U}$ and $\Phi(u) \subseteq \mathcal{K}$;
2) $|\Phi(k)| \leq RF_k^{\max}, \forall k \in \mathcal{K}$;
3) $|\Phi(u)| \leq RF_u^{\max}, \forall u \in \mathcal{U}$;
4) $k \in \Phi(u) \Leftrightarrow u \in \Phi(k)$,

where $\Phi(u)$ and $\Phi(k)$ denote, respectively, the set of partners for the mobile user u and the set of partners for the mmWave SBS k under the mapping state Φ. Condition 1) indicates that users and mmWave SBSs are matched with partners in the set $\mathcal{U}$ and $\mathcal{K}$. Conditions 2) and 3) set the quota constraints (i.e., RF chains) of each user and mmWave SBS, respectively, corresponding to (8c) and (8d). To be specific, mmWave SBS k has a maximum quota $RF_k^{\max}$ indicating that mmWave SBS k can serve $RF_k^{\max}$ users. In turn, each user has a quota $RF_u^{\max}$ such that user u can be associated with at most $RF_u^{\max}$ mmWave SBSs with the multi-association capability. Condition 4) denotes that the establishment of a map relationship Φ is successful if and only if $\Phi(u) = k$ and $\Phi(k) = u$.

To produce the BPLs that lead to maximum backhaul energy savings, participants in the matching game - namely, mmWave SBS beam and user beam - will determine the utility function towards each other such that the utilities are calculated and used to select the set of players. Pr_u and Pr_k denote, respectively, the sets of preference values for users and mmWave SBSs. Besides, to compare the preferences, a binary preference relation "$\succ$" is utilized, which is reflexive, transitive and complete [31]–[35].

Definition 2: For the user u, the expression $k \succ_u k' \Leftrightarrow Pr_u(k) > Pr_u(k')$ indicates that user u prefers to be associated with mmWave SBS k rather than k'. Here, $k \neq k'$, $k \in \mathcal{K}$, $k' \in \mathcal{K}$. Similarly, $u \succ_k u' \Leftrightarrow Pr_k(u) > Pr_k(u')$ implies that mmWave SBS k prefers to serve user u rather than user u'.

In order to maximize the utility, each user (mmWave SBS) tends to calculate the preferences over the mmWave SBS (user) by ranking the SBS set depending on their utilities. For each user, it concerns about the download rate from associated mmWave SBSs. The utility function of the user u when it associates with mmWave SBSs over the BPLs can be represented as

$$V_u(k) = R_u^m. \tag{11}$$

Algorithm 1 Greedy-based multi-beam concurrrent transmission algorithm (DM)

Step 1 : **Initialization**

1): Each user discovers SBSs and gets the channel status information (CSI).

2): Each user/SBS constructs their own preference list based on $V_u(k)$ and $V_k(u)$.

Step 2 : **Matching process**

while quota $q_u < RF_u^{\max}$ **do**

For each user $u \in \mathcal{U}$:

1): Each user $u \in \mathcal{U}$ selects the most preferred mmWave SBS and creates the proposed matrix.

For each proposed mmWave SBS $k \in \mathcal{K}$:

2): **If** quota $q_k > RF_k^{\max}$, mmWave SBS k rejects the proposed user.

3): **else** $q_k < RF_k^{\max}$, **do**

4): **if** more than one user proposes to the same mmWave SBS $k \in \mathcal{K}$, **do**

5): The mmWave SBS k calculates the utility function of each user.

6): The mmWave SBS k selects the most preferred user u based on the utility function and rejects other users.

7): **else** only one user proposes to the mmWave SBS $k \in \mathcal{K}$, the mmWave SBS select the user, $q_u = q_u + 1$, $q_k = q_k + 1$.

8): **end if**

9): **end if**

10): MmWave SBS k accepts the proposal and rejects other users, the rejected users remove the mmWave SBS k from their preference lists.

11): The matching end until all users are matched with $RF_u^{\max}$ SBSs.

Step 3 : End of algorithm

The multi-beam concurrent transmission strategy is done.

From the mmWave SBS's perspective, it concerns about the backhaul energy saving in the content transmission process. The utility function of mmWave SBS k can be defined as

$$V_k(u) = E_{bh,\text{saving}}. \tag{12}$$

According to (4) and (6), the utility of user u depends not only on the mmWave SBS it matched with, but also on the set of users that matched to the same mmWave SBS. In another word, multi-beam concurrent transmission problem reformulated as many-to-many matching game has *peer effects* (i.e., *externalities*), which is more challenging to find the stable matching than the conventional deferred-acceptance algorithm [30]. Considering the NP-hardness of the problem $\mathcal{P}_\mathcal{D}$, we firstly propose a greedy-based many-to-many matching algorithm (DM). The main idea is to start from the users with larger download rate (i.e., R_u^m) and preferentially select the users that result in more backhaul energy savings.

As shown in Algorithm 1, we first sort the preference lists of users and mmWave SBSs. At each round, each mmWave SBS k receives requests from users that rank k in their preferences.

Algorithm 2 Swap matching for multi-beam concurrent transmission algorithm (SM)

Phase I : **DM algorithm-based initialization**

1): Obtain the matching state based on DM algorithm, set $E_{\max} = E_{\text{total}}(\Phi)$;

Phase II : **Swap matching evaluation**

1): $\forall u \in \mathcal{U}$, it searches for another user $u' \in \{\mathcal{U} \backslash u, \mathcal{H}\}$, where $\mathcal{H}$ is a "hole" of available mmWave SBSs;

2): Execute swap matching $\Phi_u^{u'}$

If $E_{\text{total}}(\Phi_u^{u'}) > E_{\max}$

then $E_{\max} = E_{\text{total}}(\Phi_u^{u'})$

else User u keeps the current matching state.

3): Repeat 1) and 2) until there doesn't exist swap blocking pair (u, u').

Phase III : End of algorithm

Based on the proposal of the users, each mmWave SBS chooses its most preferred user in its preferences until the RF chains of the SBS (i.e., $RF_k^{\max}$) are all occupied by the users. If the number of proposed users exceed the RF chains of the SBS, the SBS will reject the less favorite users. Meanwhile, rejected users remove the mmWave SBS from their preference lists. Finally, the matching process ends until all users are matched with maximum $RF_u^{\max}$ mmWave SBSs.

Considering the matching approach of Algorithm 1 may yield lower utilities. In the following, motivated by the many-to-one housing assignment problem in [34], we also introduce the definition of *swap matching* into our many-to-many matching model.

Definition 3 : Given a matching Φ with $u \in \Phi(k)$, $u \notin \Phi(k')$, and $u' \in \Phi(k')$, $u' \notin \Phi(k)$, a *swap matching* can be defined as $\Phi_u^{u'} = \Phi \backslash \{(k,u),(k',u')\} \cup \{(k,u'),(k',u)\}$, where $u' \in \Phi_u^{u'}(k), u \in \Phi_u^{u'}(k')$ and $u \notin \Phi_u^{u'}(k), u' \notin \Phi_u^{u'}(k')$.

A *swap matching* allows users u and u' to swap one of their parterners (i.e., mmWave SBSs), while keeping other users' matchings unchanged [31]–[35]. Accordingly, the concept of "*swap blocking pair*" is defined as follows.

Definition 4 : A user pair (u, u') is a *swap blocking pair* if and only if a) $\forall x \in \{k, k', u, u'\}, V_x(\Phi_u^{u'}) \geq V_x(\Phi)$, and b) $\exists x \in \{k, k', u, u'\}, V_x(\Phi_u^{u'}) > V_x(\Phi)$.

The definition 4 represents that the swap matching $\Phi_u^{u'}$ is approved, and (u, u') is called a *swap blocking pair* in Φ. The condition a) represents that utilities of involved two-side players should not be decreased after the swap process between (u, u'). The condition b) implies that at least one of the two-side players' utility is improved after the swap process between *swap blocking pairs*.

Inspired by the previous works [31]–[35], we propose a swap matching-based multi-beam pair selection algorithm (SM), which is described detailedly in Algorithm 2. The proposed SM algorithm consists of three main phases: Phase I initialize the matching state; Phase II excutes the swap matching procedure between users and mmWave SBSs; Phase III outputs the final matching state. Specifically, in Phase I, users and mmWave SBSs initializes the matching state via the greedy matching algorithm. In the Phase II, each user

keeps searching for all the other users and the available hole of mmWave SBSs to check whether there exsits a *swap blocking pair*. Correspondingly, users and mmWave SBSs update their utilities. The operations of swap matching end when there does not exist *swap blocking pair* (u, u'). Subsequently, the final matching state between mmWave SBSs and users is obtained.

Definition 5: A matching Φ is *two-sided exchange-stable (2ES)* [34] if there does not exsit a *swap-blocking pair*.

Note that the notion of 2ES is different from the traditional pairwise stability, but one that is relevant to our model where mmWave SBSs and users can compare notes with each other.

Theorem 1: The final matching Φ_{rmtotal} of the proposed SM algorithm is 2ES.

Proof: By swap matching, we find that the backhaul energy saving of the network increases after each approved swap operation. Specifically, after searching for all possible swap operations, the swap-matching phase terminates and there does not exist any swap-blocking pair to further improve the backhaul energy saving of the current matching. Therefore, it can be concluded that Φ_{rmtotal} is 2ES. □

In the SM algorithm, SBSs and users try to find *swap blocking pairs*, and number of potential swap operations between any two players is $\binom{RF_k^{\max}}{2}$. In addition to that, each mmWave SBS also searches for a "open spot" to form the *swap blocking pair*, and the number of potential swap operations is $RF_k^{\max} RF_u^{\max}$. During each swap process, the complexity caused by the quik sort ordering operations is $\mathcal{O}\left(RF_k^{\max} RF_u^{\max} \log_2\left(RF_u^{\max}\right)\right)$ for (11) and $\mathcal{O}\left(RF_k^{\max} RF_u^{\max} \log_2\left(RF_k^{\max}\right)\right)$for (12). Thus, the overall complexity of the SM algorithm can be calculated by $\mathcal{O}\left(\left(\binom{RF_k^{\max}}{2} + \varsigma\right)\varsigma \log_2(\varsigma)\right)$, where $\varsigma = RF_k^{\max} RF_u^{\max}$.

In the DM algorithm, with the (10), each user needs to have a sorted list of mmWave SBSs, and the complexity is $\mathcal{O}\left(U \log_2(K)\right)$, Moreover, since each user proposes to every mmWave SBS in its list, it introduces a complexity of $\mathcal{O}(UK)$. On the other hand, since each mmWave SBS can accept a finite number of users, i.e., $RF_k^{\max}$, mmWave SBSs need to maintain their sorted user list, which introduces a complexity of $\mathcal{O}\left(K \log_2\left(RF_k^{\max}\right)\right)$. Therefore, the complexity of the proposed DM algorithm is $\mathcal{O}\left(U \log_2(K) + UK \log_2\left(RF_k^{\max}\right)\right)$.

For the exhaustive search method, it should be noted that the optimal result of exhaustive search algorithm is exhaustively searched over all possible results in the optimization function (9) and computed for each user at the SBSs. Specifically, the number of possible results of (9) for one user is $\left(2^K - 1\right)$ and subsequently the number of all possible states for all U users is $\left(2^K - 1\right)^U$. For each of these results, the exhaustive search method computes (7) which consists of $U(KU + 3)$ operations. Accordingly, the overall complexity of the exhaustive search method is $\mathcal{O}_{es} = \left(2^K - 1\right)^U U(KU + 3)$. Obviously, the computing complexity of the proposed SM algorithm (may attain a local optimal solution [35]) is significantly lower than that of the exhaustive searching method.

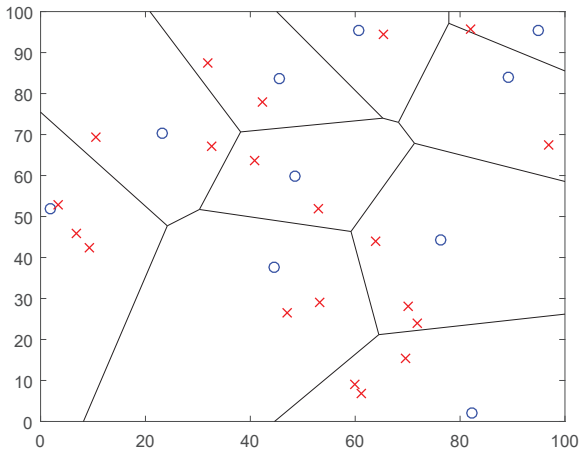


Fig. 3. Users (red cross) and mmWave SBS (blue circles) randomly distributed in the μW MBS coverage area.

V. Simulation Results

In this section, numerical results are provided to evaluate the performance of the proposed algorithms based on MATLAB. An illustration of the considered mmWave HetNet is shown in Fig. 2. We consider a 100 m $\times$ 100 m square area with one μW MBS located at the center, multiple mmWave SBSs and users randomly distributed within the MBS coverage area. The rate of energy consumption for the transmissions from MBS (i.e., e_{MBS}) is $0.5 * 10^{-8}$ (J/bit) [36], and the maximum power of the SBS is set to 35 dBm. The mmWave channel is constructed according to (2) and (3). The complex gain $\alpha_{i\ell}$ is generated according to a complex Gaussian distribution $\alpha_{i\ell} \sim \mathcal{CN}\left(0, 10^{-0.1\kappa}\right)$, where κ is the free space path loss [21]. For the LOS path, $\kappa_{\mathrm{LOS}} = 32.4 + 17.3\log_{10}(d_{\mathrm{TR}}) + 20\log_{10}(f_c)$, and for the NLOS path, $\kappa_{\mathrm{NLOS}} = \max\left(\kappa_{\mathrm{LOS}}, 38.3\log_{10}\left(d_{\mathrm{TR}}\right) + 17.30 + 24.9\log_{10}\left(f_c\right)\right)$ [37], where d_{TR} is the distance between the transmitter and the receiver, and $f_c = 60$ GHz is the carrier center frequency. The inter-antenna distance d is assumed to be half-wavelength. The bandwidth W and the noise power spectral density are set to 100 MHz and -174 dBm/Hz, respectively. Additionally, we set $RF_k^{\max} = RF_u^{\max} = 3$ for tractability. The QoS requirements of users are over [10, 100] Mbps [38]. Besides, suppose the cache capacity of each mmWave SBS is the same, which sets to 15% of the entire file set. The performance of proposed algorithms is evaluated with comparisons of the other three algorithms, including max-SINR (IM), best channel gain (GM) and min-distance (DM) algorithms. For fair comparison, these three multi-beam pair selection algorithms also have completed matching stage to ensure that all users are associated with multi-mmWave SBSs as its serving SBSs.

Since the exhausive search is of high complexity in our scenario, simulation with small networks size is possible. Here, we would like to analyze the gap between our local optimal solution and the optimal solution (i.e., exhaustive search method) in Fig. 4. To be specific, we consider a mmWave-μW HetNet consisting of 5 mmWave SBSs randomly distributed in the MBS coverage area, and each user can be associated with $RF_u^{\max} = 3$ mmWave SBSs. Fig. 4 plots the backhaul energy saving versus different number of

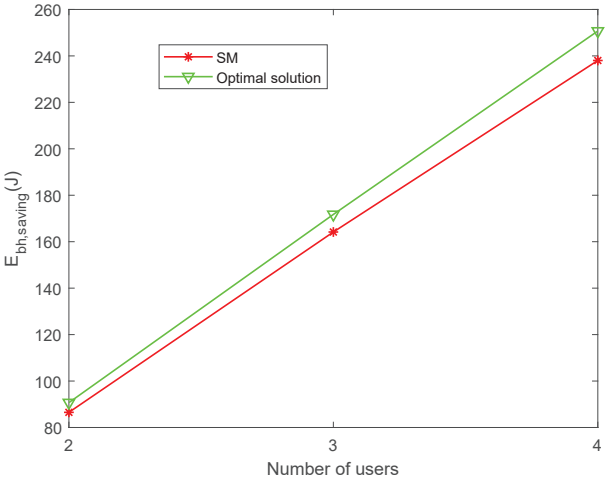


Fig. 4. Backhaul energy savings versus number of users.

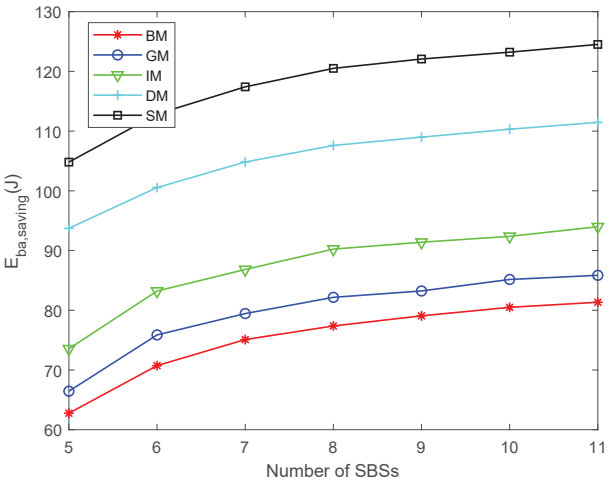


Fig. 5. Backhaul energy savings versus number of SBSs, where $U = 3$.

users (e.g., $U = 2; 3; 4$). We can observe that the backhaul energy saving increases with the number of users increasing. It is noted from Fig. 4 that the proposed SM algorithm with swap matching model catches up with the optimum solution (i.e., exhausive method). As shown in Fig. 4, when the number of users U = 2, 3, and 4, the proposed SM algorithm can reach 95.59%, 95.38%, and 94.94% of the exhaustive optimal result, unequivocally substantiating the plausibility of the proposed SM algorithm. In another word, compared with the optimal solution by exhaustive method with total complexity of $\mathcal{O}_{es} = \left(2^K - 1\right)^U U\left(KU + 3\right)$, which increases exponentially over the number of users and SBSs, the proposed SM algorithm is quite simple and effective. In addition to that, searching for optimal scheme imposes extremely high complexity as well as signaling overhead, and thus limits its practical applications.

In Fig. 5, the performance of the proposed methods under different number of mmWave SBSs is investigated. Obviously, with the increasing of the SBSs, the proposed SM and DM algorithms yield significant performance gains relative to the IM, GM, and BM-based matching algorithms. When the

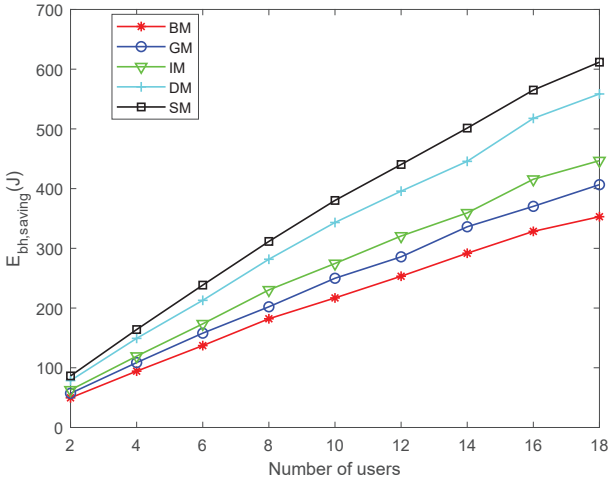


Fig. 6. Backhaul energy saving versus number of users, where $K = 20$.

number of SBSs $K = 5$, the backhaul energy savings of our proposed SM and DM algorithms are almost 1.42/1.58/1.67 and 1.27/1.41/1.49 times more over IM, GM, and DM algorithms. However, the growth trend of all five algorithms is gradually slowing down. The main reason is that under the optimized caching variables, with the increasing of SBSs, the interference of mmWave SBSs becomes more severe, which limits the growth trend of all algorithms.

Fig. 6 shows the bakhaul energy savings with the variations of users. When u increases from 2 to 18, we notice that the backhaul energy savings of all methods are rising up. The reason can be described as follows. For one thing, the optimal cache variable is obtained by the available software packages CVX, i.e., more cached file fragments can meet more user demands. For another thing, since multi-beam concurrent transmissions are combined with coded caching to increase the mmWave throughput and reduce the packet loss, the backhaul energy savings of all methods grow. Moreover, it is observed that compared to the conventional IM matching algrithm, proposed SM and DM algorithms improve backhaul energy savings by around 36.92% and 25.01% when the number of users $U = 18$. That is to say, compared with other three algorithms, our proposed algorithms can save higher backhaul energy in case of more users scenarios.

Fig. 7 plots the impact of skewness α on backhaul energy savings. It can be observed that for the given α, when $U = 8$, our proposed SM algorithm saves the maximal backhaul energy. It is worthy noting that α indicates the similarity in content requests of different users. A smaller α indicates lower similarity. For example, if $\alpha = 0$, the probability that each video content is requested has a uniform distribution. As α increases, different users' requests have a higher similarity. Thus, we model the scenarios in terms of user similarity: low and high similarity, i.e., by setting α from 0.5 to 1 in our simulation. From Fig. 7, it can be seen that no matter the similarity is, our proposed SM algorithm obtains the maximum backhaul energy saving. This implies that our proposed method is robust to the Zipf distribution.

Fig. 8 reveals how the number of files N affects the

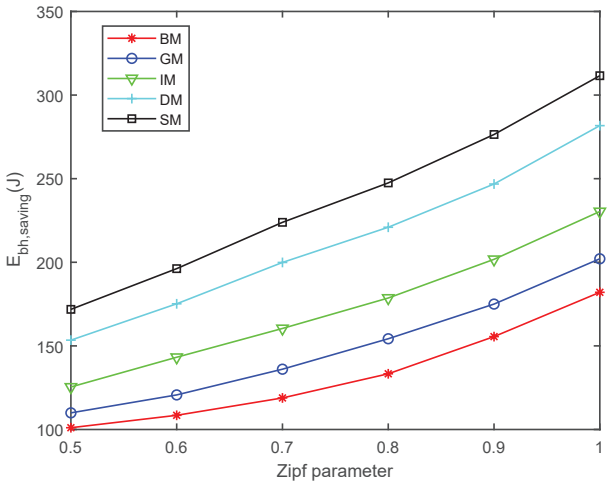


Fig. 7. Backhaul energy savings versus Zipf parameter α, where $K = 20$.

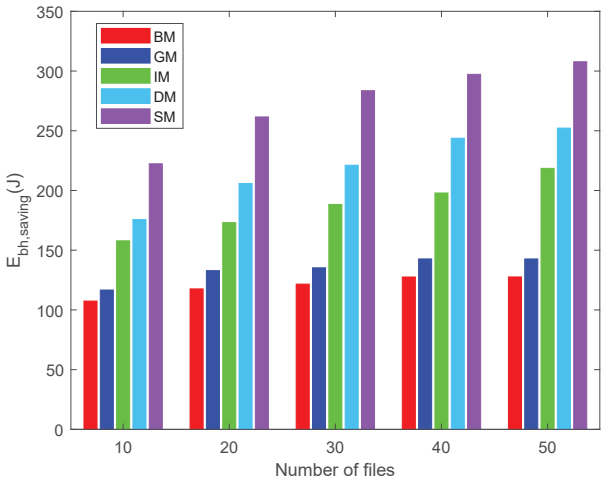


Fig. 8. Backhaul energy savings versus number of files, where $K = 20$.

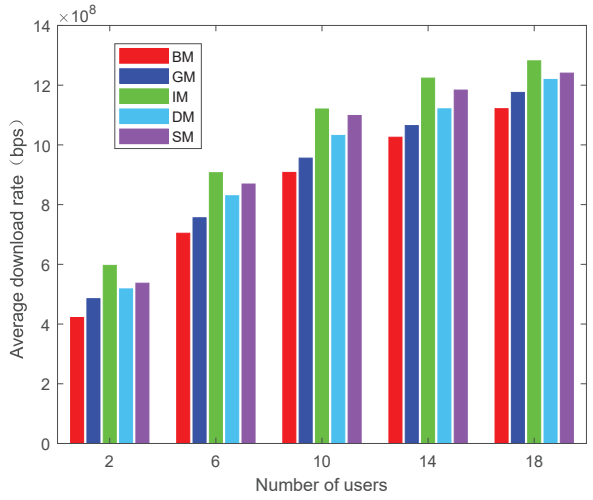


Fig. 9. Average download rate versus number of users, where $K = 20$.

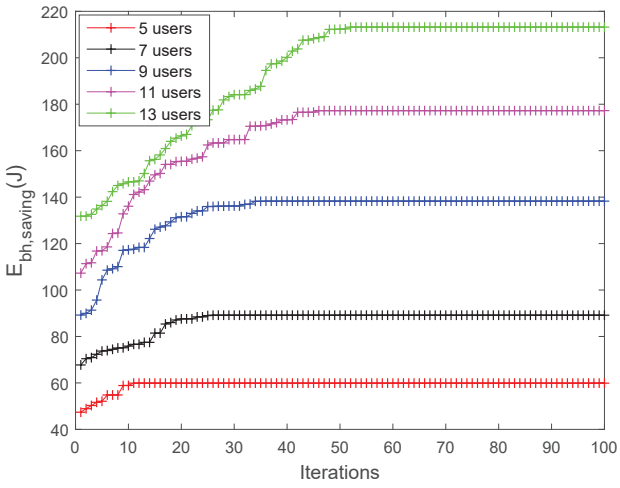


Fig. 10. Convergence of SM algorithm, where $K = 20$ and $\alpha = 1$.

backhaul energy savings under the five methods with $U = 8$. It can be seen that the backhaul energy savings of the five algorithms increase with the number of contents. This is because that the proposed caching algorithm based on the convex optimization achieves much higher cache hit radio. However, the proposed SM and DM algorithms still save more backhaul energy than the other three algorithms. For example, when $N = 50$, the backhaul energy saving improvement of our proposed SM algorithm is 40.85%, 116.73%, and 140.18% over the IM, GM, and BM algorithm, respectively. Meanwhile, the proposed DM algorithm has 15.43%, 76.77%, and 97.64% gains over the three benchmarks. Thus, we can conclude that with the increasing number of contents, the performance of both network throughput and the number of associated users for both proposed algorithms have significant improvements. It again validates the importance of combination coded caching and multi-beam concurrent transmissions between mmWave SBSs and users.

Fig. 9 illustrates the sum rate versus the number of users, where the transceiver antennas are set to be 32×16. It can be seen that the sum rate increases with the number of users. We also observe that all of five methods can enhance the rate of users. The main reason is that the proposed multi-beam concurrent transmissions support multi-connectivity which can provide more access opportunities for users. However, it also can be seen that lower sum rate can be achieved by the proposed algorithms than the conventional IM algorithm. This is because that our proposed algorithms mainly optimizes the objective function (i.e., the backhaul energy saving $E_{bh,\text{saving}}$), to a certain extent, which would weaken the download rate of all users. In another word, in the proposed algorithms, the optimum rule of the multi-beam pair selection is mainly maximizing the backhaul energy savings other than the total throughput of users. In fact, this also implies a performance trade-off between the backhaul energy saving and the download rate of users. In summary, the results demonstrate that our proposed algorithms can effectively save the backhaul energy at the expense of the limited download rate for users.

Fig. 10 shows the convergence performance of the proposed SM algorithm. In this simulation, we set $K = 20$ and $\alpha = 1$. Fig. 10 depicts the convergence of the proposed swap

matching-based algorithm with different number of users, e.g., $U = 5, 7, 9, 11, 13$. As shown in Fig. 10, the objective values exhibit an overall trend of increasing, and converge gradually with the number of iterations increasing. Then, the final backhaul energy savings reach to the maximum objective values after no more than 100 iterations. Moreover, Fig. 9 implies that the backhaul energy savings with different U have different convergence speed and converged value. It is observed that the backhaul energy saving of the proposed SM algorithm with $U = 13$ increases slowly. This is because that although the swap matching is generously explored, the backhaul energy saving increases slowly with more swap blocking pairs. In another word, because of more potential swap blocking pairs, the objective value calculating of the proposed SM algorithm with $U = 13$ is more complex and time consuming. For the above reasons, the converged speed of the proposed SM algorithm with $U = 13$ is relatively slower.

VI. Conclusion

This paper investigates the coded caching and transmission strategy to overcome the blockage and provide robustness for multimedia delivery in mmWave HetNets. A joint optimization of MDS coded caching and multi-beam concurrent transmissions is considered to maximize the backhaul energy saving of the mmWave-μW HetNet, satisfying users' diverse QoS requirements, wireless resourece and storage limitations of mmWave SBSs. Then, the greedy and swap-based many to many matching algorithms are proposed to solve the multi-beam concurrent transmissions, while the convex optimization is applied to solve the optimal coded caching placement. Simulation results show that the proposed algorithms significantly outperform the traditional schemes.

References

[1] Z. Gu, H. Lu, P. Hong, and Y. Zhang, "Reliability enhancement for VR delivery in mobile-edge empowered dual-connectivity sub-6 GHz and mmWave HetNets," *IEEE Trans. Wireless Commun.*, vol. 21, no. 4, pp. 2210–2226, Apr. 2022.

[2] C. Fan, T. Zhang, Y. Liu, and Z. Zeng, "Cache-enabled HetNets with limited backhaul: A stochastic geometry model," *IEEE Trans. Commun.*, vol. 68, no. 11, pp. 7007–7022, Nov. 2020.

[3] X. Wu *et al.*, "Joint long-term cache updating and short-term content delivery in cloud-based small cell networks," *IEEE Trans. Commun.*, vol. 68, no. 5, pp. 3173–3186, May. 2020.

[4] R. Liu, M. Lee, G. Yu, and G. Y. Li, "User association for millimeter-wave networks: A machine learning approach," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4162–4174, Jul. 2020.

[5] M. Feng, S. Mao, and T. Jiang, "Dealing with link blockage in mmWave networks: A combination of D2D relaying, multi-beam reflection, and handover," *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 6746–6759, Aug. 2022.

[6] Y. Liu, X. Fang, M. Xiao, and S. Mumtaz, "Decentralized beam pair selection in multi-beam millimeter-wave networks," *IEEE Trans. Commun.*, vol. 66, no. 6, pp. 2722–2737, Jun. 2018.

[7] I. Aykin *et al.*, "Multi-beam transmissions for blockage resilience and reliability in millimeter-wave systems," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 12, pp. 2772–2785, Dec. 2019.

[8] P. Zhou, X. Fang, X. Wang, and L. Yan, "Multi-beam transmission and dual-band cooperation for control/data plane decoupled WLANs," *IEEE Trans. Veh. Technol.*, vol. 68, no. 10, pp. 9806–9819, Oct. 2019.

[9] L. Liu, S. Zhang, and R. Zhang, "Exploiting NOMA for multi-beam UAV communication in cellular uplink," in *Proc. IEEE ICC*. pp. 1–6, May. 2019.

[10] Z. Chen, Z. Zhou and C. Chen, "Code caching-assisted computation offloading and resource allocation for multi-user mobile edge xomputing," *IEEE Trans. Netw. Service Manag.*, vol. 18, no. 4, pp. 4517–4530, Dec. 2021.

[11] S. Gu, X. Sun, Z. Yang, T. Huang, W. Xiang, and K. Yu, "Energy-aware coded caching strategy design with resource optimization for satellite-UAV-vehicle-integrated networks," *IEEE Internet Things J.*, vol. 9, no. 8, pp. 5799–5811, Apr. 2022.

[12] Y. Fu, Y. Zhang, Q. Zhu, M. Chen and T. Q. S. Quek, "Joint content caching, recommendation, and transmission optimization for next generation multiple access networks," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 5, pp. 1600–1614, May. 2022.

[13] Z. Hajiakhondi-Meybodi, A. Mohammadi, J. Abouei, M. Hou, and K. N. Plataniotis, "Joint transmission scheme and coded content placement in cluster-centric UAV-aided cellular networks," *IEEE Internet Things J.*, vol. 9, no. 13, pp. 11098–11114, Jul. 2022.

[14] R. Deng, "Resource allocation for multipath cooperative video transmission over 5G networks," in *Proc. IEEE ICSPCC*, pp. 1–6, Aug. 2021.

[15] M. N. Dani, D. K. C. So, J. Tang, and Z. Ding, "NOMA and coded multicasting in cache-aided wireless networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 4, pp. 2506–2520, Apr. 2022.

[16] X. Li *et al.*, "Reliability and robust resource allocation for cache enabled HetNets: QoS-aware mobile edge computing," *Reliability Eng. Sys. Safety*, 220, 108272, 2022.

[17] D. Han, W. Liao, H. Peng, H. Wu, W. Wu and X. Shen, "Joint cache placement and cooperative multicast beamforming in integrated satellite-terrestrial networks," *IEEE Trans. Veh. Technol.*, vol. 71, no. 3, pp. 3131–3143, Mar. 2022.

[18] F. Yang *et al.*, "A partially dynamic subarrays structure for wideband mmWave MIMO systems," *IEEE Trans. Commun.*, vol. 68, no. 12, pp. 7578–7592, Dec. 2020.

[19] S. Park, A. Alkhateeb and R. W. Heath, "Dynamic subarrays for hybrid precoding in wideband mmWave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 2907–2920, May. 2017.

[20] O. El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath Jr., "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.

[21] D. Zhao, H. Lu, Y. Wang, H. Sun, and Y. Gui, "Joint power allocation and user association optimization for IRS-assisted mmWave systems," *IEEE Trans. Wireless Commun.*, vol. 21, no. 1, pp. 577–590, Jan. 2022.

[22] Y. Zhang, X. Dong, and Z. Zhang, "Machine learning-based hybrid precoding with low-resolution analog phase shifters," *IEEE Commun. Lett.*, vol. 25, no. 1, pp. 186–190, Jan. 2021.

[23] Y. Zhang *et al.*, "Tree-coding-aided adaptive-cross-entropy algorithm for hybrid precoding with low-resolution analog phase shifters," *IEEE Trans. Veh. Technol.*, vol. 71, no. 6, pp. 6807–6812, Jun. 2022.

[24] Q. Wei *et al.*, "Hierarchical coded caching for multiscale content sharing in heterogeneous vehicular networks,"*IEEE Trans. Veh. Technol.*, vol. 71, no. 6, pp. 5770–5786, Jun. 2022.

[25] K. Shanmugam *et al.*, "FemtoCaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory.*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

[26] F. Gabry, V. Bioglio, and I. Land, "On energy-efficient edge caching in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3288–3298, Dec. 2016.

[27] X. Wu, Q. Li, V. C. M. Leung, and P. C. Ching, "Joint fronthaul multicast and cooperative beamforming for cache-enabled cloud-based small cell networks: An MDS codes-aided approach," *IEEE Trans. Wireless Commun.*, vol. 18, no. 10, pp. 4970–4982, Oct. 2019.

[28] CVX Research Inc, "CVX: Matlab software for disciplined convex programming, version 3.0 beta," http://cvxr.com/cvx, 2015.

[29] Karp, R. M, " Reducibility among combinatorial problems," *Complexity of computer computations.*, Springer, Boston, MA, 1972: 85–103.

[30] Y. Gu, W. Saad, M. Bennis, M. Debbah, and Z. Han, "Matching theory for future wireless networks: fundamentals and applications," *IEEE Commun. Mag.*, vol. 53, no. 5, pp. 52–59, May. 2015.

[31] W. Ni, X. Liu, Y. Liu, H. Tian, and Y. Chen, "Resource allocation for multi-cell IRS-aided NOMA networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4253–4268, Jul. 2021.

[32] L. Zhao *et al.*, "Radio resource allocation for integrated sensing, communication, and computation networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 10, pp. 8675–8687, Oct. 2022.

[33] T. Zhang, Y. Wang, Y. Liu, W. Xu, and A. Nallanathan, "Cache-enabling UAV communications: Network deployment and resource allocation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7470–7483, Nov. 2020.

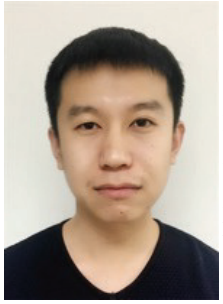
[34] E. Bodine-Baron, C. Lee, A. Chong, B. Hassibi, and A. Wierman, "Peer effects and stability in matching markets," in *Proc. SAGT*, pp. 117–129, Oct. 2011.
[35] S. Zeng, H. Zhang, B. Di and L. Song, "Trajectory optimization and resource allocation for OFDMA UAV relay networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 10, pp. 6634–6647, Oct. 2021.
[36] Y. Zhang *et al.*, "A mobility-aware vehicular caching scheme in content centric networks: Model and optimization," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3100–3112, Apr. 2019.
[37] *Study on channel model for frequencies from 0.5 to 100 GHz*, document TR 38.901 V17.0.0, 3GPP, Mar. 2022.
[38] S. A. Busari, K. M. S. Huq, S. Mumtaz, L. Dai, and J. Rodriguez, "Millimeter-wave massive MIMO communication for future wireless systems: A survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 2, pp. 836–869, 2018.

Danpu Liu received the Ph.D. degree in Communication and Electrical Systems from Beijing University of Posts and Telecommunications, Beijing, China in 1998. She was a Visiting Scholar at City University of Hong Kong in 2002, University of Manchester in 2005, and Georgia Institute of Technology in 2014. She is currently working at the Beijing Key Laboratory of Network System Architecture and Convergence, Beijing University of Posts and Telecommunications, Beijing, China. Her research involved MIMO, OFDM as well as broadband wireless access systems. She has published over 100 papers and 3 teaching books, and submitted 26 patent applications. Her research involved MIMO, OFDM, and broadband wireless access systems. Her recent research interests include B5G/6G mobile communications and air-space-ground integrated networks.

Fangfang Yin received the Ph.D. degree in Information and Communication Engineering from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2020. She is currently working with the State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing, China. Her current research interests include resource management for B5G/6G HetNets, multimedia communications, multi-access edge computing, millimeter wave communications.

Libiao Jin received the B.S. degree in Electronic Information Engineering from Minzu University of China in 2000, and the M.S. and Ph.D. degrees from Communication University of China in 2003 and 2008. He is currently a professor with the School of Information and Communication Engineering, Communication University of China, Beijing, China. He is the author of three books, and has published more than 100 articles. His research interests mainly include intelligent network, artificial intelligence and wireless communication.

An Wang received his Bachelor degree in Electronic Information Engineering from Dalian University of Technology (DLUT), Dalian, China, in 2021. He is currently working toward the M.S. degree with Beijing University of Posts and Telecommunications, Beijing, China. His research interests include wireless video transmission, wireless resource management, matching theory and machine learning.



Yu Zhang received the Ph.D. degree in Information and Communication Engineering from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2022. He is currently an Assistant Researcher with the Wireless Communication Research Center, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. His current research interests include massive MIMO, hybrid beamforming, deep learning and covert communication.