Edge Caching in Dense Heterogeneous Cellular Networks with Massive MIMO Aided Self-backhaul

Lifeng Wang, *Member, IEEE,* Kai-Kit Wong, *Fellow, IEEE,* Sangarapillai Lambotharan, *Senior Member, IEEE,* Arumugam Nallanathan, *Fellow, IEEE,* and Maged Elkashlan, *Member, IEEE*

Abstract

This paper studies edge caching in dense heterogeneous cellular networks, in which small base stations (SBSs) with limited cache size store the popular contents, and massive multiple-input multiple-output (MIMO) aided macro base stations provide wireless self-backhaul when SBSs require the non-cached contents. We address the effects of cell load and hit probability on the successful content delivery (SCD), and evaluate the minimum required base station density for avoiding the access overload in a small cell and backhaul overload in a macrocell. We demonstrate that hit probability needs to be appropriately selected, in order to achieve SCD. We derive the massive MIMO backhaul achievable rate without downlink channel estimation, to calculate the backhaul time. We provide important insights on the interplay between cache size and SCD, and analyze the latency in such networks. We demonstrate that when non-cached contents are requested, the average delay of the non-cached content delivery could be comparable to the cached content delivery with the help of massive MIMO aided self-backhaul, if the average access rate of cached content delivery is lower than that of self-backhauled content delivery. Simulation results are presented to validate our analysis.

Index Terms

Edge caching, dense small cell, massive MIMO, self-backhaul.

L. Wang and K.-K. Wong are with the Department of Electronic and Electrical Engineering, University College London, London, UK (Email: {lifeng.wang, kai-kit.wong}@ucl.ac.uk).

S. Lambotharan is with School of Electronic, Electrical and System Engineering, Loughborough University, Loughborough Leicestershire, UK (Email: {s.lambotharan}@lboro.ac.uk)

A. Nallanathan and M. Elkashlan are with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK. (Email: {arumugam.nallanathan, maged.elkashlan}@qmul.ac.uk)

I. Introduction

A. Motivation and Background

New findings from Cisco [1] indicate that mobile video traffic accounts for the majority of mobile data traffic. To offload the traffic of the core networks and reduce the backhaul cost and latency, caching the popular contents at the edge of wireless networks becomes a promising solution [2–4]. The latest 3GPP standard has required that the fifth generation (5G) system shall support content caching applications and operators need to place the content caches close to mobile terminals [5]. In addition, the emerging radio-access technologies and wireless network architectures provide edge caching with new opportunities [6].

Recent works have focused on the caching design and analysis in various scenarios. In [7], a probabilistic caching model was considered in single-tier cellular networks and the optimal content placement was designed to maximize the total hit probability. In [8], a stochastic content multicast scheduling problem was formulated to jointly minimize the average network delay and power costs in heterogeneous cellular networks (HetNets), and a structure-aware optimal algorithm was proposed to solve this problem. Caching cooperation in multi-tier HetNets was studied in [9], where a low-complexity suboptimal solution was developed to maximize the capacity in such networks. Caching in device-to-device (D2D) networks was investigated in the literature such as [10, 11]. In [10], a holistic design on D2D caching at multi-frequency band including sub-6 GHz and millimeter wave (mmWave) was presented. In [11], the performance difference between maximizing hit probability and maximizing cache-aided throughput in D2D caching networks was evaluated. The work of [12] showed that in multi-hop relaying systems, the efficiency of caching could be further improved by using collaborative cache-enabled relaying. Joint design of cloud and edge caching in fog radio access networks were introduced in [13, 14], where the popular contents were cached at the remote radio heads. However, prior works [7–13] did not present design and insights involving edge caching in the future dense/ultra-dense cellular networks (e.g., 5G) with backhaul concerns, where wireless self-backhauling shall be supported [4].

Cache-enabled small cell networks with stochastic models have been investigated in the literature such as [15–19]. Cluster-centric caching with base station (BS) cooperation was studied in [15], where the tradeoff between transmission diversity and content diversity was revealed. In [16], two cache-enabled BS modes were considered, i.e., always-on and dynamic on-off, and it was assumed that the intensity of BSs is much larger than the intensity of mobile terminals. The work of [17–19] concentrated on the cache-enabled multi-tier HetNets.

Specifically, [17] and [18] studied optimal content placement under probabilistic caching strategy, and [19] considered the joint BS caching and cooperation, in contrast to the single-tier case in [15]. However, [15–19] only aimed to maximize the probability that the requested content is not only cached but also successfully delivered. In realistic networks, when users' requested contents are not cached at their associated BSs, they will obtain their requested contents from the core networks via wired/wireless backhaul, which also needs to be studied in cache-enabled cellular networks.

In fact, existing contributions such as [20–22] have studied the effects of backhaul on content delivery in cache-enabled networks. The work of [20] considered that non-cached contents were obtained via backhaul, and designed a downlink content-centric sparse multicast beamforming in the cache-enabled cloud radio access network (Cloud-RAN) architecture, to minimize the weighted sum of backhaul cost and transmit power. In [21], the network successful content delivery consisting of cached content delivery and backhauled content delivery was studied, and the optimization problem was formulated to minimize the cache size under quality-of-service constraint. The work of [22] analyzed the capacity scaling law when there are limited number of wired backhaul in single-tier networks, and showed that cache size needs to be large enough to achieve linear capacity scaling. However, none of [20–22] has studied the cache-enabled cellular networks with specified wireless backhaul transmission, such as massive multiple-input multiple-output (MIMO) aided self-backhaul.

B. Novelty and Contributions

In this paper, we focus on the edge caching in dense HetNets with massive MIMO aided self-backhaul, which has not been understood yet. Considering massive MIMO aided self-backhaul is motivated by the fact that it is challenging to let each backhaul link be fiber-optic in such networks and massive MIMO can support high-speed transmissions thanks to large array gains [4]. Our contributions are summarized as follows:

- In contrast to the prior works such as [15–22], we consider cache-enabled HetNets, in which randomly located small BSs (SBSs) cache finite popular contents, and macro BSs (MBSs) equipped with massive MIMO antennas provide wireless backhaul to deliver the non-cached requested contents to the SBSs. Moreover, we also consider the resource allocation when multiple users request the contents from the same SBS, which has not been studied in a cache-enabled stochastic model.
- We first derive the successful content delivery probability when the requested content is cached at the SBS. The maximum small cell load is calculated, and the minimum required

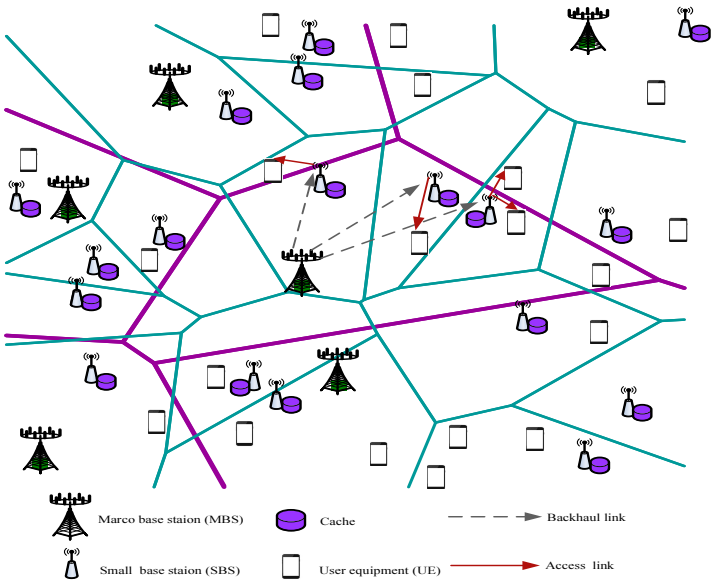

Fig. 1. An illustration of cache-enabled heterogeneous cellular network with massive MIMO backhaul.

density of SBSs for avoiding access overload is obtained. We show that hit probability needs to be lower than a certain value, to guarantee successful cached content delivery.

- We derive the successful content delivery probability when the requested content is not cached and has to be obtained via massive MIMO backhaul. We analyze the massive MIMO backhaul achievable rate when downlink channel estimation is not available, to evaluate the backhaul time. The minimum required density of MBSs for avoiding backhaul overload is obtained. We show that hit probability needs to be higher than a certain value, to guarantee successful self-backhauled content delivery.
- We analyze the effects of cache size on the successful content delivery, and provide important insights on the interplay between time-frequency resource allocation and cache size from the perspective of successful content delivery probability. We characterize the latency in terms of average delay in such networks, and confirm that when the requested contents are not cached, the average delay of the non-cached content delivery could be comparable to the cached content delivery with the assistance of massive MIMO backhaul, if the average access rate of cached content delivery is lower than that of self-backhauled content delivery.

II. Network Model

As shown in Fig. 1, we consider a two-tier self-backhauled HetNet, in which each single-antenna SBS with finite cache size can store popular contents to serve user equipment (UEs),

and each massive MIMO aided MBS equipped with N antennas accesses to the core networks via optical fiber and delivers the non-cached contents to the SBSs via wireless backhaul. UEs, SBSs, and MBSs are assumed to be distributed following independent homogeneous Poisson point processes (HPPPs) denoted by Φ_{U} with the density λ_{U}, Φ_{S} with the density λ_{S}, and Φ_{M} with the density λ_{M}, respectively. It is assumed that UEs are associated with the SBSs that can provide the maximum average received power, which is also utilized in 4G networks [6]. In addition, each channel undergoes independent and identically distributed (i.i.d.) quasi-static Rayleigh fading.

A. Content Placement

Content placement mechanism is mainly designed based on content popularity [4]. We assume that there is a finite content library denoted as $\mathcal{F} := \{f_1, \ldots, f_j, \ldots, f_J\}$, where f_j is the j-th most popular content and the number of contents is J. The request probability for the j-th most popular content is commonly-modeled by following the Zipf distribution, which is expressed as [23]

$$a_j = j^{-\varsigma} / \sum\nolimits_{m=1}^{J} m^{-\varsigma}, \tag{1}$$

where ς is the Zipf exponent to represent the popularity skewness [23]. Each content is assumed to be unit size and each SBS can only cache $L\,(L \ll J)$ contents. We employ the probabilistic caching strategy [7], i.e., the probability that the content j is cached at an arbitrary SBS is $q_j (0 \leq q_j \leq 1)$, and $\sum\limits_{j=1}^{J} q_j \leq L$.

B. Self-backhaul Load

We assume that the access and backhaul links orthogonally share the sub-6 GHz spectrum, and the bandwidths allocated to the access and backhaul links are ηW and $(1-\eta) W$, respectively, where η is the fraction factor and W is the system bandwidth. The number of UEs that is associated with an SBS is denoted as K, and UEs in the same small cell are time-dividedly served with equal-time sharing. Thus, the fraction of time-frequency resources allocated to each access link is $\eta W/K$ during the cached content delivery. When an associated SBS does not cache the requested content, it has to be connected to an MBS that provides the strongest wireless backhaul link such that the requested content can be obtained from core networks. Let $S_j\,(N \gg S_j)$ denote the number of SBSs served by the j-th MBS $(j \in \Phi_{\mathrm{M}})$ for wireless backhaul.

Since the hit probability that UE's requested content file is stored at an SBS is $q_{\mathrm{hit}} = \sum_{j=1}^{J} a_j q_j$, the set of SBSs can be partitioned into two independent HPPPs $\Phi_{\mathrm{S}}^{\mathrm{a}}$ and $\Phi_{\mathrm{S}}^{\mathrm{b}}$ based on the thinning theorem [24], where $\Phi_{\mathrm{S}}^{\mathrm{a}}$ with the density $\lambda_{\mathrm{S}} q_{\mathrm{hit}}$ denotes the point process of SBSs with access links, and $\Phi_{\mathrm{S}}^{\mathrm{b}}$ with the density $\lambda_{\mathrm{S}}\left(1-q_{\mathrm{hit}}\right)$ denotes the point process of SBSs with backhaul links. Let $\omega_{\mathrm{b}} = \lambda_{\mathrm{S}}\left(1-q_{\mathrm{hit}}\right)/\lambda_{\mathrm{M}}$ represent the average number of SBSs served by an MBS for wireless backhaul.

C. Resource Allocation Model

We consider the saturated traffic condition, i.e., all the SBSs keep active to serve their associated UEs.

1) Access: When the requested content is stored at a typical SBS, the rate for a typical access link is given by

$$R_{\mathrm{a}} = \frac{\eta W}{K}\log_2\left(1+\frac{P_{\mathrm{a}} h_o L\left(|X_o|\right)}{\underbrace{\sum_{i\in\Phi_{\mathrm{S}}^{\mathrm{a}}\setminus\{o\}} P_{\mathrm{a}} h_i L\left(|X_{o,i}|\right)}_{I_{\mathrm{a}}}+\sigma_{\mathrm{a}}^2}\right), \tag{2}$$

where I_{a} denotes the total interference power from other SBSs, P_{a} is the SBS's transmit power, $L\left(|X|\right)=\beta\left(|X|\right)^{-\alpha_{\mathrm{a}}}$ denotes the path loss with frequency dependent constant value β, distance $|X|$ and path loss exponent α_{a}, $h_o \sim \exp(1)$ and $|X_o|$ are the small-scale fading channel power gain and distance between the typical UE and its associated SBS respectively, $h_i \sim \exp(1)$ and $|X_{o,i}|$ are the small-scale fading interfering channel power gain and distance between the typical UE and the interfering SBS $i \in \Phi_{\mathrm{S}}^{\mathrm{a}}\setminus\{o\}$ (except the typical SBS o) respectively, and σ_a^2 is the noise power at the typical UE.

2) Self-Backhaul: When the requested content is not stored at SBSs, it is obtained through massive MIMO backhaul. For massive MIMO backhaul link, we consider that massive MIMO enabled MBS adopts zero-forcing beamforming with equal power allocation [25]. In such a massive MIMO self-backhauled network, SBSs will not perform any channel estimation, and we adopt an achievable backhaul transmission rate as confirmed in [26, 27]. Therefore, given a typical distance $|Y_o|$ between the typical SBS and its associated MBS, the rate for a typical massive MIMO backhaul link is given by

$$R_{\mathrm{b}} = (1-\eta)\, W \log_2\left(1+\mathrm{SINR}_{\mathrm{b}}\right) \tag{3}$$

with

$$\mathrm{SINR_b}=\frac{\frac{P_\mathrm{b}}{S_o}\left(\mathbb{E}\left\{\sqrt{g_o}\right\}\right)^2 L\left(|Y_o|\right)}{\frac{P_\mathrm{b}}{S_o}\left(\sqrt{g_o}-\mathbb{E}\left\{\sqrt{g_o}\right\}\right)^2 L\left(|Y_o|\right)+\underbrace{\sum_{j\in\Phi_\mathrm{M}\setminus\{o\}}\frac{P_\mathrm{b}}{S_j}g_j L\left(|Y_{o,j}|\right)}_{I_\mathrm{b}}+\sigma_\mathrm{b}^2},$$

where $\mathbb{E}\{\cdot\}$ is the expectation operator, I_b denotes the total interference power from other MBSs, P_b is the MBS's transmit power, $L\left(|Y|\right)=\beta\left(|Y|\right)^{-\alpha_\mathrm{b}}$ denotes the path loss with the distance $|Y|$ and path loss exponent α_b, $g_o\sim\Gamma\left(N-S_o+1,1\right)$ is the small-scale fading channel power gain between the typical SBS and its associated MBS, $g_j\sim\Gamma\left(S_j,1\right)$[1] and $|Y_{o,j}|$ are the small-scale fading interfering channel power gain and distance between the typical SBS and interfering MBS j, respectively, and σ_b^2 is the noise power at the typical SBS.

After obtaining the requested content via backhaul, the associated SBS delivers it to the corresponding UE. In this case, the corresponding access-link rate is expressed as

$$R_{\mathrm{a}'}=\frac{(1-\eta)W}{K}\times \log_2\left(1+\frac{P_\mathrm{a}h_o L\left(|X_o|\right)}{\underbrace{\sum_{i'\in\Phi_\mathrm{S}^\mathrm{b}\setminus\{o\}}P_\mathrm{a}h_{i'}L\left(\left|X_{o,i'}\right|\right)+\sigma_{\mathrm{a}'}^2}_{I_{\mathrm{a}'}}}\right), \tag{4}$$

where $I_{\mathrm{a}'}$ is the total interference power, $h_{i'}\sim\exp(1)$ and $L\left(\left|X_{o,i'}\right|\right)=\beta\left(\left|X_{o,i'}\right|\right)^{-\alpha_\mathrm{a}}$ are the small-scale fading channel power gain and pathloss between the typical SBS and interfering SBS $i'\in\Phi_\mathrm{S}^\mathrm{b}\setminus\{o\}$, respectively, and $\sigma_{\mathrm{a}'}^2$ is the noise power at the typical UE.

From (3) and (4), we see that to cut latency, massive MIMO backhaul link needs to be of high-speed, which can be achieved by using large array gains and large bandwidths via carrier aggregations (CA). In the following section, we will further examine how much backhaul time is needed at an achievable backhaul rate.

III. Content Delivery Efficiency

In this paper, there are two cases for successful content delivery (SCD), i.e., 1) when the associated BS has cached the requested content, SCD occurs if the time for successfully delivering Q bits will not exceed the threshold T_th; and 2) when the requested content is not cached at the associated BS and needs to be obtained via massive MIMO backhaul, SCD occurs if the total time for successfully delivering Q bits to the UE is less than T_th.

[1]$\Gamma\left(\cdot,\cdot\right)$ is the upper incomplete gamma function [28, (8.350)].

A. Cached Content Delivery

Different from [15, 16, 18] where it is assumed that each small cell has only one active UE, we evaluate SCD probability by considering multiple UEs served by an SBS in practice, and analyze the effect of resource allocation on SCD probability. We first have the following important theorem.

***Theorem** 1:* When a requested content is stored at the typical SBS, the SCD probability is derived as

$$\Psi_{\mathrm{SCD}}^{\mathrm{a}}\left(Q, T_{\mathrm{th}}\right)=\sum_{k=1}^{K_{\max}^{\mathrm{a}}} \mathcal{P}_{\frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}}}\left(k\right), \tag{5}$$

where $\mathcal{P}_{\frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}}}\left(k\right)$ is the probability mass function (PMF) that there are other $k-1$ UEs (except typical UE) served by the typical SBS, and is given by $\mathcal{P}_{\frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}}}\left(k\right)=\frac{\gamma^{\gamma}}{(k-1)!}\frac{\Gamma(k+\gamma)}{\Gamma(\gamma)}\frac{\left(\frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}}\right)^{k-1}}{\left(\gamma+\frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}}\right)^{k+\gamma}}$ with $\gamma=3.5$ [29]. In (5), $K=K_{\max}^{\mathrm{a}}$ is the maximum load in a typical small cell, and can be quickly obtained by using **Algorithm 1** to solve the following equation

$$\frac{2^{\frac{K_{\max}^{\mathrm{a}}Q}{\eta W T_{\mathrm{th}}}+1}-2}{\alpha_{\mathrm{a}}-2}\chi_{k}^{\mathrm{a}}\left(K_{\max}^{\mathrm{a}}\right)=\frac{1-\epsilon}{q_{\mathrm{hit}}\epsilon}, \tag{6}$$

where $\chi_{k}^{\mathrm{a}}\left(K_{\max}^{\mathrm{a}}\right)={}_2F_1\left[1,1-\frac{2}{\alpha_{\mathrm{a}}},2-\frac{2}{\alpha_{\mathrm{a}}},1-2^{\frac{K_{\max}^{\mathrm{a}}Q}{\eta W T_{\mathrm{th}}}}\right]$, ${}_2F_1\left[\cdot,\cdot;\cdot;\cdot\right]$ is the Gauss hypergeometric function [28, (9.142)], and ϵ is the predefined threshold, i.e., SCD occurs when the probability that R_{a} is larger than $\frac{Q}{T_{\mathrm{th}}}$ is above ϵ.

***Proof** 1:* See Appendix A.

It is implied from **Theorem 1** that in the dense small cell networks (i.e., interference-limited)[2], the SCD probability depends on the ratio of UE density to SBS density and hit probability given the time-frequency resource allocation. Based on **Theorem 1**, we have

***Corollary** 1:* From (6), we see that to achieve the load $K=K_{\max}^{\mathrm{a}}\geq 1$ in a small cell, the hit probability should satisfy

$$q_{\mathrm{hit}}\leq\min\left\{\Xi_{\mathrm{a}}\frac{1-\epsilon}{\epsilon},1\right\}, \tag{7}$$

where $\Xi_{\mathrm{a}}=\left(\frac{2^{\frac{Q}{\eta W T_{\mathrm{th}}}+1}-2}{\alpha_{\mathrm{a}}-2}\chi_{k}^{\mathrm{a}}\left(1\right)\right)^{-1}$.

It is indicated from (7) that there is an upper-bound on the hit probability, which can be explained by the fact that when more UEs can obtain their requested contents from their associated SBSs in dense cellular networks with large hit probability, there will also be more interference from nearby SBSs that degrades the cached content delivery.

[2]The near-field pathloss exponent is assumed to be larger than 2 [4].

Algorithm 1 One-dimension Search

1: **if** $t=0$

2: Initialize $\varphi=\frac{1-\epsilon}{q_{\rm hit}\epsilon}$, $k^l=1, k^h=10\times\frac{\lambda_{\rm U}}{\lambda_{\rm S}}$, and calculate

$$F^l=\frac{2^{\frac{k^lQ}{\eta WT_{\rm th}}+1}-2}{\alpha_{\rm a}-2}{}_2F_1\left[1,1-\frac{2}{\alpha_{\rm a}},2-\frac{2}{\alpha_{\rm a}},1-2^{\frac{k^lQ}{\eta WT_{\rm th}}}\right]$$

and

$$F^h=\frac{2^{\frac{k^hQ}{\eta WT_{\rm th}}+1}-2}{\alpha_{\rm a}-2}{}_2F_1\left[1,1-\frac{2}{\alpha_{\rm a}},2-\frac{2}{\alpha_{\rm a}},1-2^{\frac{k^hQ}{\eta WT_{\rm th}}}\right]$$

3: **else**

4: **While** $F^l\neq\varphi$ and $F^h\neq\varphi$

5: Let $k=\frac{k^l+k^h}{2}$, and compute F_k.

6: **if** $F_k=\varphi$

7: The optimal k^* is obtained, i.e., $K^{\rm a}_{\max}=round\,(k^*)$.

8: **break**

9: **elseif** $F_k<\varphi$

10: $k^l=k$.

11: **else** $F_k>\varphi$

12: $k^h=k$.

13: **end if**

14: **end while**

15: **end if**

In realistic networks, there may be overload issues when the scale of small cells is not adequate to support large level of connectivity, which needs to be addressed. Therefore, given a specified scale of UEs $\lambda_{\rm U}$, we evaluate the minimum required scale of small cells as follows.

***Corollary** 2:* To mitigate the harm of overloading, the minimum required SBS density needs to satisfy

$$\lambda_{\rm S}=\begin{cases}\frac{\lambda_{\rm U}}{K^{\rm a}_{\max}+1},\ if\ \mathcal{P}_{\frac{\lambda_{\rm U}}{\lambda_{\rm S}}=K^{\rm a}_{\max}+1}\left(k=K^{\rm a}_{\max}+1\right)\leq\rho,\\ \frac{\lambda_{\rm U}}{\mu_{\rm a}},\ if\ \mathcal{P}_{\frac{\lambda_{\rm U}}{\lambda_{\rm S}}=K^{\rm a}_{\max}+1}\left(k=K^{\rm a}_{\max}+1\right)>\rho,\end{cases}\tag{8}$$

where $\mu_{\rm a}\in\left(0,\frac{K^{\rm a}_{\max}\gamma}{\gamma+1}\right]$ is the solution of $\mathcal{P}_{\frac{\lambda_{\rm U}}{\lambda_{\rm S}}=\mu_{\rm a}}\left(k=K^{\rm a}_{\max}+1\right)=\rho$ with arbitrary small $\rho>0$, and can be easily obtained via one-dimension search, similar to **Algorithm 1**. Such network deployment given in (8) can guarantee $\mathcal{P}_{\frac{\lambda_{\rm U}}{\lambda_{\rm S}}}\left(k\right)\leq\rho,\forall k>K^{\rm a}_{\max}$.

***Proof** 2:* See Appendix B.

From (8), we see that the minimum required density of SBSs only depends on the maximum load of a small cell and the density of UEs in dense cache-enabled cellular networks.

B. Self-backhauled Content Delivery

1) Massive MIMO Backhaul: When the required content is not stored at the typical SBS, SBS has to obtain it from the core networks via massive MIMO backhaul. Therefore, we need to evaluate the backhaul time for delivering the requested content to the typical SBS. Given the load S_o in a typical macrocell, the achievable transmission rate for a typical backhaul link is given by

$$\overline{R}_{\mathrm{b}}\left(S_o\right)=\left(1-\eta\right)W\int_{r_{\mathrm{b}}}^{\infty}C_{\mathrm{b}}\left(y\right)\frac{2\pi\lambda_{\mathrm{M}}y\exp\left(-\pi\lambda_{\mathrm{M}}y^2\right)}{\exp\left(-\pi\lambda_{\mathrm{M}}r_{\mathrm{b}}^2\right)}dy, \tag{9}$$

where $C_{\mathrm{b}}\left(y\right)=\log_2\left(1+\frac{\frac{P_{\mathrm{b}}}{S_o}\Xi_1(y)}{\frac{P_{\mathrm{b}}}{S_o}\Xi_2(y)+\Xi_3(y)+\sigma_{\mathrm{b}}^2}\right)$ with $\Xi_1\left(y\right)=L\left(y\right)\left(\frac{\Gamma\left(N-S_o+\frac{3}{2}\right)}{\Gamma(N-S_o+1)}\right)^2$, $\Xi_2\left(y\right)=(N-S_o+1)L\left(y\right)-\Xi_1$, and $\Xi_3\left(y\right)=P_{\mathrm{b}}2\pi\lambda_{\mathrm{M}}\beta\frac{y^{2-\alpha_{\mathrm{b}}}}{\alpha_{\mathrm{b}}-2}$, and r_{b} is the minimum distance between the typical MBS and its associated SBS. A detailed derivation of (9) is provided in Appendix C. Therefore, the time for delivering Q bits to the typical SBS via wireless backhaul is $T_1=\frac{Q}{\overline{R}_{\mathrm{b}}}$. When the number of antennas at the MBS grows large, we have the following corollary.

***Corollary** 3:* For large N, the achievable transmission rate for a typical backhaul link is tightly lower-bounded as

$$\overline{R}_{\mathrm{b}}^{\mathrm{Low}}\left(S_o\right)=\left(1-\eta\right)W\log_2\left(1+P_{\mathrm{b}}\beta\frac{N-S_o+\frac{1}{2}}{S_o}e^{\Delta_1-\Delta_2}\right), \tag{10}$$

where

$$\begin{cases}\overline{\Delta}_1=-\alpha_{\mathrm{b}}e^{\pi\lambda_{\mathrm{M}}r_{\mathrm{b}}^2}\Big(-\dfrac{Ei\big(-r_{\mathrm{b}}^2\pi\lambda_{\mathrm{M}}\big)}{2}+e^{-r_{\mathrm{b}}^2\pi\lambda_{\mathrm{M}}}\ln r_{\mathrm{b}}\Big),\\ \overline{\Delta}_2=\displaystyle\int_{r_{\mathrm{b}}}^{\infty}\ln\left(\frac{P_{\mathrm{b}}\beta}{2S_o}y^{-r_{\mathrm{b}}}+P_{\mathrm{b}}2\pi\lambda_{\mathrm{M}}\beta\frac{y^{2-\alpha_{\mathrm{b}}}}{\alpha_{\mathrm{b}}-2}+\sigma_{\mathrm{b}}^2\right)\\ \qquad\qquad\times\dfrac{2\pi\lambda_{\mathrm{M}}y}{\exp\left(-\pi\lambda_{\mathrm{M}}r_{\mathrm{b}}^2\right)}\exp\left(-\pi\lambda_{\mathrm{M}}y^2\right)dy,\end{cases}$$

in which $Ei\left(z\right)$ is the exponential integral given by $Ei\left(z\right)=-\int_{-z}^{\infty}\frac{e^{-t}}{t}dt$ [28]. Based on (10), the typical MBS's required time for delivering Q bits to its associated SBS satisfies

$$T_1\leq\frac{Q(1-\eta)^{-1}W^{-1}}{\log_2\left(1+P_{\mathrm{b}}\beta\frac{\left(N-S_o+\frac{1}{2}\right)}{S_o}e^{\overline{\Delta}_1-\overline{\Delta}_2}\right)}. \tag{11}$$

***Proof** 3:* See Appendix D.

It is explicitly shown from **Corollary 3** that large number of antennas and bandwidths are required, in order to significantly cut the wireless backhaul time. From (11), we see that the backhaul time can at least be cut proportionally to $1/\ln N$.

In the self-backhauled networks, the number of SBSs being simultaneously served by an MBS for wireless backhaul should not exceed the maximum value denoted by $S_{\max}$,

i.e., $S_o \leq S_{\max}$; otherwise high-speed massive MIMO aided backhaul transmission cannot be guaranteed. Hence, given the minimum required backhaul transmission rate $R_{\rm b}^{\min}$, the maximum backhaul load of a typical massive MIMO MBS is the solution of $\overline{R}_{\rm b}\left(S_{\max}\right) = R_{\rm b}^{\min}$, which can be quickly obtained by using one-dimension search since $\overline{R}_{\rm b}\left(S_o\right)$ is a decreasing function of S_o for large N, as suggested in Appendix D. After obtaining $S_{\max}$, we can obtain the minimum number of massive MIMO aided MBSs that needs to be deployed, in order to mitigate the backhaul overload.

***Corollary** 4:* Similar to **Corollary 2**, the minimum required density of MBSs is given by

$$\lambda_{\rm M} = \begin{cases} \frac{\lambda_{\rm S}(1-q_{\rm hit})}{S_{\max}+1}, \ if \ \mathcal{P}_{\omega_{\rm b}=S_{\max}+1}\left(\ell = S_{\max}+1\right) \leq \rho, \\ \frac{\lambda_{\rm S}(1-q_{\rm hit})}{\mu_{\rm b}}, \ if \ \mathcal{P}_{\omega_{\rm b}=S_{\max}+1}\left(\ell = S_{\max}+1\right) > \rho, \end{cases} \tag{12}$$

where $\mathcal{P}_{\omega_{\rm b}}\left(\ell\right) = \frac{\gamma^{\gamma}}{(\ell-1)!}\frac{\Gamma(\ell+\gamma)}{\Gamma(\gamma)}\frac{\left(\omega_{\rm b}\right)^{\ell-1}}{\left(\gamma+\omega_{\rm b}\right)^{\ell+\gamma}}$, $\mu_{\rm b} \in \left(0, \frac{S_{\max}\gamma}{\gamma+1}\right]$ is the solution of $\mathcal{P}_{\omega_{\rm b}=\mu_{\rm b}}\left(\ell = S_{\max}+1\right) = \rho$ with arbitrary small $\rho > 0$, and can be easily obtained via one-dimension search.

It is explicitly shown in (12) that higher hit probability can significantly reduce the scale of MBSs because of less backhaul.

2) Access: After obtaining the required content via backhaul, the typical SBS transmits it to the associated UE. Thus, we have the following important theorem.

***Theorem** 2:* When the required content is not stored at the typical SBS and has to be obtained via massive MIMO self-backhaul, the SCD probability is derived as

$$\Psi_{\rm SCD}^{\rm b}\left(Q, T_{\rm th}\right) = \sum_{k=1}^{K_{\max}^{\rm b}} \mathcal{P}_{\frac{\lambda_{\rm U}}{\lambda_{\rm S}}}\left(k\right), \tag{13}$$

where $K_{\max}^{\rm b}$ is the maximum number of UEs that a typical small cell can serve when the typical UE's content needs to be attained via backhaul, and $K_{\max}^{\rm b}$ can be obtained by solving the following equation[3]

$$\frac{2^{\frac{K_{\max}^{\rm b}Q}{(1-\eta)W\left(T_{\rm th}-T_1\right)}+1}-2}{\alpha_{\rm a}-2}\chi_k^{\rm b}\left(K_{\max}^{\rm b}\right) = \frac{1-\epsilon}{\left(1-q_{\rm hit}\right)\epsilon} \tag{14}$$

with $\chi_k^{\rm b}\left(K_{\max}^{\rm b}\right) = {}_2F_1\left[1, \frac{\alpha_{\rm a}-2}{\alpha_{\rm a}}, \frac{2\alpha_{\rm a}-2}{\alpha_{\rm a}}, 1-2^{\frac{K_{\max}^{\rm b}Q}{(1-\eta)W\left(T_{\rm th}-T_1\right)}}\right]$, and the minimum required SBS density for mitigating overload is given from (8) by interchanging $K_{\max}^{\rm a} \to K_{\max}^{\rm b}$.

***Proof** 4:* See Appendix E.

It is indicated from (14) that when a typical UE's requested content is not stored at the typical SBS, the number of UEs that can be served by the typical SBS decreases with increasing backhaul time. Based on **Theorem 2**, we have the following corollary

[3]It can be solved by following **Algorithm 1**.

***Corollary** 5:* From (14), we see that to achieve the load $K = K^{\mathrm{b}}_{\max} \geq 1$ in a small cell, the hit probability should satisfy

$$q_{\mathrm{hit}} \geq \left[1 - \Xi_{\mathrm{b}} \frac{1-\epsilon}{\epsilon}\right]^{+}, \tag{15}$$

where $\Xi_{\mathrm{b}} = \left(\frac{Q}{2^{\frac{Q}{(1-\eta)W\left(T_{\mathrm{th}}-T_1\right)}+1}-2} \chi^{\mathrm{b}}_{k}(1) \right)^{-1}$, and $[x]^{+} = \max\{x, 0\}$.

From (15), we see that there is a lower-bound on the hit probability, i.e., minimum cache capacity is demanded at the SBS, since more backhaul results in more interference, which will degrade the self-backhauled content delivery.

***Corollary** 6:* After obtaining the maximum load $K^{\mathrm{b}}_{\max}$, we can calculate the minimum required SBS density given from (8) by interchanging $K^{\mathrm{a}}_{\max} \to K^{\mathrm{b}}_{\max}$, to overcome overload.

Based on **Theorem 1** and **Theorem 2**, the SCD probability in dense cellular networks with massive MIMO self-backhaul for a typical UE is calculated as

$$\begin{aligned}
\Psi_{\mathrm{SCD}}(Q, T_{\mathrm{th}}) &= q_{\mathrm{hit}} \Psi^{\mathrm{a}}_{\mathrm{SCD}}(Q, T_{\mathrm{th}}) + (1 - q_{\mathrm{hit}}) \Psi^{\mathrm{b}}_{\mathrm{SCD}}(Q, T_{\mathrm{th}}) \\
&= \begin{cases}
\sum\limits_{k=1}^{K^{\mathrm{b}}_{\max}} \mathcal{P}_{\frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}}}(k) + q_{\mathrm{hit}} \sum\limits_{k=K^{\mathrm{b}}_{\max}+1}^{K^{\mathrm{a}}_{\max}} \mathcal{P}_{\frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}}}(k), \ K^{\mathrm{a}}_{\max} \geq K^{\mathrm{b}}_{\max}, \\
\sum\limits_{k=1}^{K^{\mathrm{a}}_{\max}} \mathcal{P}_{\frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}}}(k) + (1 - q_{\mathrm{hit}}) \\
\qquad \times \sum\limits_{k=K^{\mathrm{a}}_{\max}+1}^{K^{\mathrm{b}}_{\max}} \mathcal{P}_{\frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}}}(k), \ K^{\mathrm{a}}_{\max} < K^{\mathrm{b}}_{\max},
\end{cases}
\end{aligned} \tag{16}$$

where $K^{\mathrm{a}}_{\max}$ and $K^{\mathrm{b}}_{\max}$ are given by (6) and (14), respectively.

The SCD probability given in (16) can be intuitively understood based on the fact that when the small cell load is light, UEs' requested contents can be successfully delivered whether they are cached or obtained from the core networks via massive MIMO backhaul. However, after a critical value of cell load, UEs can only obtain their requested contents that are cached by the SBSs or via backhaul, which depends on the maximum cell load in cached content delivery and self-backhauled content delivery cases.

IV. Content Placement, Cache Size and Latency

In this section, we study the effects of content placement and cache size on the content delivery performance. Then, we evaluate the latency in such networks.

A. Content Placement and Cache Size

As shown in (16), hit probability plays an important role in content delivery. Since hit probability depends on the cache size and content placement, SBSs with large storage capacity

can cache more popular contents, to avoid frequent backhaul and cut backhaul cost and latency. Therefore, higher hit probability is meaningful to cut the network's operational and capital expenditures (OPEX, CAPEX). Given the SBS's cache size, different content placement strategies may result in various hit probability, and caching the most popular contents (MPC) can achieve the highest hit probability, which is commonly-considered in the literature involving edge caching such as [13, 30]. Therefore, we consider MPC caching and analyze the appropriate cache size in such networks. Considering the fact that for large J with MPC caching, $q_{\rm hit} = \sum_{j=1}^{L} a_j \approx \left(\frac{L}{J}\right)^{1-\varsigma}$, we have

***Corollary** 7:* Given $\frac{T_{\rm th}-T_{\rm l}}{T_{\rm th}} \leq \frac{\eta}{1-\eta}$ (i.e., more time-frequency resources are allocated to the cached content delivery), the SCD probability is

$$\Psi_{\rm SCD}\left(Q, T_{\rm th}\right) = \sum_{k=1}^{K_{\max}^{\rm b}} \mathcal{P}_{\frac{\lambda_{\rm U}}{\lambda_{\rm S}}}\left(k\right), \tag{17}$$

and it is an increasing function of the cache size, if the cache size $L \in \left[J\left(\left[1-\Xi_{\rm b}\frac{1-\epsilon}{\epsilon}\right]^{+}\right)^{\frac{1}{1-\varsigma}}, J\left(\frac{1}{2}\right)^{\frac{1}{1-\varsigma}}\right]$ and the minimum SBS density satisfies the condition given in **Corollary 6**; Given $\frac{T_{\rm th}-T_{\rm l}}{T_{\rm th}} > \frac{\eta}{1-\eta}$, the SCD probability is

$$\Psi_{\rm SCD}\left(Q, T_{\rm th}\right) = \sum_{k=1}^{K_{\max}^{\rm a}} \mathcal{P}_{\frac{\lambda_{\rm U}}{\lambda_{\rm S}}}\left(k\right), \tag{18}$$

if $L \in \left[J\left(\frac{1}{2}\right)^{\frac{1}{1-\varsigma}}, \left(\min\left\{\Xi_{\rm a}\frac{1-\epsilon}{\epsilon}, 1\right\}\right)^{\frac{1}{1-\varsigma}}\right]$, and the minimum SBS density satisfies the condition given in **Corollary 4**.

***Proof** 5:* See Appendix F.

The above corollary provides some important insights into the interplay between time-frequency resource allocation and cache size in cache-enabled dense cellular networks with massive MIMO backhaul, which plays a key role in the content delivery performance.

B. Latency

To evaluate the latency in such networks, we consider the average delay for successfully obtaining the requested content in such networks. It should be noted that when the small cells are overloaded, UEs may suffer longer delay. There are many approaches to solve the overload issue such as deploying enough small cells following the rule of **Corollary 2** and **Corollary 6** or advanced multi-antenna techniques. Moreover, it may be more lightly loaded in realistic small cell networks [31]. For tractability, we assume that the load of a small

cell will not exceed its maximum load $K_{\max}$. As suggested in [32], the average delay for requesting a content from a typical small cell can be expressed as

$$\mathcal{D} = \sum_{k=1}^{K_{\max}} \mathcal{P}_{\frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}}}(k) \left(q_{\mathrm{hit}} \frac{Q}{\mathbb{E}\{R_{\mathrm{a}}\}} + (1-q_{\mathrm{hit}}) \left(T_1 + \frac{Q}{\mathbb{E}\{R_{\mathrm{a}'}\}} \right) \right), \tag{19}$$

where T_1 is the massive MIMO backhaul time detailed in Section III-B, and $\mathbb{E}\{R_{\mathrm{a}}\}$ and $\mathbb{E}\{R_{\mathrm{a}'}\}$ are the average access rate of the cached and self-backhauled content delivery, respectively, which are given by

$$\begin{cases} \mathbb{E}\{R_{\mathrm{a}}\} = \int_0^\infty \varphi(x, q_{\mathrm{hit}}, \eta) dx, \\ \mathbb{E}\{R_{\mathrm{a}'}\} = \int_0^\infty \varphi(x, 1-q_{\mathrm{hit}}, 1-\eta) dx, \end{cases} \tag{20}$$

where $\varphi(x, \theta_1, \theta_2) = \left(1 + \theta_1 \frac{2^{\frac{kx}{\theta_2 W}+1}-2}{\alpha_{\mathrm{a}}-2} \chi(k)\right)^{-1}$ with $\chi(k) = {}_2F_1\left[1, 1-\frac{2}{\alpha_{\mathrm{a}}}, 2-\frac{2}{\alpha_{\mathrm{a}}}, 1-2^{\frac{kx}{\theta_2 W}}\right]$ is the complementary cumulative distribution function of the R_{a} or $R_{\mathrm{a}'}$, respectively, which is obtained by using the approach in Appendix A.

Given the hit probability, i.e., the cache size is fixed, the spectrum fraction $\eta = \eta_o$ for meeting $\mathbb{E}\{R_{\mathrm{a}}\} = \mathbb{E}\{R_{\mathrm{a}'}\}$ can be easily obtained by using one-dimension search, considering the fact that $\mathbb{E}\{R_{\mathrm{a}}\} - \mathbb{E}\{R_{\mathrm{a}'}\}$ is an increasing function of η.

***Corollary** 8:* When $\eta < \eta_o$, the average delay of self-backhauled content delivery could be lower than cached content delivery if massive MIMO antennas meet

$$N \geq \left(\frac{2^{\Theta(\eta_o)} - 1}{P_{\mathrm{b}} \beta e^{\bar{\Delta}_1 - \bar{\Delta}_2}} + 1 \right) S_o - \frac{1}{2} \tag{21}$$

with $\Theta(\eta_o) = \frac{(1-\eta_o)^{-1} W^{-1} \mathbb{E}\{R_{\mathrm{a}}\} \mathbb{E}\{R_{\mathrm{a}'}\}}{\mathbb{E}\{R_{\mathrm{a}'}\} - \mathbb{E}\{R_{\mathrm{a}}\}}$, for a specified typical backhaul load S_o.

The proof of **Corollary 8** can be easily obtained by considering $T_1 \leq \frac{Q}{\mathbb{E}\{R_{\mathrm{a}}\}} - \frac{Q}{\mathbb{E}\{R_{\mathrm{a}'}\}}$ for $\eta < \eta_o$ and **Corollary 3**. It is implied from **Corollary 8** that for the case of requesting non-cached contents, the average delay of the non-cached content delivery via massive MIMO backhaul could be comparable to the cached content delivery, if the average access rate of cached content delivery is lower than that of self-backhauled content delivery.

V. Simulation Results

In this section, simulation results are presented to validate the prior analysis and further shed light on the effects of key system parameters including cell load, cache size, BS density, and massive MIMO antennas on the performance. The basic simulation parameters are shown in Table I.

TABLE I

SIMULATION PARAMETERS

Parameter	Symbol	Value
Pathloss exponent to UE	α_{a}	3.0
Pathloss exponent to SBS	α_{b}	2.6
Transmit power of MBS	P_{b}	46 dBm
Transmit power of SBS	P_{a}	30 dBm
Carrier frequency	f_c	3.5 GHz
Frequency dependent constant value	β	$\left(\frac{3\times10^8}{4\pi f_c}\right)^2$
System bandwidth	W	100 MHz
Noise power	$\sigma_{\mathrm{a}}^2,\ \sigma_{\mathrm{b}}^2,\ \sigma_{\mathrm{a}'}^2$	$-174+10\times\log_{10}(\mathrm{Bandwidth})$ dBm
Content library size	J	10^5
Zipf exponent	ς	0.7

A. *Cached Content Delivery*

In this subsection, we illustrate the cell load, SCD probability, and minimum required SBS density when the requested content is cached at the associated SBS.

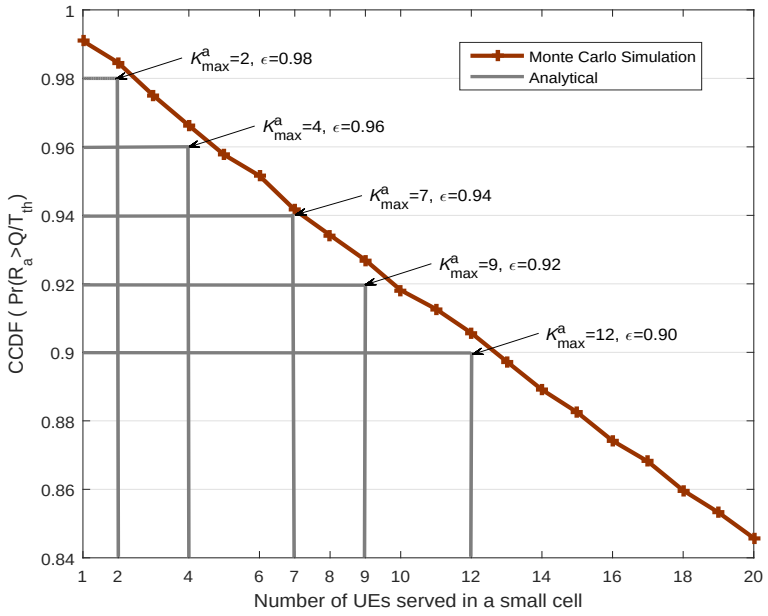

Fig. 2. The complementary cumulative distribution function (CCDF) of the R_{a}: $\frac{Q}{T_{\mathrm{th}}} = 1$ Mbps, $\lambda_{\mathrm{U}} = 3\times10^{-4}$ m^{-2}, $\lambda_{\mathrm{S}} = 10^{-4}$ m^{-2}, $\eta = 0.5$, and Cache Size$= 3\times10^3$.

Fig. 2 shows the complementary cumulative distribution function (CCDF) of the rate R_{a}

for different number of UEs served in a small cell. The analytical maximum cell load $K_{\max}^{\mathrm{a}}$ for different CCDF thresholds are obtained from (6), which has a precise match with the Monte Carlo simulations. The CCDF is a decreasing function of number of UEs served in a small cell, since resources allocated to each UE become less when serving more UEs.

Fig. 3 shows the SCD probability when the requested content is cached at the associated SBS, based on **Theorem 1** and Fig. 2. We see that for fixed cache size, the SCD probability decreases when the system requires higher SCD threshold ϵ, since higher ϵ reduces the level of maximum allowable cell load, as suggested in Fig. 2. Moreover, given ϵ, the SCD probability decreases with increasing the cache size. The reason is that hit probability increases with increasing the cache size, i.e., UEs are more likely to obtain the requested contents cached by their associated SBSs, which results in more interference at the same frequency band and reduces the maximum allowable cell load.

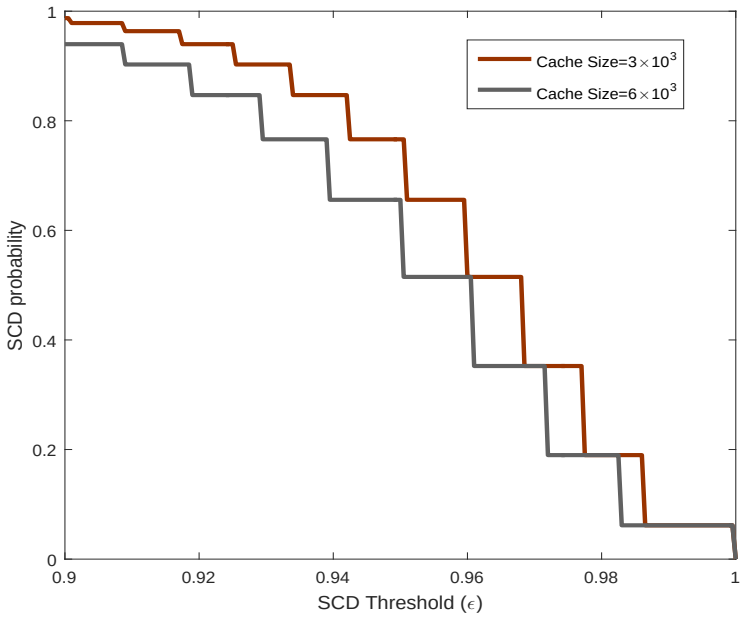

Fig. 3. The SCD probability: $\frac{Q}{T_{\mathrm{th}}} = 1$ Mbps, $\lambda_{\mathrm{U}} = 3 \times 10^{-4}$ m^{-2}, $\lambda_{\mathrm{S}} = 10^{-4}$ m^{-2}, and $\eta = 0.5$.

Fig. 4 shows the minimum required SBS density to avoid the overload issue given the UE density λ_{U}. Without loss of generality, we assume that the maximum allowable load of a small cell is $K_{\max}^{\mathrm{a}} = 5$ in this figure (Note that for specified system performance requirement, the maximum small cell load is obtained from (6), as illustrated in Fig. 2.). The numerical result has a precise match with the analysis shown in **Corollary 2**. We see that when the probability that more than $K_{\max}^{\mathrm{a}}$ UEs need to be served in a small cell is not larger than

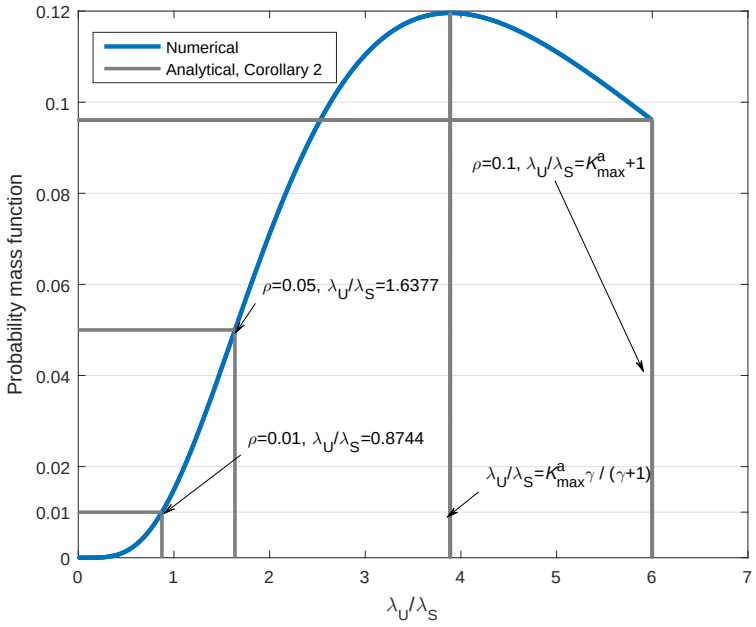

Fig. 4. The minimum required SBS density for avoiding overloading.

$\rho = 0.1$, the minimum required SBS density satisfies $\frac{\lambda_{\rm U}}{\lambda_{\rm S}} = K^{\rm a}_{\max} + 1 = 6$, as confirmed in (8). When the system requires lower $\rho = 0.1$ (i.e., lower overload probability.), the density ratio $\frac{\lambda_{\rm U}}{\lambda_{\rm S}}$ in such networks decreases, which means that more SBSs need to be deployed.

B. Massive MIMO Backhaul Transmission

In this subsection, we focus on the massive MIMO backhaul achievable rate, which determines the amount of backhaul time when an SBS obtains the requested content from its associated MBS. Note that the macrocell load and minimum required MBS density have been studied in Section III-B, which are similar to **Theorem 1** and **Corollary 2**, and numerical results can be easily obtained by following Figs. 2 and 4.

Fig. 5 shows the backhaul achievable rate for different macrocell load and massive MIMO antennas. The analytical exact and lower-bound curves are obtained from (9) and (10), respectively, which tightly matches with the simulated exact curves. We see that the backhaul achievable rate decreases when macrocell load increases, since each SBS will obtain less transmit power and array gains. Adding more massive MIMO antennas improves the achievable rate because of larger array gains.

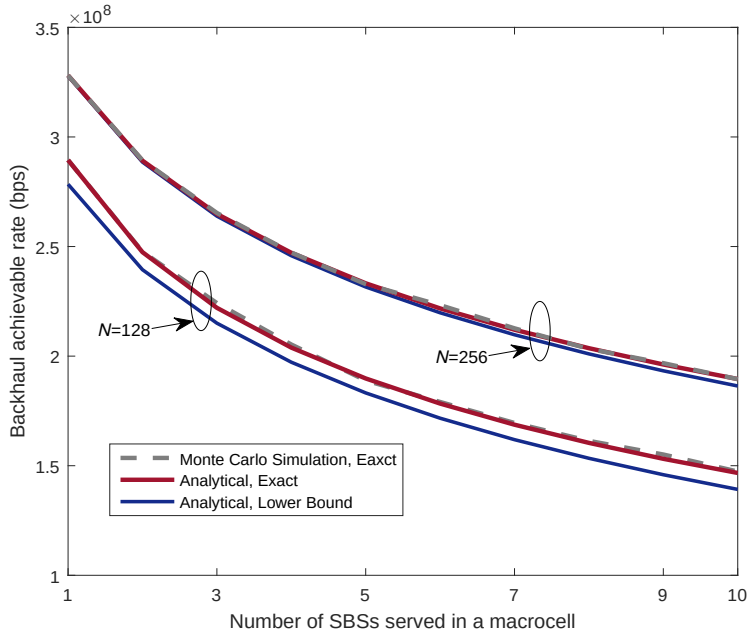

Fig. 5. Backhaul achievable rate: $\lambda_{\mathrm{M}} = 10^{-5}$ m^{-2}, $\eta = 0.5$ and $r_{\mathrm{b}} = 5$ m.

C. Latency

In this subsection, we evaluate the average delay in two scenarios, i.e., 1) the requested content is cached at the associated SBS; and 2) the requested content is not cached and needs to be obtained via massive MIMO backhaul.

Fig. 6 shows the average delay for different cache size. The analytical curves are obtained based on the average rate given by (20). We see that the average delay for cached content delivery is lower than self-backhualed content delivery. The average delay for cached content delivery increases with increasing the cache size. In contrast, the average delay for self-backhauled content delivery decreases with increasing the cache size. The reason is that larger cache size results in higher hit probability, and more SBSs can provide cached content delivery, which results in more inter-SBS interference over the frequency band allocated to the cached content delivery, and less inter-SBS interference over the frequency band allocated to the self-backhauled content delivery. In addition, the backhaul time T_1 is much lower than the access when using massive MIMO backhaul.

VI. Conclusion

In this paper, we have studied the content delivery in cache-enabled HetNets with massive MIMO backhaul. We derived the successful content delivery probability involving cached

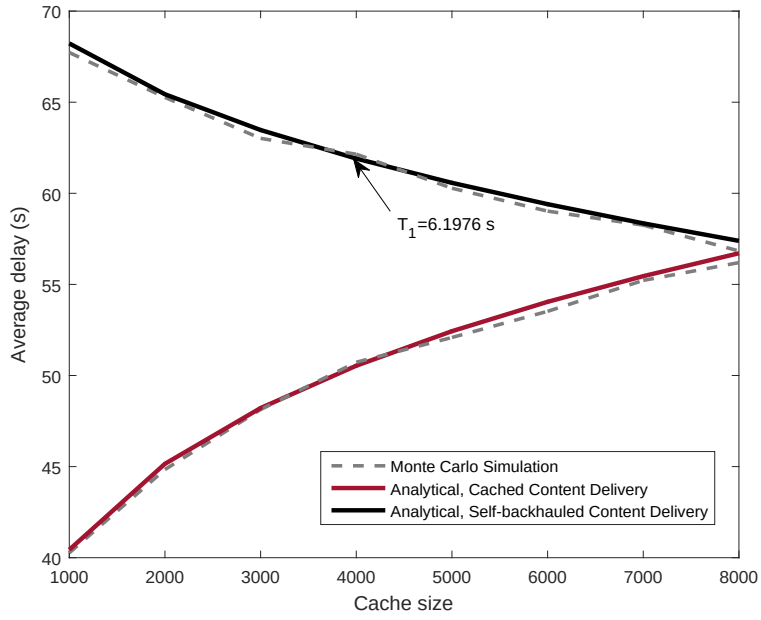

Fig. 6. Average delay: $Q = 1$ Gbit, $\lambda_{\mathrm{U}} = 3 \times 10^{-4}$ m^{-2}, $\lambda_{\mathrm{S}} = 10^{-4}$ m^{-2}, $\lambda_{\mathrm{M}} = 10^{-5}$ m^{-2}, $N = 128$, $S_o = 10$, $K = 5$, $\eta = 0.45$, and $r_{\mathrm{b}} = 5$ m.

content delivery and non-cached content delivery via massive MIMO backhaul in such networks. We addressed the effects of hit probability, UE and SBS densities on the performance. Particularly, we provided the minimum required SBS and MBS densities for avoiding overloading. We demonstrated that hit probability needs to be properly determined, in order to achieve successful content delivery. We showed the interplay between cache size and time-frequency resource allocations from the perspective of successful content delivery probability. We characterized the latency in terms of average delay in this networks, and showed that when UEs request non-cached contents, the average delay of the non-cached content delivery could be comparable to the cached content delivery with the help of massive MIMO aided self-backhaul in some cases.

APPENDIX A: PROOF OF THEOREM 1

Based on (2), SCD probability is calculated as

$$\begin{aligned}\Psi_{\mathrm{SCD}}^{\mathrm{a}}\left(Q,T_{\mathrm{th}}\right)&=\Pr\left(R_{\mathrm{a}}\geq\frac{Q}{T_{\mathrm{th}}}\right)\\&=\mathbb{E}_{K}\Bigg\{\underbrace{\Pr\left(R_{\mathrm{a}}\geq\frac{Q}{T_{\mathrm{th}}}|K=k\right)}_{\Lambda(k)}\Bigg\}\\&=\sum_{k=1}\mathcal{P}_{\frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}}}\left(k\right)\Lambda\left(k\right),\end{aligned}\tag{A.1}$$

where $\mathcal{P}_{\frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}}}\left(k\right)$ is the probability mass function (PMF) of the number of other $k-1$ UEs (except typical UE) served by the typical SBS, and Λ_{k}^{a} is the conditional SCD probability given $K=k$. According to [33], $\mathcal{P}_{\frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}}}\left(k\right)$ can be calculated as

$$\mathcal{P}_{\frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}}}\left(k\right)=\frac{\gamma^{\gamma}}{(k-1)!}\frac{\Gamma\left(k+\gamma\right)}{\Gamma\left(\gamma\right)}\frac{\left(\frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}}\right)^{k-1}}{\left(\gamma+\frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}}\right)^{k+\gamma}},\tag{A.2}$$

where $\gamma=3.5$ [29]. Given $K=k$, $\Lambda\left(k\right)$ is calculated as

$$\begin{aligned}\Lambda\left(k\right)&=\Pr\left(R_{\mathrm{a}}\geq\frac{Q}{T_{\mathrm{th}}}\right)\\&=\mathbb{E}_{|X_{o}|}\left\{\Pr\left(\frac{P_{\mathrm{a}}h_{o}L\left(|X_{o}|\right)}{I_{\mathrm{a}}+\sigma_{\mathrm{a}}^{2}}\geq2^{\frac{kQ}{\eta WT_{\mathrm{th}}}}-1\right)\right\}\\&=\int_{0}^{\infty}\underbrace{\Pr\left(\frac{P_{\mathrm{a}}h_{o}L\left(x\right)}{I_{\mathrm{a}}+\sigma_{\mathrm{a}}^{2}}\geq2^{\frac{kQ}{\eta WT_{\mathrm{th}}}}-1\right)}_{\Upsilon_{1}(x)}f_{|X_{o}|}\left(x\right)dx,\end{aligned}\tag{A.3}$$

where $f_{|X_{o}|}\left(x\right)=2\pi\lambda_{\mathrm{S}}x\exp\left(-\pi\lambda_{\mathrm{S}}x^{2}\right)$ is the probability density function (PDF) of the distance between the typical UE and its associated SBS, and $\Upsilon_{1}\left(x\right)$ is the conditional SCD probability given $K=k$ and $|X_{o}|=x$. Considering the fact that dense cellular network is interference-limited in practice, the effect of noise power on the performance is negligible. As such, we can evaluate $\Upsilon_{1}\left(x\right)$ as

$$\begin{aligned}\Upsilon_{1}\left(x\right)&=\mathbb{E}_{\Phi_{\mathrm{S}}^{\mathrm{a}}}\left\{\exp\left(-\frac{2^{\frac{kQ}{\eta WT_{\mathrm{th}}}}-1}{P_{\mathrm{a}}L\left(x\right)}I_{\mathrm{a}}\right)\right\}\\&\overset{(a)}{=}\exp\left(-2\pi\lambda_{\mathrm{S}}q_{\mathrm{hit}}\int_{x}^{\infty}\frac{\left(2^{\frac{kQ}{\eta WT_{\mathrm{th}}}}-1\right)x^{\alpha_{\mathrm{a}}}r^{1-\alpha_{\mathrm{a}}}}{1+\left(2^{\frac{kQ}{\eta WT_{\mathrm{th}}}}-1\right)x^{\alpha_{\mathrm{a}}}r^{-\alpha_{\mathrm{a}}}}dr\right)\\&=\exp\Bigg(-2\pi\lambda_{\mathrm{S}}q_{\mathrm{hit}}\frac{x^{2}}{\alpha_{\mathrm{a}}-2}\left(2^{\frac{kQ}{\eta WT_{\mathrm{th}}}}-1\right)\times\\&\qquad{}_{2}F_{1}\left[1,1-\frac{2}{\alpha_{\mathrm{a}}},2-\frac{2}{\alpha_{\mathrm{a}}},1-2^{\frac{kQ}{\eta WT_{\mathrm{th}}}}\right]\Bigg)\end{aligned}\tag{A.4}$$

where step (a) is obtained by using the generating functional of the PPP [34]. By substituting (A.4) into (A.3), $\Lambda(k)$ can be derived in closed-form as

$$\Lambda(k) = \frac{1}{1 + q_{\text{hit}} \frac{2^{\frac{kQ}{\eta W T_{\text{th}}}+1} - 2}{\alpha_{\text{a}} - 2} \chi_k^{\text{a}}(k)}, \tag{A.5}$$

where $\chi_k^{\text{a}}(k) = {}_2F_1\left[1, 1 - \frac{2}{\alpha_{\text{a}}}, 2 - \frac{2}{\alpha_{\text{a}}}, 1 - 2^{\frac{kQ}{\eta W T_{\text{th}}}}\right]$. Based on (A.5), the maximum load $K_{\max}^{\text{a}}$ of a typical small cell is given by

$$\Lambda(k)|_{k=K_{\max}^{\text{a}}} = \epsilon, \tag{A.6}$$

where ϵ is the threshold that SCD occurs when $\Lambda(k) \geq \epsilon$. Although the closed-form solution with respect to (w.r.t.) $k = K_{\max}^{\text{a}}$ of (A.6) is unfeasible, it can be quickly obtained by using one-dimension search as detailed in **Algorithm 1** due to the fact that $\Lambda(k)$ is a decreasing function of k. The SCD probability in (A.1) is rewritten as

$$\Psi_{\text{SCD}}^{\text{a}}(Q, T_{\text{th}}) = \sum_{k=1}^{K_{\max}^{\text{a}}} \mathcal{P}_{\frac{\lambda_{\text{U}}}{\lambda_{\text{S}}}}(k), \tag{A.7}$$

where $\mathcal{P}_{\frac{\lambda_{\text{U}}}{\lambda_{\text{S}}}}(k)$ and $K_{\max}^{\text{a}}$ are defined by (A.2) and (A.6), respectively, and the proof of **Theorem 1** is completed.

Appendix B: Proof of Corollary 2

After obtaining $K_{\max}^{\text{a}}$, we can find out how many small cells are sufficient to serve a specified scale of UEs λ_{U}, since serving larger than $K_{\max}^{\text{a}}$ UEs in a small cell cannot achieve SCD. Assuming that $\mathcal{P}_{\frac{\lambda_{\text{U}}}{\lambda_{\text{S}}}}(k = K_{\max}^{\text{a}} + 1) = \rho$ with arbitrary small $\rho > 0$, we need to guarantee $\mathcal{P}_{\frac{\lambda_{\text{U}}}{\lambda_{\text{S}}}}(k) \leq \rho, \forall k > K_{\max}^{\text{a}}$, in order to avoid content delivery failure resulting from overloading. Let

$$\frac{\mathcal{P}_{\frac{\lambda_{\text{U}}}{\lambda_{\text{S}}}}(k+1)}{\mathcal{P}_{\frac{\lambda_{\text{U}}}{\lambda_{\text{S}}}}(k)} = \left(1 + \frac{\gamma}{k}\right) \frac{\frac{\lambda_{\text{U}}}{\lambda_{\text{S}}}}{\gamma + \frac{\lambda_{\text{U}}}{\lambda_{\text{S}}}} \leq 1, k \geq K_{\max}^{\text{a}} + 1. \tag{B.1}$$

We can intuitively interpret (B.1) based on the fact that given the maximum load $K_{\max}^{\text{a}}$, the probability that serving more than $K_{\max}^{\text{a}}$ UEs should be lower when adding more UEs. From (B.1), we get $\frac{\lambda_{\text{U}}}{\lambda_{\text{S}}} \leq K_{\max}^{\text{a}} + 1$ such that $\mathcal{P}_{\frac{\lambda_{\text{U}}}{\lambda_{\text{S}}}}(k) \leq \rho, \forall k > K_{\max}^{\text{a}}$. Then, we need to solve $\mathcal{P}_{\frac{\lambda_{\text{U}}}{\lambda_{\text{S}}}}(k = K_{\max}^{\text{a}} + 1) = \rho$ w.r.t. $\frac{\lambda_{\text{U}}}{\lambda_{\text{S}}}$ under the constraint $\frac{\lambda_{\text{U}}}{\lambda_{\text{S}}} \leq K_{\max}^{\text{a}} + 1$. The first-order partial derivative of $\mathcal{P}_{\frac{\lambda_{\text{U}}}{\lambda_{\text{S}}}}(k)$ w.r.t. $\frac{\lambda_{\text{U}}}{\lambda_{\text{S}}}$ is

$$\begin{aligned}\frac{\partial \mathcal{P}_{\frac{\lambda_{\text{U}}}{\lambda_{\text{S}}}}(k)}{\partial \frac{\lambda_{\text{U}}}{\lambda_{\text{S}}}} = \frac{\gamma^{\gamma} \Gamma(k+\gamma)}{(k-1)! \Gamma(\gamma)} \left(\frac{\lambda_{\text{U}}}{\lambda_{\text{S}}}\right)^{k-2} \left(\gamma + \frac{\lambda_{\text{U}}}{\lambda_{\text{S}}}\right)^{-(k+\gamma+1)} \\ \times \left((k-1)\gamma - (\gamma+1)\frac{\lambda_{\text{U}}}{\lambda_{\text{S}}}\right).\end{aligned} \tag{B.2}$$

From (B.2), we see that for $k = K^{\mathrm{a}}_{\max} + 1$, $\frac{\partial \mathcal{P}_{\frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}}}}{\partial \frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}}} \geq 0$ as $\frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}} \in \left(0, \frac{K^{\mathrm{a}}_{\max}\gamma}{\gamma+1}\right]$, and $\frac{\partial \mathcal{P}_{\frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}}}}{\partial \frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}}} < 0$ as $\frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}} \in \left(\frac{K^{\mathrm{a}}_{\max}\gamma}{\gamma+1}, K^{\mathrm{a}}_{\max} + 1\right]$. Therefore, the minimum required density of SBSs satisfies

$$\frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}} = \begin{cases} \left(K^{\mathrm{a}}_{\max} + 1\right),\ if\ \mathcal{P}_{\frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}} = K^{\mathrm{a}}_{\max}+1}\left(k = K^{\mathrm{a}}_{\max} + 1\right) \leq \rho, \\ \mu_{\mathrm{a}},\ if\ \mathcal{P}_{\frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}} = K^{\mathrm{a}}_{\max}+1}\left(k = K^{\mathrm{a}}_{\max} + 1\right) > \rho, \end{cases} \tag{B.3}$$

where $\mu_{\mathrm{a}} \in \left(0, \frac{K^{\mathrm{a}}_{\max}\gamma}{\gamma+1}\right]$ is the solution of $\mathcal{P}_{\frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}} = \mu_{\mathrm{a}}}\left(k = K^{\mathrm{a}}_{\max} + 1\right) = \rho$, and can be easily obtained by using one-dimension search approach, since $\mathcal{P}_{\frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}} = \mu_{\mathrm{a}}}\left(k = K^{\mathrm{a}}_{\max} + 1\right)$ is an increasing function of μ_{a} as $\mu_{\mathrm{a}} \in \left(0, \frac{K^{\mathrm{a}}_{\max}\gamma}{\gamma+1}\right]$. Thus, we obtain the minimum required SBS density, in order to avoid overloading.

Appendix C: Detailed derivation of (9)

Since the typical SBS is associated with the nearest MBS, the PDF of the typical communication distance is

$$f_{|Y_o|}\left(y\right) = \frac{2\pi\lambda_{\mathrm{M}} y}{\exp\left(-\pi\lambda_{\mathrm{M}} r_{\mathrm{b}}^2\right)} \exp\left(-\pi\lambda_{\mathrm{M}} y^2\right),\ y \geq r_{\mathrm{b}}, \tag{C.1}$$

where r_{b} is the minimum distance between the typical MBS and its associated SBS. According to (3) and [26, 27], the achievable transmission rate can be written as

$$\begin{aligned} \overline{R}_{\mathrm{b}} &= \left(1-\eta\right) W \mathbb{E}_{|Y_o|}\left\{\log_2\left(1 + \frac{\frac{P_{\mathrm{b}}}{S_o}\Xi_1}{\frac{P_{\mathrm{b}}}{S_o}\Xi_2 + \Xi_3 + \sigma_{\mathrm{b}}^2}\right)\right\} \\ &= \left(1-\eta\right) W \int_{r_{\mathrm{b}}}^{\infty} C_{\mathrm{b}}\left(y\right) f_{|Y_o|}\left(y\right) dy, \end{aligned} \tag{C.2}$$

where $C_{\mathrm{b}}\left(y\right) = \log_2\left(1 + \frac{\frac{P_{\mathrm{b}}}{S_o}\Xi_1(y)}{\frac{P_{\mathrm{b}}}{S_o}\Xi_2(y) + \Xi_3(y) + \sigma_{\mathrm{b}}^2}\right)$ with $\Xi_1(y) = L\left(y\right)\left(\mathbb{E}\left\{\sqrt{g_o}\right\}\right)^2$, $\Xi_2\left(y\right) = L\left(y\right)\mathrm{var}\left\{\sqrt{g_o}\right\}$,[4] and $\Xi_3\left(y\right) = \mathbb{E}_{|Y_o|=y}\left\{I_{\mathrm{b}}\right\}$.

We first calculate Ξ_1 as

$$\begin{aligned} \Xi_1\left(y\right) &= L\left(y\right)\left(\int_0^{\infty} \sqrt{x} \frac{x^{N-S_o} e^{-x}}{\Gamma\left(N - S_o + 1\right)} dx\right)^2 \\ &= L\left(y\right)\left(\frac{\Gamma\left(N - S_o + \frac{3}{2}\right)}{\Gamma\left(N - S_o + 1\right)}\right)^2. \end{aligned} \tag{C.3}$$

Then, Ξ_2 is given by

$$\Xi_2\left(y\right) = L\left(y\right)\mathbb{E}\left\{g_o\right\} - \Xi_1 = \left(N - S + 1\right)L\left(y\right) - \Xi_1. \tag{C.4}$$

[4] $\mathrm{var}\left\{\cdot\right\}$ is the variance operator.

By using the Campbell's theorem [24], Ξ_3 is obtained as

$$\begin{aligned}\Xi_3(y) &= \frac{P_{\rm b}}{S_j}\mathbb{E}\left\{g_j\right\}2\pi\lambda_{\rm M}\beta\int_y^{\infty} t^{1-\alpha_{\rm b}}dt \\ &= P_{\rm b}2\pi\lambda_{\rm M}\beta\frac{y^{2-\alpha_{\rm b}}}{\alpha_{\rm b}-2}.\end{aligned} \tag{C.5}$$

By substituting (C.3), (C.4) and (C.5) into (C.2), we obtain (9).

Appendix D: Proof of Corollary 3

According to the Stirling's formula, i.e., $\Gamma(x+1) \approx \left(\frac{x}{e}\right)^x\sqrt{2\pi x}$ as $x\to\infty$, we have

$$\begin{aligned}\Xi_1(y) &\approx L(y)\left(\frac{\left(\frac{N-S_o+\frac{1}{2}}{e}\right)^{N-S_o+\frac{1}{2}}\sqrt{2\pi\left(N-S_o+\frac{1}{2}\right)}}{\left(\frac{N-S}{e}\right)^{N-S_o}\sqrt{2\pi\left(N-S_o\right)}}\right)^2 \\ &\approx L(y)\frac{N-S_o+\frac{1}{2}}{e}\left(1+\frac{1}{2\left(N-S_o\right)}\right)^{2(N-S_o)} \\ &\overset{(a)}{\approx}\left(N-S_o+\frac{1}{2}\right)L(y),\end{aligned} \tag{D.1}$$

when the number of antennas at the MBS grows large. Note that step (a) is obtained by the fact that $\left(1+\frac{1}{x}\right)^x \approx e$ as $x\to\infty$. Thus, $\Xi_2(y) = \frac{L(y)}{2}$. By using Jensen's inequality [35], we derive a tight lower-bound on the achievable transmission rate (C.2) as

$$\overline{R}_{\rm b}^{\rm Low} = (1-\eta)W\log_2\left(1+\frac{P_{\rm b}}{S_o}e^{\Delta_1-\Delta_2}\right), \tag{D.2}$$

where

$$\begin{cases}\Delta_1 = \mathbb{E}_{|Y_o|}\left\{\ln\Xi_1\right\}, \\ \Delta_2 = \mathbb{E}_{|Y_o|}\left\{\ln\left(\frac{P_{\rm b}}{S_o}\Xi_2+\Xi_3+\sigma_{\rm b}^2\right)\right\}.\end{cases} \tag{D.3}$$

For large N, based on (D.1), Δ_1 can be asymptotically derived as

$$\begin{aligned}\Delta_1 &\approx \ln\left(N-S_o+\frac{1}{2}\right)+\mathbb{E}\left\{\ln L(y)\right\} \\ &= \ln\left(N-S_o+\frac{1}{2}\right)+\ln(\beta) \\ &\quad \underbrace{-\frac{\alpha_{\rm b}}{\exp\left(-\pi\lambda_{\rm M}r_{\rm b}^2\right)}\Big(-\frac{Ei\big(-r_{\rm b}^2\pi\lambda_{\rm M}\big)}{2}+e^{-r_{\rm b}^2\pi\lambda_{\rm M}}\ln r_{\rm b}\Big)}_{\overline{\Delta}_1},\end{aligned} \tag{D.4}$$

where $Ei\left(z\right)$ is the exponential integral given by $Ei\left(z\right)=-\int_{-z}^{\infty}\frac{e^{-t}}{t}dt$. Then, Δ_2 can be asymptotically calculated as

$$\begin{aligned}\Delta_2&=\int_{r_{\rm b}}^{\infty}\ln\left(\frac{P_{\rm b}}{S_o}\Xi_2\left(y\right)+\Xi_3\left(y\right)+\sigma_{\rm b}^2\right)f_{|Y_o|}\left(y\right)dy\\&\approx\int_{r_{\rm b}}^{\infty}\ln\left(\frac{P_{\rm b}\beta}{2S_o}y^{-r_{\rm b}}+P_{\rm b}2\pi\lambda_{\rm M}\beta\frac{y^{2-\alpha_{\rm b}}}{\alpha_{\rm b}-2}+\sigma_{\rm b}^2\right)\\&\quad\underbrace{\times\frac{2\pi\lambda_{\rm M}y}{\exp\left(-\pi\lambda_{\rm M}r_{\rm b}^2\right)}\exp\left(-\pi\lambda_{\rm M}y^2\right)dy}_{\overline{\Delta}_2}.\end{aligned}\tag{D.5}$$

Considering the fact that $T_1=\frac{Q}{\overline{R}_{\rm b}}\leq\frac{Q}{\overline{R}_{\rm b}^{\rm Low}}$, we obtain $T_1\leq\frac{Q}{(1-\eta)W}\left(\log_2\left(1+\frac{P_{\rm b}\beta\left(N-S_o+\frac{1}{2}\right)}{S_o}e^{\overline{\Delta}_1-\overline{\Delta}_2}\right)\right)^{-1}$, which confirms the **Corollary 3**.

Appendix E: Proof of Theorem 2

Based on (4), SCD probability is given by

$$\begin{aligned}\Psi_{\rm SCD}^{\rm b}\left(Q,T_{\rm th}\right)&=\Pr\left(R_{\rm a'}>\frac{Q}{T_{\rm th}-T_1}\right)\\&=\sum_{k\geq1}\mathcal{P}_{\frac{\lambda_{\rm U}}{\lambda_{\rm S}}}\left(k\right)\Lambda_k^{\rm b},\end{aligned}\tag{E.1}$$

where $\mathcal{P}_{\frac{\lambda_{\rm U}}{\lambda_{\rm S}}}\left(k\right)$ is given by (A.2), and $\Lambda_k^{\rm b}$ is the conditional SCD probability given $K=k$. Similar to (A.3), $\Lambda_k^{\rm b}$ is calculated as

$$\begin{aligned}\Lambda_k^{\rm b}&=\Pr\left(\frac{P_{\rm a'}h_oL\left(|X_o|\right)}{I_{\rm a'}+\sigma_{\rm a'}^2}>2^{\frac{kQ}{(1-\eta)W\left(T_{\rm th}-T_1\right)}}-1\right)\\&=\frac{1}{1+\left(1-q_{\rm hit}\right)\frac{2^{\frac{kQ}{(1-\eta)W\left(T_{\rm th}-T_1\right)}+1}-2}{\alpha_{\rm a}-2}\chi_k^{\rm b}},\end{aligned}\tag{E.2}$$

where $\chi_k^{\rm b}={}_2F_1\left[1,1-\frac{2}{\alpha_{\rm a}},2-\frac{2}{\alpha_{\rm a}},1-2^{\frac{kQ}{(1-\eta)W\left(T_{\rm th}-T_1\right)}}\right]$. Like (A.6), the maximum load $K_{\max}^{\rm b}$ of a typical small cell is the solution of $\Lambda\left(k\right)|_{k=K_{\max}^{\rm b}}=\epsilon$. Then, the SCD probability is obtained as (13).

Appendix F: Proof of Corollary 7

Based on (6) and (14), we see that $K_{\max}^{\rm a}\geq K_{\max}^{\rm b}$ if $\frac{T_{\rm th}-T_1}{T_{\rm th}}\leq\frac{\eta}{1-\eta}$ and $q_{\rm hit}\leq\frac{1}{2}$. In this case, the SCD probability in (16) reduces to (17), and the corresponding cache size is obtained by considering **Corollary 5** and $q_{\rm hit}\leq\frac{1}{2}$. Likewise, $K_{\max}^{\rm a}>K_{\max}^{\rm b}$ if $\frac{T_{\rm th}-T_1}{T_{\rm th}}>\frac{\eta}{1-\eta}$ and $q_{\rm hit}>\frac{1}{2}$, and we can obtain (18) accordingly.

REFERENCES

[1] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2016–2021 white paper," 2017.

[2] G. Paschos, E. Bastug, I. Land, G. Caire, and M. Debbah, "Wireless caching: technical misconceptions and business barriers," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 16–22, Aug. 2016.

[3] W. Han, A. Liu, and V. K. N. Lau, "PHY-caching in 5G wireless networks: design and analysis," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 30–36, Aug. 2016.

[4] L. Wang, K.-K. Wong, S. Jin, G. Zheng, and R. W. Heath Jr., "A new look at physical layer security, caching, and wireless energy harvesting for heterogeneous ultra-dense networks," *arXiv preprint arXiv:1705.09647*, May 2017.

[5] 3GPP TS 22.261:"Service requirements for the 5G system," Mar. 2017.

[6] D. Liu, L. Wang, Y. Chen, M. Elkashlan, K. K. Wong, R. Schober, and L. Hanzo, "User association in 5G networks: A survey and an outlook," *IEEE Commun. Surveys & Tutorials*, vol. 18, no. 2, pp. 1018–1044, Second Quarter 2016.

[7] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," in *IEEE Int. Conf. Commun. (ICC)*, 2015, pp. 3358–3363.

[8] B. Zhou, Y. Cui, and M. Tao, "Stochastic content-centric multicast scheduling for cache-enabled heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6284–6297, Sept. 2016.

[9] X. Li, X. Wang, K. Li, Z. Han, and V. C. Leung, "Collaborative multi-tier caching in heterogeneous networks: Modeling, analysis, and design," *IEEE Trans. Wireless Commun.*, Early Access Articles, 2017.

[10] M. Ji, G. Caire, and A. F. Molisch, "Wireless device-to-device caching networks: Basic principles and system performance," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 1, pp. 176–189, Jan. 2016.

[11] Z. Chen, N. Pappas, and M. Kountouris, "Probabilistic caching in wireless D2D networks: Cache hit optimal versus throughput optimal," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 584–587, Mar. 2017.

[12] G. Zheng, H. A. Suraweera, and I. Krikidis, "Optimization of hybrid cache placement for collaborative relaying," *IEEE Commun. Lett.*, vol. 21, no. 2, pp. 442–445, Feb. 2017.

[13] S. H. Park, O. Simeone, and S. S. Shitz, "Joint optimization of cloud and edge processing for fog radio access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7621–7632, Nov. 2016.

[14] H. Zhang, Y. Qiu, X. Chu, K. Long, and V. C. Leung, "Fog radio access networks: Mobility management, interference mitigation, and resource optimization," arXiv preprint arXiv:1707.06892, July 2017.

[15] 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.

[16] Y. Chen, M. Ding, J. Li, Z. Lin, G. Mao, and L. Hanzo, "Probabilistic small-cell caching: Performance analysis and optimization," *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 4341–4354, May 2017.

[17] K. Li, C. Yang, Z. Chen, and M. Tao, "Optimization and analysis of probabilistic caching in n-tier heterogeneous networks," arXiv preprint arXiv:1612.04030, Dec. 2016.

[18] J. Wen, K. Huang, S. Yang, and V. O. K. Li, "Cache-enabled heterogeneous cellular networks: Optimal tier-level content placement," *IEEE Trans. Wireless Commun.*, Early Access Articles 2017.

[19] W. Wen, Y. Cui, F. chun Zheng, S. Jin, and Y. Jiang, "Random caching based cooperative transmission in heterogeneous wirelesss networks," arXiv preprint arXiv:1701.05761, Jan. 2017.

[20] 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, Sept. 2016.

[21] X. Peng, J. Zhang, S. H. Song, and K. B. Letaief, "Cache size allocation in backhaul limited wireless networks," in *IEEE Int. Conf. Commun. (ICC)*, 2016, pp. 1–6.

[22] A. Liu and V. K. N. Lau, "How much cache is needed to achieve linear capacity scaling in backhaul-limited dense wireless networks?" *IEEE/ACM Trans. Netw.*, vol. 25, no. 1, pp. 179–188, Feb. 2017.

[23] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and zipf-like distributions: evidence and implications," in *Proc. IEEE INFOCOM*, 1999, pp. 126-134.

[24] F. Baccelli and B. Blaszczyszyn, *Stochastic Geometry and Wireless Networks, Volume I: Theory*. Now Publishers Inc. Hanover, MA, USA, 2009.

[25] K. Hosseini, W. Yu, and R. S. Adve, "Large-scale MIMO versus network MIMO for multicell interference mitigation," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 930–941, Oct. 2014.

[26] J. Jose, A. Ashikhmin, T. L. Marzetta, and S. Vishwanath, "Pilot contamination and precoding in multi-cell TDD systems," *IEEE Trans. Wireless Commun.*, vol. 10, no. 8, pp. 2640–2651, Aug. 2011.

[27] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 160–171, Feb. 2013.

[28] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products*, 7th ed. San Diego, C.A.: Academic Press, 2007.

[29] J.-S. Ferenc and Z. Néda, "On the size distribution of poisson voronoi cells," *Physica A: Statistical Mechanics and its Applications*, vol. 385, no. 2, pp. 518–526, Nov. 2007.

[30] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82–89, Aug. 2014.

[31] J. G. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. Soong, and J. Zhang, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, June 2014.

[32] J. Liu, B. Bai, J. Zhang, and K. B. Letaief, "Cache placement in Fog-RANs: From centralized to distributed algorithms," *IEEE Trans. Wireless Commun.*, Early Access Articles 2017.

[33] 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.

[34] M. Haenggi, J. G. Andrews, F. Baccelli, O. Dousse, and M. Franceschetti, "Stochastic geometry and random graphs for the analysis and design of wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 7, pp. 1029–1046, Sept. 2009.

[35] L. Wang, H. Q. Ngo, M. Elkashlan, T. Q. Duong, and K. K. Wong, "Massive MIMO in spectrum sharing networks: Achievable rate and power efficiency," *IEEE Systems Journal*, vol. 11, no. 1, pp. 20–31, Mar. 2017.