Loughborough University

This item was submitted to Loughborough's Research Repository by the author.
Items in Figshare are protected by copyright, with all rights reserved, unless otherwise indicated.

# Edge caching in dense heterogeneous cellular networks with massive MIMO aided self-backhaul

PLEASE CITE THE PUBLISHED VERSION

https://doi.org/10.1109/TWC.2018.2859936

PUBLISHER

Institute of Electrical and Electronics Engineers (IEEE)

VERSION

VoR (Version of Record)

PUBLISHER STATEMENT

This work is made available according to the conditions of the Creative Commons Attribution 3.0 Unported (CC BY 3.0) licence. Full details of this licence are available at: http://creativecommons.org/licenses/by/3.0/

LICENCE

CC BY 3.0

REPOSITORY RECORD

Wang, Lifeng, Kai-Kit Wong, Sangarapillai Lambotharan, Arumugam Nallanathan, and Maged Elkashlan. 2019. “Edge Caching in Dense Heterogeneous Cellular Networks with Massive MIMO Aided Self-backhaul”. figshare. https://hdl.handle.net/2134/34589.

# 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 focuses on 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. Our aim is to address the effects of cell load and hit probability on the successful content delivery (SCD) and present the minimum required base station density for avoiding the access overload in an arbitrary small cell and backhaul overload in an arbitrary macrocell. The achievable rate of massive MIMO backhaul without any downlink channel estimation is derived to calculate the backhaul time, and the latency is also evaluated in such networks. The analytical results confirm that hit probability needs to be appropriately selected in order to achieve SCD. The interplay between cache size and SCD is explicitly quantified. It is theoretically demonstrated 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.**

## 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 requires 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] and [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 limitations, 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] where it was assumed that the intensity of BSs is much larger than the intensity of mobile terminals, two cache-enabled BS modes were considered, namely always-on and dynamic on-off. The work of [17]–[19] concentrated on the cache-enabled

Manuscript received September 5, 2017; revised January 24, 2018 and April 20, 2018; accepted July 10, 2018. Date of publication August 2, 2018; date of current version September 10, 2018. This work was supported by the U.K. Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/M016005/1, Grant EP/M016145/2, and Grant EP/M015475/1. The associate editor coordinating the review of this paper and approving it for publication was S. Pollin. *(Corresponding author: Lifeng Wang.)*

L. Wang and K.-K. Wong are with the Department of Electronic and Electrical Engineering, University College London, London WC1E 6BT, U.K. (e-mail: lifeng.wang@ucl.ac.uk; kai-kit.wong@ucl.ac.uk).

S. Lambotharan is with the School of Mechanical, Electrical and Manufacturing Engineering, Loughborough University, Loughborough LE11 3TU, U.K. (e-mail: 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 E1 4NS, U.K. (e-mail: a.nallanathan@qmul.ac.uk; maged.elkashlan@qmul.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2018.2859936

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/

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 the requested contents of users are not cached at their associated BSs, they will be obtained from the core network 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 a downlink content-centric sparse multicast beamforming was proposed for the cache-enabled cloud radio access network (Cloud-RAN), 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. 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. Massive MIMO aided self-backhaul is motivated by the facts that it may not be feasible to have optical fiber for every backhaul channel and massive MIMO can support high-speed transmissions thanks to large array gains and multiplexing 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 density of SBSs for avoiding access overload is obtained. We show that hit probability needs to be lower than a critical 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 necessary, to evaluate the backhaul transmission delay. The minimum required density of MBSs for avoiding backhaul overload is obtained. We show that hit probability needs to be higher than a critical 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. We 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.

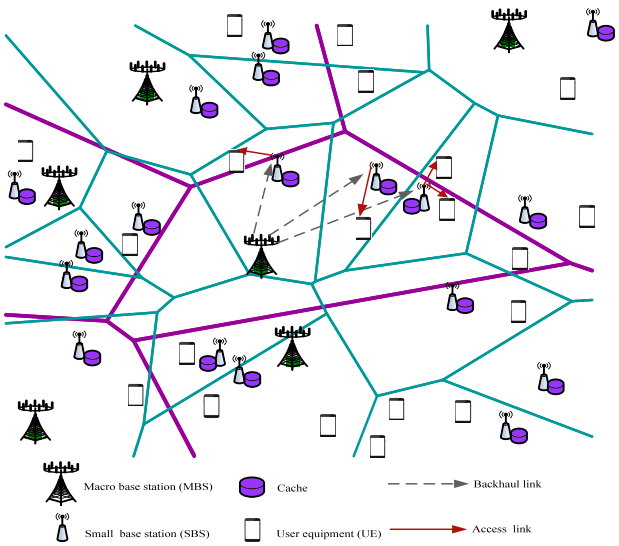



Fig. 1. An illustration of cache-enabled heterogeneous cellular network with massive MIMO backhaul.

## 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). Each massive MIMO aided MBS equipped with $N$ antennas has access to the core network 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 $\Phi_{\mathrm{U}}$ with the density $\lambda_{\mathrm{U}}$, $\Phi_{\mathrm{S}}$ with the density $\lambda_{\mathrm{S}}$, and $\Phi_{\mathrm{M}}$ with the density $\lambda_{\mathrm{M}}$, 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 [23]

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

where $\varsigma$ 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 the sum of probabilities for all the contents being cached at an arbitrary SBS should be less than the SBS's cache size (namely $\sum\limits_{j=1}^{J} q_j \leq L$) [7]. Note that based on the probabilistic caching strategy, each SBS only stores $L$ files from the content library for each caching realization, which is also illustrated in [7, Fig. 1].

### B. Self-Backhaul Load

We assume that the access and backhaul links share the same sub-6 GHz spectrum. The bandwidths allocated to the access and backhaul links are $\eta W$ and $(1-\eta) W$, respectively, where $\eta$ is the fraction factor and $W$ is the system bandwidth. The number of UEs that is associated with an SBS is denoted by $K$. The UEs in the same small cell are served in a time-division manner 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. Note that different SBSs may be served by different MBSs. 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.

Hit probability characterizes the probability that a requested content file is stored at an arbitrary SBS [4], and is calculated as $q_{\mathrm{hit}} = \sum\limits_{j=1}^{J} a_j q_j$. 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 $\lambda_{\mathrm{S}}\,(1-q_{\mathrm{hit}})$ denotes the point process of SBSs with backhaul links. Let $\omega_{\mathrm{b}} = \lambda_{\mathrm{S}}\,(1-q_{\mathrm{hit}})\,/\lambda_{\mathrm{M}}$ 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

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

where $I_{\mathrm{a}}$ denotes the total interference power from other SBSs; $P_{\mathrm{a}}$ 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 $\beta$, distance $|X|$ and path loss exponent $\alpha_{\mathrm{a}}$; $h_o \sim \exp(1)$ 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 $\sigma_a^2$ 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 time-division duplex (TDD) massive MIMO self-backhauled network, SBSs will not perform any channel estimation,[1] and we will adopt an achievable backhaul transmission rate as confirmed in [29] and [30]. Based on the instantaneous received signal expression in [29, eq. 6], given a typical distance $|Y_o|$ between the typical SBS and its associated MBS, the instantaneous 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 $\mathrm{SINR}_{\mathrm{b}}$, as shown at the bottom of the next page, where $\mathbb{E}\{\cdot\}$ is the expectation operator. Here, $I_{\mathrm{b}}$ denotes the total interference power from other MBSs; $P_{\mathrm{b}}$ 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 $\alpha_{\mathrm{b}}$; $g_o \sim \Gamma\left(N - S_o + 1, 1\right)$ is the small-scale fading channel power gain between the typical SBS and its associated MBS; $g_j \sim \Gamma\left(S_j, 1\right)$,[2] 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 $\sigma_{\mathrm{b}}^2$ 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,

[1] In TDD massive MIMO systems, downlink precoder is designed based on the uplink channel estimation, thanks to channel reciprocity [26]. Moreover, when deploying massive number of antennas at MBS, wireless channel behaves in a deterministic manner called "channel hardening," and the effect of small-scale fading could be negligible [27]. Thus, in the considered massive MIMO backhaul scenario where MBSs and SBSs are usually still, the coherence time of backhaul channel will be much longer than ever before, and the time occupied by uplink channel estimation will be much lower.

It should be noted that when high-mobility UEs are served by TDD massive MIMO BSs, downlink pilots may still be needed to estimate the fast-changing channels [28].

[2] $\Gamma(\cdot,\cdot)$ is the upper incomplete gamma function [31, eq. (8.350)].

the corresponding access-link rate is expressed as

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

where $I_{\mathrm{a}'}$ is the total interference power, $h_{i'} \sim \exp(1)$ and $L\left(|X_{o,i'}|\right) = \beta\left(|X_{o,i'}|\right)^{-\alpha_{\mathrm{a}}}$ 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 reduce latency, massive MIMO backhaul link needs to be of high-speed, which can be achieved by using large antenna arrays at the MBS. 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 $T_{\mathrm{th}}$; 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 $T_{\mathrm{th}}$.

### A. Cached Content Delivery

Different from [15], [16], and [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, 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

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

where $\mathcal{P}_{\frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}}}(k)$ 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 $\mathcal{P}_{\frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}}}(k) = \frac{\gamma^\gamma}{(k-1)!}\frac{\Gamma(k+\gamma)}{\Gamma(\gamma)}\frac{\left(\frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}}\right)^{k-1}}{\left(\gamma+\frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}}\right)^{k+\gamma}}$ with $\gamma = 3.5$ [32]. In (5), $K = K_{\max}^{\mathrm{a}}$ is the maximum load in a typical small cell, and can be efficiently obtained by using **Algorithm 1** to solve the following equation

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

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 [31, eq. (9.142)][3], and $\epsilon$ is the predefined threshold, i.e., SCD occurs when the probability that $R_{\mathrm{a}}$ is larger than $\frac{Q}{T_{\mathrm{th}}}$ is above $\epsilon$.

**Algorithm 1** One-Dimension Search

1: **if** $t=0$
2: Initialize $\varphi = \frac{1-\epsilon}{q_{\mathrm{hit}}\epsilon}$, $k^l = 1, k^h = 10\times\frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}}$, and calculate $F^l = \frac{2^{\frac{k^lQ}{\eta W T_{\mathrm{th}}}+1}-2}{\alpha_{\mathrm{a}}-2}{}_2F_1\left[1,1-\frac{2}{\alpha_{\mathrm{a}}};2-\frac{2}{\alpha_{\mathrm{a}}};1-2^{\frac{k^lQ}{\eta W T_{\mathrm{th}}}}\right]$ and $F^h = \frac{2^{\frac{k^hQ}{\eta W T_{\mathrm{th}}}+1}-2}{\alpha_{\mathrm{a}}-2}{}_2F_1\left[1,1-\frac{2}{\alpha_{\mathrm{a}}};2-\frac{2}{\alpha_{\mathrm{a}}};1-2^{\frac{k^hQ}{\eta W T_{\mathrm{th}}}}\right]$
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.,** $K_{\max}^{\mathrm{a}} = round\left(k^*\right)$**.**
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**

*Proof:* See Appendix A. ∎

It is implied from **Theorem 1** that in the dense small cell networks (i.e., interference-limited),[4] 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

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

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

[3]In MATLAB R2015b software, hypergeom([a,b],c,z) is the Gauss hypergeometric function ${}_2F_1\left[a,b;c;z\right]$.

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

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

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 hinders the cached content delivery.

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

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

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 $\mathcal{P}_{\frac{\lambda_{\rm U}}{\lambda_{\rm S}}}(k)\le\rho,\forall k>K^{\rm a}_{\max}$.

*Proof:* 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 network via massive MIMO backhaul. Therefore, we need to evaluate the backhaul time for delivering the requested content to the typical SBS. It should be noted that the load in a macrocell will not change fast, in order to deliver the requested contents to the associated SBSs. Hence, given the load $S_o$ in a typical macrocell, the achievable transmission rate for a typical backhaul link is given by

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

where $C_{\rm b}\left(y\right)=\log_2\left(1+\frac{\frac{P_{\rm b}}{S_o}\Xi_1(y)}{\frac{P_{\rm b}}{S_o}\Xi_2(y)+\Xi_3(y)+\sigma_{\rm b}^2}\right)$ 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 $\Xi_3\left(y\right)=P_{\rm b}2\pi\lambda_{\rm M}\beta\frac{y^{2-\alpha_{\rm b}}}{\alpha_{\rm b}-2}$, and $r_{\rm b}$ 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 $T_1=\frac{Q}{\overline{R}_{\rm b}}$. 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

$$\overline{R}^{\rm Low}_{\rm b}\left(S_o\right)=\left(1-\eta\right)W\log_2\left(1+P_{\rm b}\beta\frac{N-S_o+\frac{1}{2}}{S_o}e^{\overline{\Delta}_1-\overline{\Delta}_2}\right),\tag{10}$$

where

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

in which $Ei\left(z\right)$ is the exponential integral given by $Ei\left(z\right)=-\int_{-z}^{\infty}\frac{e^{-t}}{t}dt$ [31]. Based on (10), the typical MBS's required time for delivering $Q$ bits to its associated SBS satisfies

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

*Proof:* See Appendix D. ∎

It is explicitly shown from **Corollary 3** that large number of antennas and bandwidths are required, in order to significantly reduce the wireless backhaul delivery time. From (11), we see that the backhaul delivery time can at least be cut proportionally to $1/\log_2 N$.

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., $S_o\le S_{\max}$; 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 efficiently 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

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

where $\mathcal{P}_{\omega_{\rm b}}\left(\ell\right)=\frac{\gamma^{\gamma}}{(\ell-1)!}\frac{\Gamma(\ell+\gamma)}{\Gamma(\gamma)}\frac{(\omega_{\rm b})^{\ell-1}}{(\gamma+\omega_{\rm b})^{\ell+\gamma}}$, $\mu_{\rm b}\in\left(0,\frac{S_{\max}\gamma}{\gamma+1}\right]$ is the solution of $\mathcal{P}_{\omega_{\rm b}=\mu_{\rm b}}\left(S_{\max}+1\right)=\rho$ 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

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

where $K^{\mathrm{b}}_{\max}$ 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 $K^{\mathrm{b}}_{\max}$ can be obtained by solving the following equation[5]

$$\frac{2^{\frac{K^{\mathrm{b}}_{\max}Q}{(1-\eta)W(T_{\mathrm{th}}-T_1)}+1}-2}{\alpha_{\mathrm{a}}-2}\chi^{\mathrm{b}}_k\left(K^{\mathrm{b}}_{\max}\right)=\frac{1-\epsilon}{(1-q_{\mathrm{hit}})\,\epsilon} \quad (14)$$

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

*Proof:* 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

*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

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

where $\Xi_{\mathrm{b}}=\left(\frac{2^{\frac{Q}{(1-\eta)W(T_{\mathrm{th}}-T_1)}+1}-2}{\alpha_{\mathrm{a}}-2}\chi^{\mathrm{b}}_k(1)\right)^{-1}$, 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 $K^{\mathrm{a}}_{\max}\rightarrow K^{\mathrm{b}}_{\max}$, 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

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

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.

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

## 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 reduce backhaul cost and latency. Therefore, higher hit probability is meaningful to reduce 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] and [33]. Therefore, we consider MPC caching and analyze the appropriate cache size in such networks. Considering the fact that for large $J$ with MPC caching, $q_{\mathrm{hit}}=\sum_{j=1}^{L}a_j\approx\left(\frac{L}{J}\right)^{1-\varsigma}$, we have

*Corollary 7:* Given $\frac{T_{\mathrm{th}}-T_1}{T_{\mathrm{th}}}\leq\frac{\eta}{1-\eta}$ (i.e., more time-frequency resources are allocated to the cached content delivery), the SCD probability is

$$\Psi_{\mathrm{SCD}}(Q,T_{\mathrm{th}})\approx\sum_{k=1}^{K^{\mathrm{b}}_{\max}}\mathcal{P}_{\frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}}}(k), \quad (17)$$

and it is an increasing function of the cache size, if the cache size $L\in\left[J\left(\left[1-\Xi_{\mathrm{b}}\frac{1-\epsilon}{\epsilon}\right]^{+}\right)^{\frac{1}{1-\varsigma}},J\left(\frac{1}{2}\right)^{\frac{1}{1-\varsigma}}\right]$ and the minimum SBS density satisfies the condition given in **Corollary 6**; Given $\frac{T_{\mathrm{th}}-T_1}{T_{\mathrm{th}}}>\frac{\eta}{1-\eta}$, the SCD probability is

$$\Psi_{\mathrm{SCD}}(Q,T_{\mathrm{th}})\approx\sum_{k=1}^{K^{\mathrm{a}}_{\max}}\mathcal{P}_{\frac{\lambda_{\mathrm{U}}}{\lambda_{\mathrm{S}}}}(k), \quad (18)$$

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

*Proof:* 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

TABLE I
SIMULATION PARAMETERS

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

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 [34]. For tractability, we assume that the load of a small cell will not exceed its maximum load $K_{\max}$. As suggested in [35], the average delay for requesting a content from a typical small cell can be expressed as

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

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

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

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 $R_{\mathrm{a}}$ 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 $\eta$.

*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

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

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 $S_o$.

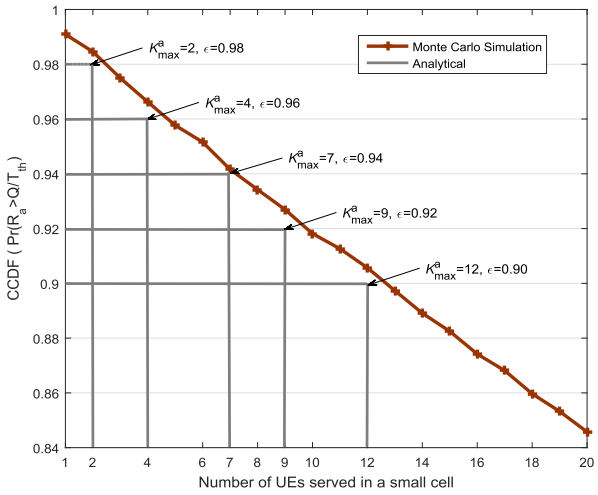



Fig. 2. The complementary cumulative distribution function (CCDF) of the $R_{\mathrm{a}}$: $\frac{Q}{T_{\mathrm{th}}} = 1$ Mbps, $\lambda_{\mathrm{U}} = 3\times10^{-4}$ m$^{-2}$, $\lambda_{\mathrm{S}} = 10^{-4}$ m$^{-2}$, $\eta = 0.5$, and Cache Size $= 3\times10^3$.

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 that of 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.

### 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 shows the complementary cumulative distribution function (CCDF) of the rate $R_{\mathrm{a}}$ for different number of UEs

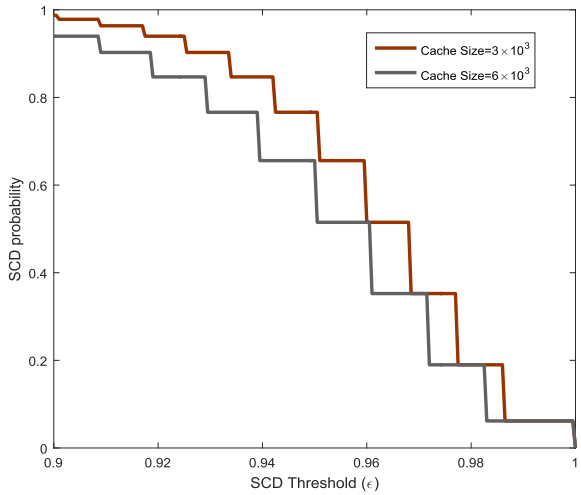



Fig. 3. The SCD probability: $\frac{Q}{T_{\rm th}} = 1$ Mbps, $\lambda_{\rm U} = 3 \times 10^{-4}$ m$^{-2}$, $\lambda_{\rm S} = 10^{-4}$ m$^{-2}$, and $\eta = 0.5$.

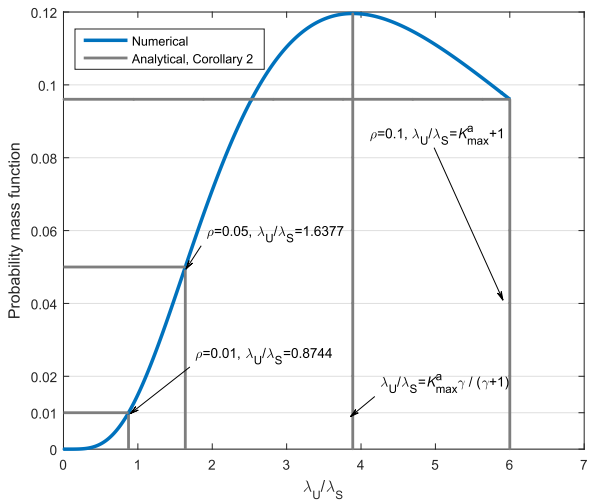



Fig. 4. The minimum required SBS density for avoiding overloading.

served in a small cell. The analytical maximum cell load $K^{\rm a}_{\max}$ 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. The stair-like curves are induced by the fact that the SCD probability given by (5) is a discrete function of maximum cell load. We see that for fixed cache size, the SCD probability decreases when the system requires higher SCD threshold $\epsilon$, since higher $\epsilon$ reduces the level of maximum allowable cell load, as suggested in Fig. 2. Moreover, for a given $\epsilon$, 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. 4 shows the minimum required SBS density to avoid the overload issue given the UE density $\lambda_{\rm U}$. Without loss of generality, we assume that the maximum allowable load of a small cell is $K^{\rm a}_{\max} = 5$ 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 precisely matches with the analysis shown in **Corollary 2**. We see that when the probability that more than $K^{\rm a}_{\max}$ UEs need to be served in a small cell is not larger than $\rho = 0.1$, 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

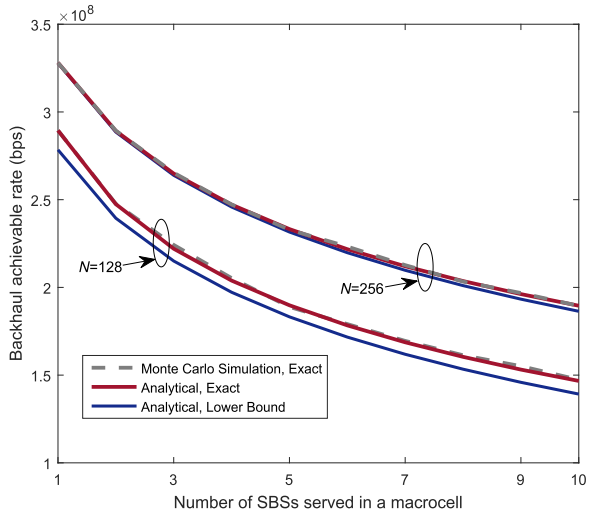



Fig. 5. Backhaul achievable rate: $\lambda_{\rm M} = 10^{-5}$ m$^{-2}$, $\eta = 0.5$ and $r_{\rm b} = 5$ m.

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.

### *C. Latency*

In this subsection, we evaluate the average delay in two scenarios: 1) The requested content is cached at the

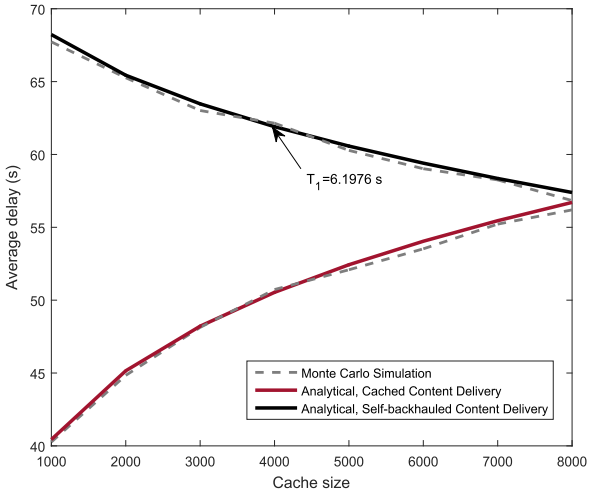



Fig. 6. Average delay: $Q = 1$ Gbit, $\lambda_{\rm U} = 3 \times 10^{-4}$ m$^{-2}$, $\lambda_{\rm S} = 10^{-4}$ m$^{-2}$, $\lambda_{\rm M} = 10^{-5}$ m$^{-2}$, $N = 128$, $S_o = 10$, $K = 5$, $\eta = 0.45$, and $r_{\rm b} = 5$ m.

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 that of the self-backhauled 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. This 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 content delivery time of massive MIMO backhaul $T_1$ is much lower than that of the access.

## VI. Conclusion

We have studied content delivery in cache-enabled HetNets with massive MIMO backhaul. In such networks, the successful content delivery probability involving cached content delivery and non-cached content delivery via massive MIMO backhaul was analyzed. The effects of hit probability, UE and SBS densities on the performance were addressed. Particularly, we provided the minimum required SBS and MBS densities for avoiding overloading. The derived results demonstrated that hit probability needs to be properly determined, in order to achieve successful content delivery. The interplay between cache size and time-frequency resource allocations was quantified from the perspective of successful content delivery probability. The latency was characterized in terms of average delay in this networks. It was proved that when UEs request non-cached contents, the average delay of the non-cached content delivery could be comparable to that of 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

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

where $\mathcal{P}_{\frac{\lambda_{\rm U}}{\lambda_{\rm S}}}(k)$ is the probability mass function (PMF) of the number of other $k-1$ UEs (except typical UE) served by the typical SBS, and $\Lambda(k)$ is the conditional SCD probability given $K = k$. According to [36], $\mathcal{P}_{\frac{\lambda_{\rm U}}{\lambda_{\rm S}}}(k)$ can be calculated as

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

where $\gamma = 3.5$ [32]. Given $K = k$, $\Lambda(k)$ is calculated as

$$\begin{aligned}\Lambda(k) &= \Pr\left(R_{\rm a} \geq \frac{Q}{T_{\rm th}}\right) \\ &= \mathbb{E}_{|X_o|}\left\{\Pr\left(\frac{P_{\rm a} h_o L(|X_o|)}{I_{\rm a} + \sigma^2_{\rm a}} \geq 2^{\frac{kQ}{\eta W T_{\rm th}}} - 1\right)\right\} \\ &= \int_0^\infty \underbrace{\Pr\left(\frac{P_{\rm a} h_o L(x)}{I_{\rm a} + \sigma^2_{\rm a}} \geq 2^{\frac{kQ}{\eta W T_{\rm th}}} - 1\right)}_{\Upsilon_1(x)} f_{|X_o|}(x)\, dx, \end{aligned} \tag{A.3}$$

where $f_{|X_o|}(x) = 2\pi\lambda_{\rm S} x \exp\left(-\pi\lambda_{\rm S} x^2\right)$ is the probability density function (PDF) of the distance between the typical UE and its associated SBS, and $\Upsilon_1(x)$ 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 $\Upsilon_1(x)$ as

$$\begin{aligned}&\Upsilon_1(x) \\ &= \mathbb{E}_{\Phi^{\rm a}_{\rm S}}\left\{\exp\left(-\frac{2^{\frac{kQ}{\eta W T_{\rm th}}} - 1}{P_{\rm a} L(x)} I_{\rm a}\right)\right\} \\ &\overset{(a)}{=} \exp\left(-2\pi\lambda_{\rm S} q_{\rm hit} \int_x^\infty \frac{\left(2^{\frac{kQ}{\eta W T_{\rm th}}} - 1\right) x^{\alpha_{\rm a}} r^{1-\alpha_{\rm a}}}{1 + \left(2^{\frac{kQ}{\eta W T_{\rm th}}} - 1\right) x^{\alpha_{\rm a}} r^{-\alpha_{\rm a}}} dr\right) \\ &= \exp\Bigg(-2\pi\lambda_{\rm S} q_{\rm hit} \frac{x^2}{\alpha_{\rm a} - 2}\left(2^{\frac{kQ}{\eta W T_{\rm th}}} - 1\right) \\ &\qquad \times {}_2F_1\left[1, 1 - \frac{2}{\alpha_{\rm a}}; 2 - \frac{2}{\alpha_{\rm a}}; 1 - 2^{\frac{kQ}{\eta W T_{\rm th}}}\right]\Bigg), \end{aligned} \tag{A.4}$$

where step (a) is obtained by using the generating functional of the PPP [37]. By substituting (A.4) into (A.3), $\Lambda(k)$ can

be derived in closed-form as

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

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

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

where $\epsilon$ is the threshold that SCD occurs when $\Lambda(k)\geq\epsilon$. Although the closed-form solution with respect to (w.r.t.) $k=K_{\max}^{\rm a}$ of (A.6) is unfeasible, it can be efficiently 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

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

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

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

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

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

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

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

where $\mu_{\rm a}\in\left(0,\frac{K_{\max}^{\rm a}\gamma}{\gamma+1}\right]$ is the solution of $\mathcal{P}_{\frac{\lambda_{\rm U}}{\lambda_{\rm S}}=\mu_{\rm a}}(K_{\max}^{\rm a}+1)=\rho$, and can be easily obtained by using one-dimension search approach, since $\mathcal{P}_{\frac{\lambda_{\rm U}}{\lambda_{\rm S}}=\mu_{\rm a}}(K_{\max}^{\rm a}+1)$ is an increasing function of $\mu_{\rm a}$ as $\mu_{\rm a}\in\left(0,\frac{K_{\max}^{\rm a}\gamma}{\gamma+1}\right]$. 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

$$f_{|Y_o|}(y)=\frac{2\pi\lambda_{\rm M}y}{e^{-\pi\lambda_{\rm M}r_{\rm b}^2}}e^{-\pi\lambda_{\rm M}y^2},\quad y\geq r_{\rm b}, \tag{C.1}$$

where $r_{\rm b}$ is the minimum distance between the typical MBS and its associated SBS. According to (3) and [29] and [30], the achievable transmission rate can be written as

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

where $C_{\rm b}(y)=\log_2\left(1+\frac{\frac{P_{\rm b}}{S_o}\Xi_1(y)}{\frac{P_{\rm b}}{S_o}\Xi_2(y)+\Xi_3(y)+\sigma_{\rm b}^2}\right)$ with $\Xi_1(y)=L(y)\left(\mathbb{E}\left\{\sqrt{g_o}\right\}\right)^2$, $\Xi_2(y)=L(y)\,\mathrm{var}\left\{\sqrt{g_o}\right\}$,[6] and $\Xi_3(y)=\mathbb{E}_{|Y_o|=y}\{I_{\rm b}\}$.

We first calculate $\Xi_1$ as

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

Then, $\Xi_2$ is given by

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

By using the Campbell's theorem [24], $\Xi_3$ is obtained as

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

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

[6] $\mathrm{var}\{\cdot\}$ is the variance operator.

## 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 [38], 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

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

For large $N$, based on (D.1), $\Delta_1$ can be asymptotically derived as

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

where $Ei(z)$ is the exponential integral given by $Ei(z) = -\int_{-z}^{\infty}\frac{e^{-t}}{t}dt$. Then, $\Delta_2$ can be asymptotically calculated as

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

Substituting (D.4) and (D.5) into (D.2), we obtain (10).

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

## APPENDIX E
## PROOF OF THEOREM 2

Based on (4), SCD probability is given by

$$\begin{aligned}\Psi_{\rm SCD}^{\rm b}(Q,T_{\rm th}) &= \Pr\left(R_{\rm a'} > \frac{Q}{T_{\rm th}-T_1}\right)\\ &= \sum_{k\ge1}\mathcal{P}_{\frac{\lambda_{\rm U}}{\lambda_{\rm S}}}(k)\,\Lambda_k^{\rm b},\end{aligned} \tag{E.1}$$

where $\mathcal{P}_{\frac{\lambda_{\rm U}}{\lambda_{\rm S}}}(k)$ 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

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

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

## APPENDIX F
## PROOF OF COROLLARY 7

Based on (6) and (14), we see that $K_{\max}^{\rm a} \ge K_{\max}^{\rm b}$ if $\frac{T_{\rm th}-T_1}{T_{\rm th}} \le \frac{\eta}{1-\eta}$ and $q_{\rm hit} \le \frac{1}{2}$. In this case, UE's requested contents are more likely to be delivered via massive MIMO self-backhaul. As such, based on (16), the SCD probability can be approximated as $\Psi_{\rm SCD}(Q,T_{\rm th}) \approx \sum_{k=1}^{K_{\max}^{\rm b}}\mathcal{P}_{\frac{\lambda_{\rm U}}{\lambda_{\rm S}}}(k)$ given in (17). Considering the fact that $q_{\rm hit} = \left(\frac{L}{J}\right)^{1-\varsigma} \ge \left[1-\Xi_{\rm b}\frac{1-\epsilon}{\epsilon}\right]^+$ in **Corollary 5** and $q_{\rm hit} = \left(\frac{L}{J}\right)^{1-\varsigma} \le \frac{1}{2}$, the corresponding cache size is obtained as $L \in \left[J\left(\left[1-\Xi_{\rm b}\frac{1-\epsilon}{\epsilon}\right]^+\right)^{\frac{1}{1-\varsigma}}, J\left(\frac{1}{2}\right)^{\frac{1}{1-\varsigma}}\right]$. Moreover, increasing the cache size boosts the hit probability, and thus enhances the maximum cell load $K_{\max}^{\rm b}$ for self-backhauled content delivery. The reason is that more SBSs can deliver the cached contents, which reduces inter-cell interference in self-backhauled content delivery.

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) and the corresponding cache size $L \in \left[J\left(\frac{1}{2}\right)^{\frac{1}{1-\varsigma}}, \left(\min\left\{\Xi_{\rm a}\frac{1-\epsilon}{\epsilon},1\right\}\right)^{\frac{1}{1-\varsigma}}\right]$.

## REFERENCES

[1] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2016–2021," Cisco, San Jose, CA, USA, 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," *IEEE Commun. Mag.*, vol. 56, no. 6, pp. 49–55, Jun. 2018.

[5] *Service Requirements for the 5G System*, document 3GPP TS 22.261, Jan. 2018.
[6] D. Liu *et al.*, "User association in 5G networks: A survey and an outlook," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1018–1044, 2nd Quart., 2016.
[7] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 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, Sep. 2016.
[9] X. Li, X. Wang, K. Li, Z. Han, and V. C. M. Leung, "Collaborative multi-tier caching in heterogeneous networks: Modeling, analysis, and design," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6926–6939, Oct. 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. 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. Shamai (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. M. Leung, "Fog radio access networks: Mobility management, interference mitigation, and resource optimization," *IEEE Wireless Commun.*, vol. 24, no. 6, pp. 120–127, Dec. 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," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 1283–1297, Feb. 2018.
[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.*, vol. 16, no. 9, pp. 5939–5952, Sep. 2017.
[19] W. Wen, Y. Cui, F.-C. Zheng, S. Jin, and Y. Jiang, "Random caching based cooperative transmission in heterogeneous wireless networks," *IEEE Trans. Commun.*, vol. 66, no. 7, pp. 2809–2825, Jul. 2018.
[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, Sep. 2016.
[21] X. Peng, J. Zhang, S. H. Song, and K. B. Letaief, "Cache size allocation in backhaul limited wireless networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 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*, Mar. 1999, pp. 126–134.
[24] F. Baccelli and B. Błaszczyszyn, "Stochastic geometry and wireless networks: Volume I theory," in *Foundations and Trends in Networking*. Hanover, MA, USA: NOW, 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] E. Björnson, E. G. Larsson, and T. L. Marzetta, "Massive MIMO: Ten myths and one critical question," *IEEE Commun. Mag.*, vol. 54, no. 2, pp. 114–123, Feb. 2016.
[27] H. Q. Ngo and E. G. Larsson, "No downlink pilots are needed in TDD massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 2921–2935, May 2017.
[28] S. Malkowsky *et al.*, "The world's first real-time testbed for massive MIMO: Design, implementation, and validation," *IEEE Access*, vol. 5, pp. 9073–9088, 2017.
[29] 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.
[30] 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.
[31] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 7th ed. San Diego, CA, USA: Academic, 2007.
[32] J.-S. Ferenc and Z. Néda, "On the size distribution of Poisson Voronoi cells," *Phys. A, Statist. Mech. Appl.*, vol. 385, no. 2, pp. 518–526, Nov. 2007.
[33] 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.
[34] J. G. Andrews *et al.*, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.
[35] J. Liu, B. Bai, J. Zhang, and K. B. Letaief, "Cache placement in fog-RANs: From centralized to distributed algorithms," *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, pp. 7039–7051, Nov. 2017.
[36] 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.
[37] 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, Sep. 2009.
[38] 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 Syst. J.*, vol. 11, no. 1, pp. 20–31, Mar. 2017.



**Lifeng Wang** (M'16) is currently a Research Associate with the Department of Electronic and Electrical Engineering, University College London. His research interests include massive MIMO, millimeter-wave communications, edge computing/caching, V2X networks, HetNets, physical layer security, and wireless energy harvesting. He received the Exemplary Editor Certificate of the IEEE COMMUNICATIONS LETTERS in 2016 and 2017, respectively. He serves as a Review Editor for the IEEE COMMUNICATIONS LETTERS.



**Kai-Kit Wong** (M'01–SM'08–F'16) received the B.Eng., M.Phil., and Ph.D. degrees in electrical and electronic engineering from The Hong Kong University of Science and Technology, Hong Kong, in 1996, 1998, and 2001, respectively. After graduation, he took up academic and research positions at The University of Hong Kong, Lucent Technologies, Bell-Labs, Holmdel, the Smart Antennas Research Group, Stanford University, and the University of Hull, U.K. He is currently the Chair in Wireless Communications at the Department of Electronic and Electrical Engineering, University College London, U.K.

His current research interests include 5G and beyond mobile communications, including topics such as massive MIMO, full-duplex communications, millimeter-wave communications, edge caching and fog networking, physical layer security, wireless power transfer and mobile computing, V2X communications, and course cognitive radios. There are also a few other unconventional research topics that he has set his heart on, including, for example, fluid antenna communications systems and remote ECG detection. He was a co-recipient of the 2013 IEEE Signal Processing Letters Best Paper Award and the 2000 IEEE VTS Japan Chapter Award at the IEEE Vehicular Technology Conference in Japan in 2000, and a few other international best paper awards.

He is a fellow of IET and is also on the editorial board of several international journals. He has been serving as a Senior Editor for the IEEE COMMUNICATIONS LETTERS since 2012 and the IEEE WIRELESS COMMUNICATIONS LETTERS since 2016. He had also previously served as an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS from 2009 to 2012 and an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2005 to 2011. He was also a Guest Editor for the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS Special Issue on Virtual MIMO in 2013 and is currently a Guest Editor for IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS Special Issue on Physical Layer Security for 5G.



**Sangarapillai Lambotharan** (SM'06) received the Ph.D. degree in signal processing from Imperial College London, U.K., in 1997. He was a Post-Doctoral Research Associate with Imperial College London until 1999. He was a Visiting Scientist with the Engineering and Theory Centre, Cornell University, NY, USA, in 1996. From 1999 to 2002, he was with the Motorola Applied Research Group, U.K., where he investigated various projects, including physical link layer modeling and performance characterization of GPRS, EGPRS, and UTRAN. He was with King's College London and Cardiff University as a Lecturer and a Senior Lecturer, respectively, from 2002 to 2007. He is currently a Professor of digital communications and the Head of the Signal Processing and Networks Research Group, Wolfson School of Mechanical, Electrical and Manufacturing Engineering, Loughborough University, U.K. His current research interests include 5G networks, MIMO, radars, smart grids, machine learning, and network security. He has authored over 200 journal and conference articles in these areas.



**Arumugam Nallanathan** (S'97–M'00–SM'05–F'17) was with the Department of Informatics, King's College London, from 2007 to 2017, where he was a Professor of wireless communications from 2013 to 2017 and has been a Visiting Professor since 2017. He was an Assistant Professor with the Department of Electrical and Computer Engineering, National University of Singapore, from 2000 to 2007. He has been a Professor of wireless communications and the Head of the Communication Systems Research Group, School of Electronic Engineering and Computer Science, Queen Mary University of London, since 2017. He has published nearly 400 technical papers in scientific journals and international conferences. His research interests include 5G wireless networks, Internet of Things, and molecular communications. He was a co-recipient of the Best Paper Awards presented at the IEEE International Conference on Communications 2016, the IEEE Global Communications Conference 2017, and the IEEE Vehicular Technology Conference 2018. He is an IEEE Distinguished Lecturer. He was selected as a Web of Science Highly Cited Researcher in 2016.



**Maged Elkashlan** (M'06) received the Ph.D. degree in electrical engineering from The University of British Columbia, Canada, in 2006. From 2007 to 2011, he was with the Wireless and Networking Technologies Laboratory, Commonwealth Scientific and Industrial Research Organization, Australia. During this time, he held an adjunct appointment at the University of Technology Sydney, Australia. In 2011, he joined the EECS, Queen Mary University of London, U.K. where he is currently an Associate Professor with the School of Electronic Engineering and Computer Science. He also holds visiting faculty appointments at the University of New South Wales, Australia, and the Beijing University of Posts and Telecommunications, China. His research interests include communication theory, wireless communications, and statistical signal processing for distributed data processing, heterogeneous networks, and massive MIMO.

Dr. Elkashlan received the Best Paper Award at the IEEE International Conference on Communications in 2014 and 2016, the International Conference on Communications and Networking in China (CHINACOM) in 2014, and the IEEE Vehicular Technology Conference (VTC-Spring) in 2013. He currently serves as an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and the IEEE COMMUNICATIONS LETTERS. He also serves as a Lead Guest Editor for the Special Issue on Green Media: The Future of Wireless Multimedia Networks of the *IEEE Wireless Communications Magazine* and the Special Issue on Millimeter Wave Communications for 5G of the *IEEE Communications Magazine*, and a Guest Editor for the Special Issue on Energy Harvesting Communications of the *IEEE Communications Magazine* and the Special Issue on Location Awareness for Radios and Networks of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS.