Analysis and Optimization of Cache-Enabled mmWave HetNets With Integrated Access and Backhaul

Chenwu Zhang, Hancheng Lu, Senior Member, IEEE, and Zhuojia Gu

Abstract—In millimeter wave heterogeneous networks with integrated access and backhaul (mABHetNets), a considerable part of spectrum resources are occupied by the backhaul link, which limits the performance of the access link. In order to overcome such backhaul “spectrum occupancy”, we introduce cache into mABHetNets. Caching popular files at small base stations (SBSs) can offload the backhaul traffic and transfer spectrum from the backhaul link to the access link. To approach the optimal performance of cache-enabled mABHetNets, we first analyze the signal-to-interference-plus-noise ratio distribution and derive the average potential throughput (APT) expression by stochastic geometric tools. Then, based on our analytical work, we formulate a joint optimization problem of cache decision and spectrum partition to maximize the APT. Inspired by the block coordinate descent (BCD) method, we propose a joint cache decision, spectrum partition and power allocation (JCSPA) algorithm to approach the optimal solution. Numerical results show the effectiveness and enhancement of the proposed algorithm. Besides, we verify the APT under different parameters and find that the introduction of cache facilitates the transfer of backhaul bandwidth to access link. Jointly deploying appropriate caching capacity at SBSs and performing specified spectrum partition can bring up about 90% APT gain in mABHetNets.

Index Terms—Millimeter wave, integrated access and backhaul, cache, spectrum transfer, average potential throughput.

I. INTRODUCTION

The number of mobile terminal equipment increases sharply in recent years and the demand for data also increases at an annual growth rate of 63% [2], while the existing scarce LTE spectrum resources [3] severely affect the high-demand communication. Fortunately, thanks to the characteristics of abundant spectrum in millimeter wave (mmWave), the high throughput communication has become possible by using mmWave in both the access and the backhaul link. However, due to the characteristic of short wavelength and blockage sensitivity, mmWave signals have limited coverage and are only suitable for short distance propagation, so the base stations (BSs) need to be deployed densely in mmWave heterogeneous networks (HetNets), which in turn increases the demand of backhaul traffic. Particularly, integrated access and backhaul (IAB) in mmWave HetNets (mABHetNets), as a promising technology, has been standardized for the fifth generation (5G) mobile communication technology in the third generation partnership project (3GPP) Rel-16 [3], [4], [5], [6], [7]. In mABHetNets, macro base stations (MBSs) are connected to the mobile core network via high-speed optical fiber backhaul, while it is not practical to connect all small base stations (SBSs) to the mobile core network via optical fiber backhaul. Thus, a lot of denser deployed low-power mmWave SBSs are connected to MBSs via wireless backhaul link. Both the MBS and SBS should provide service for the user via wireless access link.

Since the wireless access and backhaul links share the mmWave spectrum resources in mABHetNets, spectrum partition has a significant impact on the network performance and hence attracted some research attempts [12], [13], [14], [15]. Saha et al. [12] performed rate coverage analysis on the spectrum partition strategies as the total spectrum is dynamically partitioned or statically partitioned in mABHetNets. Shi et al. [13] leveraged allocated resource ratio between radio access and backhaul to study the maximization of network capacity by considering the fairness among SBSs. Saha et al. [14] performed the downlink rate coverage probability analysis and found an optimal split of access and backhaul bandwidth of integrated access and backhaul network. Diamanti et al. [15] optimized phase shift matrix, transmit power as well as spectrum partition between wireless access and backhaul to maximize the energy efficiency in reconfigurable intelligent surface aided IAB network from an optimization perspective. Alghafari et al. partitioned the bandwidth between access and backhaul links for a two-tier integrated access and backhaul topology in an iterative and decentralized way in [16]. Besides, Kwon et al. [17] designed a centralized as well as a distributed resource allocation scheme...
to improve the throughput of an out-band based integrated access and backhaul network.

However, even with the relevant research on the spectrum partition for the wireless access link and the wireless backhaul link, a considerable part of mmWave spectrum resources are still occupied by the wireless backhaul link to meet the huge data traffic demand. According to the findings in [14], up to 50% mmWave spectrum might be used in wireless backhaul link to satisfy the high speed data traffic. Such a “spectrum occupancy” phenomenon of wireless backhaul link severely limits the transmission of wireless access link, which further affects the communication quality of users and degrades the spectrum resource utilization. Besides, several reports show that a few files with high popularity are often requested by users, and this increases the transmission pressure on the wireless backhaul link [18], [19]. Repeated transmission also causes a lot of waste of power consumption and spectrum resources. Fortunately, enabling caching at the wireless edge such as SBSs is considered as a promising way to improve the energy efficiency and network throughput recently [18], [19], [20], [21], [22], [25], [26], [27], [28], especially for those files with high popularity and frequently requested. To solve such a “spectrum occupancy” problem fundamentally, we introduce the cache into BSs in mABHetNets to effectively alleviate the repeated transmission of the wireless backhaul, thereby improving the network performance. In [19], Tao et al. proposed a cache-enabled radio access network to minimize the total network cost. Xu and Tao in [20] applied coded caching into small-cell network and investigated average fractional offloaded traffic (AFOT) and average ergodic rate performance metrics, then maximized the AFOT. In [21], Fan et al. proposed a cache-enabled HetNets with limited backhaul and analyzed the successful content delivery probability, successful delivery rate as well as energy efficiency theoretically. Liu et al. [25] investigated the impact of cache on the energy efficiency of a wireless access network and found the optimal caching capacity. Gabry et al. focused on the effect of several edge caching strategies on the energy consumption of heterogeneous wireless networks in [27]. Xiao et al. [40] studied the effect of cluster size, caching policies and capacity of user terminal and SBS on service delay in the clustered cache-enabled HetNets. Zhang et al. analyzed the performance of cache-enabled hybrid satellite-aerial-terrestrial networks with stochastic geometry and revealed significant performance gain brought by cache in [28]. Caching popular files proactively during off-peak time at the edge of the network can reduce the data traffic pressure of backhaul link [22]. When the backhaul traffic is offloaded by caching popular files at SBSs in mABHetNets, a part of mmWave spectrum can be transferred from the backhaul link to the access link.

Although the introduction of cache brings the benefits of improvement the network throughput, nonetheless some new problems arise both theoretically and technically: how to deploy the cache appropriately to solve the “spectrum occupancy” problem and how much improvement the cache can bring to the throughput performance of the mABHetNets. Thus, it is essential to jointly consider the cache decision and spectrum partition in mABHetNets to improve the APT.

In this paper, average potential throughput (APT) metric is used to measure the network performance, which has become the major performance metric in mmWave HetNets [8], [9], [10], [11] and focuses on analyzing user’s average throughput with the specific rate requirement [8]. On the one hand, since each SBS in cache-enabled mABHetNets has a limited power budget and the power consumed by additional cache cannot be neglected [25], [27], caching capacity at SBS will affect the transmit power of SBS, thereby determining the user association and signal-to-interference-plus-noise ratio (SINR) distribution. There is a tradeoff between caching capacity and transmit power at SBS. On the other hand, since caching capacity as well as transmit power are limited by the maximum power constraint at SBS, this will affect the optimal spectrum partition coefficient and thus affect the APT. Therefore, these three resource variables are coupled with each other, and a more reasonable resource allocation strategy (i.e., caching capacity, spectrum partition and power allocation) in mABHetNets needs to be designed. Simultaneously considering the effect of blockage of mmWave and user association with BSs in cache-enabled mABHetNets makes it become more sophisticated, which requires subsequent thorough analysis and optimization both theoretically and technically. The joint optimization algorithm needs to be further investigated to achieve desired APT in cache-enabled mABHetNets. Besides, how much network performance improvement can be obtained through the joint optimization also needs to be investigated extensively. To address the aforementioned issues, in this paper, the main contribution can be summarized as follows.

- We develop a tractable analytical framework for a down-link cache-enabled mABHetNet from a stochastic geometry perspective. Considering the characteristic blockage sensitivity of mmWave, we give the probability distribution function (PDF) of the distance between the transmitters and receivers. Then, we analyze the association probability based on the criterion of maximum biased received power and derive the SINR distribution expression by taking the directional beamforming into account. An SINR distribution expression under noise-limited scenario is also derived utilizing the stochastic geometry tools. The APT expression of cache-enabled mABHetNet is derived under the joint consideration of caching capacity, transmit power as well as spectrum partition.

- Based on the aforementioned analytical work, an APT maximization problem is formulated with respect to caching capacity, spectrum partition and transmit power; As the formulated problem is a mixed-integer nonlinear programming (MINLP) problem, it is intractable to handle directly. Inspired by the block coordinate descent (BCD) method, the original problem is decomposed into two sub-problems, i.e., cache decision problem, spectrum partition and power allocation problem. Then, we solve them by the proposed joint cache decision, spectrum partition and power allocation (JCSAP) algorithm in a two-step alternating optimization manner to approach the optimal solution to the APT maximization problem.
• We investigate the impacts of caching capacity, transmit power and spectrum partition on APT both theoretically and experimentally. Some important insights on the interplay between caching capacity and spectrum partition are provided from the perspective of APT increment. Extensive numerical results are carried out to verify the convergence of the proposed algorithm and show the effects of other cache-related parameters on APT. Numerical results illustrate that joint optimization of cache decision, spectrum partition and power allocation is an effective method to improve APT.

The rest of the paper is organized as follows. Firstly, we introduce the system model of cache-enabled mABHetNet in Section II. Section III derives the SINR distribution of cache-enabled mABHetNet and then APT is further defined and derived based on the SINR distribution. Next, Section IV gives the APT maximization problem and solution. Performance evaluation and numerical results are provided in Section V. Finally, the paper is concluded in Section VI.

II. SYSTEM MODEL

A. Network Model

In this section, we develop a downlink cache-enabled mABHetNet with integrated access and backhaul architecture, which consists of an MBS tier and an SBS tier as shown in Fig. 1.

In this architecture, high power MBSs are connected to the core network via high speed optical fiber links and SBSs are associated with the corresponding MBSs via providing mmWave spectrum wireless backhaul transmission links. The typical user could be associated with the MBS or SBS to obtain the wireless access service [29]. By stochastic geometry tools, the locations of the MBS and SBS are modeled as the independent Poisson Point Processes (PPPs), which are denoted by $\Phi_m$ and $\Phi_s$ with densities of $\lambda_m$ and $\lambda_s$, respectively. The location of the users is modeled as another independent PPP $\Phi_u$ with density $\lambda_u$. We stipulate that the user density is large sufficiently so that each BS consists of at least one associated user in its coverage area. We select a typical user at the origin for analysis. Based on Slivnyak’s

B. Caching Model

The architecture we developed in this paper is a cache-enabled HetNet. File library is denoted by symbol $\mathcal{F}$ and the number of all files is $|\mathcal{F}| = F$. In order to derive conveniently, we assume that all the files in library have the same size and the size is expressed in file units. Such assumption can hold since those files can be divided into chunks of equal size in practical transmission [30]. In the whole file library, different files have different popularities and the popularity of the file often doesn’t change or changes slowly in a short time [31]. The popularity of the file can be predicted according to interview records or machine learning retrieval methods and this problem has been actively studied in recent years while it is outside the scope of this paper. In general, the popularity of the file obeys the Zipf distribution [25], [32]. For the files denoted by the index set $\mathcal{F} = \{1, 2, \ldots , F\}$, the popularity of the file $f \in \mathcal{F}$ is $p_f = \frac{f^{-\alpha}}{\sum_{g=1}^{F} g^{-\alpha}}$, of which $\gamma_p$ is the parameter of the Zipf distribution and it denotes the skewness of Zipf distribution, while the bigger $\gamma_p$ means that the fewer files have higher popularity [33]. Different from the probabilistic caching model in [24], in this paper, we use the highest-popularity-first cache strategy like [25], which means that files with higher popularity will be cached preferentially. We assume that the caching capacity of MBS is large enough to load all the $F$ files in the file library [36]. The partial files will be deployed proactively in the SBS in descending order of file popularity. In this case, the SBS will cache the most popular $C$ files to achieve the maximum cache hit ratio [25]. The cache hit ratio of the SBS for caching capacity $C$ is given as

$$p_h(C) = \frac{\sum_{f=1}^{C} f^{-\gamma_p}}{\sum_{g=1}^{F} g^{-\gamma_p}}.$$  (1)

C. Power Consumption Model

Like [23], the total power consumption model of the mABHetNet can be regarded as the power consumption of the MBS and SBS, where the power consumption of the MBS is determined as $P_{m}^{\text{max}} = \rho_m P_{m}^{\text{tr}} + P_{m}^{\text{fc}} + P_{m}^{\text{ca}} = \rho_m P_{m}^{\text{tr}} + P_{m}^{\text{fc}} + \omega_{ca} s F$, and the power consumption of the SBS is $P_{s}^{\text{tr}} + P_{s}^{\text{fc}} + P_{s}^{\text{ca}} = \rho_s P_{s}^{\text{tr}} + P_{s}^{\text{fc}} + \omega_{ca} s C$. $P_{m}^{\text{max}}$ is the total power consumption of MBS. $\rho_m$ and $\rho_s$ are the power consumption amplification factor of MBS and SBS transmitter. $P_{m}^{\text{tr}}$ and $P_{s}^{\text{tr}}$ are the fixed circuits-related power consumption of MBS and SBS [34]. $P_{m}^{\text{ca}}$ and $P_{s}^{\text{ca}}$ are the caching power consumption of MBS and SBS based on energy proportional model. In order to quantify the caching power consumption, we adopt the caching capacity proportional caching power model [25], [26], [27], [34], [35] in our study, which is widely used in the existing content-centric network and radio-access network. As the caching power consumption is proportional to the caching file size, we denote $P_{m}^{\text{ca}} = \omega_{ca} s F$ and $P_{s}^{\text{ca}} = \omega_{ca} s C$, where $s$ is the size of each file and $\omega_{ca}$ is the power coefficient of
caching hardware (e.g., SSD) with unit: watt/bit. Assuming that the maximum power of each SBS has been preset, each SBS can adjust the transmit power and caching power (caching capacity) within the preset power budget $P_i^{\text{max}}$ [26], namely, $\rho_s P_i^{\text{tx}} + P_i^{\text{ch}} + \omega_c a s C \leq P_i^{\text{max}}$.

D. Channel and Transmission Model

On account of the high density of MBS or SBS deployment and the linear propagation transmission characteristic of mmWave, each transmission link is assumed to be independent Rayleigh fading [37] and the signals are based on the transmission link as line-of-sight (LOS) and non-line-of-sight (NLOS). The path loss function expresses the signal attenuation relationship with distance $r$. The mathematical expression of the path loss function is as follows [38]:

$$L(r) = \begin{cases} A_r r^{-\alpha}, & \text{with LOS probability } P_L(r), \\ A_{NL} r^{-\alpha_{NL}}, & \text{with NLOS probability } P_{NL}(r), \end{cases}$$

(2)
of which the LOS probability $P_L(r) = e^{-\beta r}$, NLOS probability $P_{NL}(r) = 1 - P_L(r)$, $A_r$ ($A_{NL}$) is the LOS (NLOS) pathloss parameter, $r$ is the distance between the user and base station, $\alpha$ ($\alpha_{NL}$) is a path loss exponent in LOS (NLOS) transmission of Rayleigh fading. $\beta \geq 0$ is the blockage parameter that captures In practical transmission, LOS signals become dominant in dense mmWave network.

Besides, directional beamforming is used for all antennas at the transceivers and signals propagates along the main lobe of the antenna. We use the sectorial antenna pattern in this analysis [39]. The antenna gain pattern for the transceiver in the mABHetNet is given as

$$G_q(\phi) = \begin{cases} M, & \text{if } |\phi| \leq \theta \\ m, & \text{otherwise.} \end{cases}$$

(3)

where $q \in \{T, R\}$ denotes the antenna at the transmitter or receiver, $\phi \in [0, 2\pi]$ is the angle off boresight direction, $\theta$ is the main width of main lobe, $M$ and $m$ are the gain of the main lobe and side lobe. Then the random gain for the transmission link and its probability is given as

$$G = \begin{cases} MM, & \text{with probability } \frac{\theta^2}{4\pi^2}, \\ Mm, & \text{with probability } \frac{\theta(2\pi - \theta)}{2\pi^2}, \\ mm, & \text{with probability } \frac{(2\pi - \theta)^2}{4\pi^2}. \end{cases}$$

(4)

For tractability of analysis, the perfect beamforming is assumed between the transmitter and receiver [40]. Based on the above analysis, we can derive the SINR expression of the distance from the typical user to the associated SBS and the associated MBS via wireless access link as Eq. (5) and Eq. (6), shown at the bottom of the page, where $b_{s,0}$ and $b_{m,0}$ denote the serving SBS and MBS for the typical user, $B_s$ and $B_m$ are the association bias factor of SBS and MBS, $G_m$, $G_s$, $G_i$ and $G_l$ are the antenna gain of the corresponding transmission link ($G_m = G_s = MM$). $h_{s,i}$ and $h_{m,i}$ are the small-scale fading from $i$-th SBS or $j$-th SBS ($h_{s,i}, h_{m,i} \sim \exp(1)$). $L(r_s)$ and $L(r_m)$ are the path loss from the serving SBS or MBS to the typical user. $r_s$, $r_m$ are the distance from the typical user to the serving SBS $b_{s,0}$ or MBS $b_{m,0}$, $L(r_s, i)$ is path loss from the $i$-th SBS to the typical user. $L(r_m, l)$ is the path loss from the $l$-th MBS to the typical user. As the receiver is designed to be capable of working on the whole spectrum bandwidth and it could receive signal in the whole spectrum, the noise power is expressed as $N_0 W$. $N_0$ is the thermal noise power density (unit: dBm/Hz) and $W$ is the total system spectrum bandwidth. Besides, MBS provides the wireless backhaul service to SBS for data transmission. Similar to Eq. (5) and Eq. (6), for a typical SBS at the distance $r_{bh}$ from its associated MBS, the SINR expression of the signal from the MBS to the SBS of the wireless backhaul link is given as follows.

$$\text{SINR}_{bh}(r_{bh}) = \frac{P_{tr}^s B_m G_m h_m L(r_{bh})}{I_{bh} + N_0 W}$$

(7)

$$= \frac{P_{tr}^m B_m G_m h_m L(r_{bh})}{\sum_{i \in \Phi_m \setminus b_{m,0}} P_{tr}^m B_m G_i h_m L(r_{bh,i}) + N_0 W}.$$  [7]

E. Spectrum Partition Strategy

In this subsection, we introduce the spectrum partition strategies in the downlink mABHetNet, where the spectrum of access and backhaul link is orthogonal, so the whole spectrum needs to be divided into two parts: $W_{bh}$ for the wireless backhaul link and $W_{ac}$ for the wireless access link. Then we will introduce two spectrum allocation strategies as follows.

1) Fixed Spectrum Allocation (FSA): This default spectrum allocation strategy is widely used in traditional heterogeneous network. In order to satisfy the communication needs fairly, we allocate spectrum for the access and backhaul link equally as $W_{bh} = \frac{1}{2} W$ and $W_{ac} = \frac{1}{2} W$ in FSA. In this case, we give the transmission data rate of backhaul link and access link based on Shannon’s theorem, respectively.

$$R_i = \frac{1}{2} W \log_2 (1 + \text{SINR}_i),$$

(8)

where $i \in \{bh, m, s\}$ denotes the wireless backhaul link, MBS tier and SBS tier via wireless access link, respectively.

\begin{align*}
\text{SINR}_s(r_s) &= \frac{P_{tr}^s B_s G_s h_s L(r_s)}{I_s + I_m + N_0 W} = \frac{P_{tr}^s B_s G_s h_s L(r_s)}{\sum_{i \in \Phi_s \setminus b_{s,0}} P_{tr}^s B_s G_i h_s L(r_{s,i}) + \sum_{l \in \Phi_m} P_{tr}^m B_m G_i h_m L(r_{m,l}) + N_0 W}, \\
\text{SINR}_m(r_m) &= \frac{P_{tr}^m B_m G_m h_m L(r_m)}{I_m + I_s + N_0 W} = \frac{P_{tr}^m B_m G_m h_m L(r_m)}{\sum_{i \in \Phi_m \setminus b_{m,0}} P_{tr}^m B_m G_i h_m L(r_{m,i}) + \sum_{s \in \Phi_s, m, l} P_{tr}^s B_s G_i h_s L(r_{s,i}) + N_0 W}.
\end{align*}

(5)  (6)
2) Dynamic Spectrum Allocation (DSA): In order to alleviate “spectrum occupancy” problem, we introduce a parameter \( \eta \) to allocate spectrum dynamically according to the current caching status at SBS, where \( \eta \in [0,1] \) is the spectrum partition ratio coefficient for the access link so the spectrum for the backhaul link is denoted as \( W_{bh} = (1-\eta)W \) and the spectrum for the access link is denoted as \( W_{ac} = \eta W \). When the user is associated with the SBS and the requested files are not cached at the SBS (cache-miss scenario), file delivery needs to go through the wireless backhaul link from the MBS to the SBS and wireless access link from the SBS to the user. In other words, the transmission rate depends on both the wireless access link rate and the wireless backhaul link rate when the user is associated with the SBS under the cache-miss case. When the user is associated with the MBS, the file delivery go through the wireless access link from the MBS to the user. Similarly, the transmission rate of each link is given as follows.

\[
\begin{align*}
R_{bh} &= (1-\eta)W \log_2(1+\text{SINR}_{bh}), \\
R_s &= \eta W \log_2(1+\text{SINR}_s), \\
R_m &= \eta W \log_2(1+\text{SINR}_m).
\end{align*}
\]

### III. SINR DISTRIBUTION AND APT ANALYSIS OF MABHetNet

In this section, we will derive the expression of SINR distribution of the typical user conditioned on its association selections. As a typical user is covered by the associated SBS or MBS via wireless access link and the SBS is covered by the associated MBS via wireless backhaul link, we first derive the PDF of the distance between the transmitters and receivers. Further, SINR distribution expressions are also obtained. Finally, we derive the APT expression of cache-enabled mABHetNet.

#### A. The PDF of Distance to Base Station

First of all, we need to derive the PDF of the distance between the typical user and its nearest SBS or MBS. For the typical user at the origin, when the typical user associates with the closest SBS or MBS at a distance \( r \), no other SBS or MBS can be closer than \( r \).

Since the typical user is associated with the closest SBS or MBS via LOS and NLOS link, we derive the PDF in the following Lemma.

**Lemma 1:** The PDFs of the distance \( r \) between the typical user and the associated SBS or MBS via LOS/NLOS link are written as

\[
f_{R_s}(r) = \mathcal{P}_k(r) \times \exp \left( -\pi r^2 \lambda_s \right) \times 2\pi r \lambda_s, \quad (12)
\]

\[
f_{R_m}(r) = \mathcal{P}_k(r) \times \exp \left( -\pi r^2 \lambda_m \right) \times 2\pi r \lambda_m, \quad (13)
\]

where \( s \) and \( m \) denote the index of SBS tier and MBS tier, \( k \in \{ L, N \} \) denotes the transmission link of LOS or NLOS in wireless access link.

**Proof:** Due to space limitation, the detailed proof process can be found in the online version [48]-Appendix A.

**Remark 1:** Similarly to Lemma 1, the cache-miss file data delivery will be delivered from the MBS to the SBS via the wireless backhaul link. The PDF of the distance \( r \) between the SBS and the associated MBS via LOS/NLOS link can be written as

\[
f_{R_{bh}}(r) = \mathcal{P}_k(r) \times \exp \left( -\pi r^2 \lambda_m \right) \times 2\pi r \lambda_m, \quad (14)
\]

where \( bh \) denotes the index of backhaul link and \( k \) is the same as Eq. (12).

#### B. Association Probability

In the mABHetNet, we will first analyze the probability that a user is associated with SBS tier or MBS tier via different links due to the different transmission links and the densities of SBS and MBS. Besides, since SBS may be associated with MBS via different transmission links in the SBS backhaul association, different SBS backhaul association probabilities should be derived in the meanwhile.

1) **User Association Probability:** We consider user association based on the maximum biased received power criterion, where a typical user is associated with the strongest BS in terms of the received power at the user. Then, for a typical user associated with the BS via LOS link and NLOS link, the received powers are \( P_{tr} B_i G_i h_i A_{L}, r^{-\alpha_L} \) and \( P_{tr} B_i G_i h_i A_{NL}, r^{-\alpha_N} \) (\( i \in \{ s, m \} \)), respectively.

Based on the maximum biased received power association criterion, the BS density and transmit power as well as transmission link determine the probability that a typical user is associated with an SBS or an MBS. The following lemma provides the user association probability with SBS or MBS via LOS and NLOS link, respectively.

**Lemma 2:** For the given distance \( r \), the probabilities that a typical user is associated with the SBS tier via LOS link and...
where exist two backhaul association probabilities, which are given backhaul transmission includes LOS link and NLOS link, there that the SBS is associated with the MBS in LOS link and the Other notations have similar definitions.

2) SBS Backhaul Association Probability: The SBS will be associated with the MBS via the wireless backhaul link. The SBS backhaul association strategy is also based on the maximum biased received power from the MBS. Since the backhaul transmission includes LOS link and NLOS link, there exist two backhaul association probabilities, which are given as below.

Remark 2: Similar to Lemma 2, the probabilities that a typical SBS is associated with the MBS tier via LOS and NLOS link are

\[
F_{bh}^L(r) = p_{bh}^s(r) f_{Rh_m}^L(r),
\]

(19)

\[
F_{bh}^{NL}(r) = p_{bh}^s(r) f_{Rh_m}^{NL}(r),
\]

(20)

where \( p_{bh}^s(r) = e^{-\lambda_s \pi \left( \frac{A_m}{\pi r^2} \right) \frac{\alpha}{\pi r^2} \frac{A_N}{r^2}} \) is the probability that the SBS is associated with the MBS in LOS link and the interference is from MBS in NLOS link, \( p_{bh}^s(r) \) has the similar definition.

C. SINR Distribution

To study the APT performance of mABHetNet, we need to first investigate the SINR distribution of the user covered by SBS tier via access link and the SINR distribution of the SBS covered by MBS via backhaul link. This SINR distribution is defined as the probability that the received SINR is above a pre-designated threshold \( \gamma \).

\[
P_{s,i}^{cov}(\gamma) = \Pr[SINR_s \geq \gamma].
\]

(21)

Since the user is covered by SBS or MBS, we first give the SINR distribution of the typical user. Then we give the SINR distribution of the typical SBS covered by MBS.

Proposition 1: 1) SINR distribution of user covered by the SBS or the MBS:

The SINR distribution that the typical user is associated with the SBS via access link is:

\[
P_{s,i}^{cov}(\gamma) = \sum_{i \in \{s, L\}} \Pi_{s,i}^{cov}(\gamma),
\]

(22)

\[
P_{s,i}^{cov}(\gamma) = \int_{0}^{\infty} \exp \left( -\gamma N_0 \right) \mathcal{L}_{i,s,m}^L F_{bh}^s(r) dr,
\]

(23)

where \( s \) denotes the SBS, \( P_{s,i}^{cov}(\gamma) = \mathbb{P} \left[ \text{SINR}_s(r) \geq \gamma \right] \) is the probability that the user is covered by the SBS, where \( s \) denotes SBS, \( \gamma \) is the default threshold for successful demodulation and decoding at the receiver and the Laplace transform of interference from SBS or MBS via LOS link is given as Eq. (24), shown at the bottom of the page, of which \( G_i, G_i \in \{MM, Mm, mm\}, p_{G_i}(p_{G_i}) \) is the probability of the antenna gain taking the corresponding

\[
\mathcal{L}_{i,s,m}^L(\gamma_{r^L}) = \prod_{G_i} \left( -2\pi \lambda_s p_{G_i} \left( \int_{r}^{\infty} \frac{P_{L}(u)u}{1 + \frac{P_{tr} B_{G_i} A_r}{\gamma P_{tr} B_{G_i} A_r} + \frac{P_{NL}(u)u}{1 + \frac{P_{tr} B_{G_i} A_r}{\gamma P_{tr} B_{G_i} A_r}}} du + \int_{r}^{\infty} \frac{P_{L}(u)u}{1 + \frac{P_{tr} B_{G_i} A_r}{\gamma P_{tr} B_{G_i} A_r} + \frac{P_{NL}(u)u}{1 + \frac{P_{tr} B_{G_i} A_r}{\gamma P_{tr} B_{G_i} A_r}}} du \right) \right)
\]

(24)
value from SBS interference tier and MBS interference tier. 
\[ d_1 = \frac{P_{tr}^s P_m}{P_{tr}^s B_m} \text{ and } d_2 = \frac{P_{tr}^m P_m}{P_{tr}^m B_m A_{NL}}. \]
Following the same logic, other Laplace transform of interference have similar expressions.

From the Laplace transform of interference, the interference from the SBS tier is independent of the transmit power \( P_{tr}^s \) and even the interference from the MBS layer is monotonically decreasing about \( P_{tr}^m \), which also reveals that increasing the transmit power \( P_{tr} \) is beneficial to improve the SINR distribution of the SBS without considering other constraints. Such a proof process can be done by substituting the interference term into the Laplace transform.

Similarly, the SINR distribution that the typical user is associated with the MBS via access link is
\[
P_{cov}^{\text{m}}(\gamma) = \sum_{i \in \{L, NL\}} \frac{P_{cov}^{\text{m,}i}(\gamma)}{C},
\]
\[
P_{cov}^{\text{m,}i}(\gamma) = \int_0^\infty \frac{-\gamma N_0}{P_{tr}^s B_m G_m A_{ir-r}} \exp \left( \frac{-\gamma N_0}{P_{tr}^s B_m G_m A_{ir-r}} \right) d\gamma.
\]

2) SINR distribution of SBS covered by MBS:

The SINR distribution that the typical SBS is covered by the MBS via backhaul link is:
\[
P_{cov}^{\text{hb}}(\gamma) = \sum_{i \in \{L, NL\}} \frac{P_{cov}^{\text{hb,}i}(\gamma)}{C},
\]
\[
P_{cov}^{\text{hb,}i}(\gamma) = \int_0^\infty \frac{-\gamma N_0}{P_{tr}^s B_m G_m A_{ir-r}} \exp \left( \frac{-\gamma N_0}{P_{tr}^s B_m G_m A_{ir-r}} \right) d\gamma.
\]

Proof: The proof process is provided in Appendix A.

Proposition 2: In noise-limited scenario, the SINR distribution can be reduced into a tractable form as Eq. (29), shown at the bottom of the page, of which \( Y(\xi_0) = \int_0^\xi \frac{A_0}{C^2} \exp(-\beta \phi) \frac{\phi}{\xi} d\phi d\xi - \int_0^\xi \frac{A_0}{C^2} \exp(-\beta \phi) \frac{\phi}{\xi} d\phi d\xi, k \in \{m, s\}, \) denotes the MBS tier or the SBS tier.

Proof: The detailed proof process of Proposition 2 is provided in Appendix B.

Proposition 2 implies that SINR distribution will mainly depends on the transmit power of BS, so improving the transmit power helps improve the APT without considering power constraints.

D. APT of Cache-Enable mABHetNet

APT is a significant metric to measure the heterogeneous network performance, which mainly focuses on the average user QoS requirement in terms of data rate. Next, we will derive the APT expressions based on the above analysis.

1) Definition: APT captures the average number of bits that can be received by user in unit area for a given pre-designated threshold \( \gamma_0 \). The definition of APT is
\[
\mathcal{R} = \lambda_k W \log_2(1 + \gamma_0) P \{ \text{SINR} \geq \gamma_0 \},
\]
where \( \lambda_k \) is the density of MBS or SBS and \( W_0 \) is the allocated spectrum bandwidth. \( \gamma_0 \) is SINR threshold for the signal demodulation and represents the receiver’s communication requirement.

2) The Detailed APT Expression: APT of mABHetNet depends on both the backhaul link and the access link, then we will give the detailed expression in the following analysis. For the user associated with the SBS, the transmission link includes the backhaul link between the SBS and SBS and the access link between the SBS and the user. Besides, in this cache-enabled mABHetNet, the cache in SBS also influences the file delivery in the transmission path. When the requested files are cached at the SBS, the files can be delivered to user directly without transmitting through the wireless backhaul link. In other words, the wireless backhaul link will not be occupied under cache-hit circumstance. Besides, if the files requested by the user are not cached at the SBS, the missing files need to be delivered from the MBS, which increases the wireless backhaul traffic. In this case, the user’s APT depends on the minimum throughput of the backhaul link and the access link. For the user associated with MBS, APT only depends on the access link. Given that partial files are cached in SBS, we give the APT expression of mABHetNet in the following corollary.

Corollary 1: Since the transmission can be LOS or NLOS in wireless access and backhaul link for user associated with SBS tier or MBS tier, we can get the APT of SBS tier, MBS tier and mABHetNet, respectively. First, we give the APT expression of SBS tier as Eq. (30), shown at the bottom of the page.

The symbol \( \min\{\cdot\} \) means that the minimum value of throughput in the wireless access link of SBS tier and throughput in the wireless backhaul link. Similarly, APT of MBS tier is given as follows.

\[
\mathcal{R}_m(\eta) = \lambda_m \eta W \log_2(1 + \gamma_0) P_{cov}^{\text{m,}s}(P_{tr}^m).
\]

Then, APT of cache-enabled mABHetNet can be obtained as follows.

\[
\mathcal{R}(\eta, C, P_{tr}^s) = \mathcal{R}_s(\eta, C, P_{tr}^s) + \mathcal{R}_m(\eta).
\]

Note that, \( C \) is the caching capacity of SBS and \( p_h(C) = \sum_{i=1}^{C} \frac{1}{i} \) is the cache hit ratio of SBS. \( (1 - p_h(C)) \) reflects the probability that the files are not cached in SBS tier and need to be delivered through the backhaul link.

\[
\mathcal{R}_s(\eta, C, P_{tr}^s) = \min\{(1 - p_h(C)) \lambda_s \eta W \log_2(1 + \gamma_0) P_{cov}^{\text{m,}s}(P_{tr}^m),
(1 - p_h(C)) \lambda_m W (1 - \eta) \log_2(1 + \gamma_0) P_{cov}^{\text{m,}s}(P_{tr}^m) + p_h(C) \lambda_s \eta W \log_2(1 + \gamma_0) P_{cov}^{\text{m,}s}(P_{tr}^m).
\]

\[P_{cov}^{\text{m,}s}(\gamma) = 1 - \exp \left( -\pi \lambda_k A_{NL}^2 \int_0^\infty \frac{1}{C_{\eta}} + 1 \left( \frac{P_k B_k G_k}{\sigma^2} \right)^{\frac{2}{\sigma^2}} 2\pi \lambda_k Y \left( \frac{P_k B_k G_k}{\sigma^2} \right)^{\frac{1}{\sigma^2}} \right) \]

\[\mathcal{R}_s(\eta, C, P_{tr}^s) = \min\{(1 - p_h(C)) \lambda_s \eta W \log_2(1 + \gamma_0) P_{cov}^{\text{m,}s}(P_{tr}^m),
(1 - p_h(C)) \lambda_m W (1 - \eta) \log_2(1 + \gamma_0) P_{cov}^{\text{m,}s}(P_{tr}^m) + p_h(C) \lambda_s \eta W \log_2(1 + \gamma_0) P_{cov}^{\text{m,}s}(P_{tr}^m)\).\]
IV. PROBLEM FORMULATION AND SOLUTION

In this section, we formulate an APT maximization problem with multivariate function and complicated integral component. However, the formulated APT maximization problem is an MINLP problem and hard to handle directly. Inspired by the BCD method, we decompose the joint optimization problem into two sub-problems and propose a two-step alternating optimization approach by solving them alternately, aiming to approach the optimal cache decision, spectrum partition coefficient as well as power allocation. Therefore, the subsequent optimization is devoted to dealing with these two small-scale sub-problems iteratively.

A. APT Maximization Problem

Based on the above theoretical analysis with stochastic geometry tools, in order to further explore the potential performance superiority of the cache-enabled mABHetNet, we formulate a joint optimization problem of cache decision, transmit power and spectrum partition to maximize the APT of cache-enabled mABHetNet as \( P \), shown at the bottom of the page.

Constraint (34b) guarantees that the power consumed by the SBS will not exceed the preset power budget \( P_{\text{max}} \). Constraint (34c) ensures that the transmit power is a non-negative value. Constraint (34d) denotes that the number of the files cached at the SBS is a discrete variable from 0 to maximum caching capacity \( C_{\text{max}} \). Constraint (34e) determines that spectrum partition coefficient is a continuous variable from 0 to 1.

From the above expression, we can easily get that the proposed problem \( P \) is a max-min and an MINLP problem, which is intractable to handle directly. To simplify the problem and facilitate the solution, we first introduce an auxiliary variable \( Y \) to convert the problem \( P \) into a more tractable form as \( P_1 \), shown at the bottom of the page.

Then, when the power consumption satisfies \( \rho_s P_{\text{str}} + P_{\text{sc}} + \omega_{\text{ca}} s C < P_{\text{max}} \), the SBS can increase its caching power to \( P_{\text{max}} - \rho_s P_{\text{str}} - P_{\text{sc}} \) for a higher cache hit ratio, so introducing the caching without reducing the transmit power can increase the APT. Increasing the transmit power also facilitates the increment of APT of associated user without the total power constraint at the SBS. Since both transmit power and caching capacity contributes to the throughput of the associated user, the SBS will make full use of total power budget for caching capacity and transmit power in the proposed algorithm to approach a better APT performance, i.e.,

\[
\rho_s P_{\text{str}} + P_{\text{sc}} + \omega_{\text{ca}} s C = P_{\text{max}}. \tag{36}
\]

Thus, we denote that the actual transmit power consumption of the SBS is \( P_{\text{str}} = \frac{P_{\text{max}} - P_{\text{sc}} - \omega_{\text{ca}} s C}{\rho_s} = P_{\text{str}} - \frac{\omega_{\text{ca}} s C}{\rho_s} \), of which \( P_{\text{str}} = \frac{P_{\text{max}} - P_{\text{sc}}}{\rho_s} \) and \( \omega_{\text{ca}} s C = \frac{\omega_{\text{ca}} s C}{\rho_s} \). Then, we will investigate the trade-off between caching capacity and transmit power at SBS.

Remark 3: From the above analysis, increasing the caching capacity leads to the spectrum transfer and increasing the APT. However, from constraint (34b), we can see that APT is not a monotonically increasing function of \( C \). Transmit power and caching capacity are coupled variables. This reveals that increasing the caching capacity is adverse to improving the APT when APT is limited by the transmit power.

B. Sub-Problem: Spectrum Partition Problem

In this subsection, we will give the solution to the spectrum partition problem when the cache decision and transmit power are given. To simplify the notation, we stipulate that \( A_1 = \lambda mW \log_2(1 + \gamma_0), A_2 = \lambda sW \log_2(1 + \gamma_0), A_3 = \lambda mW \log_2(1 + \gamma_0) \).

Then the original problem can be transformed to the spectrum partition problem as \( P_2 \).

\[
P_2: \max_{Y, \eta, C, P_{\text{str}}} Y \tag{37a}
\]

s.t. \( A_1 (1 - p_h)(1 - \eta) P_{\text{cov}}^{bh} + A_2 p_h P_{\text{cov}}^{\eta} + A_3 \eta P_{\text{cov}}^{m} \geq Y, \) \( A_2 \eta P_{\text{cov}}^{m} + A_3 \eta P_{\text{cov}}^{m} \geq Y, \) \( (34c), (34d), (34e) \).

\[
P_1: \max_{Y, \eta, C, P_{\text{str}}} Y \tag{35a}
\]

s.t. \( (1 - p_h) \lambda mW \log_2(1 + \gamma_0) P_{\text{cov}}^{m} \)
\[
+ p_h \lambda sW \log_2(1 + \gamma_0) P_{\text{cov}}^{s} \geq Y, \tag{35b}
\]
\[
(1 - p_h) \lambda sW \log_2(1 + \gamma_0) P_{\text{cov}}^{s} + p_h \lambda sW \log_2(1 + \gamma_0) P_{\text{cov}}^{s} \geq Y, \tag{35c}
\]

(34b), (34c), (34d), (34e).
of which \( P_{k}^{cov} (k \in \{s, m, bh \}) \) is an integral term containing the transmit power of the SBS or MBS. It is obvious that this sub-problem becomes a linear constraint programming problem of variable \( Y \) and \( \eta \), of which \( Y \) is the corresponding \( \text{APT} \) and \( \eta \) is the solution of spectrum partition coefficient. Utilizing interior point method by fmincon function or CVX tools in MATLAB software, the optimal solution \( \eta^* \) and corresponding objective value \( Y^* \) of the linear programming problem \( P2 \) can be obtained. The complexity of obtaining the optimal spectrum partition coefficient (one variable) based on interior-point method is \( O(1) \).

C. Sub-Problems: Cache Decision and Power Allocation Problem

In this subsection, we will optimize the cache decision and power allocation problem for the given spectrum partition coefficient \( \eta \), which is still a discrete, nonlinear and non-convex optimization problem. Therefore, the cache decision and power allocation problem is reformulated as \( P3 \).

\[
\begin{align*}
P3 : & \quad \max_{Y, C, P^T} Y \\
\text{s.t.} & \quad (37b), (37c), (34e), (3Ad), (34e). \quad (38a)
\end{align*}
\]

To further simplify the problem, we stipulate that \( f_1(C) = A_1(1-p_h(C)(1-\eta)P_{bh}^{cov} + A_2p_h(C)\eta P_{bh}^{cov}(P'_{s} - \omega \eta c) + A_3\eta P_{m}^{cov} \), \( f_2(C) = A_2\eta P_{s}^{cov}(P'_{s} - \omega \eta c) + A_3\eta P_{m}^{cov}. \) It is obvious that maximizing the minimum value of these two constraint functions is equivalent to maximizing \( Y \) in \( P3 \). Based on genetic algorithm, we propose a genetic-based cache decision and power allocation algorithm to obtain the approximate optimal solution. As an adaptive stochastic global search optimization algorithm, genetic algorithm has the potential to solve this integer and non-linear programming problem \( P3 \) effectively. Then a Genetic-based Cache Decision and Power Allocation (GCDPA) algorithm is elaborated in Algorithm 1, where the fitness function is set as the objective function as \( f(C) = \min\{f_1(C), f_2(C)\} = f_1(C) + f_2(C) - f_1(C) - f_2(C) \). Thus, the cache decision and transmit power are designed for the SBS. The worst case for cache decision and power allocation algorithm at SBS is to traverse all feasible solutions with complexity \( O(C_{max}) \).

D. Joint Optimization Solution

In view of the fact that this joint optimization problem is a non-convex and multivariate problem, we propose an alternating optimization algorithm to optimize spectrum partition and cache decision, which is elaborated in Algorithm 2. The idea of iterative optimization is inspired by the BCD method proposed in [46]. BCD is a method for optimizing multivariate function even if it is neither necessarily strictly convex nor differentiable [47]. In each iteration, it takes the extremum along the direction of multiple coordinate axes (variables), and one of the coordinates or the coordinate blocks is fixed to optimize another coordinate or coordinate block. Then the new result is immediately substituted into the next iteration. Particularly, the BCD method will converge in finite number of times as long as the two variables have no strong correlation.

The JCSPA algorithm is based on the BCD method to optimize sub-problem of spectrum partition and sub-problem of cache decision and power allocation alternately, so the worst case is to iterate to the maximum iteration times Iter\( \text{max} \). The complexity of entire JCSPA algorithm is upper bound by \( O(1 + C_{max}) \). We further illustrate the convergence and optimality of JCSPA algorithm in the following theorem.

**Theorem 1:** JCSPA algorithm based on the BCD method will converge in finite number of iterations and the convergent output \((\eta^*, C^*, P^T)\) is the approach to the optimal solution to the original problem \( P \).

This theorem can be proved that these two sub-problems have a unique optimal solution. Then based on basic cyclic rule, the consequences from the BCD method is defined and bounded and the constraint set is a Cartesian product of closed convex sets. In proposed JCSPA algorithm, after Proposition 3, the spectrum partition coefficient \( \eta \) and caching decision \( C \) are optimized in turn while keeping the other fixed, with the goal of maximizing APT. After spectrum partition solution for fixed \( C_t \), it will hold that

\[
\text{APT}(\eta^{(t)}, C^{(t)}) \leq \text{APT}(\eta^{(t+1)}, C^{(t)}) \tag{39}
\]

Then after cache decision in Algorithm 1 for given \( \eta^{(t+1)} \), we can get

\[
\text{APT}(\eta^{(t+1)}, C^{(t)}) \leq \text{APT}(\eta^{(t+1)}, C^{(t+1)}) \tag{40}
\]

Combining (39) with (40), it holds that

\[
\text{APT}(\eta^{(t)}, C^{(t)}) \leq \text{APT}(\eta^{(t+1)}, C^{(t+1)}) \tag{41}
\]

This illustrates that APT is non-decreasing in each iteration. The upper bound of the APT is limited to a finite value, so the proposed JCSPA algorithm will converge.

---

**Algorithm 1 Genetic-Based Cache Decision and Power Allocation Algorithm (GCDPA)**

**Input:** Initial \( C = C_1 \); Spectrum partition ratio \( \eta \) from first sub-problem;

Generation size: \( S^P \); Chromosome size: \( S^c \); Population size: \( S^p \); Crossover rate: \( \rho_c \); Mutation rate: \( \rho_m \);

Fitness function: \( f(C) \);

**Output:** Best individual \( C^* \); Best fitness value \( f^*(C) \);

1. Optimize fitness function;
2. \( C^1 = \text{Initialization}(S^c, S^p) \);
3. for \( k = 1; k < S^p; k + + \) do
   4. \( C^*, \text{Calculate fitness value as} \ f^*(C) = f_1(C) + f_2(C) - f_1(C) - f_2(C) \);
5. \( C^k = \text{Selection}(C^k, f^2(C)) \);
6. \( C^k = \text{Crossover}(C^k, \rho_c) \);
7. \( C^k = \text{Mutation}(C^k, \rho_m) \);
8. Return \( C^*, f^*(C) \);
9. end
10. Output \( C^*, f^*(C) \);
| Parameters | Physics meaning | Values |
|------------|-----------------|--------|
| $W$ | Total mmWave bandwidth | 250 MHz |
| $\lambda_s$, $\lambda_m$ | The density of SBS and MBS | $10^{-3}$ BSs/m$^2$, $4 \times 10^{-5}$ BSs/m$^2$ |
| $\alpha_{UL}$, $\alpha_{L}$ | Pathloss parameters of LOS | $10^{-10}$, 2 |
| $\alpha_{NUL}$, $\alpha_{NL}$ | Pathloss parameters of NLOS | $10^{-14}$, 4 |
| $N_0$ | Thermal noise density | $-174$ dBm/Hz |
| $P_{\text{max}}^{S}$, $P_{\text{max}}^{M}$ | Power budget of SBS and MBS | 38.2 dBm, 60 dBm |
| $\gamma_0$ | SINR threshold | 10 dB |
| $P_{\text{APC}}^{S}$, $P_{\text{APC}}^{M}$ | Fixed circuit power at SBS and MBS | 20 dBm, 46 dBm |
| $\mu_s$, $\mu_m$ | Power amplification factor of SBS and MBS | 1, 1.5 |
| $B_s$, $B_m$ | Association biases of SBS and MBS | 10, 5 |
| $\theta$ | Blockage density | $2 \times 10^{-3}$ |
| $\theta$ | Main lobe beamwidth | 30° |
| $M$, $m$ | Main lobe, side lobe antenna gain | 10dB, -10dB |
| $\omega_{CA}$ | Caching power coefficient | $6.25 \times 10^{-12}$ W/bit |
| $s$ | Size of each file | 100 MB |
| $F$ | The file units in the file library | 1000 |
| $C_{\text{max}}$ | Maximum caching capacity/file units of SBS | 800 |
| $\gamma_F$ | Zipf parameter | 1.0 |

Algorithm 2 Joint Cache Decision, Spectrum Partition and Power Allocation Alternate Optimization Algorithm (JCSPA)

**Input:** Parameter of simulation scenario; Maximum number of iterations $\text{Iter}_{\text{max}}$, $\epsilon$

**Output:** Optimal cache decision $C^*$, spectrum partition $\eta^*$ and transmit power $P_{tr}^*$.

1. **Initial procedure:**
   2. Set $t = 0$;
   3. Starting point $C^{(0)}$, $A^P{(t)} = 0$;
   4. **end Initial procedure:**
   5. **repeat**
   6. Set $t = t + 1$;
   7. Update $\eta^{(t)}$ for fixed $C^{(t)}$ according to Eq. (37a) and send it to next sub-problem, then calculate the corresponding $A^P{(t)}$ according to Eq. (34a);
   8. Solve sub-problem of cache decision for given $\eta^{(t)}$ based on Algorithm 1 to get $C^{(t)}$, then update the corresponding $A^P{(t)}$ according to Eq. (34a), and send $C^{(t)}$ to the $(t + 1)$th iteration;
   9. **until** $t = \text{Iter}_{\text{max}}$ or $\|A^P{(t)} - A^P{(t-1)}\| \leq \epsilon$;
   10. $C^* \leftarrow C^{(t)}$, $\eta^* \leftarrow \eta^{(t)}$, compute $P_{tr}^*$ according to Eq. (36);
   11. **Return** $C^*, \eta^*, P_{tr}^*$.

**V. PERFORMANCE EVALUATION**

In this section, we use numerical simulation results to validate and evaluate of APT of the cache-enabled mABHetNet. Particularly, we compare APT under different scenarios.

**A. Parameter Setting**

Unless otherwise specified, the parameters of the simulation scenario are set as follows. The BSs and users are distributed in a square area $1000m \times 1000m$. Note that the density of the users is assumed to be sufficiently larger than that of the BS, we assume that the BS is active in the downlink. In other words, each SBS is in a backhaul association cell and has at least one associated user in its coverage. Some default simulation configurations are listed in Table II, based on 3GPP specification and literature [13], [25], [41], [42], [43], [44]. All the above settings can be changed according to different scenarios.

**B. Convergence of JCSPA**

In this subsection, simulations are carried out to validate the convergence performance and maximum APT of the proposed JCSPA algorithm for different skewness of caching file popularity (from 0.8 to 1.4) as Fig. 3. As expected, we can see that the joint optimization problem converges to maximum APT after approximately no more than five iterations. It can be concluded that our proposed algorithm is fast-convergent.

**C. Performance of APT Under Different Caching Capacity and Spectrum Partitions**

In Fig. 4(a), the performance of APT with respect to both the caching capacity and spectrum partition is shown. From this three-dimensional APT, both caching file capacity and spectrum partition ratio have a prominent impact on APT.
APT is not monotonically increasing or decreasing with respect to these two variables.

In order to further observe their quantitative relationship, Fig. 4(b) shows the APT with respect to different caching capacity under four spectrum partition coefficients. For a lower spectrum partition ratio, APT will decrease with increasing caching capacity. This is because the smaller spectrum can only satisfy the partial data transmission of the access link, and the power consumption by caching reduces the transmit power of the SBS, which further reduces the APT. While for higher spectrum partition ratio, APT will first increase with increasing caching capacity and then decrease. The reason is that as more caching files can improve the cache hit ratio of files at SBS and less files’ delivery occupy backhaul resources, more files can be obtained through access link directly without backhaul link. However, APT will decrease when the caching capacity is more than about 200 file units. This is mainly because maximum power of SBS is limited and too much caching capacity consumes too much power, therefore the transmit power of SBS will be reduced, which limits the performance of APT. This illustrates that increasing the caching capacity does not necessarily increase the APT when the maximum power of SBS is limited.

Fig. 4(c) shows the APT of different spectrum partitions under four given caching decisions. In the case of no cache, APT will increase and then decrease with the increasing spectrum partition coefficient. With DSA strategy, there are about 60% spectrum resources allocated for the wireless backhaul link and 40% spectrum resources allocated for the wireless access link in the case of no cache. When the cache is introduced into the mABHetNet, the spectrum resources of the wireless backhaul link begin to be transferred to the wireless access link and about 80% spectrum resources are allocated for the access link. Thus, the introduction of cache at the SBS also increases APT accordingly. And the optimal caching scheme approaches the best performance. This reveals that caching files at SBS can transfer the spectrum from the backhaul link to the access link, which further improves the APT.

D. Comparison With Baseline Algorithms

In this subsection, we compare the proposed JCSPA algorithm with three basic algorithms.

- **No Cache, Full transmit Power with DSA (NCFP+DSA):** This algorithm is used to perform the dynamic spectrum allocation while the SBS does not cache any files and transmits with full power in the network. This means that all files are cache-miss and the file delivery needs to go through both the access and backhaul link.

- **Optimal Cache and transmit Power with FSA (OCP+FSA):** This algorithm makes the optimal caching capacity and power allocation decisions at the SBS for the fixed spectrum allocation scheme that the wireless backhaul and access link accounts for the half of the total spectrum. The SBS needs to decide how many most popular files in file library to cache and how much power to transmit.

- **Full Cache, Minimum transmit Power with DSA (FCMP+DSA):** In this algorithm, the SBS makes use of all the caching capacity to store the $C_{\text{max}}$ most popular file units and transmits with minimum power. The spectrum partition is based on the DSA strategy.

- **Uniform Cache, Optimal transmit Power with DSA (UCOP+DSA):** This baseline algorithm is based on caching files uniformly with the same caching capacity and power allocation of (OCP), where each file is cached with identical popularity. The spectrum partition scheme is DSA.

In Fig. 5(a), we begin evaluating APT in terms of different skewness of file popularity among proposed JCSPA and other three benchmark algorithms. In the cases of no cache and full transmit power, APT will not change with $\gamma_p$ but the DSA exceeds the FSA algorithm. And APT of full cache combined with DSA is better than that of optimal cache with FSA. Such result reveals the significance of the dynamic spectrum partition. Besides, FCMP with DSA scheme performs better than OCP with FSA scheme, this phenomenon reveals the significant gain of dynamic spectrum allocation on APT performance. In contrast, our proposed JCSPA algorithm achieves the best APT performance compared with other baseline algorithms. This illustrates the effectiveness of proposed JCSPA algorithm.

Fig. 5(b) shows the APT with different SINR thresholds under different algorithms. Note that APT increases as SINR threshold increases when the SINR threshold is at a low level,
this is because increasing the SINR threshold $\gamma_0$ will increase the transmission rate dominates. While for larger SINR threshold, it improves the demodulation threshold of the received signal and reduces the coverage probability, so that APT will decrease with the increasing SINR threshold. Besides, the performance of JCSPA algorithm still significantly outperforms other basic algorithms.

From Fig. 5(c) we can see that the optimal caching capacity is reduced with the increasing the skewness while the maximum APT will increase. Comparing the two cases where $\gamma_p = 1.4$ and $\gamma_p = 0.8$, APT under JCSPA algorithm is increased by nearly 40% while the optimal caching capacity is reduced by more than 70%. In other words, only a smaller caching capacity is needed to reach the maximum APT at a higher skewness. The reason is that higher skewness helps to improve cache hit ratio, so only fewer caching file units are needed to achieve higher APT.

**E. Impact of Other Cache Parameters on APT**

In order to obtain the impact of cache-related parameters on APT, we further give the APT with respect to the caching capacity of different skewness of file popularity and caching power coefficients. The skewness $\gamma_p$ and caching power coefficient reflect the main characteristics of the caching. Fig. 6(a) shows the APT under different caching capacity for four skewness of caching files. It can be observed that APT first increases with caching capacity and then decreases, as well as higher skewness can lead to higher APT. This is because that higher skewness increases the cache hit ratio $p_h$ then increases the APT. Fig. 6(b) shows the APT under different caching capacity for four caching power coefficients from $5.25 \times 10^{-12}$ to $8.25 \times 10^{-12}$. When the caching capacity is small, APT is almost equal under different cache power consumption coefficients, but with the increasing of caching capacity, for the case of high cache power consumption coefficient, APT will decrease rapidly as the cache increases, while APT with lower cache power consumption coefficient decreases slowly. This is because the maximum power of SBS is limited and higher $\omega_{ca}$ will lead to more caching power consumption for the same number of caching files, thereby reducing the transmit power and decreasing the APT.

This prompts us to use storage technologies with lower caching power coefficient to improve network performance.

**VI. CONCLUSION**

In this paper, to cope with the “spectrum occupancy” problem, we develop a tractable cache-enabled mABHetNet by stochastic geometry tools and investigate the impact of spectrum partition between the access link and the backhaul link as well as caching capacity on network performance. Considering the effect of blockage and association probability,
we analyze the SINR distribution and derive the expression of APT with respect to caching capacity, spectrum partition and transmit power. Based on the analytical work, we optimize the cache decision, power allocation and spectrum partition problem jointly to obtain the maximum APT by the proposed algorithm. Through numerical simulation, we first evaluate the convergence of our proposed algorithm. Then we find that there exist the optimal caching capacity and spectrum partition to maximize APT, which verifies the effectiveness of our proposed algorithm. Besides, we explore the impact of some cache-related factors on APT. The results show up to 90% APT gain of appropriate caching capacity and spectrum partition compared with traditional mABHetNets.

Cooperative caching is a promising technology for content storage and file delivery. When the cooperative caching is applied, the user may be associated with MBS and SBS simultaneously. The number of serving BSs is undetermined and multiple BSs need to cooperate with each other for the typical user, which needs to be explored comprehensively. Thus, cooperative caching scenario can be expanded to our future work.

APPENDIX

A. Proof of Proposition 1

Then we first focus on the SINR distribution of a user covered by SBS:

\[ P_{s}^{\text{cov}}(\gamma) = P_{s,L}^{\text{cov}}(\gamma) + P_{s,NL}^{\text{cov}}(\gamma), \]

where the SINR distribution of a user covered by LOS/NLOS SBS:

\[ P_{s,k}^{\text{cov}}(\gamma) = \mathbb{E}_{r} \left[ \mathbb{P} \left[ \text{SINR}^k_s(r) > \gamma \right] \right] = \int_{0}^{\infty} \mathbb{P} \left[ \text{SINR}^k_s(r) > \gamma \right] F^k_s(r)dr, \]

where \( k \in \{L, NL\} \) denotes the LOS or NLOS transmission link. \( \gamma \) is the threshold for successful demodulation and decoding at the receiver. \( \mathbb{P} \left[ \text{SINR}^L_s(r) > \gamma \right] \) means the probability of the event that the SINR of the user covered by SBS is over \( \gamma \) via the LOS path at distance \( r \):

\[
\mathbb{P} \left[ \text{SINR}^k_s(r) > \gamma \right] = \mathbb{P} \left[ \frac{P_s^r B_s G_s h_s A_k r^{-\alpha_k}}{I_s + I_m + N_0} > \gamma \right] = \mathbb{P} \left[ h_s > \gamma (I_s + I_m + N_0) \right] \frac{\exp \left( -\frac{\gamma N_0}{P_s^r B_s G_s A_k r^{-\alpha_k}} \right)}{\mathcal{L}^{k}_{s,m}(\gamma^{\alpha_k})},
\]

where step (a) follows from small fading \( h \sim \exp(1) \). \( \mathcal{L}^{k}_{s,m}(\gamma^{\alpha_k}) \) is the Laplace transform of the cumulative interference from the SBS tier and the MBS tier, shown at the bottom of the page, where \( p_{G_i} \) and \( p_{G_i} \) is the probability of the antenna gain taking corresponding value from SBS interference tier and MBS interference tier. Step (b) is based on the PGFL of PPP [45], \( d_1 = \frac{P_s^r B_s A_s}{P_m^r B_m A_m} \) and \( d_2 = \frac{P_s^r B_s A_s}{P_m^r B_m A_m A_s} \). Following the same logic, other Laplace transforms of cumulative interference \( \mathcal{L}^{L}_{CPU}(\gamma^{\alpha_{NL}}) \), \( \mathcal{L}^{L}_{L_{h,b}}(\gamma^{\alpha_{NL}}) \), \( \mathcal{L}^{NL}_{L_{h,b}}(\gamma^{\alpha_{NL}}) \) can also be obtained.

Similarly, the SINR distribution of SBS is covered by MBS:

\[
\mathbb{P} \left[ \text{SINR}^L_{bb}(r) > \gamma \right] = \mathbb{P} \left[ \frac{P_s^r B_s G_s h_s A_k r^{-\alpha_k}}{I_s + I_m + N_0} > \gamma \right] \mathcal{L}^{L}_{L_{h,b}}(\gamma^{\alpha_{NL}}) + \mathbb{P} \left[ \text{SINR}^L_{bb}(r) > \gamma \right] \mathcal{L}^{NL}_{L_{h,b}}(\gamma^{\alpha_{NL}}).
\]

B. Proof of Proposition 2

In noise-limited scenario, the interference is close to zero, so the SINR distribution can be reduced to the SNR distribution as

\[
P_{k}^{\text{cov}}(\gamma) = \mathbb{P}[\frac{P_k B_k G_k A r^{-\alpha}}{\sigma^2} > \gamma] = \mathbb{P}[\frac{\gamma}{\sigma^2} < \frac{P_k B_k G_k}{\sigma^2}],
\]

where \( A \in \{A_L, A_NL\} \), \( A \in \{A_L, A_NL\} \). Considering the effect of blockage, we define a point process

\[
\mathcal{L}^{L}_{L_{s,m}}(\gamma^{\alpha_{NL}}) = \prod_{G_i} \exp \left( -2\pi \lambda_s p_{G_i} \int_{r} \left( \frac{P_L(\gamma u)}{1 + \frac{P_s^r B_s G_s A_k r^{-\alpha_k}}{\gamma P_s^r B_s G_s A_k r^{-\alpha_k}}} \right) du \right) \times \prod_{G_i} \exp \left( -2\pi \lambda_s p_{G_i} \int_{(d_1)^{+1}} \left( \frac{\mathcal{P}_{NL}(\gamma u)}{1 + \frac{P_s^r B_s G_s A_k r^{-\alpha_k}}{\gamma P_s^r B_s G_s A_k r^{-\alpha_k}}} \right) du \right) \times \prod_{G_i} \exp \left( -2\pi \lambda_s p_{G_i} \int_{(d_2)^{+1}} \left( \frac{\mathcal{P}_{NL}(\gamma u)}{1 + \frac{P_s^r B_s G_s A_k r^{-\alpha_k}}{\gamma P_s^r B_s G_s A_k r^{-\alpha_k}}} \right) du \right),
\]
\[ P_k^{\text{cov}}(\gamma) = \mathbb{P}\left[ \frac{\pi^2}{\alpha} < \frac{P_k B_k G_k}{\sigma^2} \right] = \mathcal{F}_\Xi\left( \frac{P_k B_k G_k}{\sigma^2} \right) \]

\[ = 1 - \exp \left( -\pi \lambda_k A_{\text{NL}} \frac{2}{\alpha} \Gamma \left( \frac{1}{\alpha} + 1 \right) \left( \frac{P_k B_k G_k}{\sigma^2} \right) \frac{\alpha}{\alpha_{\text{NL}}} - 2\pi \lambda_k Y \left( \frac{P_k B_k G_k}{\sigma^2} \right) \right) . \]  

(53)

\[ \Lambda = \{ \frac{\phi}{\alpha} \} = \{ \phi \}. \] The intensity measure of this point process can be calculated according to the Mapping theorem as

\[ \Lambda(\phi) = \int_0^{(A_{\text{NL}} \phi)^{\frac{1}{\alpha}}} 2\pi \lambda_k u e^{-\beta u} du \]

\[ + \int_0^{(A_{\text{NL}} \phi)^{\frac{1}{\alpha}}} 2\pi \lambda_k u (1 - e^{-\beta u}) du, \]  

(47)

where \( k \in \{ m, s \} \) denotes the MBS tier or SBS tier. Then, the density function is

\[ \lambda(\phi) = \frac{d\Lambda(\phi)}{d\phi} = 2\pi \lambda_k A_{\text{NL}} \frac{\alpha_{\text{NL}}}{\alpha} (A_{\text{NL}} \phi)^{\frac{1}{\alpha} - 1} \]

\[ + 2\pi \lambda_k (A_{\text{NL}} \phi)^{\frac{1}{\alpha} - 1} - 2\pi \lambda_k (A_{\text{NL}} \phi)^{\frac{1}{\alpha} - 1} \frac{\alpha_{\text{NL}}}{\alpha} e^{(A_{\text{NL}} \phi)^{\frac{1}{\alpha}}} \frac{1}{\alpha}. \]  

(48)

Then, the joint distribution function and probability density function of \( \{ \phi, g \} \) can be given as

\[ \mathbb{P}\left[ \frac{\phi}{g} \leq \xi \right] = \mathbb{P}[g \geq \frac{\phi}{\xi}] = 1 - F_\phi\left( \frac{\phi}{\xi} \right), \]  

(49)

\[ \rho(\phi, \xi) = \frac{d(1 - F_\phi(\frac{\phi}{\xi}))}{d\xi} = \frac{\phi}{\xi^2} f_\phi(\frac{\phi}{\xi}) = \frac{\phi}{\xi^2}, \]  

(50)

where step (a) is based on the Rayleigh fading channel with exponential distribution (i.e., \( h \sim \exp(1) \)). Based on the displacement theorem, the density function of process \( \Xi = \{ \frac{\phi}{\alpha} \} = \{ \xi \} \) can be obtained as

\[ \lambda_\Xi(\xi) = \int_0^\infty \lambda(\phi) \rho(\phi, \xi) d\phi \]

\[ = \frac{2\pi \lambda_k A_{\text{NL}}^{\frac{1}{\alpha}}}{\alpha_{\text{NL}}} \xi^{\frac{1}{\alpha} - 1} \Gamma\left( \frac{2}{\alpha} + 1 \right) \frac{2}{\alpha_{\text{NL}}} + \frac{2\pi \lambda_k}{\xi^2} (H_L - H_{\text{NL}}), \]  

(51)

where \( H_i = \frac{A_{\text{NL}}^{\frac{1}{\alpha}}}{\alpha_{\text{NL}}} \int_0^\infty \phi^{\frac{1}{\alpha}} \exp\left( -\beta(A_{\text{NL}} \phi)^{\frac{1}{\alpha}} - \frac{\phi}{\xi} \right) d\phi, i \in \{ L, \text{NL} \} \). Now, based on the complementary void function of PPP, the CDF of \( \xi \) is given by

\[ F_\xi(\xi_0) = \mathbb{P}[\xi < \xi_0] = 1 - \mathbb{P}[\Xi(\xi_0) = 0] \]

\[ = 1 - \exp \left( -\pi \lambda_k A_{\text{NL}}^{\frac{1}{\alpha}} \Gamma\left( \frac{1}{\alpha} + 1 \right) \frac{\alpha_{\text{NL}}}{\alpha} \frac{\xi_0}{\alpha_{\text{NL}}} - 2\pi \lambda_k Y \left( \frac{P_k B_k G_k}{\sigma^2} \right) \right), \]  

(52)

where \( Y(\xi_0) = \int_0^{\xi_0} \frac{A_{\text{NL}}^{\frac{1}{\alpha}}}{\alpha_{\text{NL}} \xi^2} \int_0^\infty \phi^{\frac{1}{\alpha}} \exp\left( -\beta(\phi)^{\frac{1}{\alpha}} - \frac{\phi}{\xi} \right) d\phi d\xi - \int_0^{\xi_0} \frac{A_{\text{NL}}^{\frac{1}{\alpha}}}{\alpha_{\text{NL}} \xi^2} \int_0^\infty \phi^{\frac{1}{\alpha}} \exp\left( -\beta(\phi)^{\frac{1}{\alpha}} - \frac{\phi}{\xi} \right) d\phi d\xi. \) Then we can obtain Eq. (53), shown at the top of the page.

\section*{ACKNOWLEDGMENT}

Compared with conference version, the authors have formulated an APT maximization problem and solved it effectively. Extensive numerical results have been carried out to illustrate the effectiveness and feasibility of the proposed algorithm.

\section*{REFERENCES}

[1] C. Zhang, H. Wu, H. Lu, and J. Liu, “Throughput analysis in cache-enabled millimeter wave HetNets with access and backhaul integration,” in \textit{Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)}, May 2020, pp. 1–6.

[2] \textit{Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017–2022}, Cisco, San Jose, CA, USA, Feb. 2019.

[3] C. Madapatha et al., “On integrated access and backhaul networks: Current status and potentials,” \textit{IEEE Open J. Commun. Soc.}, vol. 1, pp. 1374–1389, 2020.

[4] C. Dehos, J. L. González, A. De Domenico, D. Kiéns, and L. Dussopt, “Millimeter-wave access and backhauling: The solution to the exponential data traffic increase in 5G mobile communications systems?” \textit{IEEE Commun. Mag.}, vol. 52, no. 9, pp. 88–95, Sep. 2014.

[5] R. Taori and A. Sridharan, “Point-to-multipoint in-band mmWave backhaul for 5G networks,” \textit{IEEE Commun. Mag.}, vol. 53, no. 1, pp. 195–201, Jan. 2015.

[6] R. J. Weiler et al., “Enabling 5G backhaul and access with millimeter-waves,” in \textit{Proc. Eur. Conf. Netw. Commun. (EUCNC)}, Bologna, Italy, Jun. 2014, pp. 1–5.

[7] \textit{NR: Study on Integrated Access and Backhaul}, document TR 38.874, 3GPP, 2017.

[8] A. Alammouri, J. G. Andrews, and F. Baccelli, “A unified asymptotic analysis of area spectral efficiency in ultradense cellular networks,” \textit{IEEE Trans. Inf. Theory}, vol. 62, no. 2, pp. 1236–1248, Feb. 2019.

[9] X. Zhang and J. G. Andrews, “Downlink cellular network analysis with multi-slope path loss models,” \textit{IEEE Trans. Commun.}, vol. 63, no. 5, pp. 1881–1894, May 2015.

[10] I. Atzeni, J. Arnau, and M. Kountouris, “Downlink cellular network analysis with LOS/NLOS propagation and elevated base stations,” \textit{IEEE Trans. Wireless Commun.}, vol. 17, no. 1, pp. 142–156, Jan. 2018.

[11] W. Yi, Y. Liu, and A. Nallanathan, “Modeling and analysis of D2D millimeter-wave networks with Poisson cluster processes,” \textit{IEEE Trans. Commun.}, vol. 65, no. 12, pp. 5574–5588, Dec. 2017.

[12] C. Saha and H. Dhillon, “Millimeter wave integrated access and backhaul in 5G: Performance analysis and design insights,” \textit{IEEE J. Sel. Areas Commun.}, vol. 37, no. 12, pp. 2669–2684, Dec. 2019.

[13] Z. Shi, Y. Wang, L. Huang, and T. Wang, “Dynamic resource allocation in mmWave unified access and backhaul network,” in \textit{Proc. PMRC}, Hong Kong, 2015, pp. 2260–2264.

[14] C. Saha, M. Afshang, and H. S. Dhillon, “Bandwidth partitioning and downlink analysis in millimeter wave integrated access and backhaul for 5G,” \textit{IEEE Trans. Wireless Commun.}, vol. 17, no. 12, pp. 8195–8210, Dec. 2018.

[15] M. Diamanti, P. Charatsaris, E. E. Tsiropoulou, and S. Papavassiliou, “The prospect of reconfigurable intelligent surfaces in integrated access and backhaul networks,” \textit{IEEE Trans. Green Commun. Netw.}, vol. 6, no. 2, pp. 859–872, Jun. 2022.
ZHANG et al.: ANALYSIS AND OPTIMIZATION OF CACHE-ENABLED mmWave HetNets WITH IAB

[16] H. Alghafari, M. S. Haghighi, and A. Jolfaei, “High bandwidth green communication with vehicles by decentralized resource optimization in integrated access backhaul 5G networks,” IEEE Trans. Green Commun. Netw., vol. 6, no. 3, pp. 1438–1447, Sep. 2022.

[17] J.-H. Kwon, B. Lim, and Y.-C. Ko, “Resource allocation and system design of out-band based integrated access and backhaul network at mmWave band,” IEEE Trans. Veh. Technol., vol. 71, no. 6, pp. 6503–6517, Jun. 2022.

[18] D. Liu, B. Chen, C. Yang, and A. F. Molisch, “Caching at the wireless edge: Design aspects, challenges, and future directions,” IEEE Commun. Mag., vol. 54, no. 9, pp. 22–28, Sep. 2016.

[19] M. Tao, E. Chen, H. Zhou, and W. Yu, “Content-centric sparse multicast beamforming for cache-enabled cloud RAN,” IEEE Trans. Wireless Commun., vol. 15, no. 9, pp. 6118–6131, Sep. 2016.

[20] X. Xu and M. Tao, “Modeling, analysis, and optimization of coded caching in small cell networks,” IEEE Trans. Commun., vol. 65, no. 8, pp. 3415–3428, Aug. 2017.

[21] 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.

[22] Y. Chiang and W. Liao, “ENCORE: An energy-aware multicell cooperation in heterogeneous networks with content caching,” in Proc. IEEE INFOCOM, San Francisco, CA, USA, 2016, pp. 1–9.

[23] G. Auer et al., “How much energy is needed to run a wireless network?” IEEE Trans. Wireless Commun., vol. 18, no. 5, pp. 40–49, Oct. 2011.

[24] L. Shi, L. Zhao, G. Zheng, Z. Han, and Y. Ye, “Incentive design for cache-enabled D2D underlaid cellular networks using Stackelberg game,” IEEE Trans. Veh. Technol., vol. 68, no. 1, pp. 765–779, Jan. 2019.

[25] D. Liu and C. Yang, “Energy efficiency of downlink networks with caching at base stations,” IEEE J. Sel. Areas Commun., vol. 34, no. 4, pp. 907–922, Apr. 2016.

[26] Z. Gu, H. Lu, and Z. Zhu, “On throughput optimization and bound analysis in cache-enabled fiber-wireless networks,” IEEE Trans. Veh. Technol., vol. 69, no. 8, pp. 9068–9082, Aug. 2020.

[27] 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.

[28] X. Zhang, B. Zhang, K. An, G. Zheng, S. Chatzinotas, and D. Guo, “Stochastic geometry-based analysis of cache-enabled hybrid satellite-terrestrial networks with non-orthogonal multiple access,” IEEE Trans. Wireless Commun., vol. 21, no. 2, pp. 1272–1287, Feb. 2022.

[29] B. Dai, Y.-F. Liu, and W. Yu, “Optimized base-station cache allocation for cloud radio access network with multicast backhaul,” IEEE J. Sel. Areas Commun., vol. 36, no. 8, pp. 1737–1750, Aug. 2018.

[30] J. Llorca et al., “Dynamic in-network caching for energy efficient content delivery,” in Proc. ACM INFOCOM, Apr. 2013, pp. 243–249.

[31] P. Gill, M. Afridi, Z. Li, and A. Mahanti, “Youtube traffic characterization: A view from the edge,” in Proc. 7th ACM SIGCOMM Conf. Internet Meas., San Diego, CA, USA, Oct. 2007, pp. 15–28.

[32] L. Wang, K.-K. Wong, S. Lambbotharan, A. Nallanathan, and M. Elkhashlan, “Edge caching in dense heterogeneous cellular networks with massive MIMO-aided self-backhaul,” IEEE Trans. Wireless Commun., vol. 17, no. 9, pp. 6360–6372, Sep. 2018.

[33] M. Cha, P. Rodriguez, J. Crowcroft, S. Moon, and X. Amatriain, “Watching television over an IP network,” in Proc. 8th ACM SIGCOMM Conf. Internet Meas., Oct. 2008, pp. 126–134.

[34] X. Zhang, T. Lv, W. Ni, J. M. Cioffi, N. C. Beaulieu, and Y. J. Guo, “Energy-efficient caching for scalable videos in heterogeneous networks,” IEEE J. Sel. Areas Commun., vol. 36, no. 8, pp. 1802–1815, Aug. 2018.

[35] J. Llorca, A. M. Tulino, M. Varvello, J. Esteban, and D. Perino, “Energy efficient dynamic content distribution,” IEEE J. Sel. Areas Commun., vol. 33, no. 12, pp. 2826–2836, Dec. 2015.

[36] G. Quer, I. Pappalardo, B. D. Rao, and M. Zorzi, “Proactive caching strategies in heterogeneous networks with device-to-device communications,” IEEE Trans. Wireless Commun., vol. 17, no. 8, pp. 5270–5281, Aug. 2018.

[37] A. K. Gupta, J. G. Andrews, and R. W. J. Heath, “On the feasibility of sharing spectrum licenses in mmWave cellular systems,” IEEE Trans. Commun., vol. 64, no. 9, pp. 3981–3995, Sep. 2016.

[38] T. Bai, R. Vaze, and R. W. Heath Jr., “Analysis of blockage effects on urban cellular networks,” IEEE Trans. Wireless Commun., vol. 13, no. 9, pp. 5070–5083, Sep. 2014.

[39] S. Singh, M. N. Kulkarni, A. Ghosh, and J. G. Andrews, “Tractable model for rate in self-backhauled millimeter wave cellular networks,” IEEE J. Sel. Areas Commun., vol. 33, no. 10, pp. 2196–2211, Oct. 2015.

[40] Y. Ye, S. Huang, M. Xiao, Z. Ma, and M. Skoglund, “Cache-enabled millimeter wave cellular networks with clusters,” IEEE Trans. Commun., vol. 68, no. 12, pp. 7752–7745, Dec. 2020.

[41] Radio Frequency (RF) System Scenarios (Release 12), document TR 36.942, V12.0.0, 3GPP, Sep. 2010.

[42] M. Peng, K. Zhang, J. Jiang, J. Wang, and W. Wang, “Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks,” IEEE Trans. Veh. Technol., vol. 64, no. 11, pp. 5275–5287, Nov. 2015.

[43] C.-H. Liu and K. L. Feng, “Fundamentals of the downlink green coverage and energy efficiency in heterogeneous networks,” IEEE J. Sel. Areas Commun., vol. 34, no. 12, pp. 3271–3287, Dec. 2016.

[44] J. G. Andrews, F. Baccelli, and R. K. Ganti, “A tractable approach to coverage and rate in cellular networks,” IEEE J. Sel. Areas Commun., vol. 59, no. 4, pp. 3122–3134, Nov. 2011.

[45] D. P. Bertsekas, Nonlinear Programming, 2nd ed. Belmont, MA, USA: Athena Scientific, 1999.

[46] P. Tseng, “Convergence of a block coordinate descent method for nondifferentiable minimization,” J. Optim. Theory Appl., vol. 109, no. 3, pp. 475–494, 2001.

[47] C. Zhang, H. Lu, and Z. Gu, “Analysis and optimization of cache-enabled mmWave HetNets with integrated access and backhaul,” 2022, arXiv:2201.01037.

Chenwu Zhang received the B.E. degree from the School of Information Engineering, Wuhan University of Technology, Wuhan, China, in 2019. He is currently pursuing the Ph.D. degree with the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China. His research interests include user-centric networks and communication resource allocation in wireless networks.

Hancheng Lu (Senior Member, IEEE) received the Ph.D. degree in communication and information systems from the University of Science and Technology of China (USTC), Hefei, China, in 2005. He is currently a Professor with the Department of Electronic Engineering and Information Science, USTC. His research interests include multimedia communication and networking and resource optimization in wireless heterogeneous networks.

Zhongia Gu received the B.S. degree from the School of Communication and Information Engineering, Shanghai University (SHU), Shanghai, China, in 2017, and the Ph.D. degree from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, in 2022. His research interests include resource allocation, millimeter-wave communications, and stochastic geometry.