

Joint Frequency Reuse and Cache Optimization in Backhaul-Limited Small-Cell Wireless Networks

Wei Han, An Liu, *SMIEEE*, Wei Yu, *FIEEE*, and Vincent K. N. Lau, *FIEEE*

Abstract—Caching at base stations (BSs) is a promising approach for supporting the tremendous traffic growth of content delivery over future small-cell wireless networks with limited backhaul. This paper considers exploiting spatial caching diversity (i.e., caching different subsets of popular content files at neighboring BSs) that can greatly improve the cache hit probability, thereby leading to a better overall system performance. A key issue in exploiting spatial caching diversity is that the cached content may not be located at the nearest BS, which means that to access such content, a user needs to overcome strong interference from the nearby BSs; this significantly limits the gain of spatial caching diversity. In this paper, we consider a joint design of frequency reuse and caching, such that the benefit of an improved cache hit probability induced by spatial caching diversity and the benefit of interference coordination induced by frequency reuse can be achieved simultaneously. We obtain a closed-form characterization of the approximate successful transmission probability for the proposed scheme and analyze the impact of key operating parameters on the performance. We design a low-complexity algorithm to optimize the frequency reuse factor and the cache storage allocation. Simulations show that the proposed scheme achieves a higher successful transmission probability than existing caching schemes.

Index Terms—Frequency reuse, Cache, Poisson point process

I. INTRODUCTION

It is predicted that there will be a 1000X increase in capacity demand for mobile data traffic in future 5G wireless networks. To meet the rapid data traffic growth, small-cell wireless networks have been proposed as an effective approach. By increasing the density of small-cell base stations (BSs) deployed per unit area, the spectral efficiency of a network can be improved. However, due to the large number of BSs per unit area in small-cell wireless networks, allocating a high-speed backhaul to each BS will lead to both high CAPEX and OPEX [1]. In practice, the backhaul capacity of small-cell BSs is limited, and this significantly limits the potential spectral efficiency gain provided by small-cell networks.

Recent works show that caches can be used in small-cell wireless networks to alleviate the high-speed backhaul

capacity requirement by moving the content closer to users [2]–[11]. For example, in [6]–[11], the benefits of caching are characterized by considering the stochastic natures of channel fading and the geographic locations of BSs and users. The caching performance can be analyzed and optimized using the theory of stochastic geometry. Specifically, in [6], the authors consider a cache placement scheme in which all BSs store the same set of the most popular content files, and then analyze the outage probability and average rate. The uncached files are served using the data obtained from the backhaul, so the service rates are limited by the backhaul capacity. Likewise, the authors of [7] analyze the average ergodic rate and the outage probability in a three-tier heterogeneous network with backhaul capacity constraints and with the caching of the most popular files. Caching the same subset of the most popular files at every BS is, however, not optimal in general. In [9], the authors consider random caching at BSs and analyze the cache hit probability under general popularity distribution (but without considering backhaul constraints), and show that it is not always optimal to cache the most popular content files in every BS. The reason behind this is that placing different contents in different BSs provides *spatial caching diversity*, which brings better overall performance.

Spatial caching diversity is typically achieved by random caching strategies in the existing literature. For example, in [10], the authors study caching in a wireless network with uniform content popularity distribution but without a backhaul constraint, and analyze cache hit probability and content outage probability for random caching with a uniform distribution. In [11], the authors consider a heterogeneous wireless network which caches the same subset of the most popular files at the macro-BSs but uses random caching at the pico-BSs; they analyze and optimize the successful transmission probability in the high signal-to-noise ratio (SNR) and user density regions with a backhaul capacity constraint, where the uncached files are served by macro BSs using the data obtained from the backhaul. Note that in [9]–[11], spatial caching diversity is achieved by randomly caching different files at different BSs, as illustrated in Fig. 1. In such cases, a user may be served by a BS which has its requested content file but is not the geographically nearest BS. This may result in strong interference coming from the geographically nearest BS for the target user. Hence, the benefit of spatial caching diversity may be overwhelmed by excessive inter-cell interference.

Spatial caching diversity can also be achieved with coded caching by encoding each content file into coded bits and caching different portions of the coded bits in different BSs [12]–[14]. Specifically, in [12] and [13], a maximum distance separable (MDS)-coded caching scheme is considered. How-

This work was supported by Science and Technology Program of Shenzhen, China (Grant No. JCYJ20170818113908577), and RGC 16204814. The work of An Liu was supported by the China Recruitment Program of Global Young Experts. Wei Yu is supported by a Hong Kong Telecom Institute of Information Technology Visiting Fellowship, and in part by an E.W.R. Steacie Memorial Fellowship. (Corresponding author: An Liu.)

Wei Han is with the HKUST Shenzhen Research Institute (email: whan@connect.ust.hk).

An Liu is with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: anliu@zju.edu.cn).

Wei Yu is with the Electrical and Computer Engineering Department, University of Toronto (email: weiyu@ece.utoronto.ca).

Vincent K. N. Lau is with the Hong Kong University of Science and Technology (email: eeknau@ust.hk).

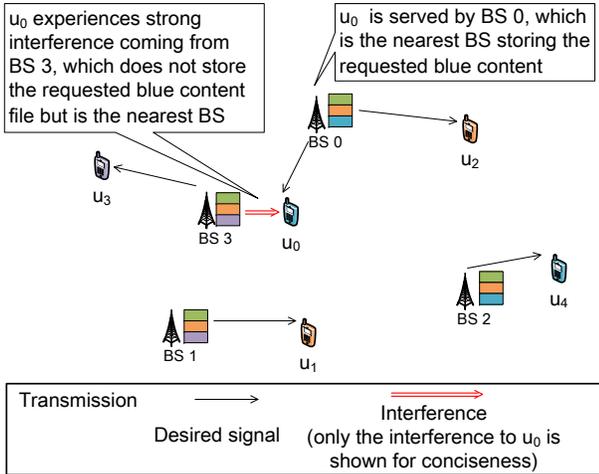


Fig. 1. Illustration of the strong interference in single-band random caching schemes. u_0 is served by BS 0, which has its requested blue content file, but is not the geographically nearest BS of u_0 . In this case, u_0 experiences strong interference coming from BS 3, which is the geographically nearest BS of u_0 .

ever, the physical layer is modeled by error-free links between users and their associated BSs, and the effect of interference is ignored. In [14], multiple BSs cache different coded packets of each file, and each user employs successive interference cancellation (SIC) to cancel the strong interference from the nearest BSs before decoding the desired signal. Although SIC removes strong interference from the nearest BSs, the complexity of the receiver at the user side increases. Moreover, the analysis in [14] is obtained under the simplified assumption that each BS transmits signals all the time, and only one typical user is considered. Also, the resource allocation for multi-user transmission at each BS and the effect of system loading are not studied in [14].

This paper proposes to address the interference induced by spatial caching diversity by joint design of frequency reuse and caching. Frequency reuse is a well-studied inter-cell interference coordination technique in conventional link-based wireless networks [15]. In a system with frequency reuse, two adjacent cells may use different frequencies to reduce the strong interference experienced by the cell edge users. In this way, both coverage and capacity are improved [16].

In this paper, we propose to explore the joint design of frequency reuse and caching, such that the benefit of spatial caching diversity and frequency reuse can be achieved at the same time. We consider a content delivery application with a fixed data rate requirement for each user. In such an application, the performance is characterized by the successful transmission probability. The main contributions of this paper are summarized as follows:

- **A joint design of frequency reuse and caching scheme:** In this paper, we propose a joint frequency reuse and caching scheme such that the subset of BSs allocated the same frequency also caches the same subset of content files. By such joint design, the strong interference caused by spatial caching diversity can be removed and a higher successful transmission probability can be achieved.

- **Closed-form characterization of the approximate successful transmission probability:** To analyze the impact of key operating parameters (such as number of subbands and cache storage capacity allocation) on the system performance, we derive a closed-form characterization of the approximate successful transmission probability under the joint frequency reuse and caching scheme.
- **Optimization of the frequency reuse factor and cache placement:** The problem of optimizing the number of subbands and the cache storage capacity allocation is a complex integer optimization problem. We propose a low-complexity algorithm and show that the proposed scheme achieves a large gain over existing caching schemes in terms of the successful transmission probability.

The rest of the paper is organized as follows. The model of caching in backhaul-limited small-cell wireless network under study is presented in Section II. A joint frequency reuse and caching scheme is proposed in Section III. In Section IV, we define the average successful transmission probability as the performance metric, and analyze the performance of the proposed scheme for a given cache storage allocation and frequency reuse factor. In Section V, we formulate and solve the joint cache storage allocation and frequency reuse optimization problem. Numerical evaluation of the proposed scheme is presented in Section VI. We conclude the paper in Section VII.

II. SYSTEM MODEL

We consider a backhaul-limited small-cell wireless network. The locations of the BSs are spatially distributed as a homogeneous Poisson point process (PPP) Φ^b with density λ^b . The locations of the users are also spatially distributed as a homogeneous PPP Φ^u with density λ^u . There is a content library $\mathcal{X} = \{X_1, X_2, \dots, X_L\}$ that contains L files, where the size of each content file is F bits. Each content file is independently requested with probability ρ_l , satisfying $\sum_{l=1}^L \rho_l = 1$. Without loss of generality, we assume $\rho_1 \geq \rho_2 \geq \dots \geq \rho_L$.

We consider the downlink transmission, where the content file requested by each user is transmitted at a fixed rate of τ bits per second. Each BS has one transmit antenna with transmission power P , a cache of storage capacity $B_C F$ bits, and a backhaul with limited capacity of $B_B \tau$ bits per second (bps). Each user has one receive antenna. The total bandwidth is W Hz. We consider a discrete-time system, with time being slotted with duration ν , and study one time slot of the network. We consider both large-scale fading (path loss) and small-scale fading. Specifically, the channel coefficient between a BS and a user with distance D is modeled by $D^{-\alpha} h$, where $\alpha > 2$ is the path loss exponent, $h \stackrel{d}{\sim} \mathcal{CN}(0, 1)$ is the small scale fading factor (i.e., we assume Rayleigh fading channels).

III. JOINT FREQUENCY REUSE AND CACHING SCHEME

In this section, we propose a *joint frequency reuse and caching scheme* which exploits the benefit of interference coordination induced by frequency reuse and the benefit of backhaul offloading induced by spatial caching diversity.

A. Joint Frequency Reuse and Cache Placement

The BSs are randomly divided into M non-overlapping BS groups indexed by $\{0, \dots, M-1\}$. Specifically, each BS independently and randomly generates a number from $\{0, \dots, M-1\}$, say m , and then joins the m -th BS group. Denote Φ_m^b with $m \in \{0, \dots, M-1\}$ as the BSs in the m -th BS group. For analysis purposes, we assume random BS grouping and independent thinning [17, p. 230] so that Φ_m^b follows a homogeneous PPP with density $\frac{\lambda_b}{M}$. The total bandwidth W is also divided into M equal-size subbands denoted as W_0, W_1, \dots, W_{M-1} , where the bandwidth of each subband is $\frac{W}{M}$. The BSs in Φ_m^b are designed to transmit in subband W_m for $m = 0, \dots, M-1$.

The proposed joint frequency reuse and cache placement design helps mitigate inter-cell interference in the network. Note that in [9], the authors propose a probabilistic cache placement policy, which sets the probability of storing each content at a given BS to an optimized target value. Such a design does not consider the strong interference induced by spatial caching diversity, because in the scheme proposed in [9], a user may be served by a BS that has its requested content file but is not the geographically nearest BS, which may result in the user experiencing strong interference from the geographically nearest BS, as illustrated in Fig. 1. In this paper, we propose a joint frequency reuse and cache placement strategy that can mitigate inter-cell interference in the network. In our scheme, the BSs in one BS group store the same subset of content files. As a user is served by the nearest BS that stores the requested content file, the serving BS must also be the geographically nearest BS in its transmitting subband (BS group), which leads to a higher receiving signal-to-interference-plus-noise ratio (SINR) at the user, as illustrated in Fig. 2.

Note that a naive combination of the cache placement scheme in [9] and frequency reuse cannot completely address the strong interference issue. If the frequency reuse and cache placement are designed separately, the nearest BS storing the requested content file would not necessarily be the geographically nearest BS in its transmitting subband. In this case, the user may experience strong interference coming from the geographically nearest BS in the transmitting subband, as illustrated in Fig. 3. Numerical results in Section VI also show that our proposed joint scheme outperforms naive combination of the cache placement scheme in [9] and frequency reuse.

A main contribution of this paper is to design the optimal caching policy of the content files. Each content file may be stored in multiple BS groups. Denote $q_l \in \{0, 1, \dots, M\}$ for $l \in \{1, \dots, L\}$ as the *cache storage allocation factor*, which indicates that the l -th content file is stored in a total of q_l BS groups. We assume that $q_l \geq q_{l+1}$, for $l = 1, \dots, L-1$, i.e., a content file with higher popularity is stored in more BSs. Denote $\mathbf{q} = [q_1, \dots, q_L]$ as the *cache storage allocation vector*, which must satisfy the following cache storage capacity constraint:

$$\sum_{l=1}^L q_l \leq MB_C. \quad (1)$$

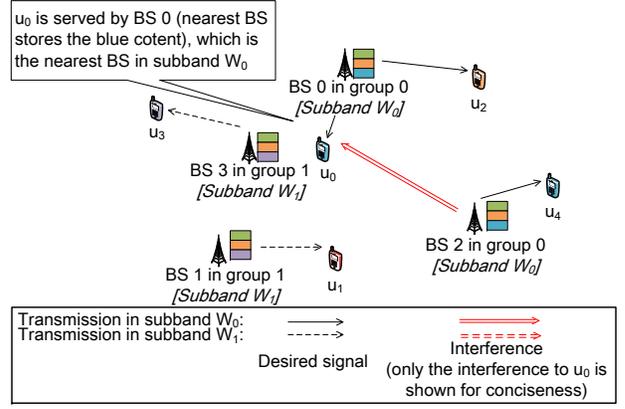


Fig. 2. Illustration of the joint frequency reuse and cache placement scheme. The user u_0 is served by BS 0, which has u_0 's requested blue content file, and BS 0 is the geographically nearest BS in the transmitting subband W_0 . The geographically nearest BS (BS 3) is transmitting in subband W_1 , and does not cause interference to u_0 .

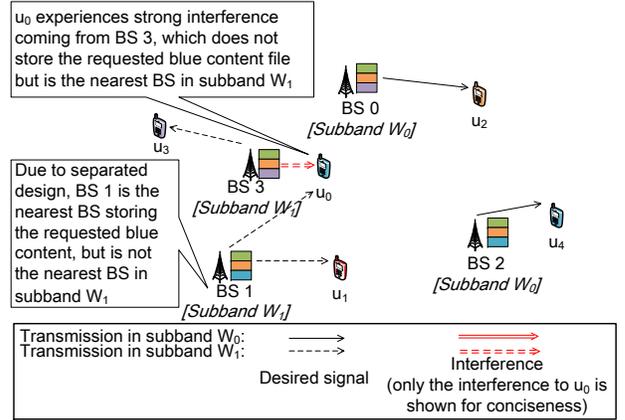


Fig. 3. Illustration of the strong interference caused by separated design of frequency reuse and cache placement. The frequency reuse factor is $\frac{1}{2}$, and the cache placement follows random caching. Due to separated design, u_0 is served by BS 1, which has its requested blue content file, but is not the geographically nearest BS in the transmitting subband W_1 . In this case, u_0 experiences strong interference coming from BS 3, which is the geographically nearest BS in the transmitting subband W_1 .

The proposed cache placement is illustrated in Fig. 4 for a network with $M = 3$ BS groups and cache storage capacity $B_C = 3$ files at each BS. The detailed cache data structure is elaborated below. Since all BSs in the same BS group cache the same subset of content files, we can use a single cache memory with B_C memory blocks to represent the cache data structure for each BS group. First, the cache memory for each BS group is divided into B_C memory blocks of size F bits and each memory block caches one content file. The cached content for all BS groups can be arranged in a matrix form, where the (n, m) -th entry is the n -th memory block for the m -th BS group, as illustrated in Fig. 4. For a given cache storage allocation vector \mathbf{q} satisfying (1), the L content files are placed one after another to fill in the cache memory sequentially from left to right and top to bottom in the matrix of cache memory blocks, where the l -th content file fills a total number of q_l cache memory blocks. Specifically, if $q_l = 0$, then the l -th content file is not stored in any of the BS caches, and will be

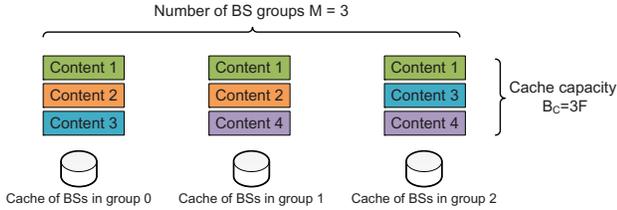


Fig. 4. Illustration of the cache placement and cache data structure in Example 1.

served via the backhaul at the BSs. If $q_l \neq 0$, then the l -th content file is stored in the cache of the BSs in $\bigcup_{m \in \mathcal{M}_l} \Phi_m^b$, where \mathcal{M}_l is the index of the BS groups that store the l -th content file, given by

$$\mathcal{M}_l = \left\{ \text{mod} \left(\sum_{l'=1}^{l-1} q_{l'}, M \right), \text{mod} \left(\sum_{l'=1}^{l-1} q_{l'} + 1, M \right), \dots, \text{mod} \left(\sum_{l'=1}^{l-1} q_{l'} + q_l - 1, M \right) \right\}. \quad (2)$$

The proposed scheme ensures that for a given \mathbf{q} satisfying (1), the number of content files stored in each cache is the same, and each BS caches B_C distinct content files. The proposed cache placement scheme is illustrated in the following example.

Example 1 (Cache placement): Consider a network with $M = 3$ BS groups, cache size $B_C F = 3F$, and total number of content files $L = 6$. The cache storage allocation vector is given by $\mathbf{q} = [3, 2, 2, 2, 0, 0]$, satisfying the cache storage capacity constraint given in (1). As illustrated in Fig. 4, the first content file is stored in three BS groups indexed by $\mathcal{M}_1 = \{0, 1, 2\}$. Each of the other content files is stored in two BS groups, given by $\mathcal{M}_2 = \{0, 1\}$, $\mathcal{M}_3 = \{0, 2\}$, $\mathcal{M}_4 = \{1, 2\}$, and $\mathcal{M}_5 = \mathcal{M}_6 = \emptyset$. It can be easily seen that each BS caches three distinct content files, which satisfies the cache storage capacity constraint.

Note that for practical consideration, the initialized cache content at the BSs does not adapt to the instantaneous realization of the user request at fast timescale. Instead, \mathbf{q} is adaptive only to the content popularity statistics. As a result, the BS cache update is done over a slow timescale when the network is lightly loaded.

B. Content Delivery

1) *Content-Centric User Scheduling:* We adopt a content-centric user scheduling scheme. Different from the conventional connection-based user scheduling scheme, which is based on physical layer parameters, this content-centric user scheduling scheme jointly considers both the physical layer and content status of BS caches. Consider a user which requests the l -th content file. If the requested l -th content file is stored in some of the BS caches in the network, i.e., $q_l \neq 0$, then the user is associated with the nearest BS which stores the l -th content file in its cache. Otherwise, if the requested l -th content file is not stored in any BS caches, i.e., $q_l = 0$, then the user is associated with the geographically nearest

BS in Φ^b (and the content file is fetched via the backhaul). The proposed user association scheme is illustrated using the example in Fig. 2. The blue content file requested by u_0 is stored in the cache of BS 0 and BS 2; in this case u_0 is associated with BS 0, which is the nearest BS which stores the blue content file. The red content file requested by u_1 is not stored in any BS caches, then u_1 is associated with BS 1, which is the geographically nearest BS, and the red content file is fetched via the backhaul. A similar content-centric user association has also been adopted in existing works on cached wireless networks [9] and [11]. Compared with the conventional nearest BS association, the content-centric user association schemes in our paper, [9] and [11] require some additional information about user requests and content placement at the BSs. Since the user requests change at a much longer timescale compared with the duration of a time slot, the induced additional overhead is low, and is practically feasible.

Without loss of generality, we study the performance of a typical user, which is located at the origin. Denote u_0 as the typical user, and denote B_0 as the serving BS of u_0 . Denote K_l as the number of users associated with B_0 which request the l -th content file, and denote $\mathbf{K} = [K_1, \dots, K_L]$ as the BS loading vector. \mathbf{K} may take value $\mathbf{k} = [k_1, \dots, k_L]$, where $k_l \in \{0, 1, \dots\}$. Note that not all associated users of B_0 can always be served by B_0 at the same time. Due to the limited backhaul transmission capacity, if more than B_B uncached content files are requested from one BS, then B_B users are randomly selected to be served with the content files obtained from the backhaul.¹ Denote S as the user scheduling state, where $S = 1$ represents the event that u_0 is scheduled to be served by B_0 , and $S = 0$ represents the event that u_0 is not scheduled to be served. The user scheduling is illustrated in the following example.

Example 2 (User scheduling): Consider a network with $M = 3$ BS groups, backhaul transmission capacity $B_B \tau = 2\tau$, cache capacity $B_C F = 3F$ at each BS, and the total number of content files $L = 6$. The cache storage allocation vector is given by $\mathbf{q} = [3, 2, 2, 2, 0, 0]$, and the BS loading vector is given by $[5, 4, 4, 3, 3, 2]$, which indicates that among the users associated with B_0 , the number of users requesting the first content file is 5, the number of users requesting the second content file is 4, and so on. Since $q_l \neq 0$, $l \in \{1, \dots, 4\}$, the users requesting the first four content files can be served using cached data. However, since $q_5 = q_6 = 0$, users requesting the fifth and sixth content files are served via the data obtained from the backhaul. Due to the backhaul transmission capacity constraint, two users will be randomly selected from the five users requesting the fifth and sixth content files to be served using the data obtained from the backhaul.

2) *PHY Transmission:* We adopt unicast and frequency division multiple access (FDMA) with uniform bandwidth and transmit power allocation for the users associated with each

¹Note that even when two users request the same content file from the backhaul, the probability that the two users request the same portion of the content file within the current time slot is very small for the typical content file size and slot duration, and thus they still need to consume a backhaul capacity of 2τ bits/s.

BS.² Consider one BS which simultaneously transmits to a total number of G_0 associated users. The BS transmits each of the associated users at a rate of τ bps over bandwidth $\frac{W}{MG_0}$. The transmit power is proportional to the allocated bandwidth, given by $\frac{P}{G_0}$. We assume that all BSs are active. When u_0 is served with file l_0 , the received signal of u_0 is given by

$$y_0 = D_{0,0}^{-\frac{\alpha}{2}} h_{0,0} x_0 + \sum_{n \in \Phi_{m_0}^b \setminus B_0} D_{n,0}^{-\frac{\alpha}{2}} h_{n,0} x_n + z_0, \quad (3)$$

where $D_{0,0}$ is the distance between B_0 and u_0 , $h_{0,0} \sim \mathcal{CN}(0,1)$ is the small-scale channel fading between B_0 and u_0 , x_0 is the transmit signal from B_0 to u_0 satisfying the transmit power constraint $\mathbb{E}(\|x_0\|) = \frac{P}{G_0}$, $\Phi_{m_0}^b$ is the group of BSs which B_0 belongs to (i.e., $B_0 \in \Phi_{m_0}^b$), $D_{n,0}$ is the distance between BS n and u_0 , $h_{n,0} \sim \mathcal{CN}(0,1)$ is the small-scale channel fading between BS n and the typical user u_0 , x_n is the transmit signal from BS n to its associated user in the m_0 -th frequency band satisfying transmit power constraint $\mathbb{E}(\|x_n\|) = \frac{P}{G_0}$, z_0 is the complex additive white Gaussian noise of power $\frac{WN_0}{MK_0}$, and N_0 is the noise spectral density. In this paper, we consider the high SINR regime where $P/W \gg N_0$. The signal-to-interference ratio (SIR) of u_0 is given by

$$\text{SIR} = \frac{D_{0,0}^{-\alpha} |h_{0,0}|^2}{\sum_{n \in \Phi_{m_0}^b \setminus B_0} D_{n,0}^{-\alpha} |h_{n,0}|^2}. \quad (4)$$

In the interference-limited regime, the achievable rate of u_0 is given by

$$C = \frac{W}{MG_0} \log_2(1 + \text{SIR}). \quad (5)$$

IV. PERFORMANCE METRIC AND ANALYSIS

The requested content file can be decoded correctly at u_0 only when u_0 is scheduled to be served by B_0 (i.e., $S = 1$) and the physical layer achievable rate is larger than the target rate (i.e., $C \geq \tau$). Therefore, the average successful transmission probability is defined as

$$p(M, \mathbf{q}) \triangleq \Pr[C \geq \tau, S = 1] \quad (6)$$

$$= \Pr[S = 1] \Pr[C \geq \tau | S = 1] \quad (7)$$

$$= \mathbb{E}_{\mathbf{K}, L_0} \Pr[S = 1 | \mathbf{K}, L_0] \Pr[C \geq \tau | \mathbf{K}, L_0, S = 1], \quad (8)$$

where L_0 is the random user request from the typical user u_0 , and $L_0 = l_0$ represents that the l_0 -th content file is requested by u_0 . The probability is with respect to the distribution of the random user requests L_0 , index of the BS group m_0 that B_0 belongs to, BS loading \mathbf{K} , large-scale channel fading $D_{0,0}^{-\frac{\alpha}{2}}$, and small-scale channel fading

²In practice, each file consists of a large number of segments, and the probability of two users requesting the same segment at the same time is small. As pointed out in [18], even though users keep requesting the same few popular files, the asynchronism of their requests is usually large with respect to the duration of the file (e.g., video) itself, such that the probability that a single transmission from the source nodes is useful for more than one user (i.e., multicasting) is essentially zero. This phenomenon is called “asynchronous content reuse” in [18]. As a result, “naive” multicasting due to repeated requests for the same segment of the same file at the same time from different users is unlikely to occur in practice.

$h_{0,0}$. Note that the number of BS groups and subbands M and cache storage capacity allocation vector \mathbf{q} fundamentally determine the average successful transmission probability that can be achieved. As a result, we explicitly write $p(M, \mathbf{q})$ as a function of M and \mathbf{q} . We will analyze the *conditional user scheduling probability* $\Pr[S = 1 | \mathbf{K} = \mathbf{k}, L_0 = l_0]$ and *conditional physical layer successful transmission probability* $\Pr[C \geq \tau | \mathbf{K} = \mathbf{k}, L_0 = l_0, S = 1]$, respectively, in the following two subsections.

A. Conditional User Scheduling Probability

Under the proposed content-centric user scheduling scheme, the number of users associated with B_0 whose requested content files exist in the cache is $\sum_{l \in \{l | q_l \neq 0\}} K_l$, and the number of users associated with B_0 whose requested content files do not exist in the cache and have to be fetched from the backhaul is $\sum_{l \in \{l | q_l = 0\}} K_l$. Therefore, the total number of users simultaneously served by B_0 is given by

$$G_0 = \sum_{l \in \{l | q_l \neq 0\}} K_l + \min \left\{ \sum_{l \in \{l | q_l = 0\}} K_l, B_B \right\}, \quad (9)$$

where the minimum in the second term is due to the limited backhaul transmission capacity. In general, we have $G_0 \leq \sum_{l=1}^L K_l$, and if $G_0 < \sum_{l=1}^L K_l$, the $\sum_{l=1}^L K_l - G_0$ unserved users will suffer from outage caused by the limited backhaul capacity. As a result, the conditional user scheduling probability is given by

$$\Pr[S = 1 | \mathbf{K} = \mathbf{k}, L_0 = l_0] = \begin{cases} 1, & q_{l_0} \neq 0, \\ \min \left\{ \frac{B_B}{\sum_{l \in \{l | q_l = 0\}} k_l}, 1 \right\}, & q_{l_0} = 0. \end{cases} \quad (10)$$

Remark 1: Equation (10) illustrates the following effect of the caching allocation strategy \mathbf{q} and the number of BS groups (i.e. subbands) M on the conditional user scheduling probability:

1) Effect of \mathbf{q} :

- a) $q_{l_0} \neq 0$: The requested l_0 -th content file is stored in some of the BS caches, and the user scheduling probability is given by $\Pr[S = 1 | \mathbf{K} = \mathbf{k}, L_0 = l_0] = 1$.
- b) $q_{l_0} = 0$: The requested l_0 -th content file does not exist in the cache and has to be fetched from the backhaul. As more content files need to be fetched from the backhaul (i.e., $\sum_{l \in \{l | q_l = 0\}} k_l$ increases), the user scheduling probability decreases.

- 2) **Effect of M :** M does not directly affect the user scheduling probability. However, it indirectly affects the user scheduling probability through the feasible region of the cache storage allocation vector \mathbf{q} . As M increases, a larger cache diversity can be achieved by collectively caching more content files at all BSs (as can be seen

from the cache capacity constraint $\sum_{l=1}^L q_l \leq MB_C$). Hence, the backhaul scheduling probability can be improved.

Consider the case in Example 2, conditioned on the typical user being served with the backhaul (i.e., $l_0 \in \{5, 6\}$ and $q_{l_0} = 0$), the probability that the typical user being scheduled for transmission is given by $\Pr[S = 1 | \mathbf{K} = \mathbf{k}, L_0 = l_0] = \frac{2}{5}$.

B. Conditional Physical Layer Successful Transmission Probability

In the following, we analyze the physical layer successful transmission probability conditioned on a given user request's realization and loading of B_0 .

Lemma 1 (Conditional physical layer successful transmission probability): Conditioned on BS loading vector \mathbf{k} , the l_0 -th content file being requested by u_0 , and u_0 being scheduled for transmission (i.e., $S = 1$), the physical layer successful transmission probability is given by

$$\Pr[C \geq \tau | \mathbf{K} = \mathbf{k}, L_0 = l_0, S = 1] = \frac{\lambda_{l_0}^b / \lambda_I^b}{\lambda_{l_0}^b / \lambda_I^b + \beta(M, g_0)}, \quad (11)$$

where

$$g_0 = \sum_{l \in \{l | q_l \neq 0\}} k_l + \min \left\{ \sum_{l \in \{l | q_l = 0\}} k_l, B_B \right\}, \quad (12)$$

is the realization of G_0 (which represents the number of users simultaneously served by B_0) conditioned on $\mathbf{K} = \mathbf{k}$,

$$\beta(M, g_0) = \frac{2}{\alpha} \left(2^{\frac{Mg_0\tau}{W}} - 1 \right)^{\frac{2}{\alpha}} B' \left(\frac{2}{\alpha}, 1 - \frac{2}{\alpha}, 2^{-\frac{Mg_0\tau}{W}} \right), \quad (13)$$

$B'(x, y, z) \triangleq \int_z^1 u^{x-1} (1-u)^{y-1} du$ is the complementary incomplete Beta function,

$$\lambda_{l_0}^b = \begin{cases} \lambda^b, & q_{l_0} = 0, \\ \frac{q_{l_0}}{M} \lambda^b & q_{l_0} \neq 0, \end{cases} \quad (14)$$

is the density of BSs that have access to the l_0 -th content file, and $\lambda_I^b = \lambda_b/M$ is the density of interfering BS in the transmitting subband of u_0 .

The proof can be found in Appendix A.

Remark 2: Lemma 1 shows the following effect of the number of BS groups M and the caching allocation strategy \mathbf{q} on the conditional physical layer successful transmission probability:

- 1) **Effect of M for given \mathbf{q} :** As M increases, the physical layer achievable rate decreases due to the lower spectral efficiency caused by the smaller frequency reuse factor $1/M$. Hence, the conditional physical layer successful transmission probability decreases. As a result, there is a tradeoff between spectral efficiency and user scheduling probability and we shall derive the optimal number of subbands and BS groups M to maximize the successful transmission probability in Section V. Note that when there is only one subband and one BS group ($M = 1$), each user is always served by

the geographically nearest BS, and the system model reduces to the conventional cellular network with PPP distributed BSs and users, as considered in [19]. In this case, $\lambda_{l_0}^b = \lambda_I^b = \lambda_b$, and the conditional physical layer successful transmission probability degenerates to $\Pr[C \geq \tau | \mathbf{K} = \mathbf{k}, L_0 = l_0, S = 1] = \frac{1}{1 + \beta(1, g_0)}$, which is consistent with the result in [19].

2) Effect of \mathbf{q} for given M :

- a) $q_{l_0} \neq 0$: The file requested by u_0 is stored in the cache of BSs, and the density of BSs that have access to the l_0 -th content file is given by $\lambda_{l_0}^b = \frac{q_{l_0}}{M} \lambda_b$. As q_{l_0} increases, the density of BSs that have access to l_0 -th content file also increases, and hence, the distance between u_0 and B_0 decreases. Meanwhile, the interference of u_0 only comes from the BSs in the same group as B_0 with BS density $\lambda_I^b = \frac{\lambda_b}{M}$, which is not affected by q_{l_0} . As a result, when q_{l_0} increases, the conditional physical layer successful transmission probability increases.
- b) $q_{l_0} = 0$: Since u_0 is associated with the geographically nearest BS with BS density $\lambda_{l_0}^b = \lambda_b$, the conditional physical layer successful transmission probability is given by $\Pr[C \geq \tau | \mathbf{K} = \mathbf{k}, L_0 = l_0, S = 1] = \frac{M}{M + \beta(M, g_0)}$, which is the same as the case when $q_{l_0} = M$.

C. Average Successful Transmission Probability

The average successful transmission probability is a function of the BS loading \mathbf{K} , which is a random vector. To simplify the analysis, we first compute the expectation of the BS loading vector \mathbf{K} as follows.

Lemma 2 (Expectation of the BS loading vector \mathbf{K}): The expectation of the BS loading vector \mathbf{K} is given by

$$\tilde{k}_l \triangleq \mathbb{E}[K_l] = \rho_l + \frac{9\lambda_u \rho_l}{7\lambda_b}. \quad (15)$$

The proof can be found in Appendix B.

Now instead of considering the distribution of \mathbf{K} , we approximate the BS loading \mathbf{K} using its expectations in Lemma 2. The approximate successful transmission probability conditioned on the l_0 -th content file being requested by u_0 is given by

$$\begin{aligned} & \Pr[C \geq \tau, S = 1 | L_0 = l_0] \\ &= \mathbb{E}_{\mathbf{K}} \Pr[S = 1 | \mathbf{K} = \mathbf{k}, L_0 = l_0] \\ & \quad \times \Pr[C \geq \tau | \mathbf{K} = \mathbf{k}, L_0 = l_0, S = 1] \end{aligned} \quad (16)$$

$$\begin{aligned} & \approx \Pr[S = 1 | \mathbf{K} = \mathbb{E}[\mathbf{K}], L_0 = l_0] \\ & \quad \times \Pr[C \geq \tau | \mathbf{K} = \mathbb{E}[\mathbf{K}], L_0 = l_0, S = 1], \end{aligned} \quad (17)$$

where the approximation in (17) is obtained by replacing the probability density function of \mathbf{K} with a delta function $\delta(x - \mathbb{E}[\mathbf{K}])$.

Using (10), (11), (15), and (17), the average successful transmission probability can then be approximated as

$$\tilde{p}(M, \mathbf{q}) \triangleq \sum_{l_0=1}^L \tilde{p}_{l_0} \rho_{l_0} \approx p(M, \mathbf{q}), \quad (18)$$

where

$$\tilde{p}_{l_0} = \begin{cases} \frac{q_{l_0}}{q_{l_0} + \beta(M, \tilde{g}_0)}, & q_{l_0} \neq 0, \\ \frac{M}{M + \beta(M, \tilde{g}_0)} \min \left\{ \frac{B_B}{\sum_{l \in \{l | q_l = 0\}} \tilde{k}_l}, 1 \right\}, & \text{otherwise.} \end{cases} \quad (19)$$

and

$$\tilde{g}_0 = \sum_{l \in \{l | q_l \neq 0\}} \tilde{k}_l + \min \left\{ \sum_{l \in \{l | q_l = 0\}} \tilde{k}_l, B_B \right\}. \quad (20)$$

In Section VI, simulations show that the approximate gap between \tilde{p} and the simulated successful transmission probability is quite small under various scenarios. Therefore, in the rest of the paper, the optimization of the frequency reuse factor $1/M$ and cache storage capacity allocation vector \mathbf{q} will be based on the approximate average successful transmission probability in (18). The accurate expression of the successful transmission probability is also provided, but it is too complicated to provide any useful insight. Interested readers should please refer to Appendix C for details. In the following section, we formulate and solve an optimization problem to find the optimal M and \mathbf{q} that maximize the approximate average successful transmission probability.

V. OPTIMIZATION OF CACHE STORAGE ALLOCATION AND FREQUENCY REUSE

A. Problem Formulation

The problem of finding the optimal frequency reuse factor and cache storage capacity allocation vector that maximize the approximate average successful transmission probability is formulated as:

$$\mathcal{P} : \max_{M \in \mathbb{N}^+, \mathbf{q}_l \in \{0, 1, \dots, M\}, \forall l} \tilde{p}(M, \mathbf{q}) \quad (21)$$

$$\text{s.t.} \quad q_l \geq q_{l+1}, l = 1, \dots, L-1, \quad (22)$$

$$\sum_{l=1}^L q_l \leq MB_C. \quad (23)$$

Denote M^* and \mathbf{q}^* as the optimal solution of \mathcal{P} . Constraint (22) is used to simplify the optimization algorithm design. Simulation results show that our proposed scheme with constraint (22) achieves a reasonably large average successful transmission probability gain over existing caching schemes in typical scenarios.

Problem \mathcal{P} is an integer optimization problem, and the objective function is neither convex nor concave. Even if we fix M and relax the integer constraint on \mathbf{q} to allow it to be a real vector, the relaxed problem is still very difficult to solve due to the indicator function w.r.t. q_l in (19) and (20), and the complicated function $\beta(M, \tilde{g}_0)$ w.r.t. \tilde{g}_0 (recall that \tilde{g}_0 also depends on \mathbf{q}) in (13). As a result, it is highly non-trivial to even design a low-complexity algorithm for Problem \mathcal{P} by solving the above relaxed problem.

B. Problem Transformation and Optimization

For a fixed M , the primary difficulty in solving Problem \mathcal{P} is how to deal with the user scheduling probability (10) and expectation of the BS loading (15), in which \mathbf{q} appears in the indication function in the subscript of summation. To address this challenge, we introduce an auxiliary variable L' , which is the number of content files that can be found in BS caches. Under (22), we have

$$q_l \geq 1, \forall l \leq L', \quad (24)$$

$$q_l = 0, \forall l > L'. \quad (25)$$

Note that for a given L' , the set of content files that need to be fetched from backhaul (whose indexes are given by $\{L' + 1, \dots, L\}$) is fixed. As a result, the backhaul success probability given by (10) is fixed. If we further assume M is given, then for the content files indexed by $\{L' + 1, \dots, L\}$, the successful transmission probability is also fixed. In this case, to find the optimal \mathbf{q} that maximizes the average successful transmission probability, we only need to minimize the average physical layer outage probability over the content files that are stored in BS caches. Specifically, after relaxing the integer constraint on \mathbf{q} , \mathcal{P} can be decomposed into a set of sub-problems that minimize the average physical layer outage probability for a given M and L' , which is given by

$$\tilde{\mathcal{P}}(M, L') : \min_{\mathbf{q}} \sum_{l=1}^{L'} \frac{\rho_l \beta(M, \tilde{g}_0)}{q_l + \beta(M, \tilde{g}_0)} \quad (26)$$

$$\text{s.t.} \quad \sum_{l=1}^{L'} q_l = MB_C, \quad (27)$$

$$q_l \geq q_{l+1}, \forall l \in \{1, \dots, L' - 1\}, \quad (28)$$

$$1 \leq q_l \leq M, \forall l \in \{1, \dots, L'\} \quad (29)$$

for $M \in \mathbb{N}^+$ and $L' \in [B_C, \min\{MB_C, L\}]$. Denote $\tilde{\mathbf{q}}^*(M, L')$ as the optimal solution of the sub-problem $\tilde{\mathcal{P}}(M, L')$. Note that for the given M and L' , both \tilde{K}_0 and $\beta(M, \tilde{g}_0)$ are fixed. It can be easily seen that $\tilde{\mathcal{P}}(M, L')$ is a convex minimization problem, and $\tilde{\mathbf{q}}^*(M, L')$ can be obtained using Karush–Kuhn–Tucker (KKT) conditions [20], as in the following theorem.

Theorem 1 (Optimal solution of $\tilde{\mathcal{P}}(M, L')$): The optimal solution $\tilde{\mathbf{q}}^*(M, L')$ of problem $\tilde{\mathcal{P}}(M, L')$ is given by

$$\tilde{q}_l^*(M, L') = \min \left\{ M, \max \left\{ 1, \sqrt{\rho_l / \lambda^*} - \beta(M, \tilde{g}_0) \right\} \right\}, \forall l \in \{1, \dots, L'\}, \quad (30)$$

where λ^* satisfies

$$\sum_{l=1}^{L'} \min \left\{ M, \max \left\{ 1, \sqrt{\rho_l / \lambda^*} - \beta(M, \tilde{g}_0) \right\} \right\} = MB_C. \quad (31)$$

The proof can be found in Appendix D.

The content popularity distribution ρ and physical layer parameter represented by $\beta(M, \tilde{g}_0)$ jointly affect $\tilde{\mathbf{q}}^*(M, L')$. Note that content file with higher popularity is allocated more cache storage resources. For a heavy-tailed popularity

Algorithm 1 Subbands decision and cache storage capacity allocation

```

1: for  $M = 1, \dots, M_{\max}$  do
2:   for  $L' = B_C, \dots, \min\{MB_C, L\}$  do
3:     Calculate optimal solution  $\tilde{\mathbf{q}}^*(M, L')$  of problem
        $\tilde{\mathcal{P}}(M, L')$  using Theorem 1
4:   end for
5: end for
6:  $(\tilde{M}^*, L'^*) = \arg \max \tilde{p}(M, \mathbf{q})$  and  $\tilde{\mathbf{q}}^* = \tilde{\mathbf{q}}^*(\tilde{M}^*, L'^*)$ 
7:  $\hat{\mathbf{q}}^* = \lfloor \tilde{\mathbf{q}}^* \rfloor$ , where  $\lfloor \cdot \rfloor$  is the rounding down function
8: while  $\sum_{l=1}^L \hat{q}_l^* < \tilde{M}^* B_C$  do
9:    $l' = \arg \min_{l \in \{1, \dots, L'^*\}} \left( \frac{\rho_l}{\hat{q}_l^* + 1 + \beta(\tilde{M}^*, \tilde{g}_0)} - \frac{\rho_l}{\hat{q}_l^* + \beta(\tilde{M}^*, \tilde{g}_0)} \right)$ 
10:   $\hat{q}_{l'}^* = \hat{q}_{l'}^* + 1$ 
11: end while

```

distribution, the differences between ρ_l 's are small, and the cache capacity is allocated to more content files instead of concentrating on a few most popular files.

To calculate a solution of \mathcal{P} , we first enumerate L' and M to find the best solution $(\tilde{M}^*, \tilde{\mathbf{q}}^*(\tilde{M}^*, L'^*))$ that maximizes the objective function $\tilde{p}(M, \mathbf{q})$ of \mathcal{P} . Then $\tilde{\mathbf{q}}^*(\tilde{M}^*, L'^*)$ is rounded down such that it becomes a feasible integer solution of \mathcal{P} . Finally, the residue cache storage capacity induced by the rounding down operation is allocated to the content files that minimizes the average physical layer outage probability in a greedy manner. The detailed algorithm is given in Algorithm 1. Note that the optimal objective value $\tilde{p}(\tilde{M}^*, \tilde{\mathbf{q}}^*(\tilde{M}^*, L'^*))$ for the relaxed problem provides an upper bound of the optimal objective value of the original problem \mathcal{P} . In Section VI, we show that this upper bound is quite close to the objective value achieved by the integer solution $(\tilde{M}^*, \hat{\mathbf{q}}^*(\tilde{M}^*, L'^*))$ obtained using Algorithm 1, which shows that the proposed low-complexity algorithm is close-to-optimal for the original integer optimization problem.

In practice, we can set a limit M_{\max} for the maximum number of subbands searched by Algorithm 1, and M_{\max} can be used to control the tradeoff between performance and complexity. Note that the optimal number of subbands \tilde{M}^* is usually small. Otherwise the bandwidth would become insufficient to support the transmission rate τ . L' is upper bounded by MB_C , which is usually much smaller than the total number of content files. Algorithm 1 needs to solve $M_{\max}^2 B_C$ sub-problems (26), and each sub-problem can be efficiently solved using bisection. Note that the cache capacity B_C is usually much smaller than the total number of content files L . As a result, the computational complexity induced by enumeration is low. In Section VI, we shall show that Algorithm 1 is quite efficient and achieves a large average successful transmission probability gain over conventional single-band caching schemes. In Section VI, we shall also provide numerical insight on \tilde{M}^* and L'^* , i.e., the optimal number of subbands and BS groups, and how many content files to cache.

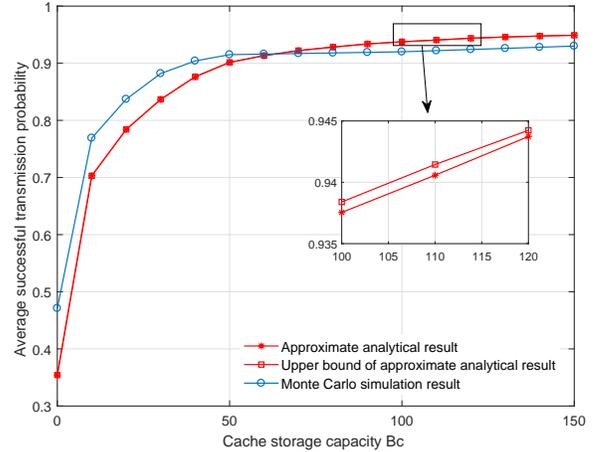


Fig. 5. Illustration of the approximate gap between \tilde{p} and p when $\gamma = 0.8$ and $B_B = 5$.

VI. SIMULATION RESULTS

In this section, we show that the simulation results are consistent with the theoretical results. We also demonstrate the performance gain of our scheme over the following baselines:

- **Baseline 1:** A standard policy which caches the most popular content (MPC) at each BS [6]. The frequency reuse factor is one, and orthogonal frequency-division multiple access (OFDMA) is applied at each BS to serve multiple users.
- **Baseline 2:** Geographic caching problem (GCP) proposed in [9], which exploits spatial caching diversity by random caching. The frequency reuse factor is one, and OFDMA is applied at each BS to serve multiple users. The corresponding cache placement is optimized to maximize the average successful transmission probability (the performance metric considered in this paper), using an algorithm similar to Algorithm 1, which combines enumeration and convex optimization.
- **Baseline 3:** Separated design of MPC and frequency reuse. The frequency reuse factor is optimized to maximize the average successful transmission probability.
- **Baseline 4:** Separated design of GCP and frequency reuse. The cache placement and frequency reuse factor are separately optimized to maximize the average successful transmission probability.

For the proposed scheme, the number of subbands \tilde{M}^* and cache storage allocation vector $\hat{\mathbf{q}}^*$ is obtained using Algorithm 1, where the maximum number of subbands is given by $M_{\max} = 5$. The system parameters are set as follows:

- **Geometric parameters:** BS density $\lambda^b = 3 \times 10^{-5}$, user density $\lambda^u = 3 \times 10^{-4}$.
- **Channel parameters:** Path loss exponent $\alpha = 4$, total bandwidth $W = 20\text{MHz}$, target rate $\tau = 0.1\text{ Mbps}$, length of each time slot $\nu = 1\text{ ms}$.
- **Content parameters:** Content library size $L = 1000$, content popularity follows Zipf distribution with exponent γ [21], [22].

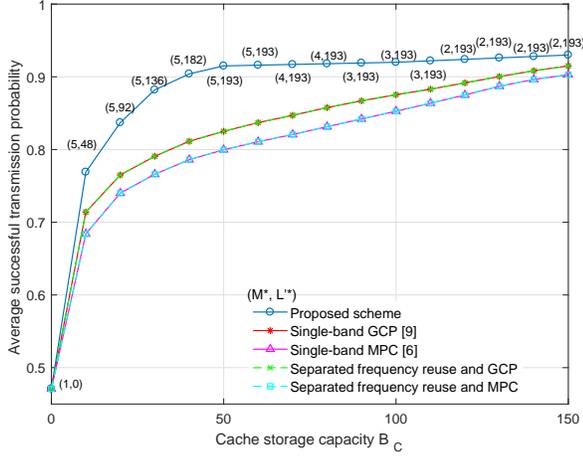


Fig. 6. Average successful transmission probability versus cache capacity when $\gamma = 0.8$ and $B_B = 5$.

In Fig. 5, we plot the average successful transmission probability versus the cache storage capacity when $\gamma = 0.8$ and $B_B = 5$. It can be observed that the simulation results largely match the theory; the relative approximation error is large only when the average successful transmission probability is very low, which is not a desirable operating regime for practical systems. In a practical regime when the success probability is high, the approximate error is small. In this case, the theoretical approximation (18) can capture the first-order behavior of the proposed scheme and can be used to optimize the cache design. The results also show that the objective value achieved using Algorithm 1 is quite close to the upper bound given by $\tilde{p}(\tilde{M}^*, \tilde{\mathbf{q}}^*(\tilde{M}^*, L^*))$ (it is an upper bound since the integer constraint on \mathbf{q} is relaxed), which indicates that the proposed low-complexity algorithm is close to optimum for the original integer optimization problem.

In Fig. 6 – Fig. 8, we compare the performance between the proposed scheme and the baselines. It is observed that for separated design and optimization of frequency reuse with either GCP or MPC, the optimal frequency reuse strategy is usually to let all BSs use the entire bandwidth (i.e., optimal frequency reuse factor is one). This is because a separated design of frequency reuse and GCP cannot completely address the strong interference issue in random caching. Meanwhile, for separated design and optimization of frequency reuse and MPC, each scheduled user is always served by the geographically nearest BS, and thus the inter-cell interference is weaker compared to the case with random caching. In both cases, a frequency reuse factor less than one would lead to lower physical layer successful transmission probability due to less bandwidth being allocated to each BS.

• **Impact of cache capacity B_C (Fig. 6):**

- **On the optimal number of subbands \tilde{M}^* :** When B_C is small, the optimal number of subbands \tilde{M}^* is large, so that the proposed scheme can achieve a lower backhaul outage probability by exploiting spatial caching diversity. As B_C increases, due to

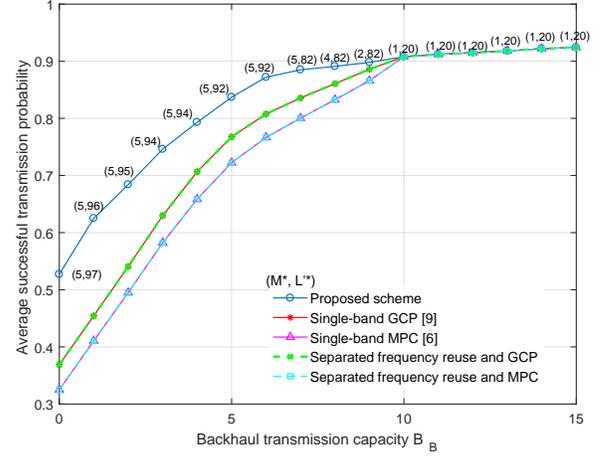


Fig. 7. Average successful transmission probability versus backhaul transmission capacity when $\gamma = 0.8$ and $B_C = 20$.

sufficient cache storage capacity, the importance of spatial caching diversity decreases. As a result, \tilde{M}^* decreases, so that the distance between the user and BS decreases and the physical layer achieves larger spectral efficiency.

- **On the optimal caching strategy:** When B_C is small, the optimal cache storage allocation of the proposed scheme is not to use the cache to store only the most popular content files in every BS (i.e., $L^* \neq B_C$). As B_C increases, L^* increases since more content files can be stored in the cache, which leads to a lower backhaul outage probability.

It can be seen that single-band MPC [6] and single-band GCP [9] are not optimal when the cache capacity B_C is limited.

• **Impact of backhaul transmission capacity B_B (Fig. 7):**

- **On the optimal number of subbands \tilde{M}^* :** When the backhaul capacity B_B is small, the backhaul can handle fewer user requests. In this case, it is better to increase M to improve the spatial caching diversity gain and cache hit probability. When B_B is large, the backhaul can handle more user requests and the importance of spatial caching diversity decreases. In this case, it is better to decrease M to achieve larger physical layer spectral efficiency. Therefore, the optimal number of subbands \tilde{M}^* decreases with the cache capacity B_B .
- **On the optimal caching strategy:** When cache capacity B_B is small, it is important to exploit the spatial caching diversity to improve the cache hit probability. In this case, the optimal cache storage allocation is not to use the cache to store only the most popular content files in every BS, but to use some cache capacity to store some less popular content files as well. On the other hand, as B_B increases, more content files can be fetched from the backhaul without causing backhaul outage. In this case, it is not necessary to store many less popular

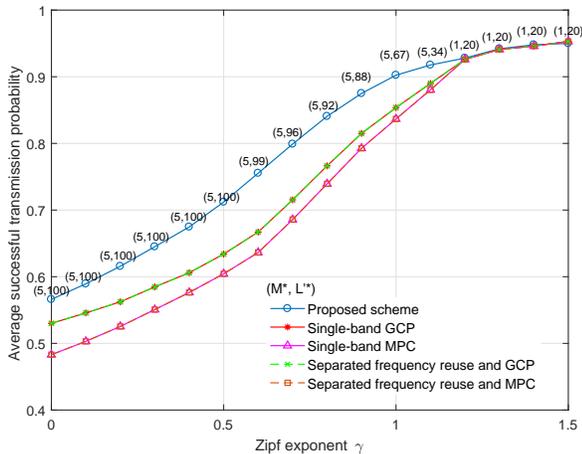


Fig. 8. Average successful transmission probability versus Zipf distribution exponent when $B_C = 20$ and $B_B = 5$.

content files; i.e., L^* will decrease.

It can be seen that single-band MPC [6] and single-band GCP [9] are not optimal when the backhaul transmission capacity is limited.

• **Impact of Zipf exponents γ (Fig. 8):**

- **On the optimal number of subbands \tilde{M}^* :** When γ is small (i.e., the popularity distribution is flat), it is important to exploit the spatial caching diversity to improve the cache hit probability. In this case, it is better to increase M to improve the spatial caching diversity gain and cache hit probability. On the other hand, as γ increases, the user requests concentrate on a few content files, and hence the benefit of spatial caching diversity decreases. Therefore, the optimal number of subbands \tilde{M}^* decreases.
- **On the optimal caching strategy:** When γ is small, it is important to exploit the spatial caching diversity to improve the cache hit probability. In this case, the optimal cache storage allocation is not to use the cache to store only the most popular content files in every BS. As γ increases, the user requests concentrate on a few content files, hence the benefit of spatial caching diversity decreases. In this case, it is not desirable to store too many less popular content files, i.e., L^* will decrease.

Based on the simulation results, our proposed scheme outperforms single-band MPC [6] and single-band GCP [9], especially for the values of γ from 0 to 1, with is typical for general applications [21].

Additionally, in Fig. 9, we plot the successful transmission probability under a given realization of PPP, which shows that the proposed design also works well in this case.

VII. CONCLUSION

In this paper, we propose a joint frequency reuse and caching scheme to achieve both the spatial cache diversity and

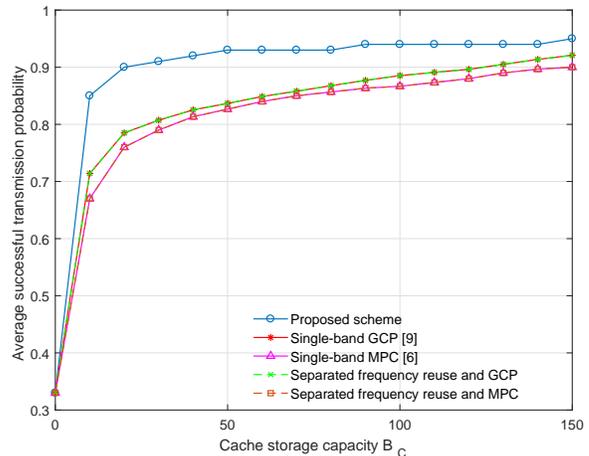


Fig. 9. Successful transmission probability versus cache capacity under a given realization of PPP when $\gamma = 0.8$ and $B_B = 5$.

interference mitigation in small-cell backhaul-limited wireless networks. We first derive a closed-form expression of the approximate successful transmission probability using the tools of stochastic geometry, and analyze the impact of key operating parameters. We then propose a low-complexity algorithm which combines enumeration and convex optimization to optimize the frequency reuse factor and the cache storage allocation vector. Finally, by simulations, we show that by exploiting the spatial cache diversity and interference mitigation benefits provided by the joint optimization of frequency reuse and caching, the proposed scheme achieves a large performance gain over the typical single-band MPC scheme [6] and random caching scheme [9], especially when the cache capacity and backhaul capacity at each BS are limited.

APPENDIX

A. Proof of Lemma 1

First, we calculate the physical layer successful transmission probability conditioned on file l_0 requested by u_0 , BS loading $\mathbf{K} = \mathbf{k}$, u_0 being scheduled for transmission (i.e., $S = 1$) and distance $D_{0,0} = d$, which is given by

$$\begin{aligned}
 & \Pr [C \geq \tau | \mathbf{K} = \mathbf{k}, L_0 = l_0, S = 1, D_{0,0} = d] \\
 & \stackrel{(a)}{=} \mathbb{E}_{\Phi_{m_0}^b} \left[\exp \left(- \left(2^{\frac{Mg_0\tau}{W}} - 1 \right) d^\alpha \sum_{n \in \Phi_{m_0}^b \setminus B_0} D_{n,0}^{-\alpha} |h_{n,0}|^2 \right) \right] \\
 & = \mathbb{E}_{\Phi_{m_0}^b} \left[\prod_{n \in \Phi_{m_0}^b \setminus B_0} \exp \left(- \left(2^{\frac{Mg_0\tau}{W}} - 1 \right) d^\alpha D_{n,0}^{-\alpha} |h_{n,0}|^2 \right) \right] \\
 & \stackrel{(b)}{=} \exp \left(- 2\pi\lambda_I^b \int_d^\infty \left(1 - \frac{1}{1 + \left(2^{\frac{Mg_0\tau}{W}} - 1 \right) d^{\alpha} r^{-\alpha}} \right) r dr \right) \\
 & \stackrel{(c)}{=} \exp \left(- \frac{2\pi}{\alpha} \lambda_I^b \left(2^{\frac{Mg_0\tau}{W}} - 1 \right)^{\frac{2}{\alpha}} B' \left(\frac{2}{\alpha}, 1 - \frac{2}{\alpha}, 2^{-\frac{Mg_0\tau}{W}} \right) d^2 \right) \\
 & = 1 \exp \left(- \pi \lambda_I^b \beta (M, g_0) d^2 \right), \tag{32}
 \end{aligned}$$

where (a) is due to Rayleigh fading channel $|h_{0,0}|^2 \stackrel{d}{\sim} \text{Exp}(1)$, (b) is obtained using the probability generating function of PPP [17, Page 235], and (c) is obtained by replacing $\left(2 \frac{Mg_0\tau}{w} - 1\right)^{-\frac{1}{\alpha}} d^{-1}r$ with t , and then replacing $\frac{1}{1+t-\alpha}$ with w .

Then, we calculate $\Pr[C < \tau | \mathbf{K} = \mathbf{k}, L_0 = l_0, S = 1]$ by removing the condition of $D_{0,0} = d$. The probability density function of $D_{0,0}$ is given by

$$f_{D_{0,0},l_0}(d) = 2\pi\lambda_{l_0}^b d \exp(-\pi\lambda_{l_0}^b d^2), \quad (33)$$

as the BSs storing the l_0 -th content file form a homogeneous PPP with density $\lambda_{l_0}^b$. By (32) and (33), we have

$$\begin{aligned} & \Pr[C \geq \tau | l_0, K_0, S = 1] \\ &= \int_0^\infty \Pr[C < \tau | \mathbf{K} = \mathbf{k}, L_0 = l_0, S = 1, D_{0,0} = d] \\ & \quad \times f_{D_{0,0},l_0}(d) dd \end{aligned} \quad (34)$$

$$= 2\pi\lambda_{l_0}^b \int_0^\infty d \exp(-\pi(\lambda_{l_0}^b + \lambda_l^b \beta(M, g_0)) d^2) dd \quad (35)$$

$$\stackrel{(a)}{=} \frac{\frac{\lambda_{l_0}^b}{\lambda_l^b}}{\frac{\lambda_{l_0}^b}{\lambda_l^b} + \beta(M, g_0)}, \quad (36)$$

where (a) is obtained using $\int_0^\infty d \exp(-cd^2) dd = \frac{1}{2c}$ (c is a constant).

B. Proof of Lemma 2

In the following, we first derive the probability mass function of \mathbf{k} . Note that due to the content-centric user scheduling scheme, each file $X_l \in \{X_1, X_2, \dots, X_L\}$ corresponds to a Voronoi tessellation, which is determined by the locations of all BSs which have access to file X_l . To calculate the probability mass function of \mathbf{k} , we need the probability density function of the size of the Voronoi cell which B_0 belongs to. Based on a widely used approximated form of this probability density function given in [19], the probability mass function of \mathbf{K} is given in the following lemma.

Lemma 3 (Probability mass function of \mathbf{K}): The probability mass function of \mathbf{K} conditioned on the l_0 -th content file being requested by u_0 is given by

$$\begin{aligned} & \Pr[\mathbf{K} = \mathbf{k} | L_0 = l_0] \\ &= \frac{1}{q_{l_0}} \sum_{m_0 \in \mathcal{M}_{l_0}} \Pr[\mathbf{K} = \mathbf{k} | L_0 = l_0, M_0 = m_0] \end{aligned} \quad (37)$$

for $q_{l_0} \neq 0$, and

$$\begin{aligned} & \Pr[\mathbf{K} = \mathbf{k} | L_0 = l_0] \\ &= \frac{1}{M} \sum_{m_0 \in \{0, \dots, M-1\}} \Pr[\mathbf{K} = \mathbf{k} | L_0 = l_0, M_0 = m_0] \end{aligned} \quad (38)$$

for $q_{l_0} = 0$, where \mathcal{M}_{l_0} is the set of indexes of BS groups that stores the l_0 -th content file, $\Pr[\mathbf{K} = \mathbf{k} | L_0 = l_0, M_0 = m_0]$

is the probability mass function of \mathbf{K} conditioned on the l_0 -th content file being requested by u_0 and $B_0 \in \Phi_{m_0}^b$, and $\Pr[K_l = k_l | L_0 = l_0, M_0 = m_0]$ is given by

$$\begin{aligned} & \Pr[K_l = k_l | L_0 = l_0, M_0 = m_0] \\ &= \begin{cases} \Psi\left(\lambda^u \rho_l, \frac{q_l \lambda^b}{M}\right), & l = l_0, q_{l_0} \neq 0, \\ \Psi\left(\lambda^u \rho_l, \lambda^b\right), & l = l_0, q_{l_0} = 0, \\ 1(k_l = 0), & l \neq l_0, q_l \neq 0, l \notin \mathcal{C}_{m_0} \\ \bar{\Psi}\left(\lambda^u \rho_l, \frac{q_l \lambda^b}{M}\right), & l \neq l_0, q_l \neq 0, l \in \mathcal{C}_{m_0}, \\ \bar{\Psi}\left(\lambda^u \rho_l, \lambda^b\right), & l \neq l_0, q_l = 0, \end{cases} \end{aligned} \quad (39)$$

where \mathcal{C}_{m_0} is the set of content files stored in the cache of the BSs in $\Phi_{m_0}^b$,

$$\Psi(x, y) = \frac{3.5^{4.5}}{\Gamma(4.5)} \frac{x^{k-1}}{y^{k-1} (k-1)!} \frac{\Gamma(k+3.5)}{\left(\frac{x}{y} + 3.5\right)^{k+3.5}}, \quad (40)$$

$$\bar{\Psi}(x, y) = \frac{3.5^{4.5}}{\Gamma(4.5)} \frac{x^k}{y^k k!} \frac{\Gamma(k+4.5)}{\left(\frac{x}{y} + 3.5\right)^{k+4.5}}. \quad (41)$$

The proof can be found in Appendix B1.

Using (39), we calculate the expectation of K_l conditioned on l_0 and m_0 , which is given by

$$\begin{aligned} & \mathbb{E}[K_l | L_0 = l_0, M_0 = m_0] \\ &= \sum_{k_l=0}^{\infty} \Pr[K_l = k_l | L_0 = l_0, M_0 = m_0] k_l \\ &= \begin{cases} 1 + \frac{9}{7} \frac{\lambda^u \rho_l}{q_l \lambda^b}, & l = l_0, q_{l_0} \neq 0, \\ 1 + \frac{9}{7} \frac{\lambda^u \rho_l}{\lambda^b}, & l = l_0, q_{l_0} = 0, \\ 0, & l \neq l_0, q_l \neq 0, l \notin \mathcal{C}_{m_0} \\ \frac{9}{7} \frac{\lambda^u \rho_l}{q_l \lambda^b}, & l \neq l_0, q_l \neq 0, l \in \mathcal{C}_{m_0}, \\ \frac{9}{7} \frac{\lambda^u \rho_l}{\lambda^b}, & l \neq l_0, q_l = 0. \end{cases} \end{aligned} \quad (42)$$

Then we calculate $\mathbb{E}[K_l | L_0 = l_0]$ by removing the condition on m_0 , given by

$$\mathbb{E}[K_l | L_0 = l_0] = \begin{cases} 1 + \frac{9}{7} \frac{\lambda^u \rho_l}{\lambda^b}, & l = l_0, \\ \frac{9}{7} \frac{\lambda^u \rho_l}{\lambda^b}, & l \neq l_0. \end{cases} \quad (44)$$

Finally, we calculate $\mathbb{E}[K_l]$ by removing the condition on l_0 , given by

$$\mathbb{E}[K_l] = \rho_l + \frac{9}{7} \frac{\lambda^u \rho_l}{\lambda^b}. \quad (45)$$

1) Probability Mass Function of \mathbf{K} : The probability mass function of \mathbf{K} depends on the probability density function of the size of the Voronoi cell of BS B_0 w.r.t. content file $l \in \{1, \dots, L\}$. Denote $f_Z(z)$ as the probability density function of the size of the Voronoi cell to which a randomly chosen user belongs, where Z is a random variable that denotes the size of the Voronoi cell normalized by the inverse of the density of BSs. A widely used approximated form of this probability density function is given by [19]

$$f_Z(z) = \frac{3.5^{4.5}}{\Gamma(4.5)} z^{3.5} \exp(-3.5z). \quad (46)$$

We first prove $\Pr [K_l = k_l | L_0 = l_0, M_0 = m_0]$ in (39):

- $l = l_0$ and $q_{l_0} \neq 0$: The users requesting the l -th content file form a homogeneous PPP with density $\lambda_u \rho_l$, and the BSs that store the l -th content file form a homogeneous PPP with density $\lambda^b q_l / M$. The probability mass function of K_l conditioned on l_0, m_0 and z is given by

$$\begin{aligned} & \Pr [K_l = k_l | L_0 = l_0, M_0 = m_0, Z = z] \\ &= \frac{\left(\frac{\lambda^u \rho_l}{\lambda^b q_l / M} z\right)^{k_l}}{k_l!} e^{-\frac{\lambda^u \rho_l}{\lambda^b q_l / M} z}. \end{aligned} \quad (47)$$

Then we calculate $\Pr [K_l = k_l | L_0 = l_0, M_0 = m_0]$ by removing the condition on z , given by

$$\begin{aligned} & \Pr [K_l = k_l | L_0 = l_0, M_0 = m_0] \\ &= \int_0^\infty \Pr [K_l = k_l - 1 | L_0 = l_0, M_0 = m_0, Z = z] \\ & \quad \times f_Z(z) dz \end{aligned} \quad (48)$$

$$= \int_0^\infty \frac{\left(\frac{\lambda^u \rho_l}{\lambda^b q_l / M} z\right)^{k_l - 1}}{(k_l - 1)!} e^{-\frac{\lambda^u \rho_l}{\lambda^b q_l / M} z} f_Z(z) dz \quad (49)$$

$$= \frac{3.5^{4.5}}{\Gamma(4.5)} \frac{(\lambda^u \rho_l)^{k_l - 1}}{(\lambda^b q_l / M)^{k_l - 1} (k_l - 1)!} \times \int_0^\infty z^{k_l + 2.5} \exp\left(-\left(\frac{\lambda^u \rho_l}{\lambda^b q_l / M} + 3.5\right) z\right) dz \quad (50)$$

$$= \Psi\left(\lambda^u \rho_l, \frac{q_l \lambda^b}{M}\right). \quad (51)$$

- $l = l_0$ and $q_{l_0} = 0$: The users requesting the l -th content file form a homogeneous PPP with density $\lambda^u \rho_l$. All BSs access to l -th content file via the backhaul, and they form a homogeneous PPP with density λ^b . Similar to (51), we have

$$\begin{aligned} & \Pr [K_l = k_l | L_0 = l_0, M_0 = m_0] \\ &= \int_0^\infty \Pr [K_l = k_l - 1 | L_0 = l_0, M_0 = m_0, Z = z] \\ & \quad \times f_Z(z) dz \end{aligned} \quad (52)$$

$$= \Psi(\lambda^u \rho_l, \lambda^b). \quad (53)$$

- $l \neq l_0, q_{l_0} \neq 0$ and $l \notin \mathcal{C}_{m_0}$: Note the $q_{l_0} \neq 0$ indicates that the l -th content file is stored in some of the BSs. Meanwhile, $l \notin \mathcal{C}_{m_0}$ indicates the BSs in $\Phi_{m_0}^b$ do not cache the l -th content file. As a result, the BSs in $\Phi_{m_0}^b$ do not serve the users requesting the l -th content file, which means that $k_l = 0$ is always satisfied.
- $l \neq l_0, q_{l_0} \neq 0$ and $l \in \mathcal{C}_{m_0}$: The users requesting the l -th content file form a homogeneous PPP with density $\lambda^u \rho_l$, and the BSs that store the l -th content file form a homogeneous PPP with density $\lambda^b q_l / M$. The conditional

probability mass function is given by

$$\begin{aligned} & \Pr [K_l = k_l | L_0 = l_0, M_0 = m_0] \\ &= \int_0^\infty \Pr [K_l = k_l | L_0 = l_0, M_0 = m_0, Z = z] \\ & \quad \times f_Z(z) dz \end{aligned} \quad (54)$$

$$= \int_0^\infty \frac{\left(\frac{\lambda^u \rho_l}{\lambda^b q_l / M} z\right)^{k_l}}{k_l!} e^{-\frac{\lambda^u \rho_l}{\lambda^b q_l / M} z} f_Z(z) dz \quad (55)$$

$$= \frac{3.5^{4.5}}{\Gamma(4.5)} \frac{(\lambda^u \rho_l)^{k_l}}{(\lambda^b q_l / M)^{k_l} k_l!} \times \int_0^\infty z^{k_l + 3.5} \exp\left(-\left(\frac{\lambda^u \rho_l}{\lambda^b q_l / M} + 3.5\right) z\right) dz \quad (56)$$

$$= \bar{\Psi}\left(\lambda^u \rho_l, \frac{q_l \lambda^b}{M}\right). \quad (57)$$

- $l \neq l_0, q_{l_0} = 0$: The users requesting the l -th content file form a homogeneous PPP with density $\lambda^u \rho_l$. All BSs access to l -th content file via the backhaul, and they form a homogeneous PPP with density λ^b . Similar to (57), we have

$$\begin{aligned} & \Pr [K_l = k_l | L_0 = l_0, M_0 = m_0] \\ &= \int_0^\infty \Pr [K_l = k_l | L_0 = l_0, M_0 = m_0, Z = z] \\ & \quad \times f_Z(z) dz \end{aligned} \quad (58)$$

$$= \bar{\Psi}(\lambda^u \rho_l, \lambda^b). \quad (59)$$

Finally, we calculate $\Pr [\mathbf{K} = \mathbf{k} | L_0 = l_0]$ by removing the condition on m_0 . For $q_{l_0} \neq 0, m_0$ is selected from the BS groups \mathcal{M}_{l_0} with equal probability. For $q_{l_0} = 0, m_0$ is selected from all the BS groups $\{0, \dots, M-1\}$ with equal probability. Hence, $\Pr [\mathbf{K} = \mathbf{k} | L_0 = l_0]$ is given by (37)–(38) by removing the condition on m_0 in $\Pr [\mathbf{K} = \mathbf{k} | L_0 = l_0, M_0 = m_0]$.

C. Expression of Successful Transmission Probability p

The BSs loading \mathbf{K} and SIR are correlated, since BSs with a larger cell size have higher loading and lower SIR [23]. However, the exact relationship between \mathbf{K} and SIR is very complex and is still not known. For tractability of the analysis, as in [23], the dependence is ignored. Hence, the successful transmission probability conditioned on the l_0 -th content file requested by u_0 and distance d is given by

$$\begin{aligned} p_{l_0, d} &= \sum_{\mathbf{k} \in \mathbb{N}^L} \Pr [\mathbf{K} = \mathbf{k} | L_0 = l_0] \\ & \quad \times \Pr [C \geq \tau | \mathbf{K} = \mathbf{k}, L_0 = l_0, S = 1] \\ & \quad \times \Pr [S = 1 | \mathbf{K} = \mathbf{k}, L_0 = l_0], \end{aligned} \quad (60)$$

where $\Pr [C \geq \tau | \mathbf{K} = \mathbf{k}, L_0 = l_0, S = 1]$ follows (11), and $\Pr [S = 1 | \mathbf{K} = \mathbf{k}, L_0 = l_0]$ follows (10). Then we calculate p_{l_0} by removing the condition on d :

$$p_{l_0} = \int_0^\infty p_{l_0, d}(d) f_{D_{0,0,l_0}}(d) dd. \quad (61)$$

Finally, by the total probability theorem, the average successful transmission probability is given by

$$p = \sum_{l_0=1}^L p_{l_0} \rho_{l_0}. \quad (62)$$

Note that the above expression of p is complicated, since the expression of the probability mass function of \mathbf{k} is complicated. As a result, it is hard to find the optimal number of subbands M and cache storage allocation vector \mathbf{q} that maximize the average successful transmission probability p .

D. Proof of Theorem 1

By removing constraint (28) in Problem $\tilde{\mathcal{P}}(M, L')$, we obtain the following problem:

$$\tilde{\mathcal{P}}(M, L') : \min_{\mathbf{q}_l} \sum_{l=1}^{L'} \rho_l \frac{\beta(M, \tilde{g}_0)}{q_{l_0} + \beta(M, \tilde{g}_0)} \quad (63)$$

$$\text{s.t.} \quad \sum_{l=1}^{L'} q_l = MB_C, \quad (64)$$

$$q_l \geq 1, \forall l \in \{1, \dots, L'\}, \quad (65)$$

$$q_l \leq M, \forall l \in \{1, \dots, L'\}. \quad (66)$$

Denote $\tilde{\mathbf{q}}^* = [\tilde{q}_1^*, \dots, \tilde{q}_{L'}^*]$ as the optimal solution of $\tilde{\mathcal{P}}(M, L')$. We will show later that an optimal solution of $\tilde{\mathcal{P}}(M, L')$ is also an optimal solution of $\mathcal{P}(M, L')$. Problem $\tilde{\mathcal{P}}(M, L')$ is a convex optimization problem. Introducing a Lagrange multiplier λ^* for equality constraint (64), multipliers ν_l^* for constraint (65), and multipliers ω_l^* for constraint (66), we obtain the KKT conditions:

$$\sum_{l=1}^{L'} \tilde{q}_l^* = MB_C, \tilde{q}_l^* \geq 1, \tilde{q}_l^* \leq M. \quad (67)$$

$$\nu^* \geq 0, \omega^* \geq 0. \quad (68)$$

$$\nu_l^* (1 - \tilde{q}_l^*) = 0, \omega_l^* (\tilde{q}_l^* - M) = 0. \quad (69)$$

$$-\frac{\rho_l}{(\tilde{q}_l^* + \beta(M, \tilde{g}_0))^2} + \lambda^* - \nu_l^* + \omega_l^* = 0, \forall l. \quad (70)$$

By eliminating ν , we have

$$\sum_{l=1}^{L'} \tilde{q}_l^* = MB_C, \tilde{q}_l^* \geq 1, \tilde{q}_l^* \leq M. \quad (71)$$

$$\omega^* \geq 0. \quad (72)$$

$$\left(\lambda^* - \frac{\rho_l}{(\tilde{q}_l^* + \beta(M, \tilde{g}_0))^2} + \omega_l^* \right) (1 - \tilde{q}_l^*) = 0, \quad (73)$$

$$\omega_l^* (\tilde{q}_l^* - M) = 0. \quad (74)$$

$$\lambda^* \geq \frac{\rho_l}{(\tilde{q}_l^* + \beta(M, \tilde{g}_0))^2} - \omega_l^*, \forall l. \quad (75)$$

- If $\lambda^* < \frac{\rho_l}{(1+\beta(M, \tilde{g}_0))^2} - \omega_l^*$, the last condition can only hold if $\tilde{q}_l^* > 1$, which by the third condition implies that $\lambda^* = \frac{\rho_l}{(\tilde{q}_l^* + \beta(M, \tilde{g}_0))^2} - \omega_l^*$, i.e., $\tilde{q}_l^* = \sqrt{\frac{\rho_l}{\lambda^* + \omega_l^*}} - \beta(M, \tilde{g}_0)$.

- If $\sqrt{\frac{\rho_l}{\lambda^* + \omega_l^*}} - \beta(M, \tilde{g}_0) \geq M$, since $\tilde{q}_l^* > M$ is impossible, we have $\tilde{q}_l^* = M$ and $\frac{\rho_l}{\lambda^* + \omega_l^*} = (M + \beta(M, \tilde{g}_0))^2$, i.e., $\omega_l^* = \frac{\rho_l}{(M + \beta(M, \tilde{g}_0))^2} - \lambda^*$. Since $\lambda^* \leq \frac{\rho_l}{(M + \beta(M, \tilde{g}_0))^2}$, $\omega_l^* \geq 0$ is feasible.
- If $\sqrt{\frac{\rho_l}{\lambda^* + \omega_l^*}} - \beta(M, \tilde{g}_0) < M$, then $\omega_l^* = 0$. We then have $\sqrt{\frac{\rho_l}{\lambda^*}} - \beta(M, \tilde{g}_0) < M$, which indicates $\lambda^* > \frac{\rho_l}{(M + \beta(M, \tilde{g}_0))^2}$, $\tilde{q}_l^* = \sqrt{\frac{\rho_l}{\lambda^*}} - \beta(M, \tilde{g}_0)$.
- If $\lambda^* \geq \frac{\rho_l}{(1+\beta(M, \tilde{g}_0))^2} - \omega_l^*$, then $\tilde{q}_l^* > 1$ is impossible, because it would imply $\lambda^* \geq \frac{\rho_l}{(1+\beta(M, \tilde{g}_0))^2} - \omega_l^* > \frac{\rho_l}{(\tilde{q}_l^* + \beta(M, \tilde{g}_0))^2} - \omega_l^*$, which violates the complementary slackness condition. Therefore, we have $\tilde{q}_l^* = 1$.

In summary, the optimal solution is given by

$$\tilde{q}_l^* = \begin{cases} M, & \lambda^* \leq \frac{\rho_l}{(M + \beta(M, \tilde{g}_0))^2}, \\ \sqrt{\frac{\rho_l}{\lambda^*}} - \beta(M, \tilde{g}_0), & \frac{\rho_l}{(M + \beta(M, \tilde{g}_0))^2} < \lambda^* < \frac{\rho_l}{(1 + \beta(M, \tilde{g}_0))^2}, \\ 1, & \lambda^* \geq \frac{\rho_l}{(1 + \beta(M, \tilde{g}_0))^2}, \end{cases} \quad (76)$$

which is equivalent to

$$\tilde{q}_l^* = \min \left\{ M, \max \left\{ 1, \sqrt{\rho_l / \lambda^*} - \beta(M, \tilde{g}_0) \right\} \right\}. \quad (77)$$

Substituting this expression for \tilde{q}_l^* into $\sum_{l=1}^{L'} \tilde{q}_l^* = MB_C$, we obtain

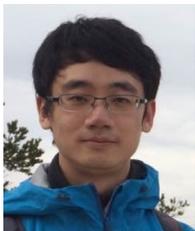
$$\sum_{l=1}^{L'} \min \left\{ M, \max \left\{ 1, \sqrt{\rho_l / \lambda^*} - \beta(M, \tilde{g}_0) \right\} \right\} = MB_C. \quad (78)$$

Note that since $\rho_l \geq \rho_{l+1}$ for $l = 1, \dots, L' - 1$, \tilde{q}_l^* satisfies $\tilde{q}_l^* \geq \tilde{q}_{l+1}^*$ for $l = 1, \dots, L' - 1$. As a result, $\tilde{\mathbf{q}}$ is also an optimal solution of $\tilde{\mathcal{P}}(M, L')$.

REFERENCES

- [1] M. Paolini, "Crucial economics for mobile data backhaul," *Senza Fili Consulting*, 2011.
- [2] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. Leung, "Cache in the air: Exploiting content caching and delivery techniques for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 131–139, 2014.
- [3] W. Han, A. Liu, and V. K. Lau, "PHY-caching in 5G wireless networks: Design and analysis," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 30–36, 2016.
- [4] —, "Degrees of freedom in cached MIMO relay networks," *IEEE Trans. Signal Process.*, vol. 63, no. 15, pp. 3986–3997, 2015.
- [5] A. Liu and V. K. Lau, "Cache-enabled opportunistic cooperative MIMO for video streaming in wireless systems," *IEEE Trans. Signal Process.*, vol. 62, no. 2, pp. 390–402, 2014.
- [6] E. Baştug, M. Bennis, M. Kountouris, and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs," *EURASIP J. Wireless Commun. Networking*, vol. 2015, no. 1, pp. 1–11, 2015.
- [7] C. Yang, Y. Yao, Z. Chen, and B. Xia, "Analysis on cache-enabled wireless heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 131–145, 2016.
- [8] D. Liu and C. Yang, "Cache-enabled heterogeneous cellular networks: Comparison and tradeoffs," in *Proc. IEEE Int. Conf. Commun. (ICC)*, pp. 1–6.
- [9] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," in *Proc. IEEE Int. Conf. Communications (ICC)*, pp. 3358–3363.

- [10] S. Tamoor-ul Hassan, M. Bennis, P. H. Nardelli, and M. Latva-Aho, "Modeling and analysis of content caching in wireless small cell networks," in *2015 Int. Symp. Wireless Commun. Sys. (ISWCS)*. IEEE, pp. 765–769.
- [11] Y. Cui and D. Jiang, "Analysis and optimization of caching and multicasting in large-scale cache-enabled heterogeneous wireless networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 1, pp. 250–264, 2017.
- [12] V. Bioglio, F. Gabry, and I. Land, "Optimizing MDS codes for caching at the edge," in *Proc. IEEE GLOBECOM*, 2015, pp. 1–6.
- [13] J. Liao, K.-K. Wong, M. R. Khandaker, and Z. Zheng, "Optimizing cache placement for heterogeneous small cell networks," *IEEE Commun. Lett.*, vol. 21, no. 1, pp. 120–123, 2017.
- [14] X. Xu and M. Tao, "Modeling, analysis, and optimization of coded caching in small-cell networks," *IEEE Trans. Commun.*, 2017.
- [15] I. Katzela and M. Naghshineh, "Channel assignment schemes for cellular mobile telecommunication systems: A comprehensive survey," *IEEE Personal Commun.*, vol. 3, no. 3, pp. 10–31, 1996.
- [16] G. Miao, J. Zander, K. W. Sung, and S. B. Slimane, *Fundamentals of Mobile Data Networks*. Cambridge Univ. Press, 2016.
- [17] M. Haenggi, R. K. Ganti *et al.*, "Interference in large wireless networks," *Foundations and Trends in Networking*, vol. 3, no. 2, pp. 127–248, 2009.
- [18] M. Ji, G. Caire, and A. F. Molisch, "The throughput-outage tradeoff of wireless one-hop caching networks," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6833–6859, 2015.
- [19] S. M. Yu and S.-L. Kim, "Downlink capacity and base station density in cellular networks," in *11th Int. Symp. on Modeling & Optimization in Mobile, Ad Hoc & Wireless Networks (WiOpt)*. IEEE, 2013, pp. 119–124.
- [20] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge Univ. Press, 2004.
- [21] P. Olivier and A. Simonian, "Performance of a cache with random replacement and Zipf document popularity," in *Proc. of the 7th Int. Conf. Performance Evaluation Methodologies and Tools*, 2013, pp. 233–242.
- [22] T. Yamakami, "A Zipf-like distribution of popularity and hits in the mobile web pages with short life time," in *Proc. Parallel Distributed Computing, Applications and Technologies*. IEEE, 2006, pp. 240–243.
- [23] S. Singh and J. G. Andrews, "Joint resource partitioning and offloading in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, pp. 888–901, 2014.

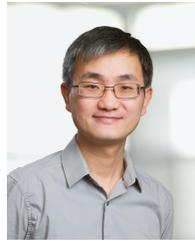


Wei Han (S'12) received the B.S. from Tsinghua University (2007-2011) and Ph.D. from the Hong Kong University of Science and Technology (HKUST) (2012-2018). He is currently a researcher with Future Network Theory Lab, 2012 Labs, Huawei Tech. Investment Co., Ltd.. His research interests include wireless caching and compressive sensing.



An Liu (S'07-M'09-SM'17) received the Ph.D. and the B.S. degree in Electrical Engineering from Peking University, China, in 2011 and 2004 respectively. From 2008 to 2010, he was a visiting scholar at the Department of ECEE, University of Colorado at Boulder. He has been a Postdoctoral Research Fellow in 2011-2013, Visiting Assistant Professor in 2014, and Research Assistant Professor in 2015-2017, with the Department of ECE, HKUST. He is currently a Distinguished Research Fellow with the College of Information Science and Electronic

Engineering, Zhejiang University. His research interests include wireless communications, stochastic optimization and compressive sensing.



Wei Yu (S'97-M'02-SM'08-F'14) received the B.A.Sc. degree in Computer Engineering and Mathematics from the University of Waterloo, Waterloo, Ontario, Canada in 1997 and M.S. and Ph.D. degrees in Electrical Engineering from Stanford University, Stanford, CA, in 1998 and 2002, respectively. Since 2002, he has been with the Electrical and Computer Engineering Department at the University of Toronto, Toronto, Ontario, Canada, where he is now Professor and holds a Canada Research Chair (Tier 1) in Information Theory and Wireless Communica-

tions. His main research interests include information theory, optimization, wireless communications and broadband access networks.

Prof. Wei Yu currently serves on the IEEE Information Theory Society Board of Governors (2015-20). He was an IEEE Communications Society Distinguished Lecturer (2015-16). He is currently an Area Editor for the IEEE Transactions on Wireless Communications (2017-20). He served as an Associate Editor for IEEE Transactions on Information Theory (2010-2013), as an Editor for IEEE Transactions on Communications (2009-2011), and as an Editor for IEEE Transactions on Wireless Communications (2004-2007). He is currently the Chair of the Signal Processing for Communications and Networking Technical Committee of the IEEE Signal Processing Society (2017-18) and served as a member in 2008-2013. Prof. Wei Yu received the Steacie Memorial Fellowship in 2015, the IEEE Signal Processing Society Best Paper Award in 2017 and 2008, an Journal of Communications and Networks Best Paper Award in 2017, an IEEE Communications Society Best Tutorial Paper Award in 2015, an IEEE ICC Best Paper Award in 2013, the McCharles Prize for Early Career Research Distinction in 2008, the Early Career Teaching Award from the Faculty of Applied Science and Engineering, University of Toronto in 2007, and an Early Researcher Award from Ontario in 2006. Prof. Wei Yu is a Fellow of the Canadian Academy of Engineering, and a member of the College of New Scholars, Artists and Scientists of the Royal Society of Canada. He is recognized as a Highly Cited Researcher.



Vincent K. N. Lau (SM'04-F'12) obtained B.Eng (Distinction 1st Hons) from the University of Hong Kong (1989-1992) and Ph.D. from the Cambridge University (1995-1997). He joined Bell Labs from 1997-2004 and the Department of ECE, Hong Kong University of Science and Technology (HKUST) in 2004. He is currently a Chair Professor and the Founding Director of Huawei-HKUST Joint Innovation Lab at HKUST. His current research focus includes robust and delay-optimal cross layer optimization for MIMO/OFDM wireless systems,

interference mitigation techniques for wireless networks, massive MIMO, M2M and network control systems.