

# Cache-enabled HetNets with Limited Backhaul: A Stochastic Geometry Model

Congshan Fan, *Student Member, IEEE*, Tiankui Zhang, *Senior Member, IEEE*,  
Yuanwei Liu, *Member, IEEE* and Zhiming Zeng, *Member, IEEE*

**Abstract**—With the rapid explosion of data volume from mobile networks, edge caching has received significant attentions as an efficient approach to boost content delivery efficiency by bringing contents near users. In this article, cache-enabled heterogeneous networks (HetNets) considering the limited backhaul is analyzed with the aid of the stochastic geometry approach. A hybrid caching policy, in which the most popular contents are cached in the macro BSs tier with the deterministic caching strategy and the less popular contents are cached in the helpers tier with the probabilistic caching strategy, is proposed. Correspondingly, the content-centric association strategy is designed based on the comprehensive state of the access link, the cache and the backhaul link. Under the hybrid caching policy, new analytical results for successful content delivery probability, average successful delivery rate and energy efficiency are derived in the general scenario, the interference-limited scenario and the mean load scenario. The simulation results show that the proposed caching policy outperforms the most popular caching policy in HetNets with the limited backhaul. The performance gain is dramatically improved when the content popularity is less skewed, the cache capacity is sufficient and the helper density is relatively large. Furthermore, it is confirmed that there exists an optimal helper density to maximize the energy efficiency of the cache-enabled HetNets.

**Index Terms**—Edge caching, heterogeneous networks, limited backhaul, stochastic geometry

## I. INTRODUCTION

Heterogeneous networks (HetNets), where small base stations (SBSs) are embed into the existing macro cells, are widely deployed as an effective method to increase the data rate of the radio links. However exiting backhaul links fail to provide large capacity at a relatively affordable cost. The heavy traffic cause severe congestion in the backhaul link, making it a bottleneck in improving the system throughput. Numerous of research contributions reveal that the vast majority of the content requests are generated by duplicating downloads of a few popular contents [1]. By taking advantage of the redundancy of the content requests and the abundance of the cache resource, edge caching where content is cached in the base stations [2, 3] or user equipments [4, 5] has been proposed for backhaul traffic releasing and content access delay reducing [6]. In cache-enabled HetNets, contents are proactively stored at BSs during off-peak time and users can

get the content from the local cache. By densely deploying the SBSs and bringing the content closer to users, the cache-enabled HetNets can greatly improve the system performance. Meanwhile, the combination of the HetNets and the BSs caching make the access protocol, user association, resource allocation, caching strategy and content delivery great change.

For an in-depth sight into the cache-enabled HetNets, extensive contributions have been carried out. Femto-caching system was proposed in [7] for the first time. The author analyzed the expected downloading time under two types of the coded and the uncoded content placement and solved the optimum content assignment problem. The potential of the energy efficiency (EE) in the cache-enabled wireless access networks was explored in [8]. The condition when the EE can benefit from caching, the relationship between the EE and the memory and the maximal EE gain were analyzed sequentially. By effectively exploiting the multicast opportunities, [9] designed a novel caching scheme to significantly reduce the traffic. However, the above research are conducted based on the regular hexagonal or the grid topology. The regular hexagonal topology and the grid topology are idealistic compared with the actual scenario and not suitable for modeling the large-scale networks. Stochastic geometry is an effective method to capture the randomness and the complexity of node distributions in the HetNets. In stochastic geometry models, the nodes are typically distributed according to the Poisson point processes (PPPs) in the two dimensional plane, enabling to characterize various system performance analytically, such as coverage probability, average rate, delay and so on.

### A. Related Works

The cache-based content delivery in the three-tier HetNets was proposed in [10], where BSs, relays and users cooperated to transmit contents. The outage probability and the average ergodic rate were theoretically elaborated. The energy efficiency of a cache enabled two-tier HetNet was analyzed in [11] under co-channel and orthogonal channel deployment scenarios. The authors of [12] investigated the average delay of users based on three different content popularity models.

The content placement in the aforementioned work is according to the deterministic caching strategy. An alternative strategy is the probabilistic caching strategy where a particular content is stored in the node with a caching probability. The authors in [3] derived the closed-form expressions for the coverage probability and local delay experienced by a typical user with the probabilistic caching strategy. The work in [13] focused on the interplay of caching and spectrum sharing

C. Fan, T. Zhang and Z. Zeng are with Beijing University of Posts and Telecommunications, Beijing, (email:{fcs, zhangtiankui, zengzm}@bupt.edu.cn).

Y. Liu is with Queen Mary University of London, London, (e-mail: yuanwei.liu@qmul.ac.uk).

This work is supported by the National Natural Science Foundation of China (No. 61971060).

under the probabilistic caching framework in the HetNets and characterized the outage probability in serving users' requests over the coverage area. A hybrid cache-enabled HetNet where macro BSs with the traditional sub-6 GHz are overlaid by dense millimeter wave pico BSs was considered in [14]. The success probability and the area spectral efficiency were discussed under two user association strategies, namely, the maximum received power scheme and the maximum rate scheme. With the successful transmission probability as the performance metric, [15] focused on the analysis of joint random caching and random DTX under two scenarios of the high mobility scenario and the static scenario. In addition to the static caching strategy, the dynamics of the cache replacement algorithm is incorporated into the analysis. An information centric modeling of the cache enabled cellular networks was investigated in [16] where both the MBS and the SBS can perform caching and operate under the least recently used (LRU) based content eviction policy. Based on the obtained performance indicators, caching policy can be optimized accordingly. The work in [17] investigated the successful offloading probability in the cache-enabled HetNets and obtained the optimal caching probability by maximizing the successful offloading probability. In [18], the author investigated the optimal caching policy by maximizing the success probability and the area spectral efficiency respectively and analyzed the impact of system settings. Probabilistic content placement was studied in [19] to control cache-based channel selection diversity and network interference in a wireless caching helper network. The research was extended to the most general N-tier HetNets in [20], in which probabilistic tier-level content placement was analyzed given the network performance metric of the successful delivery probability. Different from the uncoded caching, coded caching can exploit the accumulated cache size by caching different segments of a content in different nodes. The coded caching in a large-scale small-cell network (SCN) was investigated in [21] and the performance was characterized by two metrics of the average fractional offloaded traffic and the average ergodic rate. The author of [22] proposed a combined coded caching strategy in disjoint cluster-centric SCNs and analyzed the successful content delivery probability.

## B. Motivation and Contributions

The above research neglected the impact of backhaul link on the performance analysis in the cache-enabled cellular network. In the practical scenario, BSs can not store all contents due to the finite cache capacity. The uncached content should be retrieved from the core network via the backhaul link and delivered to users via the wireless link. It is necessary to analyze the performance of the cache-enabled HetNets by taking the backhaul link into account.

The backhaul link with the limited capacity makes constraint on the content delivery and introduces extra power consumption. In cache-enabled HetNets, the design of the caching policy, the access strategy and the backhaul assignment jointly determines the mode of the content delivery. The delivery rate for different delivery modes is affected by different factors. Unlike the wireless delivery rate is determined by the

channel conditions, the backhaul delivery rate depends on the allocation of the limited backhaul capacity, and is generally smaller than the wireless delivery rate, which thus limits the overall content delivery rate. Moreover, the backhaul power consumption is related to the backhaul delivery rate. In [23–26], the authors conduct the backhaul-aware performance analysis of the cache-enabled cellular network based on a simplified backhaul model in which the backhaul delivery rate is set as the backhaul capacity. The work of [27, 28] studied the average delivery rate, in which the backhaul rate was obtained by the average splitting of backhaul capacity among users. In [29], the authors analyzed the cache performance in terms of successful content delivery probability (SCDP) taking backhaul capacity assignment based on the delivery rate demand into account. By contrast, we focus on the content caching and corresponding content delivery in HetNets, which dramatically affect the cache performance. Heterogeneous BSs differ in cache capacity and the density of nodes. Joint caching policies are required to be designed to take full advantages of the different cache ability of heterogeneous BSs to meet the content requests with different content popularity. Moreover, there exist different link states for heterogeneous BSs equipped with or without backhaul in the case of cache hit and miss. The association strategy should be coordinately designed by taking the caching policy and backhaul into account to make sure that the requested content is delivered with the proper link of the heterogeneous BSs. To this end, we expand the research of the cache-enabled one-tier cellular network with the limited backhaul link in [30]. We consider a two-tier HetNet in which the macro base stations (MBSs) tier is overlaid with the helpers tier. Both MBSs and helpers are equipped with caches while only the MBSs can connect to the core network via the backhaul link, helpers have no backhaul link. We analyze the system performance in cache-enabled HetNets with the limited backhaul. The main contributions are summarized as follows:

- We propose a hybrid cache policy in which the most popular contents are cached in the MBSs tier with the deterministic caching strategy and the less popular contents are cached in the helpers tier with the probabilistic caching strategy. We design a corresponding association scheme by taking the overall consideration of the wireless channel condition, cache policy and the backhaul assignment. Based on the proposed network framework, we analyze the system performance with the aid of the stochastic geometry approach.
- We derive the tractable expression of the successful content delivery probability (SCDP) for three scenarios, specifically the general scenario, the interference-limited scenario and the mean load scenario. These expressions reveal that the SCDP is determined by three types of the network parameters which is respectively related to the access link, BSs caching and the backhaul link.
- We deduce the expression of the average successful delivery rate. The average successful delivery rate is verified to consist of two parts: the basic rate demand and the average of the extra rate exceeding the rate demand. With the results of the average successful delivery rate,

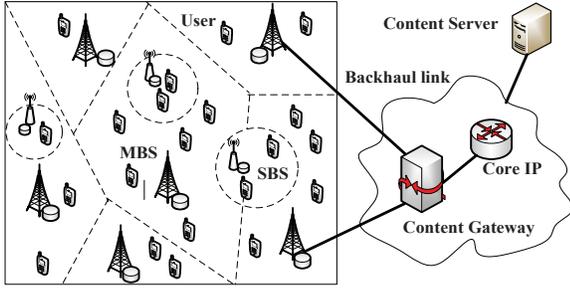


Fig. 1: An example of cache-enabled HetNets structure.

we obtain the expression of the energy efficiency.

- Numerical results demonstrate that 1) the hybrid cache policy is able to improve the system performance effectively compared with the most popular caching strategy; 2) the performance gain is more obvious under the condition that the content popularity is less skewed, the cache capacity is sufficient and the helpers density is relatively large; 3) for a fixed cache capacity, there exists an optimal BS density ratio to maximize the energy efficiency.

The rest of the paper is organized as follows. Section II gives the system model for cache-enabled HetNet with the limited backhaul. In Section III, new analytical expressions for the successful content delivery probability, average successful delivery rate and energy efficiency are derived. Numerical results are presented in Section IV, which is followed by the conclusions in Section V.

## II. SYSTEM MODEL

The system model, including cellular network model, cache model and association model, is described in this section.

### A. Cellular Network Model

A two-tier cache-enabled HetNet is considered in this paper, where a macro base stations (MBSs) tier is overlaid with a dense helpers tier. Denoting  $x_{i,j}$  as the location of the  $j$ th BS in the  $i$ th tier, the spatial distribution of the BSs in the  $i$ th tier  $\Phi_i = \{x_{i,j}, j = 0, 1, 2, \dots\}$  obeys independent Poisson Point Process (PPP) in the two dimensional Euclidean plane, and the intensity is  $\lambda_i$ . The locations of users (UEs) are also spatially distributed according to an independent PPP  $\Phi_u$  and the intensity is  $\lambda_u$ . MBSs are equipped with caches and connected to the core network via the backhaul link with limited capacity. Helpers have caches but no backhaul link, contents can be prefetched during off-peak times by broadcasting [31]. Fig. 1 shows an illustration of the network topology. According to the Palm theory, the statistical properties of UE at any position coincide with that of a typical UE at a fixed position [32]. Without loss of generality, the analysis is conducted on a typical UE at the origin, namely the tagged UE.  $k \in \{1, 2\}$  is denoted as the index of the tier that the tagged UE is associated with.

### B. Channel Model

The wireless channel gain consists of two types of propagation effect, i.e., path loss and Rayleigh fading. Denoting  $z_{i,j}$  as

the distance of the tagged UE from BS  $x_{i,j}$  in the  $i$ th tier, the path loss is calculated with the widely used power-law model as  $z_{i,j}^{-\alpha}$ , where  $2 < \alpha \leq 4$  is the path loss exponent. Rayleigh fading follows the independent and identically distributed (i.i.d.) exponential distribution with mean 1,  $h_{i,j} \sim \exp(1)$ . Setting the transmission power of the BSs as  $P_i$ , the receive power at the tagged UE can be expressed as  $P_i h_{i,j} z_{i,j}^{-\alpha}$ .

### C. Cache Model

The content library is denoted as  $\mathcal{F} = \{f_1, f_2 \dots f_{N-1}, f_N\}$  and contains  $N = |\mathcal{F}|$  contents. All contents are assumed to have the same size  $F$ . For the scenarios of different content sizes, the same analysis is still applicable by splitting the content into chunks of equal size [22]. A large number of statistical results show that content popularity distribution changes slowly over time and can be approximated as static [33]. In this paper, the popularity of the content library follows Zipf distribution. Sorting the content in the descending order of popularity, the popularity of the  $n$ -ranked content is written as

$$q_n = \frac{n^{-\gamma}}{\sum_{k=1}^N k^{-\gamma}}, \quad (1)$$

where  $\gamma \geq 0$  is the shape parameter, reflecting the skewness of the popularity distribution. The larger the value of  $\gamma$  is, the more uneven the content popularity distribution.

In HetNets, MBSs have large transmission power and can ensure wide coverage range. Correspondingly, helpers are densely deployed to boost the network capacity. In order to ensure high cache hit ratio and increase content diversity at the same time, we propose a hybrid caching policy in this paper. Contents library is divided into two non-overlapping groups: the first group  $\mathcal{F}_{Mp} = \{f_1, f_2 \dots f_{N_{Mp}-1}, f_{N_{Mp}}\}$  contains the most popular  $N_{Mp} = |\mathcal{F}_{Mp}|$  contents, the second group  $\mathcal{F}_{Lp} = \{f_{N_{Mp}+1}, f_{N_{Mp}+2} \dots f_{N-1}, f_N\}$  contains the remaining less popular  $N_{Lp} = |\mathcal{F}_{Lp}|$  contents. MBSs and helpers are equipped with cache of different size to store  $N_1$  and  $N_2$  contents respectively. MBSs tier employs the deterministic caching strategy in which all MBSs store the same first contents group  $\mathcal{F}_{Mp}$  and  $N_1 = N_{Mp}$ . As such, the most popular contents requested frequently by major users can obtain high cache hit ratio. Probabilistic caching strategy is performed in helpers tier. Each helper independently selects  $N_2$  contents from the second content group  $\mathcal{F}_{Lp}$  to store in a random way and the caching probabilities of the contents in  $\mathcal{F}_{Lp}$  are the same, equal to  $p_{Lp} = \frac{N_2}{N_{Lp}}$ . Since the helpers density is comparatively large and each UE may be served by multiple helpers, probabilistic caching can make effective use of spatial content diversity to increase UEs' chance of obtaining the huge less popular contents.

The probability that the user requests the content from the first group is calculated as

$$Q_{Mp} = \sum_{n=1}^{N_{Mp}} q_n. \quad (2)$$

The probability that the user requests the content from the second group is calculated as

$$Q_{Lp} = 1 - Q_{Mp} = \sum_{n=N_{Mp}+1}^N q_n. \quad (3)$$

#### D. Cell Association

Based on the hybrid caching policy, we adopt a new content-centric association strategy, where the user is associated with the strongest BS capable of accessing the requested content in the ways of local cache or backhaul link. For different content groups, the association strategies are specified as follows:

##### 1) Association Strategy of First Content Group

In case the requested content belongs to the first group, user connects to the MBSs tier since the content is only cached in MBSs. In the MBSs tier, all MBSs store the same content group, then user is associated with the MBS with the maximum received power,

$$x_0 = \arg \max_{x_{1,j} \in \Phi_1} P_1 z_{1,j}^{-\alpha}. \quad (4)$$

In cache-enabled HetNets, MBSs have large transmission power and sufficient cache capacity, the combination of the maximum received power association and the deterministic caching strategy on the one hand greatly increases the delivery rate, and on the other hand improves the cache hit ratio.

Universal frequency reuse is adopted to improving the spectrum efficiency. When the user requests content from the first group, user is associated with the closet MBS. The total interference consists of all MBSs except the serving MBS and all helpers. The signal to interference plus noise ratio (SINR) is specified as:

$$\text{SINR}_n = \frac{P_1 h z^{-\alpha}}{I_1 + I_2 + \sigma^2}, f_n \in \mathcal{F}_{Mp}, \quad (5)$$

where  $I_1 = \sum_{x_{1,j} \in \Phi_1 \setminus x_0} P_1 h_{1,j} z_{1,j}^{-\alpha}$ ,  $I_2 = \sum_{j \in \Phi_2} P_2 h_{2,j} z_{2,j}^{-\alpha}$  denote the interference from MBSs tier, helpers tier respectively.  $\sigma^2$  is the additive white Gaussian noise (AWGN).

##### 2) Association Strategy of Second Content Group

In case the requested content belongs to the second group, user can connect to either MBSs tier or helpers tier. In MBSs tier, no MBSs store the second contents group and the user needs to retrieve the requested content from the core network through the backhaul link. In the helpers tier, each helper independently selects contents to store according to the specific caching probability and the user can promptly get the requested content from the local cache. According to the properties of PPP, the distribution of the helpers caching and not caching the content  $f_n$  can be modeled as the thinning PPP  $\Phi_{n+,2}$ ,  $\Phi_{n-,2}$ , the density are  $p_{Lp} \lambda_2$ ,  $(1 - p_{Lp}) \lambda_2$ . Considering channel characteristics, BSs caching and backhaul link comprehensively, the user is assumed to be associated with the BS which has the maximum received power and is capable of accessing the requested content. The associated BS can be expressed as

$$x_0 = \arg \max_{x_{i,j} \in \Phi_1 \cup \Phi_{n+,2}} P_i z_{i,j}^{-\alpha}. \quad (6)$$

Due to the ultra dense deployment of helpers and the overall low popularity of the second content group, the probabilistic caching and the corresponding association strategy can guarantee users' Quality of Service (QoS) and improve the cache hit ratio at the same time.

When the user requests content from the second group, user connects to either the MBSs tier or the helpers tier. The serving BS may not be the closet BS and the interference can be divided into two types according to the content availability. The SINR of the user connecting to  $k$  th tier is specified as:

$$\text{SINR}_n = \frac{P_k h_k z_k^{-\alpha}}{I_{n+} + I_{n-} + \sigma^2}, f_n \in \mathcal{F}_{Lp}, \quad (7)$$

where  $I_{n+} = \sum_{x_{i,j} \in \Phi_1 \cup \Phi_{n+,2} \setminus x_0} P_i h_{i,j} z_{i,j}^{-\alpha}$ ,  $I_{n-} = \sum_{x_{2,j} \in \Phi_{n-,2}} P_2 h_{2,j} z_{2,j}^{-\alpha}$  denote the interference from BSs capable of accessing the requested content and BSs not capable of accessing the requested content respectively.

### III. PERFORMANCE ANALYSIS

This section analyzes the successful content delivery probability, average successful delivery rate and energy efficiency of the cache-enabled Hetnets, which effectively quantify the QoS of users and the performance of the whole network.

#### A. Successful Content Delivery Probability

Successful content delivery probability (SCDP) is defined as the probability that the requested content of the tagged user is successful accessed and the delivery rate  $R$  exceeds the rate demand  $R_0$ . SCDP helps to measure users' satisfaction with the achievable delivery rate with respect to the rate demand, and can be regarded as a metric of the QoS. Combining the content popularity  $q_n$ , the SCDP is given as

$$C = \sum_{n=1}^N q_n \mathbb{P} \left( \frac{W}{L} \log_2(1 + \text{SINR}_n) > R_0 \right), \quad (8)$$

where  $W$  denotes the system bandwidth,  $L$  is the total number of UEs served by the tagged BS, namely the load. BSs allocate equal spectrum resource to the associated UEs.

In order to analyze the follow-up system performance, we derive some auxiliary results in advance. Since the caching policy and the cell association strategy for the first content group and the second content group is separately set irrespective of the content index, the related statistical characteristics like serving distance, association probability, load distribution and the system performance containing the SCDP, average successful delivery rate and energy efficiency are the same for different contents in two groups. For notational brevity,  $\{M_p, L_p\}$  are used to uniformly mark two groups. In addition, we define  $\hat{\lambda}_{j,k} = \frac{\lambda_j}{\lambda_k}$ ,  $\hat{P}_{j,k} = \frac{P_j}{P_k}$  as the BSs density ratio and the transmit power ratio.

#### 1) Association Probability

Users requesting content from the first group can only connect to the MBSs tier, the association probability is expressed as

$$A_{Mp,1} = 1. \quad (9)$$

Correspondingly, users requesting content from the second group can connect to either the MBSs tier or the helpers tier, the association probability is derived in the following lemma.

**Lemma 1.** *The probability that the tagged UE requesting the content from the second group connects to the  $k$  th BSs tier is given by*

$$A_{Lp,k} = \frac{\lambda'_k P_k^{2/\alpha}}{\sum_{i=1}^2 \lambda'_i P_i^{2/\alpha}}, \quad (10)$$

where  $\lambda'_1 = \lambda_1, \lambda'_2 = p_{Lp}\lambda_2$  denote the equivalent MBSs density and helpers density of the content-centric HetNets.

*Proof.* If user requests content from the second group, user can access to the content from either the backhaul link of all MBSs or the local cache of part helpers. As a result, the network capable of accessing to the requested content is equivalent to a content-centric HetNets in which the distribution of MBSs and helpers follow two thinning PPPs and the corresponding tier density are  $\lambda'_1 = \lambda_1, \lambda'_2 = p_{Lp}\lambda_2$ . Matching the content-centric HetNets with the traditional HetNets, the association probability is derived by straightforwardly modifying the Lemmas 1 in [34].  $\square$

**Remark 1.** *The tagged user prefers to connect to the tier with higher equivalent BS density  $\lambda'_k$  and transmit power  $P_k$  in the content-centric HetNets.*

According to the content-centric association, users connecting to the MBSs tier fall into two categories: MBSs cache hit users and MBSs cache miss users, namely users who request content from the first group and users who request content from the second group and associate with MBSs. To sum up, the probability to connect to the MBSs tier is expressed as

$$\begin{aligned} A_1 &= \sum_{n=1}^{N_1} q_n A_{Mp,1} + \sum_{n=N_1+1}^N q_n A_{Lp,1} \\ &= Q_{Mp} + Q_{Lp} \frac{\lambda_1 P_1^{2/\alpha}}{\lambda_1 P_1^{2/\alpha} + p_{Lp}\lambda_2 P_2^{2/\alpha}}. \end{aligned} \quad (11)$$

With the conditional probability formula, the probability that user connecting to the MBS needs to fetch the requested content from the backhaul link is given by

$$p_b = \frac{Q_{Lp} A_{Lp,1}}{A_1}. \quad (12)$$

By comparison, users connecting to the SBSs tier is users requesting content from the second group and associating with the helpers, namely the helpers cache hit users. The probability to connect to the helpers tier is expressed as

$$A_2 = Q_{Lp} A_{Lp,2}. \quad (13)$$

## 2) Load Distribution

The load is the number of users served by the tagged BS, and the tagged user is included. Based on the lemma 3 of [35], the distribution of the load of in the  $k$  th tier is expressed as

$$\begin{aligned} P_{L_k}(L_k = l_k + 1) \\ = \frac{3.5^{3.5} \Gamma(l_k + 4.5)}{l_k! \Gamma(3.5)} \left( \frac{A_k \lambda_u}{\lambda_k} \right)^{l_k} \left( 3.5 + \frac{A_k \lambda_u}{\lambda_k} \right)^{-(l_k + 4.5)}, \end{aligned} \quad (14)$$

where  $A_k$  is the association probability as (11), (13).

The mean load in the  $k$  th tier is approximated as

$$\bar{L}_k \approx \mathbb{E}(L_k) = 1 + 1.28 \frac{A_k \lambda_u}{\lambda_k}. \quad (15)$$

## 3) Active Probability

In dense Hetnets, the intensity of helpers is comparable to or even higher than the intensity of users, some helpers may have no users to serve. In order to mitigate the interference and save the power consumption, inactive helpers should be turned off. The rest helpers that have users to serve are referred as active helpers. The probability that a helper is active is derived as [35].  $f_S(x)$  is the PDF of the coverage area of the helpers, and can be approximated as

$$f_S(x) \approx \frac{3.5^{3.5}}{\Gamma(3.5)} \left( \frac{\lambda_2}{A_2} \right)^{3.5} x^{2.5} e^{-3.5\lambda_2 x/A_2}. \quad (16)$$

Active probability can be derived as

$$\begin{aligned} p_a &= 1 - \int_0^\infty e^{-\lambda_u x} f_S(x) dx \\ &= 1 - \int_0^\infty \frac{3.5^{3.5}}{\Gamma(3.5)} \left( \frac{\lambda_2}{A_2} \right)^{3.5} x^{2.5} e^{-3.5\lambda_2 x/A_2 - \lambda_u x} dx, \end{aligned} \quad (17)$$

Let  $3.5\lambda_2 x/A_2 - \lambda_u x = t$ ,

$$\begin{aligned} p_a &= 1 - \int_0^\infty \frac{3.5^{3.5}}{\Gamma(3.5)} \left( \frac{\lambda_2}{A_2} \right)^{3.5} \frac{1}{(3.5\lambda_2/A_2 - \lambda_u)^{3.5}} t^{2.5} e^{-t} dt \\ &= 1 - \left( 1 + \frac{A_2 \lambda_u}{3.5\lambda_2} \right)^{-3.5}. \end{aligned} \quad (18)$$

Based on the combination of the caching policy and association strategy, the SCDP can be specifically divided into three cases. In the first case, the user requests content from the first group and connects to the MBSs tier. Since the requested content is stored in the local cache, the user can directly get the content from the serving MBS and the delivery rate is equal to the access delivery rate of the MBS; In the second case, the user requests content from the second group and connects to the helpers tier. Due to the probabilistic caching strategy, the user can get the content from the serving helper and the delivery rate is equal to the access delivery rate of the helper; In the third case, the user requests content from the second group and connects to the MBSs tier. Since the requested content is not stored in the local cache, the user need the serving MBS to retrieve the requested content from the core network via the backhaul link and forward it to the user via the access link. The delivery rate is composed of the access delivery rate and the backhaul delivery rate. To sum up, SCDP is expressed as

$$\begin{aligned}
& \Pr(R > R_0) \\
&= Q_{M_p} \Pr(R_{M_p,1} \geq R_0) + Q_{L_p} \Pr(R_{L_p,2} \geq R_0, A_{L_p,2}) \\
&+ Q_{L_p} \Pr(R_{L_p,1}^w \geq R_0 \& R_{L_p,1}^b \geq R_0, A_{L_p,1}) \\
&\stackrel{(a)}{=} Q_{M_p} \underbrace{\Pr(R_{M_p,1} \geq R_0)}_{\text{SADP } C_{M_p}} + Q_{L_p} \underbrace{\Pr(R_{L_p,2} \geq R_0, A_{L_p,2})}_{\text{SADP } C_{L_p,2}} \\
&+ Q_{L_p} \underbrace{\Pr(R_{L_p,1}^w \geq R_0, A_{L_p,1})}_{\text{SADP } C_{L_p,1}^w} + \underbrace{\Pr(R_{L_p,1}^b \geq R_0, A_{L_p,1})}_{\text{SBDP } C_{L_p,1}^b}
\end{aligned} \tag{19}$$

Equation (a) makes sense due to the independence of the access link and the backhaul link. It is shown in (19) that SCDP depends on four factors, namely the successful access delivery probability (SADP)  $C_{M_p}$ , SADP  $C_{L_p,2}$ , SADP  $C_{L_p,1}^w$ , the successful backhaul delivery probability (SBDP)  $C_{L_p,1}^b$ .

In the following, we analyze four probabilities in sequence. To this end, two functions are defined to facilitate the description of two types of interference coming from the BSs with and without the requested content respectively:

$$G(x, y) = \frac{2x}{y-2} {}_2F_1\left(1, 1 - \frac{2}{y}; 2 - \frac{2}{y}; -x\right), \tag{20}$$

$$H(x, y) = \frac{2}{y} x^{2/\alpha} B\left(\frac{2}{y}, 1 - \frac{2}{y}\right), \tag{21}$$

where  ${}_2F_1(\cdot)$  is the Gauss hypergeometric function,  $B(\cdot)$  is the Beta function.

SADP  $C_{M_p}$  is the probability that the access delivery rate of MBSs is higher than the rate demand when user requests content from the first group.

**Lemma 2.** *The successful access delivery probability  $C_{M_p}$  is expressed as*

$$\begin{aligned}
C_{M_p} &= \int_0^\infty \sum_{l_1=0}^\infty 2\pi\lambda_1 z \exp(-z^\alpha P_1^{-1} \delta_1 \sigma^2) \\
&\times \exp\left(-\pi\lambda_1 z^2 \left(G(\delta_1, \alpha) + p_a \hat{\lambda}_{2,1} \hat{P}_{2,1}^{2/\alpha} H(\delta_1, \alpha) + 1\right)\right) \\
&\times P_{L_1}(l_1 + 1) dz,
\end{aligned} \tag{22}$$

where  $\delta_1 = 2^{l_1 \frac{R_0}{W}} - 1$ .

*Proof.* Please refer to Appendix A.  $\square$

SADP  $C_{L_p,2}$  is the probability that the access delivery rate of helpers is higher than the rate demand when user requests content from the second group and connects to the helpers tier.

**Lemma 3.** *The successful access delivery probability  $C_{L_p,2}$  is expressed as*

$$\begin{aligned}
C_{L_p,2} &= \int_0^\infty \sum_{l_2=0}^\infty 2\pi p_{L_p} \lambda_2 z \exp(-z^\alpha P_2^{-1} \delta_2 \sigma^2) \\
&\times \exp\left(-\pi p_{L_p} \lambda_2 z^2 (\xi_2 G(\delta_2, \alpha) + \varsigma_2 H(\delta_2, \alpha) + \zeta_2)\right) \\
&\times P_{L_2}(l_2 + 1) dz,
\end{aligned} \tag{23}$$

where  $\delta_2 = 2^{l_2 \frac{R_0}{W}} - 1$ , coefficients for  $k$ th tier are given as  $\xi_k = (\lambda_1 P_1 + p_a p_{L_p} \lambda_2 P_2) (\lambda'_k P_k)^{-1}$ ,  $\varsigma_k = (p_a (1 - p_{L_p}) \lambda_2 P_2) (\lambda'_k P_k)^{-1}$ ,  $\zeta_k = (\lambda_1 P_1 + p_{L_p} \lambda_2 P_2) (\lambda'_k P_k)^{-1}$ .

*Proof.* Please refer to Appendix B.  $\square$

SADP  $C_{L_p,1}$  is the probability that the access delivery rate of MBSs is higher than the rate demand when user requests content from the second group and connects to the MBSs tier.

**Lemma 4.** *The successful access delivery probability  $C_{L_p,1}$  is expressed as*

$$\begin{aligned}
C_{L_p,1}^w &= \int_0^\infty \sum_{l_1=0}^\infty 2\pi\lambda_1 z \exp(-z^\alpha P_1^{-1} \delta_1 \sigma^2) \\
&\times \exp\left(-\pi\lambda_1 z^2 (\xi_1 G(\delta_1, \alpha) + \varsigma_1 H(\delta_1, \alpha) + \zeta_1)\right) \\
&\times P_{L_1}(l_1 + 1) dz,
\end{aligned} \tag{24}$$

*Proof.* Please refer to Appendix C.  $\square$

SBDP  $C_{L_p,1}^b$  is the probability that the backhaul delivery rate of MBSs is higher than the rate demand when user requests content from the second group and connects to the MBSs tier.  $C_{L_p,1}^b$  is determined by the backhaul assignment among users. Since the backhaul capacity is limited, the maximum number of users supported to deliver the content at the required delivery rate is fixed and can be calculated as  $N_b = \frac{C_b}{R_0}$ . Supposing the number of users accessing the requested content via the backhaul link in the MBS is  $N_{miss}$ , if  $N_{miss} \leq N_b$ , all  $N_{miss}$  users can be scheduled with the backhaul link, otherwise, the backhaul link fails to support  $N_{miss}$  users, and will randomly picks  $N_b$  users.

Assuming that the requested content of the tagged user needs to be fetched through the backhaul link and there are another  $m$  MBSs cache miss users associated with the tagged MBS, the successful backhaul delivery probability is expressed as

$$\begin{aligned}
C_{L_p,1}^b(l_1 + 1) &= (R_{L_p,1}^b \geq R_0 | L_1 = l_1 + 1) \\
&= \sum_{m=0}^{N_b-1} \binom{l_1}{m} (1-p_b)^{l_1-m} (p_b)^m \\
&+ \sum_{m=N_b}^{l_1} \binom{l_1}{m} (1-p_b)^{l_1-m} (p_b)^k \frac{N_b}{(m+1)} \\
&= \sum_{m=0}^{l_1} \binom{l_1}{m} (1-p_b)^{l_1-m} (p_b)^k \min\left\{1, \frac{N_b}{(m+1)}\right\}.
\end{aligned} \tag{25}$$

**Remark 2.** *As the backhaul capacity tends to infinity, the number of the cache miss users that can be supported by the backhaul link  $N_b$  is much greater than the actual number of cache miss users  $m+1$  in Hetnets.  $C_{L_p,1} = 1$ , the backhaul link has no effect on SCDP.*

Substituting (22), (23), (24) and (25) in (19), the SCDP is obtained.

**Remark 3.** *SCDP is determined by three network parameters sets. The first set contains the access link related parameters: BSs density, transmit power and path loss parameter. The*

second set has correlation with the cache, including the content popularity and cache capacity. The third set is related to the backhaul link, namely the backhaul capacity.

Due to the combination of various operations, such as the improper integral of the serving distance, the infinite summation over the load, the judgment of the minimum and the especial use of lookup tables for  ${}_2F_1$ , the SCDP expression for the general case is extremely complicated. In order to analyze the SCDP more conveniently, we derive two approximate results containing the interference-limited scenario and the mean load scenario in the following corollaries.

**Corollary 1.** *The successful content delivery probability in the interference-limited scenario is given by*

$$\begin{aligned}\tilde{C} &= Q_{M_p} \tilde{C}_{M_p} + Q_{L_p} \tilde{C}_{L_p,2} + Q_{L_p} \tilde{C}_{L_p,1}^w C_{L_p,1}^b, \\ \tilde{C}_{M_p} &= \sum_{l_1=0}^{\infty} \left( G(\delta_1, \alpha) + p_2 \hat{\lambda}_{2,1} \hat{P}_{2,1}^{2/\alpha} H(\delta_1, \alpha) + 1 \right)^{-1} \\ &\quad \times P_{L_1}(l_1 + 1), \\ \tilde{C}_{L_p,2} &= \sum_{l_2=0}^{\infty} (\xi_2 G(\delta_2, \alpha) + \varsigma_2 H(\delta_2, \alpha) + \zeta_2)^{-1} P_{L_2}(l_2 + 1), \\ \tilde{C}_{L_p,1}^w &= \sum_{l_1=0}^{\infty} (\xi_1 G(\delta_1, \alpha) + \varsigma_1 H(\delta_1, \alpha) + \zeta_1)^{-1} P_{L_1}(l_1 + 1).\end{aligned}\quad (26)$$

*Proof.* In interference-limited scenario, the noise power is very small compared to the interference power.  $\tilde{C}_{M_p}$ ,  $\tilde{C}_{L_p,2}$  and  $\tilde{C}_{L_p,1}^w$  can be obtained by letting the noise power tend to zero  $\sigma^2 \rightarrow 0$  and calculating the integral of the serving distance in (22), (23) and (24) respectively. The noise power has no effect on SBDP  $C_{L_p,1}^b$ . Substituting  $\tilde{C}_{M_p}$ ,  $\tilde{C}_{L_p,2}$  and  $\tilde{C}_{L_p,1}^w$  and the unchanged  $C_{L_p,1}^b$  into (19), Corollary is proved.  $\square$

**Corollary 2.** *The successful content delivery probability with the mean load in the interference-limited scenario is given by*

$$\begin{aligned}\bar{C} &= Q_{M_p} \bar{C}_{M_p} + Q_{L_p} \bar{C}_{L_p,2} + Q_{L_p} \bar{C}_{L_p,1}^w \bar{C}_{L_p,1}^b, \\ \bar{C}_{M_p} &= \left( G(\bar{\delta}_1, \alpha) + p_2 \hat{\lambda}_{2,1} \hat{P}_{2,1}^{2/\alpha} H(\bar{\delta}_1, \alpha) + 1 \right)^{-1}, \\ \bar{C}_{L_p,2} &= (\xi_2 G(\bar{\delta}_2, \alpha) + \varsigma_2 H(\bar{\delta}_2, \alpha) + \zeta_2)^{-1}, \\ \bar{C}_{L_p,1}^w &= (\xi_1 G(\bar{\delta}_1, \alpha) + \varsigma_1 H(\bar{\delta}_1, \alpha) + \zeta_1)^{-1}, \\ \bar{C}_{L_p,1}^b &= \sum_{m=0}^{\bar{L}_1-1} \binom{\bar{L}_1-1}{m} (1-p_b)^{\bar{L}_1-m-1} (p_b)^m \\ &\quad \times \min \left\{ 1, \frac{N_b}{(m+1)} \right\},\end{aligned}\quad (27)$$

where  $\bar{\delta}_1 = 2^{\bar{L}_1} \frac{R_0}{W} - 1$ ,  $\bar{\delta}_2 = 2^{(\bar{L}_2+1)} \frac{R_0}{W} - 1$ ,  $\bar{L}_1$ ,  $\bar{L}_2$  denote the mean load of the MBSs tier and the helpers tier.

*Proof.*  $\bar{C}_{M_p}$ ,  $\bar{C}_{L_p,2}$ ,  $\bar{C}_{L_p,1}^w$  and  $\bar{C}_{L_p,1}^b$  can be obtained by calculating the mean load as (15) and eliminating the summation over the load  $l_k$  in (22), (23), (24) and (25) respectively. Using the approximation  $\mathbb{E}_L(C(l)) \approx C(\mathbb{E}_L)$  and substituting

$\bar{C}_{M_p}$ ,  $\bar{C}_{L_p,2}$ ,  $\bar{C}_{L_p,1}^w$  and  $\bar{C}_{L_p,1}^b$  into (18), the corollary is proved.  $\square$

## B. Average successful delivery rate

Average successful delivery rate is defined as the average delivery rate of the tagged UE under the premise of the successful content delivery  $R_{suc} = \mathbb{E}(R|R \geq R_0)$ . Similar to the SCDP, the average successful delivery rate  $R_{suc}$  has three cases.

- If the user requests content from the first group, content can be obtained directly from the MBSs and  $R_{suc}$  is equal to the average successful access delivery rate of the MBSs  $R_{M_p,1}^{suc}$ , namely the average access delivery rate of MBSs on the condition that the access delivery rate exceeds the rate demand.
- If the user requests content from the second group and connects to the helpers tier, content can be obtained from the helpers and  $R_{suc}$  is equal to the average successful access delivery rate of the helpers  $R_{L_p,2}^{suc}$ , namely the average access delivery rate of helpers on the condition that the access delivery rate exceeds the rate demand.
- If the user requests content from the second group and connects to the MBSs tier, content needs to be retrieved from the core network through the backhaul link and  $R_{suc}$  is equivalent to the average successful backhaul delivery rate  $R_{L_p,1}^{suc}$  set to the rate demand  $R_0$  on the premise that access link and backhaul link are successfully provided.

According to the total probability law,  $R_{suc}$  is expressed as

$$\begin{aligned}R^{suc} &= Q_{M_p} R_{M_p,1}^{suc} + Q_{L_p} R_{L_p,2}^{suc} \\ &\quad + Q_{L_p} A_{L_p,1} C_{L_p,1}^w C_{L_p,1}^b R_{L_p,1}^{suc}.\end{aligned}\quad (28)$$

$R_{M_p,1}^{suc}$  and  $R_{L_p,2}^{suc}$  are given in the following theorems successively.

**Lemma 5.** *The average successful access delivery rate of MBSs  $R_{M_p,1}^{suc}$  is expressed as (29) at the top of next page, where  $G^*(x_1, x_2, y) = G(x_1, y) - G(x_2, y)$ ,  $H^*(x_1, x_2, y) = H(x_1, y) - H(x_2, y)$ ,  $\delta_1^r = 2^{(l_1+1)} \frac{R_0}{W} - 1$ ,  $\delta_1 = 2^{(l_1+1)} \frac{R_0}{W} - 1$ .*

*Proof.* Please refer to Appendix D.  $\square$

**Lemma 6.** *The average successful access delivery rate of SBSs  $R_{L_p,2}^{suc}$  is expressed as (30) at the top of next page, where  $\delta_2^r = 2^{(l_2+1)} \frac{R_0}{W} - 1$ ,  $\delta_2 = 2^{(l_2+1)} \frac{R_0}{W} - 1$ .*

*Proof.* Proof is similar to Appendix D.  $\square$

Plugging (24), (25), (29) and (32) into (28), we can get the average successful delivery rate in general scenario.

**Remark 4.** *Average successful delivery rate is consist of two parts: the basic rate demand and the average of the extra rate exceeding the rate demand.*

Two special cases containing the interference-limited scenario and the mean load scenario are derived to further simplify the average successful delivery rate.

$$R_{Mp,1}^{suc} = R_0 + \int_{R_0}^{\infty} \int_0^{\infty} \sum_{l_1=0}^{\infty} 2\pi\lambda_1 z \exp(-z^\alpha P_1^{-1}(\delta_1^r - \delta_1)\sigma^2) \\ \times \exp\left(-\pi\lambda_1 z^2 \left(G^*(\delta_1^r, \delta_1, \alpha) + \hat{\lambda}_{2,1} \hat{P}_{2,1}^{2/\alpha} H^*(\delta_1^r, \delta_1, \alpha) + 1\right)\right) P_{L_1}(l_1 + 1) dz dr, \quad (29)$$

$$R_{Lp,2}^{suc} = A_{Lp,2} R_0 + \int_{R_0}^{\infty} \int_0^{\infty} \sum_{l_2=0}^{\infty} 2\pi p_{Lp} \lambda_2 z \exp(-z^\alpha P_2^{-1}(\delta_2^r - \delta_2)\sigma^2) \\ \times \exp\left(-\pi p_{Lp} \lambda_2 z^2 (\xi_2 G^*(\delta_2^r, \delta_2, \alpha) + \varsigma_2 H^*(\delta_2^r, \delta_2, \alpha) + \zeta_2)\right) P_{L_2}(l_2 + 1) dz dr, \quad (30)$$

**Corollary 3.** *The average successful delivery rate in the interference-limited scenario is given by*

$$\begin{aligned} \tilde{R}^{suc} &= Q_{Mp} \tilde{R}_{Mp,1}^{suc} + Q_{Lp} \tilde{R}_{Lp,2}^{suc} + Q_{Lp} A_{Lp,1} \tilde{C}_{Lp,1}^w C_{Lp,1}^b R_0, \\ \tilde{R}_{Mp,1}^{suc} &= R_0 \\ &+ \int_{R_0}^{\infty} \sum_{l_1=0}^{\infty} \left(G^*(\delta_1^r, \delta_1, \alpha) + \hat{\lambda}_{2,1} \hat{P}_{2,1}^{2/\alpha} H^*(\delta_1^r, \delta_1, \alpha) + 1\right)^{-1} \\ &\times P_{L_1}(l_1 + 1) dr, \\ \tilde{R}_{Lp,2}^{suc} &= A_{Lp,2} R_0 \\ &+ \int_{R_0}^{\infty} \sum_{l_2=0}^{\infty} (\xi_2 G^*(\delta_2^r, \delta_2, \alpha) + \varsigma_2 H^*(\delta_2^r, \delta_2, \alpha) + \zeta_2)^{-1} \\ &\times P_{L_2}(l_2 + 1) dr, \end{aligned} \quad (31)$$

where  $\tilde{C}_{Lp,1}^w$ ,  $C_{Lp,1}^b$  are given as (26), (25).

**Corollary 4.** *The average successful delivery rate with the mean load in the interference-limited scenario is given by*

$$\begin{aligned} \bar{R}^{suc} &= Q_{Mp} \bar{R}_{Mp,1}^{suc} + Q_{Lp} \bar{R}_{Lp,2}^{suc} + Q_{Lp} A_{Lp,1} \bar{C}_{Lp,1}^w \bar{C}_{Lp,1}^b R_0, \\ \bar{R}_{Mp,1}^{suc} &= R_0 \\ &+ \int_{R_0}^{\infty} \left(G^*(\bar{\delta}_1^r, \bar{\delta}_1, \alpha) + \hat{\lambda}_{2,1} \hat{P}_{2,1}^{2/\alpha} H^*(\bar{\delta}_1^r, \bar{\delta}_1, \alpha) + 1\right)^{-1} dr, \\ \bar{R}_{Lp,2}^{suc} &= A_{Lp,2} R_0 \\ &+ \int_{R_0}^{\infty} (\xi_2 G^*(\bar{\delta}_2^r, \bar{\delta}_2, \alpha) + \varsigma_2 H^*(\bar{\delta}_2^r, \bar{\delta}_2, \alpha) + \zeta_2)^{-1} dr, \end{aligned} \quad (32)$$

where  $\bar{\delta}_1^r = 2^{\bar{L}_1} \frac{R_0}{\bar{W}} - 1$ ,  $\bar{\delta}_2^r = 2^{\bar{L}_2} \frac{R_0}{\bar{W}} - 1$ ,  $\bar{\delta}_1 = 2^{\bar{L}_1} \frac{R_0}{\bar{W}} - 1$ ,  $\bar{\delta}_2 = 2^{(\bar{L}_2+1)} \frac{R_0}{\bar{W}} - 1$ ,  $\bar{C}_{Lp,1}^w$ ,  $\bar{C}_{Lp,1}^b$  are given as (27).

The proof process of **Corollary 3**, **Corollary 4** are similar to the **Corollary 1**, **Corollary 2**.

### C. Energy Efficiency

Throughput is defined as the successful delivery throughput in the cache-enabled HetNets and can be expressed as

$$T_{suc} = \lambda_u C R_{suc}, \quad (33)$$

where  $C$  is the SCDP,  $R_{suc}$  is the average successful delivery rate.  $\tilde{T}^{suc} = \lambda_u \tilde{C} \tilde{R}^{suc}$ ,  $T^{suc} = \lambda_u \bar{C} \bar{R}^{suc}$  are used for the interference-limited scenario and the mean load scenario.

The total power of the cache-enabled HetNets is the power summation over the active BSs of two tiers:  $Pow = \sum_{k=1}^2 \lambda_k Pow_k$ . The power consumed at a BS contains the BSs power consumption, the cache power consumption and the backhaul power consumption

$$Pow_k = Pow_k^s + Pow_k^c + Pow_k^b. \quad (34)$$

BS power consumption consists of the transmit power consumption  $P_k^t$  and the circuit power consumption  $P_k^0$ , and is expressed as

$$Pow_k^s = \varepsilon_k p_{k,a} P_k^t + P_k^0, \quad (35)$$

where  $p_{k,a}$  is the active probability of the BSs in the  $k$  th tier and satisfies  $p_{1,a} = 1$ ,  $p_{2,a} = p_a$ .  $\varepsilon_k$  is the power amplifier efficiency.

Energy-proportional model is adopted to describe the cache power consumption. In this model, the cache power consumption of the BSs in the  $k$  th tier is proportional to the cache capacity, and is expressed as

$$Pow_k^c = \rho N_k F, \quad (36)$$

where  $\rho$  is the power coefficient of cache hardware.

Backhaul power consumption depends on the successful backhaul delivery throughput. For the cases of helpers, the backhaul power consumption is zero due to the lack of the backhaul link. For the cases of MBSs, the backhaul power consumption is proportional to the successful backhaul delivery throughput calculated by multiplying the backhaul usage ratio with the backhaul capacity.

$$\begin{aligned} Pow_1^b &= \omega T_b \\ &= \omega C_b \sum_{l_1=1}^{\infty} \sum_{m=0}^{l_1} \binom{l_1}{m} (1-p_b)^{l_1-m} (p_b)^m \min\left\{1, \frac{m}{N_b}\right\}, \end{aligned} \quad (37)$$

where  $\omega$  is the power coefficient of the backhaul.

With the approximation of the MBSs mean load, backhaul power consumption can be simplified as

$$\overline{Pow_1^b} = \omega C_b \sum_{m=1}^{\bar{L}_1} \binom{\bar{L}_1}{m} (1-p_b)^{\bar{L}_1-m} (p_b)^m \min\left\{1, \frac{m}{N_b}\right\}, \quad (38)$$

where  $\bar{L}$  is shown as (15).

Substituting (35), (36) and (37) into (34), the total power is obtained. The total power with the mean load approximation replaces (38) with (37).

According to the definition, the energy efficiency of the cache-enabled cellular networks for various cases are given in the sequel by dividing the throughput by the total power consumption. The energy efficiency in the general scenario can be derived as

$$\eta_{EE} = \frac{T^{suc}}{P_{ow}} = \frac{\lambda_u C R^{suc}}{P_{ow}}. \quad (39)$$

The energy efficiency in the interference-limited scenario can be derived as

$$\tilde{\eta}_{EE} = \frac{\tilde{T}^{suc}}{P_{ow}} = \frac{\lambda_u \tilde{C} \tilde{R}^{suc}}{P_{ow}}. \quad (40)$$

The energy efficiency using the mean load approximation in the interference-limited scenario can be derived as

$$\bar{\eta}_{EE} = \frac{\bar{T}^{suc}}{P_{ow}} = \frac{\lambda_u \bar{C} \bar{R}^{suc}}{P_{ow}}. \quad (41)$$

#### IV. NUMERICAL RESULTS

In this section, both the numerical simulations and Monte Carlo simulations are presented. The scenario of the two-tier HetNets is considered, in which MBSs and helpers are distributed according to PPP in a circular area with 200 km radius. The basic simulation parameters are listed in Table I and the results are averaged over 10000 Monte Carlo trials [36] [37]. We validate the theoretical analysis in the previous section via Monte Carlo simulations and compare it with the most popular caching policy in which MBSs and helpers separately select the  $N_1$  most popular contents and the  $N_2$  following less popular contents to store. The impact of various parameters on the system performance are investigated.

TABLE I: SIMULATION PARAMETERS

Parameter	Value
UEs density	$\lambda_u = 100 \times 10^{-6} m^{-2}$
BSs density	$\lambda_1 = 2 \times 10^{-6} m^{-2}$
Bandwidth	$W = 10 MHz$
Path loss exponent	$\alpha = 4$
Rate threshold	$R_0 = 100 Kbps$
Transmit power	$P_1^t = 46 dBm, P_2^t = 21 dBm$
Static power consumption	$P_1^0 = 724.6 W, P_2^0 = 10.16 W$
BS power coefficient	$\varepsilon_1 = 3.22, \varepsilon_2 = 15.13$
Content catalog	$N = 2000$
Cache capacity	$N_1 = 400$
Content size	$F = 10 Mbit$
Caching power coefficient	$\rho = 6.25 \times 10^{-12} W/bit$
Backhaul power coefficient	$\omega = 5 \times 10^{-7} W/bps$

The performance of SCDP with respect to different system parameters including the helper density, content library size, helper cache capacity and backhaul capacity are shown in Fig. 2- Fig. 5. It can be seen from the figures that the simulation curves match with the numerical curves, which gives an effective validation of the theoretical analysis. The general result and the interference-limited result obtained respectively from (19) and **Corollary 1** match. It is confirmed that noise is not a very important factor in interference-limited HetNets.

The shape of the curves for the general scenario and the mean load scenario obtained from **Corollary 2** are consistent. The small gap under some parameter settings is caused due to the fact that the approximation error of the mean load is aggravated with the operation of summation in (25).

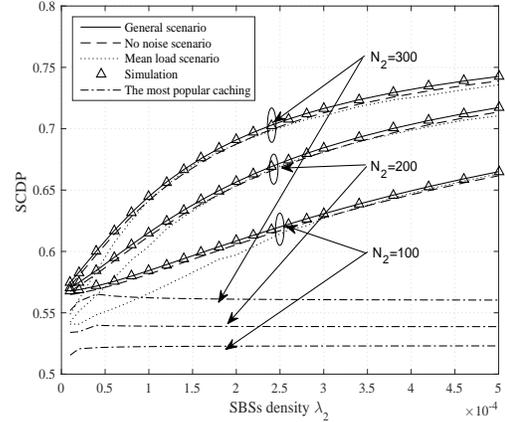


Fig. 2: SCDP versus helper density with different cache capacity,  $\delta = 0.5$ ,  $C_b = 2.5 \times 10^6$  bps.

Fig. 2 shows the evolution of the SCDP with regard to the helper density for different helper cache capacity. Hybrid caching policy outperforms the most popular caching policy and the performance gain increases with a growing helper density. The reason is that numerous users are offloaded to the helpers tier and obtain the content from the local cache as the helper density increases, which is different from the most popular caching policy whose load distribution is independent of the helper density. By avoiding the influence of the limited MBS backhaul link as much as possible, dense helpers deployment can increase the delivery rate dramatically and improve the SCDP. Comparing the curves of different helper cache capacity, caching at helpers also helps to improve the SCDP.

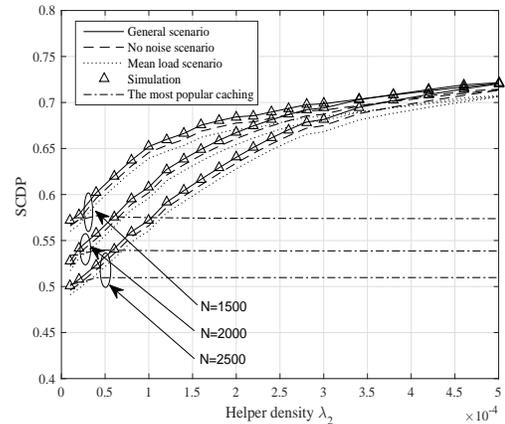


Fig. 3: SCDP versus helper density with different content library size,  $C_b = 2.5 \times 10^6$  bps,  $\delta = 0.5$ .

Fig. 3 illustrates the evolution of SCDP with respect to the helper density for different content library size  $N$ . For

the fixed content library size, the SCDP increases with the increase of the helper density. The performance gain of the hybrid caching policy over the most popular caching policy increases with the increase of the helper density as well. As content library size increases, the SCDP decreases. When the content library contains a large number of contents, the span of the content popularity distribution is large and the popularity of each content is relatively small given the fixed  $\delta$ . On the one hand, the cache hit ratio of the first content group decreases. On the other hand, the cache probability of the second content group decreases. As a result, UEs requesting content from the second group prefer to associate with the MBSs tier, the SBDP decreases accordingly. SADP  $C_{Lp,2}$  decreases due to the low helper density caching the requested content. Increasing the helper density can effectively cope with the expansion of the content library to achieve the fixed SCDP requirements.

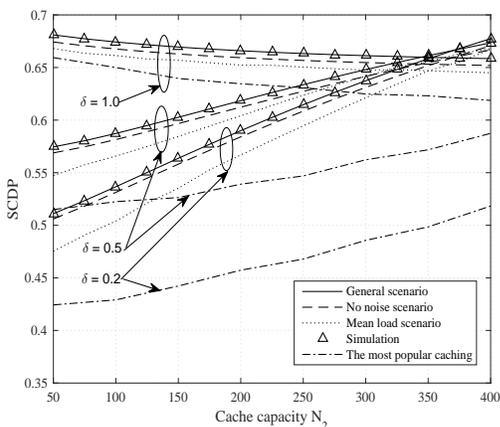


Fig. 4: SCDP versus helper cache capacity with different content popularity parameter,  $\lambda_2 = 1 \times 10^{-4} \text{ m}^{-2}$ ,  $C_b = 2.5 \times 10^6$  bps.

Fig. 4 depicts the impact of the helper cache capacity on the SCDP for different content popularity distribution parameter  $\delta$ . The cache probability of the content in the second group increases as cache capacity increases. For a small parameter indicating the more even content popularity distribution, the SCDP gain of the proposed caching policy over the most popular caching policy increases with the increasing helper cache capacity. For instance, setting  $\delta = 0.5$ , the performance gain ranges from 10% to 19%, when  $N_2$  varies from 25 to 400. The reason for this trend is that the increasement of cache capacity helps to make full use of the content diversity to improve the SCDP. For a big parameter indicating the more skewed content popularity distribution, a large number of users requests focus on the popular contents in the first group and MBS caching gives a large SCDP. The gap between two caching policy shrinks. However, as cache capacity increases, more users requesting the content from the second group associate with the helpers tier and active probability of the helpers increases accordingly. The SCDP experiences a slight decline under the influence of increasing inter-tier interference.

In Fig. 5, the SCDP variation with regard to backhaul capacity for different helper cache capacity is illustrated. It is

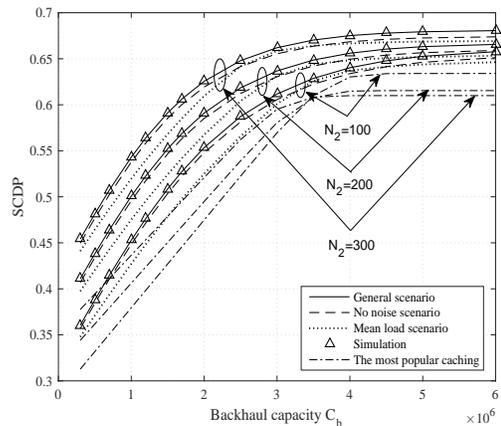


Fig. 5: SCDP versus backhaul capacity with different helper cache capacity,  $\lambda_2 = 1 \times 10^{-4} \text{ m}^{-2}$ ,  $\delta = 0.5$ .

shown that the increment of backhaul capacity can effectively improve the SCDP by increasing the SBDP of the UEs requesting content from the second group and associating with the MBSs tier. Each curve eventually converges to a point at which backhaul link is sufficient to assure the content availability for the cache miss UEs. Despite the increase in backhaul capacity, the gap between the hybrid caching policy and the most popular caching policy remains almost the same. As the helper cache capacity increases, the benefit of the BSs caching increase to make up for the limited backhaul link and the SCDP is improved. The larger the helper cache capacity, the more quickly the curve converges, causing the left shift of the MBSs backhaul capacity point to reach the maximum SCDP. The average performance gain increases with the increment of the cache capacity, ranging from 9.6% to 19.2% when  $N_2$  varies from 100 to 300.

The performance of average successful delivery rate, energy efficiency under different system parameters containing the helper density, content library size, helper cache capacity and backhaul capacity are shown in Fig. 6- Fig. 9. It can be seen from the figures that the simulation curve is in agreement with the numerical curve. The general result and the interference-limited result obtained respectively from (28), (39) and **Corollary 3**, (40) match. The general result and the mean load approximation obtained from **Corollary 4**, (41) are consistent. Small difference exists due to the fact that the operation of summation in (25) and integral in (32) aggravate the approximation error.

The variations of the average successful delivery rate and the energy efficiency with regard to the helper density for different helper cache capacity are illustrated in Fig. 6. In Fig. 6a, as helper density increases, a large number of users are offloaded to the helpers tier. Dense helpers deployment greatly increases the access delivery rate and improves the average successful delivery rate. The increment of the helper cache capacity further enhance the average successful delivery rate. In contrast, the average successful delivery rate achieved by the most popular caching policy is far less than that of the hybrid caching policy and exhibits a slight improvement as helper density increases. The reason is that the load of different BSs tier is fixed. In Fig. 6b, for small helper density, the curves

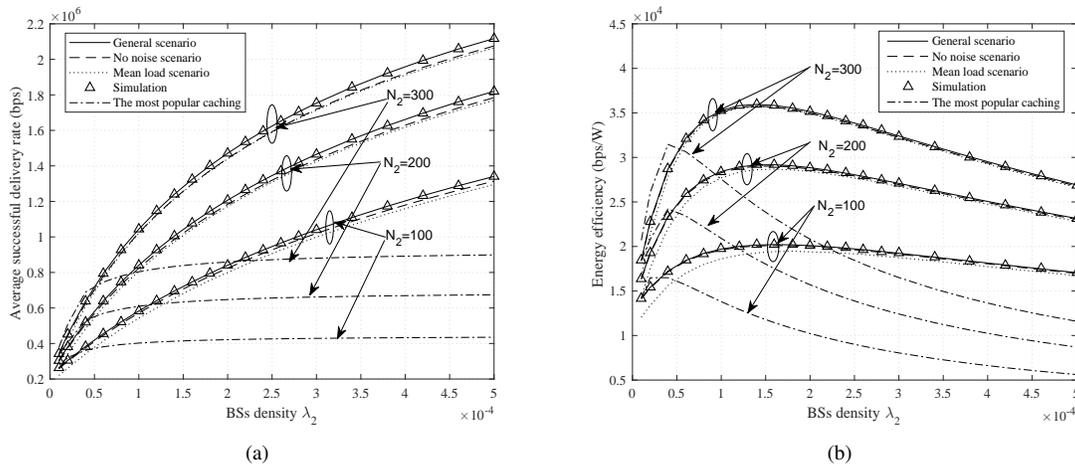


Fig. 6: Average successful delivery rate (a) and energy efficiency (b) versus helper density with different helper cache capacity,  $\delta = 0.5$ ,  $C_b = 2.5 \times 10^6$  bps.

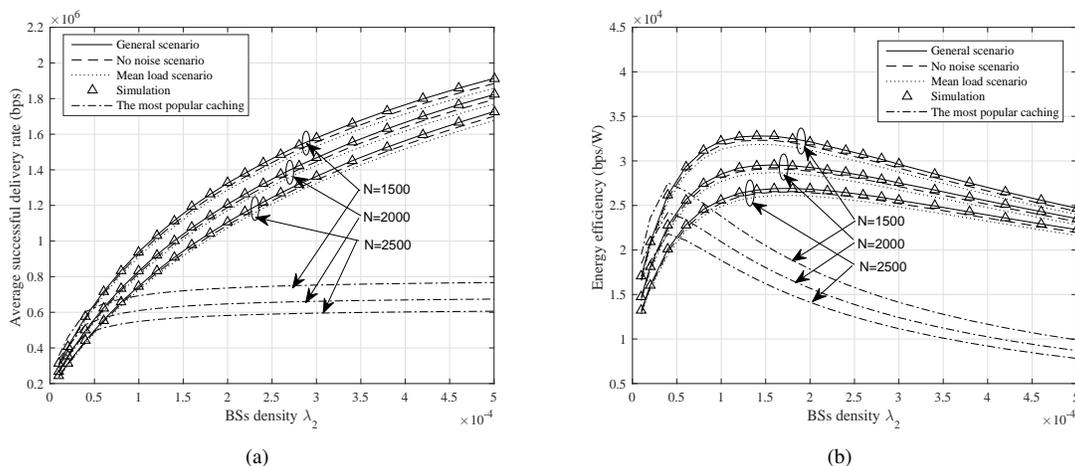


Fig. 7: Average successful delivery rate (a) and energy efficiency (b) versus helper density with different content library size,  $\delta = 0.5$ ,  $C_b = 2.5 \times 10^6$  bps.

of two caching policy almost coincide. As the helper density increases, the EE first increases rapidly benefitting from the improvement of the throughput, then decreases on account of the large power consumption, especially the most popular caching policy suffers a more severe degradation. There exists an optimal helper density to maximize the EE. BSs caching is able to improve the EE due to the increment of the delivery rate for the one side and the reduction of the backhaul power consumption for the other side. The larger the cache capacity, the smaller the helper density to get the maximum EE.

Fig. 7 illustrates the evolution of the average successful delivery rate and the energy efficiency with respect to the helper density for different content library size  $N$ . In Fig. 7a, for the fixed content library size, the average successful delivery rate increases with the increment of the helper density. Hybrid caching policy outperforms the most popular caching policy. The average successful delivery rate decreases with the increasing content library size. As content library size increases, the cache hit ratio of the first content group decreases for the reason that MBS with a fixed cache capacity can only store small parts of the contents. Moreover, the

cache probability of the second content group decreases, the probability that UEs requesting content from the second group associate with the MBSs tier increases accordingly, which result in the decrease of the average successful backhaul delivery rate. The average successful access delivery rate of the helpers decreases due to the decreasing density of the helpers capable of accessing the requested content. In Fig. 7b, as the helper density increases, the EE increases first and then decreases. EE decreases with the increasing content library size as a result of the decreasing average successful delivery rate. The larger the content library size, the bigger the helper density to get the maximum EE.

The variations of the average successful delivery rate and the energy efficiency with regard to the helper cache capacity for different content popularity distribution parameter are illustrated in Fig. 8. In Fig. 8a, the increment of the helper cache capacity lead to dramatic improvements in the average successful delivery rate owing to the overwhelming advantages of the access delivery rate obtained by the helper caching over the limited backhaul capacity installed in the MBSs. The hybrid caching policy achieve higher than the most popular

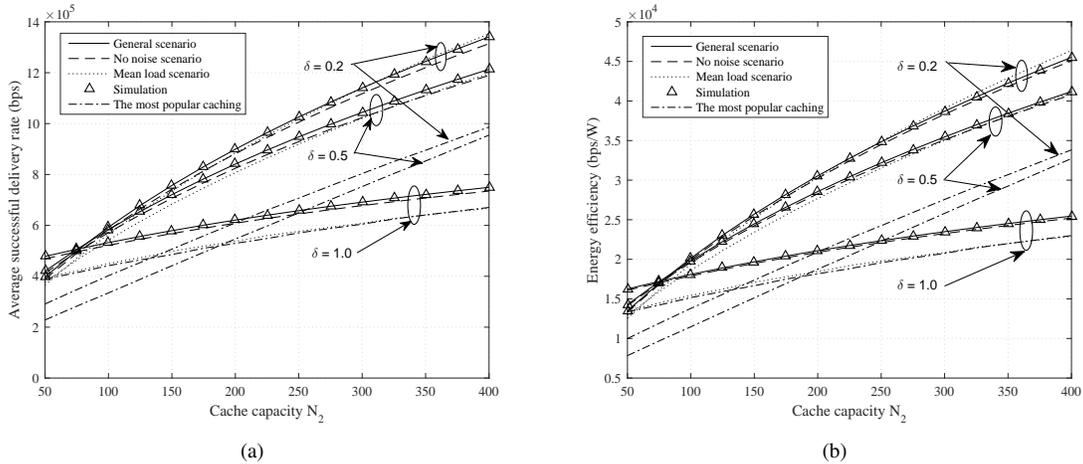


Fig. 8: Average successful delivery rate (a) and energy efficiency (b) versus helper cache capacity with different content popularity parameter,  $\lambda_2 = 1 \times 10^{-4} \text{ m}^{-2}$ ,  $C_b = 2.5 \times 10^6 \text{ bps}$ .

caching policy and the performance gain increases greatly with a growing cache capacity. A big  $\delta$  indicates that a large number of users tend to request the most popular contents in the first group and associate request with the MBSs tier. The heavy load of the MBSs decreases the access delivery rate of the MBSs and results in the serious decline of the average successful delivery rate. The trend of the EE curve in Fig. 8b is the same with average successful delivery rate curve in Fig. 6a due to the fact that the successful delivery throughput depends on the average successful delivery rate and the cache power is quite small compared with the BSs power consumption. The EE increases with the increasing helper cache capacity. The hybrid caching policy performs better than the most popular caching policy in terms of the EE and the performance gain increases with the increasing cache capacity, the maximum gain can reach 41% for  $\delta = 0.2$ . As parameter  $\delta$  increases, the EE decreases due to the sharp reduction of the throughput of the MBSs tier which makes up a large portion of the total network throughput. The EE gain over the most popular caching policy decreases.

The variations of the average successful delivery rate and the energy efficiency with regard to backhaul capacity for different helper cache capacity are illustrated in Fig. 9. In Fig. 9a, the increment of the backhaul capacity increases the backhaul delivery rate of the cache miss UEs requesting the content from the second group and improves the average successful delivery rate. All curves eventually approach the fixed value, indicating that there are enough backhaul capacity to support the cache miss UEs. Moreover, caching at helpers further increases the average successful delivery rate. The average successful delivery rate of the hybrid caching policy is higher than that of the most popular caching policy. In Fig. 9b, the EE increases with the increasing backhaul capacity until the EE gradually tends to a fixed value. Increasing the helper cache capacity enhances the EE. The larger the cache capacity, the smaller the backhaul capacity required to achieve the maximum EE. The Hybrid caching policy exhibits a significantly better EE than the most popular caching policy and the gain increases with the increasing cache capacity, reaching 32% for  $N_2 = 300$ .

## V. CONCLUSION

The cache enabled Hetnets with the limited backhaul was analyzed using the stochastic geometry. A hybrid caching policy in which the most popular contents are cached in the macro BSs tier with the deterministic caching strategy and the less popular contents are cached in the helpers tier with the probabilistic caching strategy was proposed. Taking the overall consideration of the access link, the cache and the backhaul link, the content centric association strategy is designed correspondingly. New analytical expressions of successful content delivery probability, average successful delivery rate and energy efficiency for the general scenario, the interference-limited scenario and the mean load scenario were derived. Numerical results showed that the performance can be improved greatly by the hybrid caching policy compared with the most popular caching policy. The performance gain is more obvious when the content popularity is less skewed, the cache capacity is sufficient and the helper density is relatively large. For the fixed cache capacity, there existed an optimal helper density to maximize the energy efficiency.

## APPENDIX A

According to the definition of SCDP  $C_{Mp,1}$ , we can deduce the formula as follows

$$\begin{aligned}
 C_{Mp,1} &= A_{Mp,1} \mathbb{P} \left( \frac{W}{L_1} \log_2(1 + \text{SINR}_{Mp,1}) > R_0 \right) \\
 &= \sum_{l_1=0}^{\infty} \mathbb{P} \left( \frac{P_1 h_{Mp,1} Z_{Mp,1}^{-\alpha}}{I_1 + I_2 + \sigma^2} > 2^{(l_1+1) \frac{R_0}{W}} - 1 \right) P_{L_1}(l_1 + 1),
 \end{aligned} \tag{A.1}$$

where  $L_1$  is the load of the MBS. Defining  $\delta_1 = 2^{(l_1+1) \frac{R_0}{W}} - 1$  as the equivalent SINR threshold of the MBS tier,  $C_{Mp,1}$  can be expressed as

$$\begin{aligned}
 C_{Mp,1} &\stackrel{(a)}{=} \mathbb{E}_{I_1, I_2} \exp \left( - \frac{Z_{Mp,1}^{-\alpha} (I_1 + I_2 + \sigma^2) \delta_1}{P_1} \right) \\
 &= \int_0^{\infty} \sum_{l_1=0}^{\infty} \exp \left( - \frac{z^\alpha \sigma^2 \delta_1}{P_1} \right) \mathcal{L}_{I_1} \left( \frac{z^\alpha \delta_1}{P_1} \right) \mathcal{L}_{I_2} \left( \frac{z^\alpha \delta_1}{P_1} \right)
 \end{aligned}$$

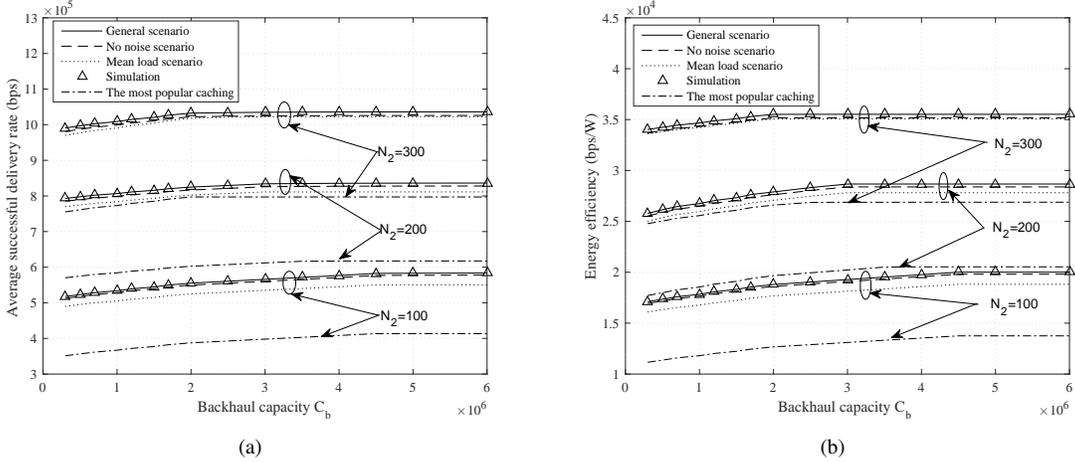


Fig. 9: Average successful delivery rate (a) and energy efficiency (b) versus backhaul capacity with different helper cache capacity,  $\lambda_2 = 1 \times 10^{-4} \text{ m}^{-2}$ ,  $\delta = 0.5$ .

$$\times P_{L_1}(l_1 + 1) f_{Z_{M_{p,1}}}(z) dz, \quad (\text{A.2})$$

where equality (a) holds due to  $h_{M_{p,1}} \sim \exp(1)$ ,  $\mathcal{L}(\cdot)$  denotes the Laplace transform,  $f_{Z_{M_{p,1}}}(z)$  is the probability density function (PDF) of the serving distance. When the tagged user requests content from the first group and connects to the MBSs tier, the serving MBSs is the closet MBSs and the PDF of the distance between the tagged user and the closet MBSs can be expressed as

$$f_{Z_{M_{p,1}}}(z) = 2\pi\lambda_1 z e^{-\pi\lambda_1 z^2}. \quad (\text{A.3})$$

The interference consists of two parts. The first part  $I_1 = \sum_{x_{1,j} \in \Phi_1 \setminus x_0} P_1 h_{1,j} z_{1,j}^{-\alpha}$  comes from all MBSs except the serving MBS located outside the circle of radius  $z$ . The second part  $I_2 = \sum_{x_{2,j} \in \Phi_2} P_2 h_{2,j} z_{2,j}^{-\alpha}$  comes from all helpers in the active mode dispersed across the entire area.

The Laplace transform of the first part of interference is derived as follows

$$\begin{aligned} \mathcal{L}_{I_1}(s) &= \mathbb{E}_{I_1}(e^{-sI_1}) \\ &= \mathbb{E}_{\Phi_1, h_{1,j}} \left( \prod_{x_{1,j} \in \Phi_1 \setminus x_0} \exp(-sP_1 h_{1,j} z_{1,j}^{-\alpha}) \right) \\ &\stackrel{(b)}{=} \mathbb{E}_{\Phi_1, h_{1,j}} \left( \prod_{x_{1,j} \in \Phi_1 \setminus x_0} \frac{1}{1 + sP_1 z_{1,j}^{-\alpha}} \right) \\ &\stackrel{(c)}{=} \exp\left(-2\pi\lambda_1 \int_z^\infty \left(1 - \frac{1}{1 + sP_1 y^{-\alpha}}\right) y dy\right), \quad (\text{A.4}) \end{aligned}$$

where (b) follows from  $h_{k,i} \sim \exp(1)$ , (c) follows from the probability generating function of the PPP.

Let  $s = \frac{z^\alpha \delta_1}{P_1}$ ,  $L_{I_1}$  is given as

$$\mathcal{L}_{I_1}\left(\frac{z^\alpha \delta_1}{P_1}\right) = \exp(-\pi\lambda_1 G(\delta_1, \alpha) z^2), \quad (\text{A.5})$$

where  $G(\delta_1, \alpha) = \frac{2\delta_1}{\alpha-2} {}_2F_1\left(1, 1 - \frac{2}{\alpha}; 2 - \frac{2}{\alpha}; -\delta_1\right)$ .

The Laplace transform of the second part of interference is derived as follows

$$\begin{aligned} \mathcal{L}_{I_2}(s) &= \mathbb{E}_{I_2}(e^{-sI_2}) = \mathbb{E}_{\Phi_2, h_{2,j}} \left( \prod_{x_{2,j} \in \Phi_{2,a}} \exp(-sP_2 h_{2,j} z_{2,j}^{-\alpha}) \right) \\ &= \mathbb{E}_{\Phi_{2,a}} \left( \prod_{i \in \Phi_{2,a}} \frac{1}{1 + sP_2 z_{2,j}^{-\alpha}} \right) \\ &= \exp\left(-2\pi p_a \lambda_2 \int_z^\infty \left(1 - \frac{1}{1 + sP_2 y^{-\alpha}}\right) y dy\right) \quad (\text{A.6}) \end{aligned}$$

Let  $s = \frac{z^\alpha \delta_1}{P_1}$ ,  $L_{I_2}$  is given as

$$\mathcal{L}_{I_2}\left(\frac{z^\alpha \delta_1}{P_1}\right) = \exp\left(-\pi p_a \lambda_2 \hat{P}_{2,1}^{2/\alpha} H(\delta_1, \alpha) z^2\right), \quad (\text{A.7})$$

where  $H(\delta_1, \alpha) = \frac{2}{\alpha} \delta_1^{2/\alpha} B\left(\frac{2}{\alpha}, 1 - \frac{2}{\alpha}\right)$ .

Integrating (A.5), (A.7) and (A.3) into (A.2) completes the proof.

## APPENDIX B

According to the definition of SCDP  $C_{L_{p,2}}$ , we can deduce the formula as follows

$$\begin{aligned} C_{L_{p,2}} &= A_{L_{p,2}} \mathbb{P}\left(\frac{W}{L_2} \log_2(1 + \text{SINR}_{L_{p,2}}) > R_0\right) \\ &= A_{L_{p,2}} \sum_{l_2=0}^{\infty} \mathbb{P}\left(\frac{P_2 h_{L_{p,2}} Z_{L_{p,2}}^{-\alpha}}{I_1 + L_{I_{L_{p^+,2}}} + L_{I_{L_{p^-,2}}} + \sigma^2} > 2^{(l_2+1)\frac{R_0}{W}} - 1\right) \\ &\times P_{L_2}(l_2 + 1) \quad (\text{B.1}) \end{aligned}$$

where  $L_2$  is the load of the helper. Defining  $\delta_2 = 2^{(l_2+1)\frac{R_0}{W}} - 1$  as the equivalent SINR threshold of the helper tier,  $C_{L_{p,2}}$  can be expressed as

$$\begin{aligned} C_{L_{p,2}} &= \mathbb{E}_{I_1, I_{L_{p^+,2}}, I_{L_{p^-,2}}} \exp\left(-\frac{Z_{L_{p,2}}^{-\alpha} (I_1 + I_{L_{p^+,2}} + I_{L_{p^-,2}} + \sigma^2) \delta_2}{P_2}\right) \\ &= A_{L_{p,2}} \int_0^\infty \sum_{l_2=0}^{\infty} \exp\left(-\frac{z^\alpha \sigma^2 \delta_2}{P_2}\right) \mathcal{L}_{I_1}\left(\frac{z^\alpha \delta_2}{P_2}\right) \mathcal{L}_{I_{L_{p^+,2}}}\left(\frac{z^\alpha \delta_2}{P_2}\right) \end{aligned}$$

$$\times \mathcal{L}_{I_{Lp^-,2}} \left( \frac{z^\alpha \delta_2}{P_2} \right) P_{L_2}(l_2 + 1) f_{Z_{Lp,2}}(z) dz, \quad (\text{B.2})$$

where  $f_{Z_{Lp,2}}(z)$  is the PDF of the serving distance when the tagged user requests content from the second group and connects to the helpers tier. Similar to the derivation of the access probability, the serving PDF can be obtained by matching the content-centric HetNets with the traditional HetNets and straightforwardly modifying the Lemmas 4 in [34] as

$$f_{Z_{Lp,2}}(z) = \frac{2\pi p_{Lp} \lambda_2 z}{A_{Lp,2}} \exp \left( -\pi \lambda_1 P_{1,2}^{2/\alpha} z^2 - \pi p_{Lp} \lambda_2 z^2 \right) \quad (\text{B.3})$$

The interference consists of three parts. The first part  $I_1 = \sum_{x_{1,j} \in \Phi_1} P_1 h_{1,j} z_{1,j}^{-\alpha}$  comes from all MBSs and the distance from the tagged user is larger than  $\hat{P}_{1,2}^{1/\alpha} z$ . The second part  $I_{Lp^+,2} = \sum_{x_{2,j} \in \Phi_{Lp^+,2,a} \setminus x_{2,0}} P_2 h_{2,j} z_{2,j}^{-\alpha}$  comes from the helpers storing the requested content in the active mode except the serving helper and the distance from the tagged user is larger than  $z$ . The third part  $I_{Lp^-,2} = \sum_{x_{2,j} \in \Phi_{Lp^-,2,a}} P_2 h_{2,j} z_{2,j}^{-\alpha}$  comes from the helpers not storing the requested content in the active mode, which are dispersed across the entire area.

Taking the Laplace transform of the three parts of the interference, and integrating (B.3) into (B.2), the proof is completed.

## APPENDIX C

Similar to the derivation in Appendix B, we can deduce  $C_{Lp,1}^w$  as follows

$$\begin{aligned} C_{Lp,1}^w &= A_{Lp,1} \int_0^\infty \sum_{l_1=0}^\infty \exp \left( -\frac{z^\alpha \sigma^2 \delta_1}{P_1} \right) \mathcal{L}_{I_1} \left( \frac{z^\alpha \delta_1}{P_1} \right) \\ &\times \mathcal{L}_{I_{Lp^+,2}} \left( \frac{z^\alpha \delta_1}{P_1} \right) \mathcal{L}_{I_{Lp^-,2}} \left( \frac{z^\alpha \delta_1}{P_1} \right) P_{L_1}(l_1 + 1) f_{Z_{Lp,1}}(z) dz \end{aligned} \quad (\text{C.1})$$

where  $f_{Z_{Lp,1}}(z)$  is the PDF of the serving distance when the tagged user requests content from the second group and connects to the MBSs tier.  $f_{Z_{Lp,1}}(z)$  can be expressed as

$$f_{Z_{Lp,1}}(z) = \frac{2\pi \lambda_1 z}{A_{Lp,1}} \exp \left( -\pi p_{Lp} \lambda_2 \hat{P}_{2,1}^{2/\alpha} z^2 - \pi \lambda_1 z^2 \right) \quad (\text{C.2})$$

The interference consists of three parts. The first part  $I_1 = \sum_{x_{1,j} \in \Phi_1 \setminus x_0} P_1 h_{1,j} z_{1,j}^{-\alpha}$  comes from all MBSs except the serving MBS and the distance from the tagged user is larger than  $z$ . The second part  $I_2 = \sum_{x_{2,j} \in \Phi_{Lp^+,2,a}} P_2 h_{2,j} z_{2,j}^{-\alpha}$  comes from the helpers storing the requested content in the active mode and the distance from the tagged user is larger than  $\hat{P}_{1,2}^{1/\alpha} z$ . The third part  $I_{Lp^-,2} = \sum_{x_{2,j} \in \Phi_{Lp^-,2,a}} P_2 h_{2,j} z_{2,j}^{-\alpha}$  comes from the helpers not storing the requested content in the active mode, which are dispersed across the entire area.

Taking the Laplace transform of the three parts of the interference, and integrating (C.2) into (C.1), the proof is completed.

## APPENDIX D

According to the definition,  $R_{Mp,1}^{suc}$  is derived as follows

$$\begin{aligned} R_{Mp,1}^{suc} &= A_{Mp,1} \mathbb{E}(R_{Mp,1} | R_{Mp,1} > R_0) \\ &= \int_0^\infty \mathbb{P} \left( \frac{W}{L_1} \log_2(1 + \text{SINR}_{Mp,1}) > r | R_{Mp,1} > R_0 \right) dr \\ &= R_0 + \int_{R_0}^\infty \int_0^\infty \sum_{l_1=0}^\infty \frac{\mathbb{P} \left( \frac{P_1 h_{Mp,1} z^{-\alpha}}{I_1 + I_2 + \sigma^2} > 2^{(l_1+1) \frac{r}{W}} - 1 \right)}{\mathbb{P} \left( \frac{P_1 h_{Mp,1} z^{-\alpha}}{I_1 + I_2 + \sigma^2} > 2^{(l_1+1) \frac{R_0}{W}} - 1 \right)} \\ &\times P_{L_1}(l_1 + 1) f_{Z_{Mp,1}}(z) dz dr \end{aligned} \quad (\text{D.1})$$

Let  $\delta_1^r = 2^{(l_1+1) \frac{r}{W}} - 1$ ,  $\delta_1 = 2^{(l_1+1) \frac{R_0}{W}} - 1$

$$\begin{aligned} R_{Mp,1}^{suc} &\stackrel{(d)}{=} R_0 + \int_{R_0}^\infty \int_0^\infty \sum_{l_1=0}^\infty \frac{\exp \left( -\frac{\delta_1^r \sigma^2 z^\alpha}{P_1} \right) \mathbb{E}_{I_1} \left( \exp \left( \frac{-\delta_1^r z^\alpha I_1}{P_1} \right) \right)}{\exp \left( -\frac{\delta_1 \sigma^2 z^\alpha}{P_1} \right) \mathbb{E}_{I_1} \left( \exp \left( \frac{-\delta_1 z^\alpha I_1}{P_1} \right) \right)} \\ &\times \frac{\mathbb{E}_{I_2} \left( \exp \left( \frac{-\delta_1^r z^\alpha I_2}{P_1} \right) \right)}{\mathbb{E}_{I_2} \left( \exp \left( \frac{-\delta_1 z^\alpha I_2}{P_1} \right) \right)} P_{L_1}(l_1 + 1) f_{Z_{Mp,1}}(z) dz dr \\ &\stackrel{(e)}{=} R_0 + \int_{R_0}^\infty \int_0^\infty \sum_{l_1=0}^\infty 2\pi \lambda_1 z \exp \left( -\frac{(\delta_1^r - \delta_1) \sigma^2 z^\alpha}{P_1} \right) \\ &\times \exp \left( -\pi \lambda_1 z^2 \left( G^*(\delta_1^r, \delta_1, \alpha) + \hat{\lambda}_{2,1} \hat{P}_{2,1}^{2/\alpha} H^*(\delta_1^r, \delta_1, \alpha) + 1 \right) \right) \\ &\times P_{L_1}(l_1 + 1) dz dr \end{aligned} \quad (\text{D.2})$$

where  $G^*(\delta_1^r, \delta_1, \alpha) = G(\delta_1^r, \alpha) - G(\delta_1, \alpha)$ ,  $H^*(\delta_1^r, \delta_1, \alpha) = H(\delta_1^r, \alpha) - H(\delta_1, \alpha)$ , (d) follows from  $h_{Mp,1} \sim \exp(1)$ , (e) follows from the Laplace transforms of the interference.

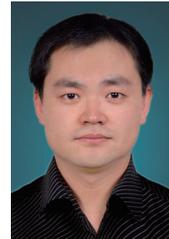
## REFERENCES

- [1] E. Amaldi, A. Capone, and F. Malucelli, "Radio planning and coverage optimization of 3G cellular networks," *Wireless Netw.*, vol. 14, no. 4, pp. 435–447, Aug. 2008.
- [2] E. Bastug and M. Bennis and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs," in *2014 11th International Symposium on Wireless Communications Systems (ISWCS)*, Aug 2014, pp. 649–653.
- [3] M. A. Abd-Elmagid and O. Ercetin and T. ElBatt, "Cache-Aided Heterogeneous Networks: Coverage and Delay Analysis," in *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*, Sep. 2017, pp. 1–5.
- [4] D. Malak and M. Al-Shalash and J. G. Andrews, "Optimizing Content Caching to Maximize the Density of Successful Receptions in Device-to-Device Networking," *IEEE Transactions on Communications*, vol. 64, no. 10, pp. 4365–4380, Oct 2016.
- [5] Z. Chen and N. Pappas and M. Kountouris, "Probabilistic Caching in Wireless D2D Networks: Cache Hit Optimal Versus Throughput Optimal," *IEEE Communications Letters*, vol. 21, no. 3, pp. 584–587, March 2017.
- [6] L. Wang, K. K. Wong, S. Jin, G. Zheng, and R. W. Heath, "A new look at physical layer security, caching, and wireless energy harvesting for heterogeneous ultra-dense networks," *IEEE Commun. Mag.*, vol. 56, no. 6, pp. 49–55, June 2018.
- [7] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.
- [8] 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.
- [9] K. Poularakis, G. Iosifidis, V. Sourlas, and L. Tassiulas, "Multicast-aware caching for small cell networks," in *2014 IEEE Wireless Communications and Networking Conference (WCNC)*, Apr. 2014, pp. 2300–2305.
- [10] C. Yang, Y. Yao, Z. Chen, and B. Xia, "Analysis on cache-enabled wireless heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 131–145, Jan. 2016.
- [11] Z. Yan, S. Chen, Y. Ou, and H. Liu, "Energy efficiency analysis of cache-enabled two-tier hetnets under different spectrum deployment strategies," *IEEE Access*, vol. 5, pp. 6791–6800, Mar. 2017.

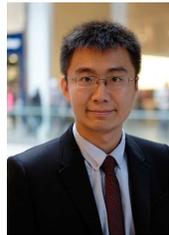
- [12] E. Bastug, M. Kountouris, M. Bennis, and M. Debbah, "On the delay of geographical caching methods in two-tiered heterogeneous networks," in *2016 IEEE 17th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, July 2016, pp. 1–5.
- [13] S. T. ul Hassan, M. Bennis, P. H. J. Nardelli, and M. Latva-aho, "Caching in wireless small cell networks: A storage-bandwidth tradeoff," *IEEE Communications Letters*, vol. 20, no. 6, pp. 1175–1178, June 2016.
- [14] W. Yi, Y. Liu, and A. Nallanathan, "Cache-enabled hetnets with millimeter wave small cells," *IEEE Trans. Commun.*, pp. 1–1, July 2018.
- [15] W. Wen, F. Zheng, Y. Cui, S. Jin, and Y. Jiang, "Cache-enabled heterogeneous wireless networks with random discontinuous transmission," in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, Apr. 2018, pp. 1–6.
- [16] S. A. R. Zaidi, M. Ghogho, and D. C. McLernon, "Information centric modeling for two-tier cache enabled cellular networks," in *2015 IEEE International Conference on Communication Workshop (ICCW)*, June 2015, pp. 80–86.
- [17] D. Liu and C. Yang, "Optimal content placement for offloading in cache-enabled heterogeneous wireless networks," in *2016 IEEE Global Communications Conference (GLOBECOM)*, Dec. 2016, pp. 1–6.
- [18] —, "Caching policy toward maximal success probability and area spectral efficiency of cache-enabled hetnets," *IEEE Transactions on Communications*, vol. 65, no. 6, pp. 2699–2714, June 2017.
- [19] S. H. Chae and W. Choi, "Caching Placement in Stochastic Wireless Caching Helper Networks: Channel Selection Diversity via Caching," *IEEE Transactions on Wireless Communications*, vol. 15, no. 10, pp. 6626–6637, Oct 2016.
- [20] K. Li and C. Yang and Z. Chen and M. Tao, "Optimization and Analysis of Probabilistic Caching in  $N$ -Tier Heterogeneous Networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 2, pp. 1283–1297, Feb 2018.
- [21] Z. Chen, J. Lee, T. Q. S. Quek, and M. Kountouris, "Cooperative caching and transmission design in cluster-centric small cell networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 3401–3415, May 2017.
- [22] 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.
- [23] D. Liu and C. Yang, "Cache-enabled heterogeneous cellular networks: Comparison and tradeoffs," in *2016 IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–6.
- [24] Z. Yang, H. Tian, S. Fan, and G. Chen, "Distributed cooperative caching in backhaul-limited small cell networks," *Electronics Letters*, vol. 53, no. 3, pp. 158–160, Feb. 2017.
- [25] D. C. Chen and T. Q. S. Quek and M. Kountouris, "Backhauling in Heterogeneous Cellular Networks: Modeling and Tradeoffs," *IEEE Transactions on Wireless Communications*, vol. 14, no. 6, pp. 3194–3206, June 2015.
- [26] G. Zhang and T. Q. S. Quek and M. Kountouris and A. Huang and H. Shan, "Fundamentals of Heterogeneous Backhaul Design: Analysis and Optimization," *IEEE Transactions on Communications*, vol. 64, no. 2, pp. 876–889, Feb 2016.
- [27] I. Atzeni and M. Maso and I. Ghamnia and E. Bastug and M. Debbah, "Flexible cache-aided networks with backhauling," in *2017 IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, July 2017, pp. 1–5.
- [28] E. Bastug and M. Bennis and M. Kountouris and M. Debbah, "Edge caching for coverage and capacity-aided heterogeneous networks," in *2016 IEEE International Symposium on Information Theory (ISIT)*, July 2016, pp. 285–289.
- [29] C. Fan, T. Zhang, Y. Liu, and Z. Zeng, "Backhaul aware analysis of cache-enabled heterogeneous networks," in *2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2019, pp. 1–7.
- [30] Fan, Congshan and Zhang, Tiankui and Zeng, Zhimin and Chen, Yue, "Energy Efficiency Analysis of Cache-Enabled Cellular Networks with Limited Backhaul," *Wireless Communications and Mobile Computing*, vol. 2018, 2018.
- [31] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Commun. Mag.*, vol. 51, no. 4, pp. 142–149, Apr. 2013.
- [32] S. N. Chiu, D. Stoyan, W. S. Kendall, and J. Mecke, *Stochastic geometry and its applications*. John Wiley & Sons, 2013.
- [33] N. Golrezaei and K. Shanmugam and A. G. Dimakis and A. F. Molisch and G. Caire, "FemtoCaching: Wireless video content delivery through distributed caching helpers," in *2012 Proceedings IEEE INFOCOM*, March 2012, pp. 1107–1115.
- [34] H. S. Jo, Y. J. Sang, P. Xia, and J. G. Andrews, "Heterogeneous cellular networks with flexible cell association: A comprehensive downlink sinr analysis," *IEEE Trans. Wireless Commun.*, vol. 11, no. 10, pp. 3484–3495, Oct. 2012.
- [35] S. Singh, H. S. Dhillon, and J. G. Andrews, "Offloading in heterogeneous networks: Modeling, analysis, and design insights," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 2484–2497, May 2013.
- [36] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. A. Imran, D. Sabella, M. J. Gonzalez, O. Blume, and A. Fehske, "How much energy is needed to run a wireless network?" *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 40–49, Oct. 2011.
- [37] N. Choi, K. Guan, D. C. Kilper, and G. Atkinson, "In-network caching effect on optimal energy consumption in content-centric networking," in *2012 IEEE International Conference on Communications (ICC)*, June 2012, pp. 2889–2894.



**Congshan Fan** received the Ph.D. degree in Information and Communication Engineering and M.S. degree in Electronic and Communication Engineering from Beijing University of Posts and Telecommunications (BUPT), China, in 2013 and 2019, respectively. Her current research interests include edge caching, UAV's network, ultra-dense networks.



**Tiankui Zhang** (M'10-SM'15) received the Ph.D. degree in Information and Communication Engineering and B.S. degree in Communication Engineering from Beijing University of Posts and Telecommunications (BUPT), China, in 2008 and 2003, respectively. Currently, he is a Professor in School of Information and Communication Engineering at BUPT. His research interests include wireless communication networks, mobile edge computing and caching, signal processing for wireless communications, content centric wireless networks. He had published more than 100 papers including journal papers on IEEE Journal on Selected Areas in Communications, IEEE Transaction on Communications, etc., and conference papers, such as IEEE GLOBECOM and IEEE ICC.



**Yuanwei Liu** (S'13-M'16-SM'19) received the B.S. and M.S. degrees from the Beijing University of Posts and Telecommunications in 2011 and 2014, respectively, and the Ph.D. degree in electrical engineering from the Queen Mary University of London, U.K., in 2016. He was with the Department of Informatics, King's College London, from 2016 to 2017, where he was a Post-Doctoral Research Fellow. He has been a Lecturer (Assistant Professor) with the School of Electronic Engineering and Computer Science, Queen Mary University of London, since 2017.

His research interests include 5G and beyond wireless networks, the Internet of Things, machine learning, and stochastic geometry. He has served as a TPC Member for many IEEE conferences, such as GLOBECOM and ICC. He received the Exemplary Reviewer Certificate of IEEE WIRELESS COMMUNICATIONS LETTERS in 2015, IEEE TRANSACTIONS ON COMMUNICATIONS in 2016 and 2017, and IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS in 2017 and 2018. He has served as the Publicity Co-Chair for VTC 2019-Fall. He is currently an Editor on the Editorial Board of the IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE COMMUNICATIONS LETTERS, and IEEE ACCESS. He also serves as a Guest Editor for IEEE JSTSP special issue on Signal Processing Advances for Non-Orthogonal Multiple Access in Next Generation Wireless Networks.



**Zhiming Zeng** received the B.S. degree in carrier communication, the M.S. degree in communication and electronic systems, and the Ph.D. degree in communication and information systems from Beijing University of Posts and Telecommunication, Beijing, China. He is currently a Professor with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications. He is a Senior Member of the China Institute of Communications, an Advanced Member of the Chinese Institute of Electronics, and a member of the Academic Committee, BUPT. His current research interests include theory and technology of next generation mobile and wireless networks.