Probabilistic Small-Cell Caching: Performance Analysis and Optimization

Youjia Chen, Ming Ding, *Member, IEEE*, Jun Li, *Senior Member, IEEE*, Zihuai Lin, *Senior Member, IEEE*, Guoqiang Mao, *Senior Member, IEEE*, and Lajos Hanzo, *Fellow, IEEE*

***Abstract*—Small-cell caching utilizes the embedded storage of small-cell base stations (SBSs) to store popular contents for the sake of reducing duplicated content transmissions in networks and for offloading the data traffic from macrocell base stations to SBSs. In this paper, we study a probabilistic small-cell caching strategy, where each SBS caches a subset of contents with a specific caching probability. We consider two kinds of network architectures: 1) The SBSs are always active, which is referred to as the always-on architecture; and 2) the SBSs are activated on demand by mobile users (MUs), which is referred to as the dynamic on–off architecture. We focus our attention on the probability that MUs can successfully download content from the storage of SBSs. First, we derive theoretical results of this successful download probability (SDP) using stochastic geometry theory. Then, we investigate the impact of the SBS parameters, such as the transmission power and deployment intensity on the SDP. Furthermore, we optimize the caching probabilities by maximizing the SDP based on our stochastic geometry analysis. The intrinsic amalgamation of optimization theory and stochastic geometry based analysis leads to our optimal caching strategy, characterized by the resultant closed-form expressions. Our results show that in the always-on architecture, the optimal caching probabilities solely depend on the content request probabilities, while in the dynamic on–off architecture, they also relate to the MU-to-SBS intensity ratio. Interestingly, in both architectures, the optimal caching probabilities are linear functions of the square root of the content request probabilities. Monte-Carlo simulations validate our theoretical analysis and show that the proposed schemes relying on the optimal caching probabilities are capable of achieving substantial SDP improvement, compared with the benchmark schemes.**

***Index Terms*—.**

Manuscript received February 11, 2016; revised June 25, 2016; accepted August 24, 2016. Date of publication; date of current version. This work was supported in part by the Fujian Provincial Natural Science Foundation (2016J01290); in part by the National Natural Science Foundation of China under Grant 61571128 and Grant 61501238; in part by the Jiangsu Provincial Science Foundation (BK20150786); in part by the China Scholarship Council, the Specially Appointed Professor Program in Jiangsu Province, 2015; and in part by the Fundamental Research Funds for the Central Universities (30916011205). The review of this paper was coordinated by Dr. B. Canberk.

Y. Chen is with the School of Electrical and Information Engineering, University of Sydney, Sydney, NSW 2006, Australia, also with Fujian Normal University, Fuzhou 350007, China, and also with the Data61, CSIRO, Canberra ACT 2600, Australia (e-mail: youjia.chen@sydney.edu.au).

M. Ding is with the Data61, CSIRO Canberra ACT, 2600, Australia (e-mail: Ming.Ding@data61.csiro.au).

J. Li is with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Beijing 100044, China (e-mail: jun.li@njust.edu.cn).

Z. Lin is with the School of Electrical and Information Engineering, University of Sydney, Sydney, NSW 2006, Australia, and also with the Data61, CSIRO Canberra ACT, 2600, Australia (e-mail: zihuai.lin@sydney.edu.au).

G. Mao is with the School of Computing and Communications, University of Technology Sydney, Ultimo, NSW 2007, Australia, and also with the Data61, CSIRO Canberra ACT, 2600, Australia (e-mail: g.mao@ieee.org).

L. Hanzo is with the Department of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K. (e-mail: lh@ecs.soton.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/TVT.2016.2606765

I. Introduction

It is forecast that at least a 100x network capacity increase will be required to meet the traffic demands in 2020 [1]. As a result, vendors and operators are now looking at using every tool at hand to improve network capacity [2].

In addition, a substantial contribution to the traffic explosion comes from the repeated download of a small portion of popular contents, such as popular movies and videos [3]. Therefore, intelligent caching in wireless networks has been proposed for effectively reducing such duplicated transmissions of popular contents, as well as for offloading the traffic from the overwhelmed macrocells to small cells [4], [5]. Caching in third-generation (3G) and fourth-generation (4G) wireless networks was shown to be able to reduce the traffic by one third to two thirds [6].

Several caching strategies have been proposed for wireless networks. Woo *et al.* [7] analyzed the strategy of caching contents in the evolved packet core of local thermal equilibrium (LTE) networks.The strategy of caching contents in the radio access network, with an aim to place contents closer to mobile users (MUs) was studied in [8] and [9]. The concept of small-cell caching, referred to as "Femtocaching" in [9] and [10], utilized small-cell base stations (SBS) in heterogeneous cellular networks as distributed caching devices. Caching strategies conceived for device-to-device (D2D) networks were investigated in [11]–[13], where the mobile terminals serve as caching devices. The coexistence of small-cell caching and D2D caching is indeed also a hot research direction. In [14], Yang *et al.* considered the joint caching in both the relays and a subset of the mobile terminals, which relies on the coexistence of small-cell caching and D2D caching. Moreover, a coded caching scheme was proposed in [15] to improve system performance.

In this paper, we focus on the small-cell caching because 1) the large number of SBSs in 4G and fifth-generation (5G) networks already provide a promising basis for caching [2]; and 2) compared with D2D caching, small-cell caching has several advantages, such as the abundance of power supply, fewer grave security issues, and more reliable data delivery. As illustrated in Fig. 1, with small-cell caching, popular contents are transmitted and cached in the storage of the SBSs during off-peak hours. Then in peak hours, if an MU can find its requested content in

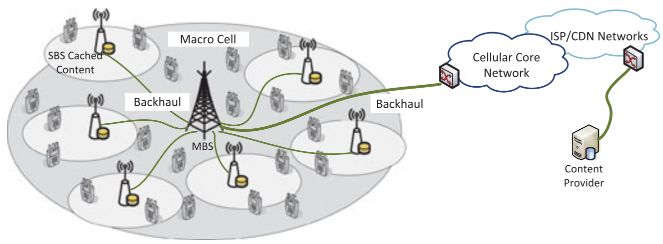


Fig. 1. Small-cell caching.

a nearby SBS, the MU can directly download the content from such SBS.

There are generally two approaches to implement the small-cell caching, i.e., the deterministic content placement and the nondeterministic content placement. In [9], [16], and [17], the deterministic contents placement was analyzed. In these works, the placement of popular contents was optimized using the information of the network node locations and the statistical or instantaneous channel states. However, in practice, the geographic distribution of MUs and the wireless channels are time variant. Thus, the optimal content placement strategy has to be frequently updated in the deterministic content placement, leading to a high complexity and fewer tractable results. On the other hand, the nondeterministic content placement permits simple implementation and has a good tractability. In [18] and [14], the distributions of SBSs and MUs were modeled as homogeneous Poisson point processes (HPPPs) to obtain a general performance analysis for the small-cell caching. However, in these works, all the SBSs were assumed to cache the same copy of certain popular contents. In [11], probabilistic content placement was proposed and analyzed in the context of D2D caching, where each mobile terminal caches a specific subset of the contents with a given caching probability. The throughput versus outage tradeoff was analyzed and the optimal caching distribution was derived for a grid network relying on a particular protocol model. The idea of probabilistic content placement was also investigated in the coded multicasting system [19]. Compared with caching the same copy of certain popular contents in all the SBSs, probabilistic content placement in small-cell caching can provide more flexibility. Therefore, in this paper, we focus on small-cell caching relying on probabilistic content placement, shortened as probabilistic small-cell caching (PSC) for brevity.

In small-cell networks, there are two network architectures, namely, the always-on architecture and the dynamic on–off architecture. The always-on architecture is a common practice in the current cellular networks, where all the SBSs are always active. By contrast, in the dynamic on–off architecture, the SBSs are only active, when they are required to provide services to nearby MUs [20]. Aiming for saving energy consumption and mitigating unnecessary intercell interference, the dynamic on–off architecture has been proposed and it is currently under investigation in 3GPP as an important candidate of 5G technologies in future dense and ultradense small-cell networks [2], [21], [22]. Energy consumption is of critical interest in future 5G systems [23], [24], especially in ultradense networks. Compared with the power-thirsty always-on architecture, where the energy consumption grows with the network's densification, the energy consumption of the ultradense network relying on the dynamic on–off architecture mainly depends on the density of MUs in the network [2]. The in-depth investigation of the associated energy consumption issues of wireless caching will constitute our future work.

Against this background, we study the PSC under the above-mentioned pair of network architectures. First, we use a stochastic geometry to develop theoretical results of the probability $\Pr(\mathcal{D})$ that MUs can successfully download contents from the storage of SBSs. Second, we investigate the impact of the SBSs' parameters on $\Pr(\mathcal{D})$, namely, that of the transmission power P and of the deployment intensity λ_s. In the always-on architecture, although $\Pr(\mathcal{D})$ monotonically increases with either P or λ_s, it approaches a constant when P or λ_s is sufficiently high. In the dynamic on–off architecture, $\Pr(\mathcal{D})$ reaches a constant when P is high enough, while it keeps on increasing as λ_s grows. Most importantly, we optimize the caching probabilities for maximizing $\Pr(\mathcal{D})$ in the pair of network architectures considered. We emphasize that it is quite a challenge to apply optimization theory to an objective function obtained from stochastic geometry analysis, especially to derive a closed-form expression for the optimal solution. Our results will demonstrate that in the always-on architecture, the optimal subset of contents to be cached depends on the content request probabilities, while in the dynamic on–off architecture, it also depends on the MU-to-SBS intensity ratio. Most interestingly, in both architectures, the optimal caching probabilities can be expressed as linear functions of the square root of the content request probabilities.

The rest of the paper is structured as follows. In Section II we describe the system model, while in Section III we present the definition of PSC and formulate the probability that MUs can successfully download contents from the storage of SBSs. The main analytical results characterizing this successful download probability (SDP) are presented in Section IV. In Section V, we optimize the caching probabilities in both of the network architectures for maximizing the derived SDP. The accuracy of the analytical results and the performance gains of optimization are characterized by simulations in Section VI. Finally, our conclusions are offered in Section VII.

II. System Model

We consider a cellular network supporting multiple MUs by the SBSs operating within the same frequency spectrum. We model the distribution of the SBSs and that of the MUs as two independent HPPPs, with the intensities of λ_s and λ_u, respectively. The transmission power of the SBSs is denoted by P. The path loss of the channel spanning from an SBS to an MU is modeled as $d^{-\alpha}$, where d denotes the distance between them, and α denotes the path-loss exponent. The multipath fading is modeled as Rayleigh fading with a unit power, and hence the channel's power gain is denoted by $h \sim \exp(1)$. All the channels are assumed to be independently and identically distributed.

A. Network Architectures

We consider two network architectures.

1) Always-On Architecture: In this architecture, all the SBSs are assumed to be active, i.e., all the SBSs are continuously transmitting signals. This architecture is commonly employed in the operational cellular networks [25]. The rationale for this architecture is that the number of SBSs is usually much lower than that of MUs, and thus each and every SBS has to be turned ON to serve the MUs in its coverage.

2) Dynamic On–Off Architecture: In this architecture, an SBS will be active only when it has to provide services to its associated MUs. In future 5G networks, the intensity of deployed SBSs is expected to be comparable to or even potentially higher than the intensity of MUs [2]. In such ultradense networks, having an adequate received signal coverage is always guaranteed, since the distance between an MU and its serving SBS is short, but the interference becomes the dominant issue. With the goal of mitigating the potentially avoidable intercell interference and saving energy, the dynamic on–off architecture has been identified as one of the key technologies in 5G networks [20]. With the dynamic on–off architecture, an SBS will switch to its idle mode, i.e., turn OFF its radio transmission, if there is no MU associated with it, otherwise, it will switch back to the active mode.

B. File Request Model

We consider a contents library consisting of M different files. Note that M does not represent the number of files available on the Internet, but the number of popular files that the MUs tend to access. We denote by q_m the probability that the mth file $\mathcal{F}_m$ will be requested. By stacking q_m into $\{q_m : m = 1, \cdots, M\}$, we can get the probability mass function (PMF) of requesting the M files. According to [26], the request- PMF of the files can be modeled as a Zipf distribution. More specifically, for $\mathcal{F}_m$, its request probability q_m is written as

$$q_m = \frac{\frac{1}{m^\beta}}{\sum_{i=1}^{M} \frac{1}{i^\beta}} \tag{1}$$

where β is the exponent of the Zipf distribution and a large β implies having an uneven popularity among those files. From (1), q_m tends to zero, as $M \to \infty$ when $\beta < 1$, while it converges to a constant value when $\beta > 1$. Note that (1) implies that the indices of the files are not randomly generated, but follow a descending order of their request probabilities.

Due to the limited storage of SBSs, an SBS is typically unable to cache the entire file library. Therefore, we assume that the library is partitioned into N nonoverlapping subsets of files, referred to as file groups (FGs), and each SBS can cache only one of the N FGs. Note that the same FG can be redundantly stored in multiple SBSs. The scenario of FGs with overlapping subsets of files will be considered later, which will be compared with the nonoverlapping scenario. We denote the nth FG, $n \in \{1, \cdots, N\}$ by $\mathcal{G}_n$. The probability Q_n that an MU requests a file in FG $\mathcal{G}_n$, is thus given by

$$Q_n = \sum_{m,\ \mathrm{for}\,\mathcal{F}_\mathrm{m} \in \mathcal{G}_\mathrm{n}} q_m. \tag{2}$$

III. Probabilistic Small-Cell Caching Strategy

In this section, we introduce the PSC strategy, and formulate the probability that MUs can successfully download contents from the storage of the SBSs, which is an important performance metric of small-cell caching.

Generally, caching consists of two phases: a contents placement phase and a contents delivery phase [27]. In the contents placement phase, popular contents are transmitted and cached in the storage units of network devices that are close to MUs. In the contents delivery phase, the popular cached contents can be promptly retrieved for serving the MUs.

A. Contents Placement Phase

In the content placement phase of PSC, each SBS independently caches FG $\mathcal{G}_n$ with a specific caching probability, denoted by S_n. Hence, from the perspective of the entire network, the fraction of the SBSs that caches $\mathcal{G}_n$ equals to S_n. Since the distribution of SBSs in the network is modeled as an HPPP with the intensity of λ_s, according to the thinning theorem of HPPP [28], we can view the distribution of SBSs that cache $\mathcal{G}_n$ as a thinned HPPP with the intensity of $S_n\lambda_s$.

We assume that at a particular time instant, an MU can only request one file, and hence, the distribution of MUs who request the files in $\mathcal{G}_n$ can also be modeled as a thinned HPPP with the intensity $Q_n\lambda_u$. We treat the SBSs that cache $\mathcal{G}_n$ together with the MUs that request the files in $\mathcal{G}_n$ as the nth tier of the network, shortened as Tier-n.

B. Contents Delivery Phase

During the contents delivery phase, an MU that requests a file in $\mathcal{G}_n$ will associate with the nearest SBS that caches $\mathcal{G}_n$, and then attempts to download the file from it. We assume that only when the received signal-to-interference-and-noise-ratio (SINR) at the MU is above a prescribed threshold, can the requested file be successfully downloaded.

If the MU cannot download the requested file from the cached SBS, the requested file would be transmitted to the MU from a remote content provider, which means the data should flow across the Internet, the cellular core network, and the backhaul network, as illustrated in Fig. 1.

C. Probability of Successful Download

Recent surveys show that 96% of the operators consider backhaul as one of the most important challenges to small-cell deployments, and this issue is exacerbated in ultradense networks [29], [30]. If an MU can successfully download a requested file from storages of SBSs, the usage of the backhaul network will be greatly reduced and the transmission latency of a requested file will be significantly shortened. Therefore, we assume that a successful download of a requested file from storages of SBSs is always beneficial to the network performance. Accordingly, we focus on our attention on this SDP as the performance metric for small-cell caching in the following.

According to Slyvnyak's theorem for HPPP [28], an existing point in the process does not change the statistical distribution of other points of the HPPP. Therefore, the probability that an MU in Tier-n can successfully download the contents from SBSs can be obtained by analyzing the probability that a *typical* MU in Tier-n, say located at the origin, can successfully download the contents from its associated SBS in Tier-n.

When the MU considered requests a file in $\mathcal{G}_n$, its received SINR from its nearest SBS in Tier-n can be formulated as

$$\gamma_n(z) = \frac{P h_{x_0} z^{-\alpha}}{\sum_{x_j \in \Phi \backslash \{x_0\}} P h_{x_j} \|x_j\|^{-\alpha} + \sigma^2} \tag{3}$$

where σ^2 denotes the Gaussian noise power, z is the distance between the typical MU and its nearest SBS in Tier-n, x_j represents the locations of the interfering SBSs, Φ denotes the set of simultaneously active SBSs, and x_0 is the location of the serving BS at a distance of z. Additionally, $\|x_j\|$ denotes the distance between x_j and the typical MU, while h_{x_0} and h_{x_j} denote the corresponding channel gains.

Since the intercell interference is the dominant factor determining the signal quality in the operational cellular networks, especially when unity frequency reuse has been adopted for improving the spectrum efficiency, the minimum received SINR is used as the metric of successful reception. Let δ be the minimum SINR required for successful transmissions and $\mathcal{D}_n$ be the event that the typical Tier-n MU successfully receives the requested file from the associated Tier-n SBS. Then, the probability of $\mathcal{D}_n$ can be formulated as

$$\Pr(\mathcal{D}_n) = \Pr[\gamma_n(z) \geq \delta]. \tag{4}$$

Considering the request probabilities of $\mathcal{G}_n$ and based on the result of $\Pr(\mathcal{D}_n)$, we obtain the average probability that the MUs can successfully download contents from the storage of the SBSs, denoted by $\Pr(\mathcal{D})$, as

$$\Pr(\mathcal{D}) = \sum_{n=1}^{N} Q_n \cdot \Pr(\mathcal{D}_n). \tag{5}$$

In essence, $\Pr(\mathcal{D})$ quantifies the weighted sum of the SDP, where the weights are the request probabilities reflecting the importance of the files.

IV. Performance Analysis of Small-Cell Caching

In this section, we derive the SDP $\Pr(\mathcal{D})$ for the pair of network architectures. Some special cases are also considered with an aim to obtain more insights into the design of PSC.

A. Always-On Architecture

Our main result on the probability $\Pr(\mathcal{D})$ for the always-on architecture is summarized in Theorem 1.

Theorem 1: In the always-on architecture, the probability $\Pr(\mathcal{D})$ is given by

$$\begin{aligned}\Pr(\mathcal{D}) &= \sum_{n=1}^{N} Q_n \Pr(\mathcal{D}_n)\\ &= \sum_{n=1}^{N} Q_n \int_0^{\infty} \pi S_n \lambda_s \exp\left(-\frac{z^{\alpha}\delta\sigma^2}{P}\right)\\ &\exp\left(-\pi\lambda_s z^2\left((1-S_n)C(\delta,\alpha) + S_n A(\delta,\alpha) + S_n\right)\right) dz^2\end{aligned} \tag{6}$$

where $A(\delta,\alpha) \triangleq \delta \frac{2}{\alpha-2}\, {}_2F_1(1, 1-\frac{2}{\alpha}; 2-\frac{2}{\alpha}; -\delta)$, and $C(\delta,\alpha) \triangleq \frac{2}{\alpha}\delta^{\frac{2}{\alpha}} B(\frac{2}{\alpha}, 1-\frac{2}{\alpha})$. Furthermore, ${}_2F_1(\cdot)$ denotes the hypergeometric function, and $B(\cdot)$ represents the beta function [31].

Proof: See Appendix A. ■

From (6), we conclude that the probability $\Pr(\mathcal{D})$ increases as the transmission power P grows, because $\exp(-\frac{z^{\alpha}\delta\sigma^2}{P})$ increases with P. Since it remains a challenge to obtain deeper insights from (6), which is not a closed-form expression, two special cases are examined in the sequel to gain deeper insight on the performance behavior of $\Pr(\mathcal{D})$.

1) Path-Loss Exponent $\alpha = 4$: According to 3GPP measurement [32], the typical value of the path-loss exponent for SBSs in practical environments is around 4. Substituting this typical value of $\alpha = 4$ into (6), we have

$$\begin{aligned}\Pr(\mathcal{D})\mid_{\alpha=4} &= \sum_{n=1}^{N} Q_n \pi S_n \sqrt{\frac{\pi}{4\delta}\frac{P\lambda_s^2}{\sigma^2}}\,\mathrm{erfc}x\left(\frac{\pi}{2}\cdot\right.\\ &\left.\sqrt{\frac{P\lambda_s^2}{\delta\sigma^2}}\left(S_n + \frac{\pi}{2}\sqrt{\delta}(1-S_n) + S_n\sqrt{\delta}\arctan\sqrt{\delta}\right)\right)\end{aligned} \tag{7}$$

where $\mathrm{erfc}x(x) \triangleq \exp(x^2)\mathrm{erfc}(x)$ is the scaled complementary error function [33].

Regarding the relationship between $\Pr(\mathcal{D})$ and λ_s, we propose Corollary 1.

Corollary 1: In the always-on architecture, for the special case of $\alpha = 4$, $\Pr(\mathcal{D})$ monotonically increases with the increase of λ_s.

Proof: See Appendix B. ■

From the results obtained in (6) that $\Pr(D)$ increases as P grows, and based on Corollary 1, we conclude that when $\alpha = 4$, the SDP $\Pr(\mathcal{D})$ can be improved by either increasing the SBSs' transmission power P or the SBSs' deployment intensity λ_s. Furthermore, since (7) can be viewed as a function of the variable $P\lambda_s^2$, the effect of increasing P to kP on $\Pr(\mathcal{D})$ is equivalent to increasing λ_s to $\sqrt{k}\lambda_s$, where k is a positive constant.

Moreover, according to the property of the function $\mathrm{erfc}x(x)$, i.e., $\lim_{x\to\infty}\mathrm{erfc}x(x) = \frac{1}{\sqrt{\pi}x}$, we have

$$\begin{aligned}&\lim_{P\to\infty}\Pr(\mathcal{D})\mid_{\alpha=4} = \lim_{\lambda_s\to\infty}\Pr(\mathcal{D})\mid_{\alpha=4}\\ &= \sum_{n=1}^{N}\frac{Q_n S_n}{\frac{\pi}{2}\sqrt{\delta} + (\sqrt{\delta}\arctan\sqrt{\delta} + 1 - \frac{\pi}{2}\sqrt{\delta})S_n}.\end{aligned} \tag{8}$$

From (8), we have Remark 1.

Remark 1: In the always-on architecture, given σ^2 and δ, the value of $\Pr(\mathcal{D})$ monotonically grows with the increase of P and λ_s, and it converges to a constant, when P or λ_s is sufficiently large.

2) Neglecting Noise, i.e., $\sigma^2 = 0$: In an interference-limited network, where the noise level is much lower than the interference, the impact of the noise can be neglected. In such cases, we assume that $\sigma^2 = 0$, and it follows that $\Pr(\mathcal{D})$ in (6) can be rewritten as

$$\Pr(\mathcal{D})\mid_{\sigma^2\to 0} = \sum_{n=1}^{N}\frac{Q_n S_n}{S_n A(\delta,\alpha) + (1-S_n)C(\delta,\alpha) + S_n}. \tag{9}$$

From (9), we have Remark 2.

Remark 2: In the always-on architecture operating in an interference-limited network, the probability of successful download depends only on the request probabilities and caching probabilities of the FGs, i.e., Q_n and S_n.

Note that in the scenario, where the different FGs may have an overlapping subset of files, the probability $\Pr(\mathcal{D})$ still has the same formulation as (6). However, all the subscripts n in (6) should be changed to m, because we should consider both the request probability and the caching probability of each file $\mathcal{F}_m$, i.e., S_m and Q_m, instead of each FG $\mathcal{G}_n$. Therefore, in this scenario, the specific SBSs that cache $\mathcal{F}_m$ and the MUs that request $\mathcal{F}_m$ are viewed as Tier-m. Since all the derivations are the same, our main results summarized in Theorem 1 as well as the aforementioned corollary and remarks, are still valid in conjunction with the subscript m. Hence we omit the analysis for this scenario with overlapping subsets of files for brevity.

B. Dynamic On–Off Architecture

As mentioned, in the dynamic on–off architecture an SBS is only active, when it has to provide services for the associated MUs. Specifically, an SBS in Tier-n is only active, when there is at least one MU in Tier-n located in its Voronoi cell. Hence, the probability that an SBS in Tier-n is active, which is denoted by $\Pr(\mathcal{A}_n)$, should be considered for the dynamic on–off architecture.

Our main result on the probability $\Pr(\mathcal{D})$ for the dynamic on–off architecture is summarized in Theorem 2.

Theorem 2: In the dynamic on–off architecture, the probability $\Pr(\mathcal{D})$ is given by

$$\Pr(\mathcal{D}) = \sum_{n=1}^{N} Q_n \Pr(\mathcal{D}_n) = \sum_{n=1}^{N} Q_n \int_0^\infty \pi S_n \lambda_s \exp\left(-\frac{z^\alpha \delta \sigma^2}{P}\right) \exp\Bigg(-\pi\lambda_s z^2 \Bigg(\sum_{i=1, i\neq n}^{N} \Pr(\mathcal{A}_i) S_i C(\delta,\alpha) + \Pr(\mathcal{A}_n) S_n A(\delta,\alpha) + S_n\Bigg)\Bigg) dz^2 \quad (10)$$

where $\Pr(\mathcal{A}_n)$ denotes the probability that an SBS in Tier-n is in the active mode, and

$$\Pr(\mathcal{A}_n) \approx 1 - \left(1 + \frac{Q_n \lambda_u}{3.5 S_n \lambda_s}\right)^{-3.5}. \quad (11)$$

Proof: See Appendix C. ■

Compared to $\Pr(\mathcal{D})$ in the always-on architecture, $\Pr(\mathcal{D})$ in the dynamic on–off architecture also depends on the intensity of the MUs λ_u. The reason behind this is that the number of active SBSs in the network depends on the number of MUs in the network.

From (10), we have Remark 3.

Remark 3: In the dynamic on–off architecture, given σ^2 and δ, the value of $\Pr(\mathcal{D})$ monotonically increases with the increase of the transmission power P.

1) Neglecting Noise, i.e., $\sigma^2 = 0$: In an interference-limited network, substituting $\sigma^2 = 0$ into (10), we have

$$\Pr(\mathcal{D})\,|_{\sigma^2 \to 0} = \sum_{n=1}^{N} \frac{Q_n S_n}{\Pr(\mathcal{A}_n) S_n A(\delta,\alpha) + \sum_{i=1, i\neq n}^{N} \Pr(\mathcal{A}_i) S_i C(\delta,\alpha) + S_n}. \quad (12)$$

From (12), we have Remark 4.

Remark 4: In the dynamic on–off architecture operating in an interference-limited network, the probability of successful download $\Pr(\mathcal{D})$ is independent of P, and depends only on Q_n, S_n as well as on the MU-to-SBS intensity ratio λ_u/λ_s.

When considering the scenario of FGs with overlapping subsets of files, the average probability $\Pr(\mathcal{D})$ cannot be formulated as the sum of $\Pr(\mathcal{D}_n)$ as in (5). Furthermore, we cannot formulate $\Pr(\mathcal{D})$ as $\Pr(\mathcal{D}) = \sum_{m=1}^{M} \Pr(\mathcal{D}_m)$, which we propose for the overlapping scenario in the always-on architecture. This is because in the dynamic on–off architecture the active probability of an SBS depends on the specific FG that it caches. Therefore, the analysis of $\Pr(\mathcal{D})$ in the dynamic on–off architecture considering the scenario with overlapping subsets of files requires further investigations as part of our future research.

V. Optimization of the Caching Probability

A larger $\Pr(\mathcal{D})$ always benefits the network because of 1) the backhaul saving and 2) the low-latency transmission of local contents from SBSs [2]. Based on such facts, in this section, we concentrate on maximizing $\Pr(\mathcal{D})$ by optimally designing the caching probabilities of the contents in the system, denoted by $\{S_n^{\mathrm{Opt}} : n = 1, \ldots, N\}$.

Note that there is a paucity of literature on applying optimization theory relying on an objective function obtained from stochastic geometry analysis, especially, when aiming for deriving a closed-form expression of the optimal solution. In order to facilitate this optimization procedure, we ensure the mathematical tractability of the objective function by using a simple user association strategy and neglect the deleterious effects of noise.

A. Always-On Architecture

From (9), we can formulate the optimization problem of maximizing $\Pr(\mathcal{D})$ as

$$\begin{aligned} \max_{\{S_n\}} \ \Pr(\mathcal{D}) &= \max_{\{S_n\}} \sum_{n=1}^{N} \frac{Q_n S_n}{(1-S_n) C(\delta,\alpha) + S_n A(\delta,\alpha) + S_n} \\ \text{s.t.} \sum_{n=1}^{N} S_n &= 1 \\ S_n &\geq 0, \ n = 1, \ldots, N. \end{aligned} \quad (13)$$

The solution of Problem (13) is presented in Theorem 3.

Theorem 3: In the always-on architecture, the optimal caching scheme, which is denoted by the file caching PMF $\{S_n^{opt}\}$, that maximizes the average probability of successful download, is given by

$$S_n^{opt} = \left\lceil \frac{\sqrt{\frac{Q_n}{\xi}} - C(\delta,\alpha)}{A(\delta,\alpha) - C(\delta,\alpha) + 1} \right\rceil^+, \; n = 1,\ldots,N \quad (14)$$

where $\sqrt{\xi} = \frac{\sum_{n=1}^{N^*}\sqrt{Q_n}}{(N^*-1)C(\delta,\alpha)+A(\delta,\alpha)+1}$, $\lceil\Omega\rceil^+ \triangleq \max\{\Omega,0\}$, and N^*, $1 \le N^* \le N$ satisfies the constraint that $S_n \ge 0\ \forall n$.

Proof: It can be shown that the optimization Problem (13) is concave and can be solved by invoking the Karush−Kuhn−Tucker conditions [34]. The conclusion then follows. ■

From (14), when the request probability obeys $Q_n > \xi C^2(\delta,\alpha)$, $\mathcal{G}_n$ is cached with a caching probability of S_n^{opt}, otherwise, it is not cached. This optimal strategy implies that ideally the SBSs should cache the specific files with high request probabilities, while those files with low request probabilities should not be cached at all due to the limited storage of SBSs in the network. Moreover, we can see that from (14) the optimal caching probability of an FG is a linear function of the square root of its request probability.

Regarding the scenario of FGs associated with overlapping subsets of files, as we mentioned before, $\Pr(\mathcal{D})$ in this scenario has the same formulation as that in the nonoverlapping scenario. Therefore, the optimal caching probability of $\mathcal{F}_m$ in the scenario of FGs having overlapping subsets of files can be formulated as

$$S_m^{\text{Opt}} = \min\left\{ \left\lceil \frac{\sqrt{\frac{Q_m}{\xi}} - C(\delta,\alpha)}{A(\delta,\alpha) - C(\delta,\alpha) + 1} \right\rceil^+, 1 \right\} \quad (15)$$

where $\sqrt{\xi} = \frac{\sum_{m=1}^{M^*}\sqrt{Q_m}}{(M^*-V)C(\delta,\alpha)+V(A(\delta,\alpha)+1)}$, and $M^*(1 \le M^* \le M)$, satisfies the constraint that $0 \le S_m \le 1\ \forall m$, and V denotes the number of files in each FG.

Compared with the nonoverlapping scenario, the presence of overlapping subsets among the FGs provides a higher grade of diversity in the system. However, based on our simulations to be discussed in the sequel, we find that the gain of maximum $\Pr(\mathcal{D})$ obtained as a benefit of this diversity is limited, while the algorithm associated with the optimal caching strategy of (15) is more complex than that of (14).

B. Dynamic On–Off Architecture

In this architecture, as shown in (11), the probability $\Pr(\mathcal{A}_n)$ that an SBS in Tier-n is in the active mode, is a function of the ratio $Q_n\lambda_u/S_n\lambda_s$. Since the intensity of SBSs is much higher than the intensity of the MUs in this architecture, i.e., we have $\lambda_s \gg \lambda_u$, the SBS activity probability $\Pr(\mathcal{A}_n)$ in (11) can be approximated as

$$\Pr(\mathcal{A}_n) \approx \frac{Q_n\lambda_u}{S_n\lambda_s}. \quad (16)$$

Substituting (16) into (12) and (5), we can formulate the optimization problem of maximizing the successful downloading probability as

$$\begin{aligned} &\max_{\{S_n,\varepsilon_n\}} \Pr(\mathcal{D}) = \\ &\max_{\{S_n,\varepsilon_n\}} \sum_{n=1}^{N} \frac{Q_n S_n}{Q_n\frac{\lambda_u}{\lambda_s}A(\delta,\alpha)\cdot\varepsilon_n + \sum_{i:i\neq n} Q_i\frac{\lambda_u}{\lambda_s}C(\delta,\alpha)\cdot\varepsilon_i + S_n} \\ &\text{s.t.} \sum_{n=1}^{N} S_n = 1 \\ &S_n \ge 0,\ n = 1,\ldots,N \\ &\varepsilon_n = \begin{cases} 1, & \text{if} \quad S_n > 0 \\ 0, & \text{if} \quad S_n = 0. \end{cases} \end{aligned} \quad (17)$$

Different from the optimization problem in (13), the variable ε_n is introduced to indicate whether $\mathcal{G}_n$ is cached. Due to the existence of ε_n, which implies 2^N hypotheses of file caching states, Problem (17) is difficult to solve. Nevertheless, we manage to find the solution and summarize it in Theorem 4.

Theorem 4: The optimal caching scheme, i.e., the optimal file caching PMF $\{S_n^{Opt}\}$, that maximizes the average probability of successful download, is given by

$$S_n^{Opt} = \begin{cases} \zeta_K\sqrt{Q_n\xi_K C(\delta,\alpha) - Q_n^2(C(\delta,\alpha) - A(\delta,\alpha))} \\ -\left(\xi_K\frac{\lambda_u}{\lambda_s}C(\delta,\alpha) - Q_n\frac{\lambda_u}{\lambda_s}(C(\delta,\alpha) - A(\delta,\alpha))\right), n \le K \\ 0, \qquad K < n \le N. \end{cases} \quad (18)$$

where

$$\begin{aligned} \xi_K &\triangleq \sum_{i=1}^{K} Q_i \\ \zeta_K &\triangleq \frac{1 + K\xi_K\frac{\lambda_u}{\lambda_s}C(\delta,\alpha) - \xi_K\frac{\lambda_u}{\lambda_s}(C(\delta,\alpha) - A(\delta,\alpha))}{\sum_{i=1}^{K}\sqrt{Q_i\xi_K C(\delta,\alpha) - Q_i^2(C(\delta,\alpha) - A(\delta,\alpha))}}. \end{aligned} \quad (19)$$

Regarding K, we have

$$K = \arg\max_k \left\{ D_k : k = 1,2,\ldots,\widehat{N} \right\} \quad (20)$$

where

$$D_k \triangleq \xi_k - \frac{\frac{\lambda_u}{\lambda_s}\left(\sum_{n=1}^{k}\sqrt{Q_n\xi_k C(\delta,\alpha) - Q_n^2(C(\delta,\alpha) - A(\delta,\alpha))}\right)^2}{1 + k\xi_k\frac{\lambda_u}{\lambda_s}C(\delta,\alpha) - \xi_k\frac{\lambda_u}{\lambda_s}(C(\delta,\alpha) - A(\delta,\alpha))} \quad (21)$$

and

$$\widehat{N} = \begin{cases} N, & \text{if } \frac{\lambda_u}{\lambda_s} < a_N \\ N-1, & \text{if } a_N \le \frac{\lambda_u}{\lambda_s} < a_{N-1} \\ \cdots \\ 1, & \text{if } a_2 \le \frac{\lambda_u}{\lambda_s}. \end{cases} \quad (22)$$

Algorithm 1: Optimal Caching Probabilities in the Dynamic On–Off Architecture.

1: Set $j = N$.
2: Compute $\xi_j = \sum_{i=1}^{j} Q_i$, and ϑ_j and a_j in (24) and (23).
3: Compare $\frac{\lambda_u}{\lambda_s}$ with a_j. If $\frac{\lambda_u}{\lambda_s} < a_j$, go to Step 4; otherwise, set $j = j-1$ and go to Step 2.
4: Set $\widehat{N} = j$.
5: Compute $\xi_k = \sum_{i=1}^{k} Q_i$ and D_k in (21), $k = 1, \cdots, \widehat{N}$.
6: Set $K = \arg\max_k \{D_k\}$.
7: Compute ξ_K and ζ_K in (19), then compute S_n^{Opt} in (18).

Furthermore, the segmentation parameter a_j, $j = 2, \ldots, N$ is given by

$$a_j = \frac{\vartheta_j}{(\vartheta_j \xi_j - Q_j)(C(\delta,\alpha) - A(\delta,\alpha)) + (1 - j\vartheta_j)\xi_j C(\delta,\alpha)} \tag{23}$$

where

$$\vartheta_j \triangleq \frac{\sqrt{Q_j \xi_j C(\delta,\alpha) - Q_j^2 (C(\delta,\alpha) - A(\delta,\alpha))}}{\sum_{i=1}^{j} \sqrt{Q_i \xi_j C(\delta,\alpha) - Q_i^2 (C(\delta,\alpha) - A(\delta,\alpha))}}. \tag{24}$$

Proof: See Appendix D. ■

To get a better understanding of Theorem 4, we propose Algorithm 1 to implement Theorem 4.

From Theorem 4, we have the following remarks.

Remark 5: In the always-on architecture, the optimal number of FGs to be cached depends only on $\{Q_n : n = 1, \cdots, N\}$. By contrast, in the dynamic on–off architecture, the optimal number of FGs to be cached depends not only on $\{Q_n\}$ but on the MU-to-SBS intensity ratio λ_u / λ_s in the network as well.

Remark 6: According to (22), given λ_u, more FGs tend to be cached in the SBSs, when λ_s becomes higher. Moreover, when the intensity of SBSs is not sufficiently high to cache all the FGs, the SBSs should cache the specific files with relatively high request probabilities, which is consistent with the conclusion for the always-on architecture.

Remark 7: In (18), with a practical region of the SINR threshold and path-loss exponent from 3GPP, i.e., for $\delta \in [0.5, 3]$ and $\alpha \in (2, 4]$, we have $\xi_K C(\delta,\alpha) \gg Q_n (C(\delta,\alpha) - A(\delta,\alpha))$, and the optimal caching probability $S_n^{Opt} \approx \zeta_K \sqrt{Q_n \frac{\lambda_u}{\lambda_s} \xi_K C(\delta,\alpha)} - \xi_K \frac{\lambda_u}{\lambda_s} C(\delta,\alpha)$. From (14) and (18), it is interesting to observe that the optimal caching scheme in both the always-on architecture and in the dynamic on–off architecture follow a square root law, i.e., S_n^{Opt} is a linear function of $\sqrt{Q_n}$.

VI. Numerical and Simulation Results

In this section, we present both our numerical and Monte-Carlo simulation results of $\Pr(\mathcal{D})$ in various scenarios. In the Monte-Carlo simulations, the performance is averaged over 1000 network deployments, where in each deployment SBSs and MUs are randomly distributed in an area of 5 × 5 km according to an HPPP distribution. The intensity of MUs in the network is 200/km^2. The transmission power of the SBSs, the noise power, the path-loss exponent, and the SINR threshold are set to 30 dBm, –104 dBm, 4 and 0.25(-6 dB), respectively [32]. In the simulations of the always-on architecture, the deployment intensity of SBSs is set to 80/km^2, while in the simulations of the dynamic on–off architecture, the intensity is set to 400/km^2.

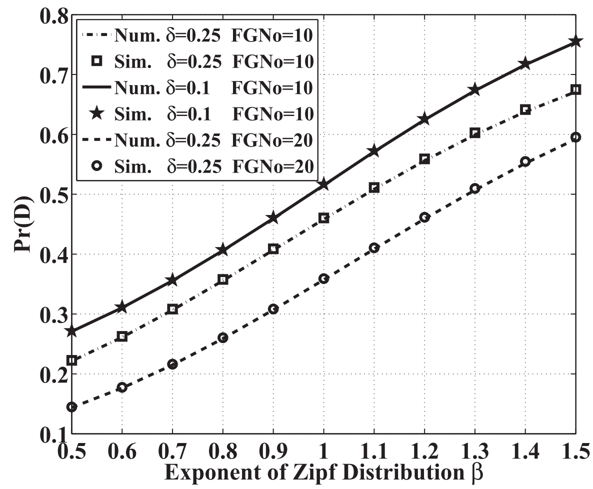


Fig. 2. Numerical and simulation results of $\Pr(\mathcal{D})$ of the O-PSC strategy in the always-on architecture.

Furthermore, we consider a file library consisting of $M = 100$ files, and we partition the file library into $N = 10$ FGs with a simple grouping strategy that the mth file belongs to $\mathcal{G}_n$ if $m \in [\frac{M}{N}(n-1)+1, \ldots, \frac{M}{N}n]\ \forall n \in \{1, \ldots, N\}$. Note that the specific choice of the file grouping strategy is beyond the scope of this paper and it does not affect our results, because it only changes the specific values of the request-PMF $\{Q_n\}$.

In addition, we consider the following two PSC strategies.

1) The request probability based PSC (RP-PSC) [12], where the caching probability of one FG equals to its request probability, i.e., $S_n = Q_n$. Intuitively, a particular FG is more popular than another, the RP-PSC strategy will designate more SBSs to cache it. This strategy is evaluated as a benchmark in our simulations.
2) The proposed optimized PSC (O-PSC) based on (14) in the always-on architecture and (18) in the dynamic on–off architecture, where $S_n = S_n^{\text{opt}}$.

A. Always-On Architecture

Fig. 2 compares the numerical and the simulation results concerning $\Pr(\mathcal{D})$ of the O-PSC strategy. First, it can be seen that the numerical results closely match the simulation results in all scenarios. In the following, we will focus on the analytical results only, due to the accuracy of our analytical results. Second, $\Pr(\mathcal{D})$ increases with the Zipf exponent β. With a larger β, the request probabilities of files are more unevenly distributed. In such cases, a few FGs dominate the requests and caching such popular FGs gives a large $\Pr(\mathcal{D})$. Third, $\Pr(\mathcal{D})$ will be lower, if the value of δ becomes higher. This is because when the SINR threshold is increased, the probability that the received SINR from the SBS storing the file exceeds this threshold is reduced. Finally, we can see that $\Pr(\mathcal{D})$ increases as the number of FGs

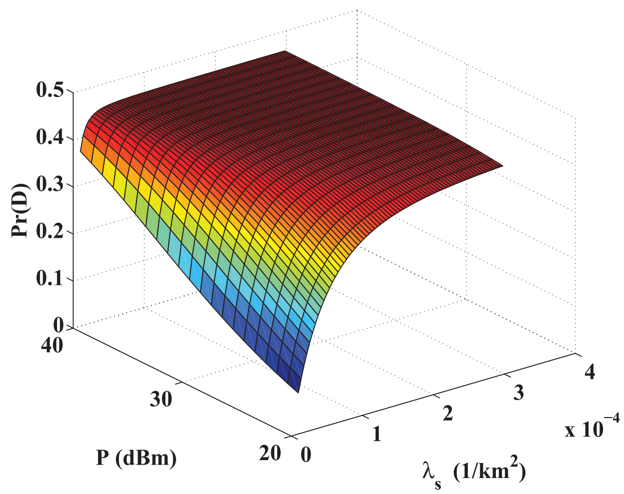


Fig. 3. $\Pr(\mathcal{D})$ of the O-PSC strategy with different P and λ_s in the always-on architecture.

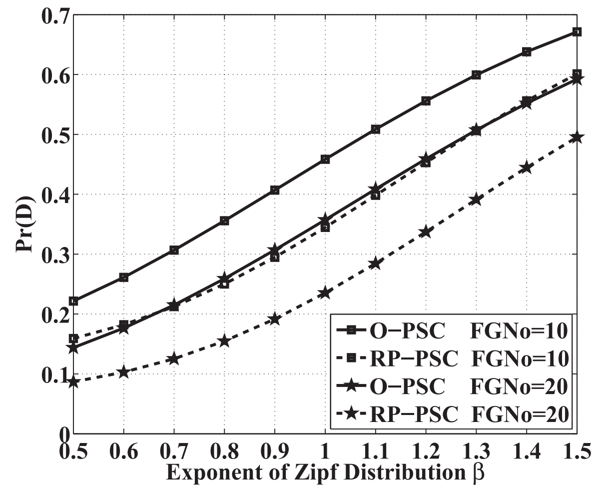


Fig. 5. Comparison of $\Pr(\mathcal{D})$ versus β of the RP-PSC and O-PSC strategies in the always-on architecture.

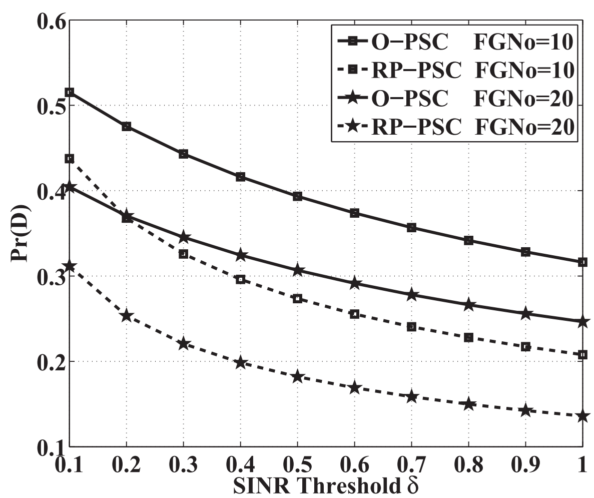


Fig. 4. Comparison of $\Pr(\mathcal{D})$ versus δ of the RP-PSC and O-PSC strategies in the always-on architecture.

decreases. Since each SBS only caches one FG, decreasing the number of FGs implies that each SBS caches more files. Hence, this $\Pr(\mathcal{D})$ improvement comes from increasing the stored contents in each SBS.

Fig. 3 shows the SDP $\Pr(\mathcal{D})$ for the O-PSC strategy when the transmission power P of SBSs varies within 20–40 dBm and the deployment intensity λ_s of SBSs varies within 10–400/km^2. To highlight the asymptotic behavior of $\Pr(\mathcal{D})$ with the growth of P, we set the noise power to -50 dBm. We can see from the figure that $\Pr(\mathcal{D})$ increases monotonically with P or λ_s. The value of $\Pr(\mathcal{D})$ remains constant, when P or λ_s is sufficiently high. This result illustrates the limit of $\Pr(\mathcal{D})$ in the always-on architecture shown in (8).

In Fig. 4, we plot $\Pr(\mathcal{D})$ versus the SINR threshold δ to compare the performances of the RP-PSC and O-PSC strategies. We can see that the proposed O-PSC strategy exhibits a significantly better performance than the RP-PSC strategy. With the number of FGs $N = 10$, the performance gain in terms of $\Pr(\mathcal{D})$ provided by the O-PSC strategy ranges from 20% to 50%, when δ varies from 0.1 to 1. When δ is high, the probability that MUs can directly download the files from the storage of SBSs becomes small. In such cases, the advantage of optimizing the caching probabilities of the FGs is more obvious.

Even more significant $\Pr(\mathcal{D})$ improvement can be observed for the case of $N = 20$ than that for $N = 10$. A larger number of FGs means that less contents can be cached in each SBS, which implies a very limited storage capacity. In such cases, the benefit of optimizing the caching probabilities is more significant.

Fig. 5 compares $\Pr(\mathcal{D})$ in the context of RP-PSC and O-PSC strategies versus the Zipf exponents β. First, we can see that the proposed O-PSC strategy greatly outperforms the RP-PSC strategy in terms of $\Pr(\mathcal{D})$. With the number of FGs $N = 20$, the performance gain of $\Pr(\mathcal{D})$ ranges from 65% to 20% when β varies from 0.5 to 1.5. In other words, the $\Pr(\mathcal{D})$ improvement decreases, as β grows. The reason behind this trend is that for a large β, a small fraction of FGs dominate the file requests. Once the SBSs cache these very popular FGs, $\Pr(\mathcal{D})$ will become sufficiently high. Thus, the additional gain given by the optimization of caching probabilities becomes smaller. Furthermore, compared with the case $N = 10$, the $\Pr(\mathcal{D})$ improvement when $N = 20$ is more significant. The reason for this phenomenon has been explained above.

Fig. 6 compares $\Pr(\mathcal{D})$ in conjunction with the O-PSC strategies in the overlapping and nonoverlapping scenarios. Since the total number of files in our simulations is 100, in the figure, the curves of "FGNo = 10" and "FGNo = 20" are compared against the curves of "FilesPerGroup = 10" and "FilesPerGroup = 5," respectively. We can see that the performance of SDP in the scenario of FGs having overlapping subsets of files is better than that of the nonoverlapping subsets of files. The reason for this observation is that allowing overlapping amongst the different FGs provides a beneficial diversity of the FGs. Furthermore, we can see that when the SINR threshold is increased, the advantage of the overlapping scenario wanes. This is because when the SINR threshold is high, the O-PSC strategy tends to cache fewer popular files and the diversity of FGs becomes of limited benefit here.

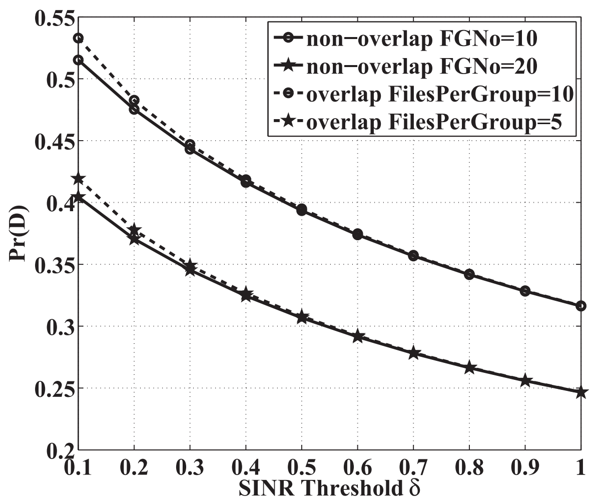


Fig. 6. Comparison of $\Pr(\mathcal{D})$ versus δ in the overlapping and nonoverlapping scenarios in the always-on architecture.

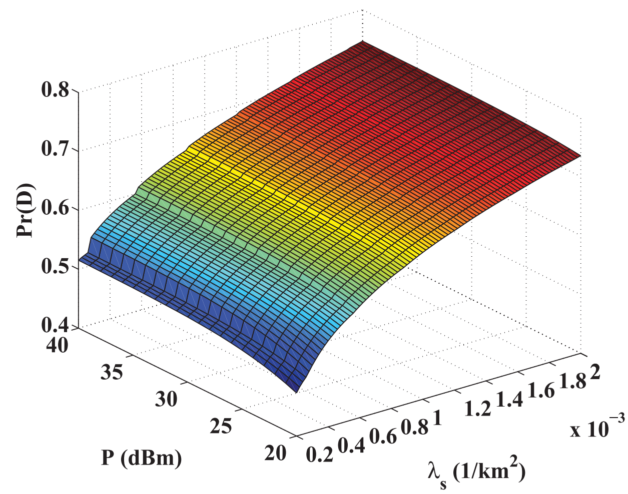


Fig. 8. $\Pr(\mathcal{D})$ with different P and λ_s in the dynamic on–off architecture.

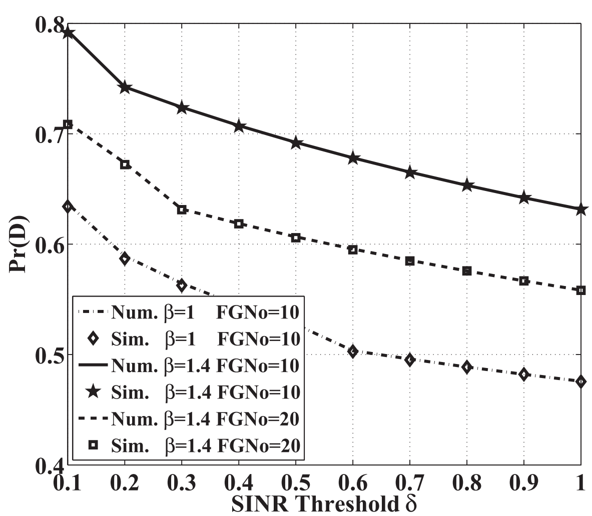


Fig. 7. Numerical and simulation results of $\Pr(\mathcal{D})$ of the O-PCP strategy in the dynamic on–off architecture.

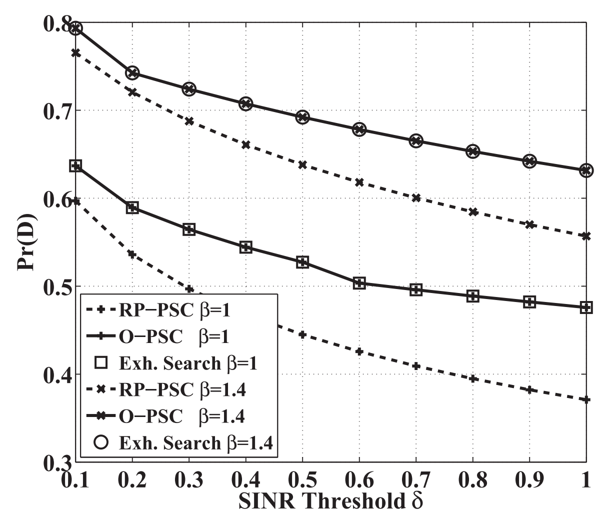


Fig. 9. Comparison of $\Pr(\mathcal{D})$ of the RP-PSC and O-PSC strategies versus δ in the dynamic on–off architecture.

B. Dynamic On–Off Architecture

Fig. 7 shows our comparison between the numerical and simulation results of $\Pr(\mathcal{D})$ for the O-PSC strategy. We can see from this figure that the numerical results closely match the simulation results in all scenarios. Similar phenomena can be observed as in the always-on architecture.

1) $\Pr(\mathcal{D})$ decreases upon increasing the SINR threshold δ.
2) $\Pr(\mathcal{D})$ increases with the Zipf exponent β.
3) $\Pr(\mathcal{D})$ increases when the number of FGs decreases.

The reasons behind these trends are the same as those discussed for the always-on architecture. Moreover, compared to Fig. 2, the value of $\Pr(\mathcal{D})$ in the dynamic on–off architecture of Fig. 7 is shown to be higher. The reason is that the dynamic on–off technique efficiently mitigates the potential avoidable interference in the network.

Fig. 8 shows the performance of $\Pr(\mathcal{D})$ for the O-PSC strategy in the dynamic on–off architecture, when the transmission power P of SBSs varies from 20 to 40 dBm and the SBS intensity λ_s varies from 200 to $2000/\text{km}^2$. We can see from this figure that $\Pr(\mathcal{D})$ increases monotonically, when either P or λ_s increases. Moreover, we can see that when P increases to a sufficiently high value, any further increase of P will no longer improve $\Pr(\mathcal{D})$. However, the increase of λ_s will always improve $\Pr(\mathcal{D})$, as seen in (12).

Fig. 9 compares $\Pr(\mathcal{D})$ of the RP-PSC and O-PSC strategies, when the SINR threshold δ varies. It can be seen from the figure that compared to the RP-PSC strategy, $\Pr(\mathcal{D})$ is obviously improved by the optimal caching PMF $\{S_n^{\text{Opt}}\}$ in the O-PSC strategy. With the Zipf exponent $\beta = 1$, the performance gain of $\Pr(\mathcal{D})$ ranges from 7% to 30%, when δ varies from 0.1 to 1. This observation is similar to that in the always-on architecture. That is, the $\Pr(\mathcal{D})$ improvement achieved by the O-PSC strategy is more pronounced, when the SINR threshold is higher. Furthermore, the $\Pr(\mathcal{D})$ improvement is higher when the Zipf exponent β is lower. The reason for this is explained above.

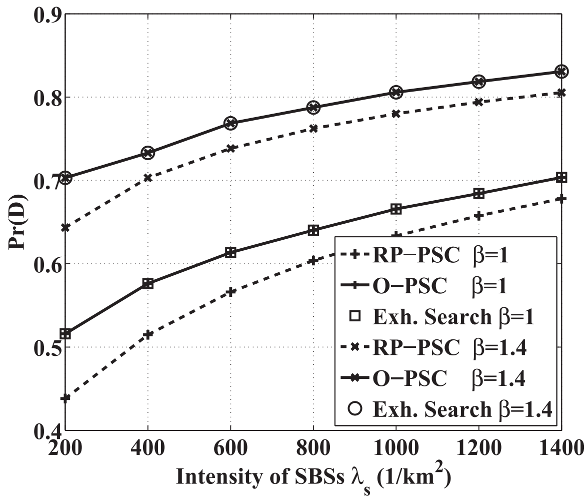


Fig. 10. Comparison of $\Pr(\mathcal{D})$ of the RP-PSC and O-PSC strategies versus λ_s in the dynamic on–off architecture.

Furthermore, in order to verify the optimality of the solution given by our algorithm, we plot the optimal solution obtained from the exhaustive search over all legitimate file caching states, denoted by "Exh. Search" in the figure. Observed from the figure that our solution exactly matches the optimal solution of "Exh. Search," which confirms our statement that the proposed solution achieves global optimality.

In Fig. 10, we portray $\Pr(\mathcal{D})$ of the RP-PSC and the O-PSC strategies versus the SBS intensity λ_s. First, it can be seen that compared with the RP-PSC strategy, the optimization of the caching probabilities in the O-PSC strategy improves $\Pr(\mathcal{D})$ in all scenarios. This $\Pr(\mathcal{D})$ improvement achieved by the O-PSC strategy wanes slightly when λ_s increases because when the SBS intensity is higher, each MU becomes capable of associating with multiple SBSs, and thus, the probability that MUs can successfully download contents from SBSs will be higher. In such a case, the $\Pr(\mathcal{D})$ improvement obtained by the optimization of the FG caching probabilities remains limited. In addition, we verify the optimality of our solution by comparing it to the optimal solution obtained from the exhaustive search.

VII. Conclusion

In this paper, based on stochastic geometry theory, we analyzed the performance of the PSC in a pair of network architectures. Specifically, we analyzed the probability $\Pr(\mathcal{D})$ that MUs can successfully download contents from the storage of SBSs. We concluded that increasing the SBSs' transmission power P or their deployment intensity λ_s is capable of increasing the SDP. However, in the always-on architecture, $\Pr(\mathcal{D})$ remains constant when P or λ_s is sufficiently high, while in the dynamic on–off architecture, $\Pr(\mathcal{D})$ always increases as λ_s grows. Furthermore, in order to maximize $\Pr(\mathcal{D})$, we optimized the caching probabilities of the FGs. Our results demonstrated that in the always-on architecture, the optimal subset of FGs depends on the contents request probabilities. In the dynamic on–off architecture, a piecewise defined function of MU-to-SBS intensity ratio λ_u/λ_s was introduced in order to find the optimal subset of FGs to be cached. Interestingly, a similar optimal caching probability law was found for both architectures, i.e., S_n^{Opt} is a linear function of $\sqrt{Q_n}$. Our simulation results showed that the proposed optimal caching probabilities of the FGs achieve a substantial gain in both architecture in terms of $\Pr(\mathcal{D})$ compared to the benchmark $S_n = Q_n$, because more caching resources are devoted to the more popular files in the proposed scheme.

Appendix A
Proof of Theorem 1

In Tier-n of the always-on architecture, where the intensity of the SBSs is $S_n\lambda_s$, the PDF of z, i.e., the distance between the typical MU and its nearest SBS, follows $f_Z(z) = 2\pi S_n\lambda_s z\exp(-\pi S_n\lambda_s z^2)$. From (3) and (4), we have

$$\begin{aligned}\Pr(\mathcal{D}_n) &= \Pr\left(\gamma_n(z) \geq \delta\right)\\ &= \int_0^\infty \Pr\left[\frac{P h_{x_0} z^{-\alpha}}{\sum_{x_j\in\Phi\backslash\{x_0\}} P h_{x_j}\left\|x_j\right\|^{-\alpha} + \sigma^2} \geqslant \delta\right] f_Z(z)\,\mathrm{d}z\\ &\overset{(a)}{=} \int_0^\infty \mathbb{E}_I\left[\exp\left(-z^\alpha\delta I\right)\right]\exp\left(-\frac{z^\alpha\delta\sigma^2}{P}\right)\\ &\qquad 2\pi S_n\lambda_s z\exp(-\pi S_n\lambda_s z^2)\mathrm{d}z \end{aligned} \tag{25}$$

where (a) is obtained by $h_{x_0}\sim\exp(1)$ and $I \triangleq \sum_{x_j\in\Phi\backslash\{x_0\}} h_{x_j}\left\|x_j\right\|^{-\alpha}$ represents the interference.

The interference I consists of two independent parts: 1) I_1: the SBSs in other tiers, which are dispersed across the entire area of the network, and 2) I_2: the SBSs in the nth tier, whose distances from the typical MU are larger than z. Due to the independence of I_1 and I_2, we have $\mathbb{E}_I\left[\exp\left(-z^\alpha\delta I\right)\right] = \mathbb{E}_{I_1}\left[\exp\left(-z^\alpha\delta I_1\right)\right]\cdot \mathbb{E}_{I_2}\left[\exp\left(-z^\alpha\delta I_2\right)\right]$.

Since the distribution of the SBSs in Tier-i is viewed as an HPPP ϕ_i with $S_i\lambda_s$ and therefore, we have

$$\begin{aligned}&\mathbb{E}_{I_1}\left[\exp\left(-z^\alpha\delta I_1\right)\right]\\ &= \mathbb{E}_{h_{x_j},x_j}\left[\prod_{x_j\in\sum_{i=1,i\neq n}^N\phi_i}\exp\left(-z^\alpha\delta h_{x_j}\left\|x_j\right\|^{-\alpha}\right)\right]\\ &\overset{(b)}{=} \mathbb{E}_{x_j}\left[\prod_{x_j\in\sum_{i=1,i\neq n}^N\phi_i}\frac{1}{1+z^\alpha\delta\left\|x_j\right\|^{-\alpha}}\right]\\ &\overset{(c)}{=} \exp\left(-\sum_{i=1,i\neq n}^N S_i\lambda_s\int_{\mathbb{R}^2}\left(1-\frac{1}{1+\delta z^\alpha\left\|x_j\right\|^{-\alpha}}\right)\mathrm{d}x_j\right)\\ &= \exp\left(-2\pi\sum_{i=1,i\neq n}^N S_i\lambda_s\frac{1}{\alpha}\delta^{\frac{2}{\alpha}}B\left(\frac{2}{\alpha},1-\frac{2}{\alpha}\right)z^2\right) \end{aligned} \tag{26}$$

where (b) uses $h_{x_j}\sim\exp(1)$, and (c) uses $\mathbb{E}\left[\prod_{v\in\Phi}\xi(v)\right] = \exp\left(-\lambda_\Phi\int\left(1-\xi(v)\right)\mathrm{d}v\right)$.

As for I_2, we have

$$\begin{aligned}
&\mathbb{E}_{I_2}\left[\exp\left(-z^\alpha\delta I_2\right)\right]\\
&=\exp\left(-S_n\lambda_s 2\pi\int_z^\infty\left(1-\frac{1}{1+z^\alpha\delta\left\|x_j\right\|^{-\alpha}}\right)\left\|x_j\right\|\mathrm{d}\left\|x_j\right\|\right)\\
&\overset{(d)}{=}\exp\left(-S_n\lambda_s\pi\delta^{\frac{2}{\alpha}}z^2\frac{2}{\alpha}\int_{\delta^{-1}}^\infty\frac{l^{\frac{2}{\alpha}-1}}{1+l}\mathrm{d}l\right)\\
&=\exp\left(-S_n\lambda_s\pi z^2\frac{2\delta}{\alpha-2}\,{}_2F_1\left(1,1-\frac{2}{\alpha};2-\frac{2}{\alpha};-\delta\right)\right)
\end{aligned}\tag{27}$$

where (d) uses $l\triangleq\delta^{-1}z^{-\alpha}\left\|x_j\right\|^\alpha$.

Our proof is completed by plugging (26) and (27) into (25). ∎

APPENDIX B
PROOF OF COROLLARY 1

Since we have $\Pr(\mathcal{D})=\sum_{n=1}^N Q_n\Pr(\mathcal{D}_n)$, to prove that $\Pr(\mathcal{D})$ increases with the increase of λ_s, we only have to prove that $\Pr(\mathcal{D}_n)$ increases monotonically upon increasing λ_s $\forall n$. Thus, in the following, we focus our attention on the proof that $\frac{\partial\Pr(\mathcal{D}_n)}{\partial\lambda_s}>0$.

To simplify our discourse, we use $C_1\triangleq\frac{\pi S_n}{2\sigma}\sqrt{\frac{\pi P}{\delta}}$, and $C_2\triangleq\frac{\pi}{2\sigma}\sqrt{\frac{P}{\delta}}\left(S_n+\frac{\pi}{2}\sqrt{\delta}(1-S_n)+S_n\sqrt{\delta}\arctan\sqrt{\delta}\right)$. Obviously, we have $C_1>0$ and $C_2>0$. Then, $\Pr(\mathcal{D}_n)$ can be rewritten as

$$\Pr(\mathcal{D}_n)=C_1\lambda_s\exp(C_2^2\lambda_s^2)\mathrm{erfc}(C_2\lambda_s).\tag{28}$$

Hence, we have

$$\begin{aligned}
\frac{\partial\Pr(\mathcal{D}_n)}{\partial\lambda_s}&=C_1\lambda_s\exp(C_2^2\lambda_s^2)\left(1-\mathrm{erf}(C_2\lambda_s)\right)\\
&=\left(C_1\exp(C_2^2\lambda_s^2)+C_1\lambda_s\exp(C_2^2\lambda_s^2)2C_2^2\lambda_s\right)\mathrm{erfc}(C_2\lambda_s)\\
&\quad-C_1\lambda_s\exp(C_2^2\lambda_s^2)\frac{2}{\sqrt{\pi}}C_2\exp(-C_2^2\lambda_s^2)\\
&=C_1\exp(C_2^2\lambda_s^2)(1+2C_2^2\lambda_s^2)\mathrm{erfc}(C_2\lambda_s)-C_1C_2\lambda_s\frac{2}{\sqrt{\pi}}.
\end{aligned}\tag{29}$$

According to [35], the continued fraction expansion of the complementary error function is

$$\mathrm{erfc}(z)=\frac{z}{\sqrt{\pi}}\exp(-z^2)\frac{1}{z^2+\frac{a_1}{1+\frac{a_2}{z^2+\frac{a_3}{1+\cdots}}}},a_m=\frac{m}{2}.\tag{30}$$

From (30), we have $\mathrm{erfc}(z)>\frac{z}{\sqrt{\pi}}\exp(-z^2)\frac{1}{z^2+\frac{1}{2}}$. Substituting $C_2\lambda_s$ for z, we have

$$\exp(C_2^2\lambda_s^2)\mathrm{erfc}(C_2\lambda_s)>\frac{C_2\lambda_s}{\sqrt{\pi}}\frac{1}{C_2^2\lambda_s^2+\frac{1}{2}}.\tag{31}$$

Substituting (31) into (29), we can prove that $\frac{\partial\Pr(\mathcal{D}_n)}{\partial\lambda_s}>0$, which implies that $\Pr(\mathcal{D})$ increases monotonically upon increasing λ_s. ∎

APPENDIX C
PROOF OF THEOREM 2

Similar to the derivation in Appendix A, in the dynamic on–off architecture, the intensity of SBSs in Tier-n is also $S_n\lambda_s$. Thus, in Tier-n the distance z between the typical MU and its nearest SBS follows the same PDF $f_Z(z)$ in the always-on architecture. It follows that we have a similar formulation for $\Pr(\mathcal{D}_n)$ in the dynamic on–off architecture, yielding

$$\begin{aligned}
\Pr(\mathcal{D}_n)=&\int_0^\infty\mathbb{E}_I\left[\exp\left(-z^\alpha\delta I\right)\right]\exp\left(-\frac{z^\alpha\delta\sigma^2}{P}\right)\\
&2\pi S_n\lambda_s z\exp(-\pi S_n\lambda_s z^2)\mathrm{d}z.
\end{aligned}\tag{32}$$

In the dynamic on–off architecture, the interference I only arrives from the SBSs in the active mode. According to [36], the activity probability $\Pr(\mathcal{A}_n)$ of the SBSs in Tier-n, can be formulated as

$$\Pr(\mathcal{A}_n)\approx1-\left(1+\frac{Q_n\lambda_u}{3.5S_n\lambda_s}\right)^{-3.5}.$$

As in Appendix A, we divide the interference into two parts: $I=I_1+I_2$. The first part of interference I_1 is inflicted by the active SBSs in any Tier-i, $i\neq n$, which can be viewed as a homogeneous PPP with the intensity of $\Pr(\mathcal{A}_i)S_i\lambda_s$. Hence, we update (26) as follows:

$$\begin{aligned}
&\mathbb{E}_{I_1}\left[\exp\left(-z^\alpha\delta I_1\right)\right]\\
&=\exp\left(-2\pi\sum_{i=1:i\neq n}^N\Pr(\mathcal{A}_i)S_i\lambda_s\frac{1}{\alpha}\delta^{\frac{2}{\alpha}}B\left(\frac{2}{\alpha},1-\frac{2}{\alpha}\right)z^2\right).
\end{aligned}\tag{33}$$

The second part of the interference I_2 comes from the active SBSs in Tier-n located in the area outside the circle with radius z. We update (27) as follows:

$$\begin{aligned}
&\mathbb{E}_{I_2}\left[\exp\left(-z^\alpha\delta I_2\right)\right]=\exp\left(-\Pr(\mathcal{A}_n)\right.\\
&\left.S_n\lambda_s\pi z^2\frac{2\delta}{\alpha-2}\,{}_2F_1\left(1,1-\frac{2}{\alpha};2-\frac{2}{\alpha};-\delta\right)\right).
\end{aligned}\tag{34}$$

Integrating (33) and (34) into (32) completes the proof. ∎

APPENDIX D
PROOF OF THEOREM 4

Note that in the following proof, we simplify the notation by introducing $a\triangleq\frac{\lambda_u}{\lambda_s}$, $C\triangleq C(\delta,\alpha)$, and $A\triangleq A(\delta,\alpha)$.

First, we investigate the optimization Problem (17) for a given indicator vector ε. Let us denote by N^* the number of ones in ε, and by $\{n_j\}$ the subscript of the ones in N^*. Then, we have a new optimization problem represented as

$$\begin{aligned}\max_{\{S_{n_j}\}} \quad & \sum_{j=1}^{N^*} \frac{Q_{n_j} S_{n_j}}{Q_{n_j} aA + \sum_{i:i\neq j} Q_{n_i} aC + S_{n_j}} \\ \text{s.t.} \quad & \sum_{j=1}^{N^*} S_{n_j} = 1 \\ & S_{n_j} > 0 \ \forall j = 1, \dots N^*. \end{aligned} \tag{35}$$

If we neglect the constraint $S_{n_j} > 0$, the solution to Problem (35) is presented in Lemma 1.

Lemma 1: Neglecting the constraint $S_{n_j} > 0$, the optimal solution for Problem (35) is given by

$$S_{n_j}^{Opt} = \zeta \sqrt{Q_{n_j} C\xi - Q_{n_j}^2 (C-A)} - \left[\xi aC - Q_{n_j} a(C-A)\right] \tag{36}$$

where we have $\zeta \triangleq \frac{1+N^*\xi aC - \xi a(C-A)}{\sum_{i=1}^{N^*} \sqrt{Q_{n_i}\xi C - Q_{n_i}^2 (C-A)}}$ and $\xi \triangleq \sum_{j=1}^{N^*} Q_{n_j}$.

Proof: See Appendix E.

From (36), we propose Lemma 2.

Lemma 2: Given the request probabilities of two FGs cached, where $Q_{n_i} > Q_{n_j}$, according to (36), we have $S_{n_i}^{Opt} > S_{n_j}^{Opt}$.

Proof: See Appendix F.

Based on Lemma 2, we have $S_{n_{j^*}}^{\text{Opt}} = \min\{S_{n_j}^{\text{Opt}}\}$ where $n_{j^*} = \arg\min_{n_j}\{Q_{n_j}\}$. Hence, the constraint $S_{n_j} > 0$, $\forall j = 1, \dots N^*$, is equivalent to $S_{n_{j^*}} > 0$. In order to ensure that $S_{n_{j^*}}^{\text{Opt}} > 0$, based on (36), we have

$$a < a_{n_{j^*}}, \ a_{n_{j^*}} \triangleq \frac{\vartheta_{n_{j^*}}}{(\vartheta_{n_{j^*}}\xi - Q_{n_{j^*}})(C-A) + (1 - N^*\vartheta_{n_{j^*}})\xi C} \tag{37}$$

where

$$\vartheta_{n_{j^*}} \triangleq \frac{\sqrt{Q_{n_{j^*}}\xi C + Q_{n_{j^*}}^2 (A-C)}}{\sum_{i=1}^{N^*} \sqrt{Q_{n_i}\xi C + Q_{n_i}^2 (A-C)}}. \tag{38}$$

Hence, (36) only becomes the optimal solution of Problem (35), when a meets the requirement (37).

Substituting the optimal solution in (36) into (35), we obtain the maximum value of $\Pr(\mathcal{D})$ for the given indicator vector $\boldsymbol{\varepsilon}$, yielding

$$D_{N^*} = \xi - \frac{a\left(\sum_{j=1}^{N^*} \sqrt{Q_{n_j} C\xi + Q_{n_j}^2 (A-C)}\right)^2}{1 + N^*\xi aC + \xi a(A-C)}. \tag{39}$$

Second, we extend the Problem (35) to Problem (17). Based on the analysis above, given the indicator vector $\boldsymbol{\varepsilon}_1$, when $a < a_{\boldsymbol{\varepsilon}_1}$ in (37), we can obtain the maximum $\Pr(\mathcal{D})$ denoted by $D_{\boldsymbol{\varepsilon}_1}$ in (39). For $\boldsymbol{\varepsilon}_2$, if we have $a_{\boldsymbol{\varepsilon}_2} > a_{\boldsymbol{\varepsilon}_1}$, then provided $a < a_{\boldsymbol{\varepsilon}_1}$ holds, we have $a < a_{\boldsymbol{\varepsilon}_2}$. Thus, $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$ are both reasonable for this optimization problem. Through the comparison of $D_{\boldsymbol{\varepsilon}_1}$ and $D_{\boldsymbol{\varepsilon}_2}$, we can find the right choice between $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$. Then obtain the optimal solution of $\{S_n\}$ in form of (36).

Using $\{Q_n\}$, we can obtain the segmentation parameters for a in (37). The smallest segmentation parameter is obtained when $\boldsymbol{\varepsilon}$ contains N ones, which is denoted by a_N. When $a < a_N$, i.e., λ_s is high enough, all FGs can be cached in SBSs. Then, with the increase of a, i.e., the decrease of λ_s, some FGs cannot be cached, where a reduced number of ones appear in $\boldsymbol{\varepsilon}$. Since we have $Q_1 > Q_2 > \cdots > Q_N$, the unpopular FGs will be discarded one by one. Accordingly, we can obtain both $\boldsymbol{\varepsilon}_i$ as well as the segmentation parameter a_i. As a result, a piecewise defined function regarding a is obtained like the number of ones in $\boldsymbol{\varepsilon}$ is shown in (20). ∎

Appendix E
Proof of Lemma1

Neglecting the constraint $S_{n_j} > 0$, it becomes plausible that Problem (35) is a concave maximization problem. Adopting the Lagrange multiplier Λ, we have

$$\Lambda(\mathbf{S}, \lambda) = \sum_{j=1}^{N^*} \frac{Q_{n_j} S_{n_j}}{Q_{n_j} aA + \sum_{i=1:i\neq j}^{N^*} Q_{n_i} aC + S_{n_j}} + \lambda\left(\sum_{j=1}^{N^*} S_{n_j} - 1\right). \tag{40}$$

Using $\xi \triangleq \sum_{j=1}^{N^*} Q_{n_j}$ and $\frac{\partial\Lambda}{\partial S_{n_j}} = 0$, we have

$$\frac{Q_{n_j} aC\xi + Q_{n_j}^2 a(A-C)}{\left(aC\xi + a(A-C)Q_{n_j} + S_{n_j}\right)^2} + \lambda = 0 \ \forall n_j. \tag{41}$$

Since $\sum_{j=1}^{N^*} S_{n_j} = 1$, we have

$$S_{n_j}^{\text{Opt}} = \zeta\sqrt{Q_{n_j} aC\xi + Q_{n_j}^2 a(A-C)} - \left[\xi aC + Q_{n_j} a(A-C)\right] \tag{42}$$

where

$$\zeta \triangleq \frac{1 + N^*\xi aC + \xi a(A-C)}{\sum_{i=1}^{N^*} \sqrt{Q_{n_i}\xi aC + Q_{n_i}^2 a(A-C)}}. \tag{43}$$

∎

Appendix F
Proof of Lemma 2

First, based on the optimal solution given in (36), we have

$$\frac{\partial S_{n_j}^{\text{Opt}}}{\partial Q_{n_j}} = \zeta\frac{\sqrt{a}}{2}\frac{C\xi + 2Q_{n_j}(A-C)}{\sqrt{Q_{n_j} C\xi + Q_{n_j}^2 (A-C)}} + a(C-A). \tag{44}$$

Since $C(\alpha,\delta) > A(\alpha,\delta) > 0$, we have $\frac{\partial S_{n_j}^{\text{Opt}}}{\partial Q_{n_j}} \geq 0$ when $Q_{n_j} \leq \frac{\xi}{2}\frac{C}{C-A}$, which means $S_{n_j}^{\text{Opt}}$ increases with the growth of Q_{n_j}, when Q_{n_j} is no bigger than $\frac{\xi}{2}\frac{C}{C-A}$.

1) Since $Q_{n_j} \leq \xi$, if $\frac{C}{C-A} \geq 2$, for all Q_{n_j}, $\frac{\partial S_{n_j}^{\text{Opt}}}{\partial Q_{n_j}} > 0$, and the proof is completed.

2) For $\frac{C}{C-A} < 2$, we consider the following case. Since $\frac{C}{C-A} > 1$, we have $\frac{\xi}{2}\frac{C}{C-A} > \frac{\xi}{2}$. Because $\sum_j Q_{n_j} = \xi$, among

the N^* FGs cached, there is only one FG associated with $Q_{n_j} > \frac{\xi}{2}\frac{C}{C-A}$. We denote the request probability of this popular file by Q_1 and its caching probability by $S_1^{\rm Opt}$. Since the request probabilities of other cached FGs must be less than $\frac{\xi}{2}\frac{C}{C-A}$, and $\frac{\partial S_{n_j}^{\rm Opt}}{\partial Q_{n_j}} > 0$ when Q_{n_j} in this region, the highest caching probability among these less popular FGs occurs when only two FGs are cached. That is, the other FG with request probability $Q_2 = \xi - Q_1$. Denoted by $S_2^{\rm Opt}$ its caching probability. We have

$$S_1^{\rm Opt} - S_2^{\rm Opt} = \zeta\sqrt{a}\left(\sqrt{Q_1 C\xi + Q_1^2(A-C)} - \sqrt{Q_2 C\xi + Q_2^2(A-C)}\right) + (Q_1 - Q_2)a(C-A). \quad (45)$$

Since $Q_1 C\xi + Q_1^2(A-C) - Q_2 C\xi - Q_2^2(A-C) = (Q_1 - Q_2)\xi a A > 0$, we have $S_1^{\rm Opt} - S_2^{\rm Opt} > 0$. Thus, for the dominate FG, its caching probability also dominates.

Combining the two parts above, we complete the proof. ■

ACKNOWLEDGMENT

The authors would like to thank Dr. H. (Henry) Chen from the University of Sydney, Australia, for his helpful discussions and suggestions.

REFERENCES

[1] CISCO, "Cisco visual networking index: Global mobile data traffic forecast update, 2014–2019 White Paper," Feb. 2014.

[2] D. Lopez-Perez, M. Ding, H. Claussen, and A. H. Jafari, "Towards 1 Gbps/UE in cellular systems: Understanding ultra-dense small cell deployments," Mar. 2015.

Q2 [3] J. Erman, A. Gerber, M. Hajiaghayi, D. Pei, S. Sen, and O. Spatscheck, "To cache or not to cache: The 3G case," *IEEE Internet Comput.*, vol. 15, no. 2, pp. 27–34, Mar. 2011.

[4] U. Niesen, D. Shah, and G. W. Wornell, "Caching in wireless networks," *IEEE Trans. Inf. Theory*, vol. 58, no. 10, pp. 6524–6540, Oct. 2012.

[5] J. Li, H. Chen, Y. Chen, Z. Lin, B. Vucetic, and L. Hanzo, "Pricing and resource allocation via game theory for a small-cell video caching system," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 8, pp. 2115–2129, Aug. 2016.

[6] 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, Feb. 2014.

[7] S. Woo, E. Jeong, S. Park, J. Lee, S. Ihm, and K. Park, "Comparison of caching strategies in modern cellular backhaul networks," in *Proc. 11th Annu. Int. Conf. Mobile Syst., Appl., Serv.*, New York, NY, USA, 2013, pp. 319–332. [Online]. Available: http://doi.acm.org/10.1145/2462456.2464442

[8] H. Ahlehagh and S. Dey, "Video-aware scheduling and caching in the radio access network," *IEEE/ACM Trans. Netw.*, vol. 22, no. 5, pp. 1444–1462, Oct. 2014.

[9] K. Shanmugam, N. Golrezaei, A. Dimakis, A. Molisch, and G. Caire, "Femtocaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

[10] N. Golrezaei, A. Molisch, A. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Commun. Mag.*, vol. 51, no. 4, pp. 142–149, Apr. 2013.

[11] 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. 2015.

[12] N. Golrezaei, P. Mansourifard, A. Molisch, and A. Dimakis, "Base-station assisted device-to-device communications for high-throughput wireless video networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, pp. 3665–3676, Jul. 2014.

[13] H. J. Kang and C. G. Kang, "Mobile device-to-device (D2D) content delivery networking: A design and optimization framework," *J. Commun. Netw.*, vol. 16, no. 5, pp. 568–577, Oct. 2014.

[14] C. Yang, Y. Yao, Z. Chen, and B. Xia, "Analysis on cache-enabled wireless heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 131–145, Jan. 2015.

[15] M. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inform. Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.

[16] K. Poularakis, G. Iosifidis, and L. Tassiulas, "Approximation algorithms for mobile data caching in small cell networks," *IEEE Trans. Commun.*, vol. 62, no. 10, pp. 3665–3677, Oct. 2014.

[17] J. Li, Y. Chen, Z. Lin, W. Chen, B. Vucetic, and L. Hanzo, "Distributed caching for data dissemination in the downlink of heterogeneous networks," *IEEE Trans. Commun.*, vol. 63, no. 10, pp. 3553–3568, Oct. 2015.

[18] E. Bastug, M. Bennis, and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs," *EURASIP J. Wireless Commun. Netw.*, vol. 2015, no. 1, p. 41, 2015. Q3

[19] G. Vettigli, M. Ji, A. Tulino, J. Llorca, and P. Festa, "An efficient coded multicasting scheme preserving the multiplicative caching gain," in *Proc. IEEE Conf. Comput. Commun. Workshops*, Apr. 2015, pp. 251–256.

[20] I. Ashraf, L. Ho, and H. Claussen, "Improving energy efficiency of femtocell base stations via user activity detection," in *Proc. IEEE Wireless Commun. Network. Conf.*, Apr. 2010, pp. 1–5.

[21] 3GPP, "Tentative 3GPP timeline for 5G," Mar. 2015.

[22] QUALCOMM, "1000x: More small cells. hyper-dense small cell deployments," Jun. 2014.

[23] C. Yang, J. Li, and M. Guizani, "Cooperation for spectral and energy efficiency in ultra-dense small cell networks," *IEEE Wireless Commun.*, vol. 23, no. 1, pp. 64–71, Feb. 2016.

[24] N. Saxena, A. Roy, and H. Kim, "Traffic-aware cloud ran: A key for green 5g networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 1010–1021, Apr. 2016.

[25] H. Dhillon, R. Ganti, F. Baccelli, and J. Andrews, "Modeling and analysis of K-tier downlink heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 550–560, Apr. 2012.

[26] M. Zinka, K. Suh, Y. Gu, and J. Kurose, "Characteristics of YouTube network traffic at a campus network—Measurements, models, and implications," *Comput. Netw.*, vol. 53, no. 4, pp. 501–514, Mar. 2009.

[27] M. Maddah-Ali and U. Niesen, "Decentralized coded caching attains order-optimal memory-rate tradeoff," *IEEE/ACM Trans. Netw.*, vol. 23, no. 4, pp. 1029–1040, Aug. 2014.

[28] D. Stoyan, W. Kendall, and J. Mecke, *Stochastic Geometry and its Applications*, 2nd ed. Hoboken, NJ, USA: Wiley, 1995.

[29] S. C. Forum, "Scf049: Backhaul technologies for small cells (release 4)," Feb. 2014. Q4

[30] C. Nicoll, "3G and 4G small cells create big challenges for MNOs," Mar. 2013.

[31] I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series, and Products*, 7th ed. Amsterdam, The Netherlands: Elsevier, 2007.

[32] 3GPP, "Further advancements for E-UTRA physical layer aspects," 3GPP, France, Tech. Rep. v.9.0.0, Mar. 2010.

[33] W. Cody, "Algorithm 715: SPECFUN—A portable FORTRAN package of special function routines and test drivers," *ACM Trans. Math. Softw.*, vol. 19, no. 1, pp. 22–30, Mar. 1993.

[34] H. Kuhn and A. Tucker, "Nonlinear programming," in *Proc. 2nd Berkeley Symp. Math. Statist. Probability*, 1951, pp. 481–492.

[35] A. Cuyt, V. Petersen, B. Verdonk, H. Waadeland, and W. Jones, *Handbook of Continued Fractions for Special Functions*. Berlin, Germany: Springer-Verlag, 2008.

[36] S. Lee and K. Huang, "Coverage and economy of cellular networks with many base stations," *IEEE Commun. Lett.*, vol. 16, no. 7, pp. 1038–1040, Jul. 2012.

Youjia Chen received the B.S. and M.S degrees in communication engineering from Nanjing University, Nanjing, China, in 2005 and 2008, respectively. She is currently working toward the Ph.D. degree in wireless engineering with the University of Sydney, Sydney, Australia.

From 2008 to 2009, she was with Alcatel Lucent Shanghai Bell. From August 2009 until now, she has been with the College of Photonic and Electrical Engineering, Fujian Normal University, China. Her research interests include resource management, load balancing, and caching strategy in heterogeneous cellular networks.

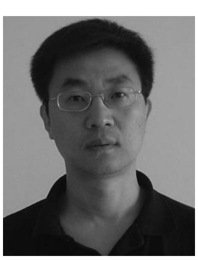
Ming Ding (M'12) received the B.S. and M.S. degrees (with first class Hons.) in electronics engineering and Ph.D. degree in signal and information processing from Shanghai Jiao Tong University (SJTU), Shanghai, China, in 2004, 2007, and 2011, respectively.

From September 2007 to September 2011, while at the same time working as a Researcher/Senior Researcher Sharp Laboratories of China (SLC), after achieving the Ph.D. degree, he continued working with SLC as a Senior Researcher/Principal Researcher until September 2014, when he joined National Information and Communications Technology Australia (NICTA). In July 2016, Commonwealth Scientific and Industrial Research Organization (CSIRO) and NICTA joined forces to create Data61, where he continued as a Senior Research Scientist in this new R&D center in Sydney, NSW, Australia. He has authored more than 30 papers in IEEE journals and conferences, all in recognized venues, and about 20 3GPP standardization contributions, as well as a Springer book entitled *Multi-point Cooperative Communication Systems: Theory and Applications* (Springer-Verlag, 2013). In addition, as the first inventor, he holds 15 CN, seven JP, three US, two KR patents, and has co-authored another 100+ patent applications on 4G/5G technologies.

Dr. Ding has been the Guest Editor/Cochair/TPC member of several IEEE top-tier journals/conferences, e.g., the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, the IEEE COMMUNICATIONS MAGAZINE, the IEEE Globecom Workshops, etc. He received the Presidents Award from the SLC in 2012 for his inventions and publications and served as one of the key members in the 4G/5G standardization team when it was awarded in 2014 as the Sharp Company Best Team: LTE 2014 Standardization Patent Portfolio.

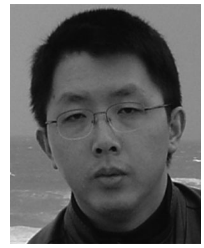
Jun Li (M'09–SM'16) received the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University, Shanghai, China in 2009.

From January 2009 to June 2009, he was a Research Scientist with the Department of Research and Innovation, Alcatel Lucent Shanghai Bell. From June 2009 to April 2012, he was a Postdoctoral Fellow with the School of Electrical Engineering and Telecommunications, University of New South Wales, Australia. From April 2012 to June 2015, he was a Research Fellow with the School of Electrical Engineering, University of Sydney, Australia. From June 2015 to the present, he has been a Professor with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, China. His research interests include network information theory, channel coding theory, wireless network coding, and cooperative communications.



Zihuai Lin (S'98–M'06–SM'10) received the Ph.D. degree in electrical engineering from Chalmers University of Technology, Gothenburg, Sweden, in 2006.

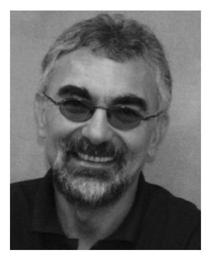
Prior to his Ph.D. degree, he has held positions at Ericsson Research, Stockholm, Sweden. Following having received the Ph.D. degree, he was a Research Associate Professor with Aalborg University, Aalborg, Denmark, and is currently with the School of Electrical and Information Engineering, University of Sydney, Sydney, Australia. His research interests include source/channel/network coding, coded modulation, MIMO, OFDMA, SC-FDMA, radio resource management, cooperative communications, small-cell networks, 5G cellular systems, etc.



Guoqiang Mao (S'98–M'02–SM'08) received the Ph.D. degree in telecommunications engineering from Edith Cowan University, Perth, Australia, in 2002.

Between 2002 and 2014, he was with the School of Electrical and Information Engineering, University of Sydney, Sydney, Australia. He joined the University of Technology Sydney in February 2014 as a Professor of wireless networking and the Director of Center for real-time information networks. The Center is among the largest university research centers in Australia in the field of wireless communications and networking. He has published about 200 papers in international conferences and journals, which have been cited more than 4000 times. His research interest includes intelligent transport systems, applied graph theory and its applications in telecommunications, Internet of Things, wireless sensor networks, wireless localization techniques, and network performance analysis.

Dr. Mao is an Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS (since 2014) and the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY (since 2010). He received Top Editor award for outstanding contributions to the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY in 2011, 2014, and 2015. He is a Cochair of the IEEE Intelligent Transport Systems Society Technical Committee on Communication Networks. He has served as a Chair, Cochair, and Technical Program Committee Member in a large number of international conferences.



Lajos Hanzo (F'08) received the M.S. degree in electronics and the Ph.D. degree from the Technical University of Budapest, Budapest, Hungary, in 1976 and 1983, respectively. He received the prestigious Doctor of Sciences research degree in wireless communications from the University of Southampton, U.K., in 2004.

In 2016, he was admitted to the Hungarian Academy of Science, Budapest, Hungary. During his 40-year career in telecommunications, he has held various research and academic posts in Hungary, Germany, and the U.K. Since 1986, he has been with the School of Electronics and Computer Science, University of Southampton, U.K., where he holds the Chair in telecommunications. He has successfully supervised 111 Ph.D. students, coauthored 20 John Wiley/IEEE Press books on mobile radio communications, totalling in excess of 10 000 pages, published 1600+ research contributions on IEEE Xplore, acted both as Technical Program Committee member and General Chair of IEEE conferences, presented keynote lectures, and received a number of distinctions. Currently he is directing a 60-strong academic research team, working on a range of research projects in the field of wireless multimedia communications sponsored by industry; the Engineering and Physical Sciences Research Council (EPSRC), U.K.; and the European Research Council.s Advanced Fellow Grant. He is an enthusiastic supporter of industrial and academic liaison, and he offers a range of industrial courses. He has 25 000+ citations and an H-index of 60. For further information on research in progress and associated publications, see http://www-mobile.ecs.soton.ac.uk. Dr. Hanzo is also a Governor of the IEEE Vehicular Technology Society. During 2008–2012, he was the Editor-in-Chief of the IEEE Press and a Chaired Professor with Tsinghua University, Beijing, China. In 2009, he received an honorary doctorate award by the Technical University of Budapest and in 2015, from the University of Edinburgh, Edinburgh, U.K., as well as the Royal Society.s Wolfson Research Merit Award. He is a Fellow of the Royal Academy of Engineering, The Institution of Engineering and Technology, and EURASIP.

QUERIES

Q1. Author: Please supply index terms/keywords for your paper. To download the IEEE Taxonomy go to http://www.ieee.org/documents/taxonomy_v101.pdf.
Q2. Author: Please update Ref. [2].
Q3. Author: Please provide page range in Ref. [18].
Q4. Author: Please provide complete bibliographic details in Refs. [29] and [30].

Probabilistic Small-Cell Caching: Performance Analysis and Optimization

Youjia Chen, Ming Ding, *Member, IEEE*, Jun Li, *Senior Member, IEEE*, Zihuai Lin, *Senior Member, IEEE*, Guoqiang Mao, *Senior Member, IEEE*, and Lajos Hanzo, *Fellow, IEEE*

***Abstract*—Small-cell caching utilizes the embedded storage of small-cell base stations (SBSs) to store popular contents for the sake of reducing duplicated content transmissions in networks and for offloading the data traffic from macrocell base stations to SBSs. In this paper, we study a probabilistic small-cell caching strategy, where each SBS caches a subset of contents with a specific caching probability. We consider two kinds of network architectures: 1) The SBSs are always active, which is referred to as the always-on architecture; and 2) the SBSs are activated on demand by mobile users (MUs), which is referred to as the dynamic on–off architecture. We focus our attention on the probability that MUs can successfully download content from the storage of SBSs. First, we derive theoretical results of this successful download probability (SDP) using stochastic geometry theory. Then, we investigate the impact of the SBS parameters, such as the transmission power and deployment intensity on the SDP. Furthermore, we optimize the caching probabilities by maximizing the SDP based on our stochastic geometry analysis. The intrinsic amalgamation of optimization theory and stochastic geometry based analysis leads to our optimal caching strategy, characterized by the resultant closed-form expressions. Our results show that in the always-on architecture, the optimal caching probabilities solely depend on the content request probabilities, while in the dynamic on–off architecture, they also relate to the MU-to-SBS intensity ratio. Interestingly, in both architectures, the optimal caching probabilities are linear functions of the square root of the content request probabilities. Monte-Carlo simulations validate our theoretical analysis and show that the proposed schemes relying on the optimal caching probabilities are capable of achieving substantial SDP improvement, compared with the benchmark schemes.**

***Index Terms*—.**

Manuscript received February 11, 2016; revised June 25, 2016; accepted August 24, 2016. Date of publication; date of current version. This work was supported in part by the Fujian Provincial Natural Science Foundation (2016J01290); in part by the National Natural Science Foundation of China under Grant 61571128 and Grant 61501238; in part by the Jiangsu Provincial Science Foundation (BK20150786); in part by the China Scholarship Council, the Specially Appointed Professor Program in Jiangsu Province, 2015; and in part by the Fundamental Research Funds for the Central Universities (30916011205). The review of this paper was coordinated by Dr. B. Canberk.

Y. Chen is with the School of Electrical and Information Engineering, University of Sydney, Sydney, NSW 2006, Australia, also with Fujian Normal University, Fuzhou 350007, China, and also with the Data61, CSIRO, Canberra ACT 2600, Australia (e-mail: youjia.chen@sydney.edu.au).

M. Ding is with the Data61, CSIRO Canberra ACT, 2600, Australia (e-mail: Ming.Ding@data61.csiro.au).

J. Li is with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Beijing 100044, China (e-mail: jun.li@njust.edu.cn).

Z. Lin is with the School of Electrical and Information Engineering, University of Sydney, Sydney, NSW 2006, Australia, and also with the Data61, CSIRO Canberra ACT, 2600, Australia (e-mail: zihuai.lin@sydney.edu.au).

G. Mao is with the School of Computing and Communications, University of Technology Sydney, Ultimo, NSW 2007, Australia, and also with the Data61, CSIRO Canberra ACT, 2600, Australia (e-mail: g.mao@ieee.org).

L. Hanzo is with the Department of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K. (e-mail: lh@ecs.soton.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/TVT.2016.2606765

I. Introduction

IT IS forecast that at least a 100x network capacity increase will be required to meet the traffic demands in 2020 [1]. As a result, vendors and operators are now looking at using every tool at hand to improve network capacity [2].

In addition, a substantial contribution to the traffic explosion comes from the repeated download of a small portion of popular contents, such as popular movies and videos [3]. Therefore, intelligent caching in wireless networks has been proposed for effectively reducing such duplicated transmissions of popular contents, as well as for offloading the traffic from the overwhelmed macrocells to small cells [4], [5]. Caching in third-generation (3G) and fourth-generation (4G) wireless networks was shown to be able to reduce the traffic by one third to two thirds [6].

Several caching strategies have been proposed for wireless networks. Woo *et al.* [7] analyzed the strategy of caching contents in the evolved packet core of local thermal equilibrium (LTE) networks.The strategy of caching contents in the radio access network, with an aim to place contents closer to mobile users (MUs) was studied in [8] and [9]. The concept of small-cell caching, referred to as “Femtocaching” in [9] and [10], utilized small-cell base stations (SBS) in heterogeneous cellular networks as distributed caching devices. Caching strategies conceived for device-to-device (D2D) networks were investigated in [11]–[13], where the mobile terminals serve as caching devices. The coexistence of small-cell caching and D2D caching is indeed also a hot research direction. In [14], Yang *et al.* considered the joint caching in both the relays and a subset of the mobile terminals, which relies on the coexistence of small-cell caching and D2D caching. Moreover, a coded caching scheme was proposed in [15] to improve system performance.

In this paper, we focus on the small-cell caching because 1) the large number of SBSs in 4G and fifth-generation (5G) networks already provide a promising basis for caching [2]; and 2) compared with D2D caching, small-cell caching has several advantages, such as the abundance of power supply, fewer grave security issues, and more reliable data delivery. As illustrated in Fig. 1, with small-cell caching, popular contents are transmitted and cached in the storage of the SBSs during off-peak hours. Then in peak hours, if an MU can find its requested content in

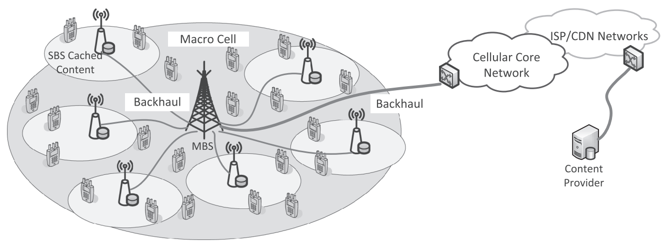


Fig. 1. Small-cell caching.

a nearby SBS, the MU can directly download the content from such SBS.

There are generally two approaches to implement the small-cell caching, i.e., the deterministic content placement and the nondeterministic content placement. In [9], [16], and [17], the deterministic contents placement was analyzed. In these works, the placement of popular contents was optimized using the information of the network node locations and the statistical or instantaneous channel states. However, in practice, the geographic distribution of MUs and the wireless channels are time variant. Thus, the optimal content placement strategy has to be frequently updated in the deterministic content placement, leading to a high complexity and fewer tractable results. On the other hand, the nondeterministic content placement permits simple implementation and has a good tractability. In [18] and [14], the distributions of SBSs and MUs were modeled as homogeneous Poisson point processes (HPPPs) to obtain a general performance analysis for the small-cell caching. However, in these works, all the SBSs were assumed to cache the same copy of certain popular contents. In [11], probabilistic content placement was proposed and analyzed in the context of D2D caching, where each mobile terminal caches a specific subset of the contents with a given caching probability. The throughput versus outage tradeoff was analyzed and the optimal caching distribution was derived for a grid network relying on a particular protocol model. The idea of probabilistic content placement was also investigated in the coded multicasting system [19]. Compared with caching the same copy of certain popular contents in all the SBSs, probabilistic content placement in small-cell caching can provide more flexibility. Therefore, in this paper, we focus on small-cell caching relying on probabilistic content placement, shortened as probabilistic small-cell caching (PSC) for brevity.

In small-cell networks, there are two network architectures, namely, the always-on architecture and the dynamic on–off architecture. The always-on architecture is a common practice in the current cellular networks, where all the SBSs are always active. By contrast, in the dynamic on–off architecture, the SBSs are only active, when they are required to provide services to nearby MUs [20]. Aiming for saving energy consumption and mitigating unnecessary intercell interference, the dynamic on–off architecture has been proposed and it is currently under investigation in 3GPP as an important candidate of 5G technologies in future dense and ultradense small-cell networks [2], [21], [22]. Energy consumption is of critical interest in future 5G systems [23], [24], especially in ultradense networks. Compared with the power-thirsty always-on architecture, where the energy consumption grows with the network's densification, the energy consumption of the ultradense network relying on the dynamic on–off architecture mainly depends on the density of MUs in the network [2]. The in-depth investigation of the associated energy consumption issues of wireless caching will constitute our future work.

Against this background, we study the PSC under the abovementioned pair of network architectures. First, we use a stochastic geometry to develop theoretical results of the probability $\Pr(\mathcal{D})$ that MUs can successfully download contents from the storage of SBSs. Second, we investigate the impact of the SBSs' parameters on $\Pr(\mathcal{D})$, namely, that of the transmission power P and of the deployment intensity λ_s. In the always-on architecture, although $\Pr(\mathcal{D})$ monotonically increases with either P or λ_s, it approaches a constant when P or λ_s is sufficiently high. In the dynamic on–off architecture, $\Pr(\mathcal{D})$ reaches a constant when P is high enough, while it keeps on increasing as λ_s grows. Most importantly, we optimize the caching probabilities for maximizing $\Pr(\mathcal{D})$ in the pair of network architectures considered. We emphasize that it is quite a challenge to apply optimization theory to an objective function obtained from stochastic geometry analysis, especially to derive a closed-form expression for the optimal solution. Our results will demonstrate that in the always-on architecture, the optimal subset of contents to be cached depends on the content request probabilities, while in the dynamic on–off architecture, it also depends on the MU-to-SBS intensity ratio. Most interestingly, in both architectures, the optimal caching probabilities can be expressed as linear functions of the square root of the content request probabilities.

The rest of the paper is structured as follows. In Section II we describe the system model, while in Section III we present the definition of PSC and formulate the probability that MUs can successfully download contents from the storage of SBSs. The main analytical results characterizing this successful download probability (SDP) are presented in Section IV. In Section V, we optimize the caching probabilities in both of the network architectures for maximizing the derived SDP. The accuracy of the analytical results and the performance gains of optimization are characterized by simulations in Section VI. Finally, our conclusions are offered in Section VII.

II. System Model

We consider a cellular network supporting multiple MUs by the SBSs operating within the same frequency spectrum. We model the distribution of the SBSs and that of the MUs as two independent HPPPs, with the intensities of λ_s and λ_u, respectively. The transmission power of the SBSs is denoted by P. The path loss of the channel spanning from an SBS to an MU is modeled as $d^{-\alpha}$, where d denotes the distance between them, and α denotes the path-loss exponent. The multipath fading is modeled as Rayleigh fading with a unit power, and hence the channel's power gain is denoted by $h \sim \exp(1)$. All the channels are assumed to be independently and identically distributed.

A. Network Architectures

We consider two network architectures.

1) Always-On Architecture: In this architecture, all the SBSs are assumed to be active, i.e., all the SBSs are continuously transmitting signals. This architecture is commonly employed in the operational cellular networks [25]. The rationale for this architecture is that the number of SBSs is usually much lower than that of MUs, and thus each and every SBS has to be turned ON to serve the MUs in its coverage.

2) Dynamic On–Off Architecture: In this architecture, an SBS will be active only when it has to provide services to its associated MUs. In future 5G networks, the intensity of deployed SBSs is expected to be comparable to or even potentially higher than the intensity of MUs [2]. In such ultradense networks, having an adequate received signal coverage is always guaranteed, since the distance between an MU and its serving SBS is short, but the interference becomes the dominant issue. With the goal of mitigating the potentially avoidable intercell interference and saving energy, the dynamic on–off architecture has been identified as one of the key technologies in 5G networks [20]. With the dynamic on–off architecture, an SBS will switch to its idle mode, i.e., turn OFF its radio transmission, if there is no MU associated with it, otherwise, it will switch back to the active mode.

B. File Request Model

We consider a contents library consisting of M different files. Note that M does not represent the number of files available on the Internet, but the number of popular files that the MUs tend to access. We denote by q_m the probability that the mth file $\mathcal{F}_m$ will be requested. By stacking q_m into $\{q_m : m = 1, \cdots, M\}$, we can get the probability mass function (PMF) of requesting the M files. According to [26], the request- PMF of the files can be modeled as a Zipf distribution. More specifically, for $\mathcal{F}_m$, its request probability q_m is written as

$$q_m = \frac{\frac{1}{m^{\beta}}}{\sum_{i=1}^{M} \frac{1}{i^{\beta}}} \tag{1}$$

where β is the exponent of the Zipf distribution and a large β implies having an uneven popularity among those files. From (1), q_m tends to zero, as $M \to \infty$ when $\beta < 1$, while it converges to a constant value when $\beta > 1$. Note that (1) implies that the indices of the files are not randomly generated, but follow a descending order of their request probabilities.

Due to the limited storage of SBSs, an SBS is typically unable to cache the entire file library. Therefore, we assume that the library is partitioned into N nonoverlapping subsets of files, referred to as file groups (FGs), and each SBS can cache only one of the N FGs. Note that the same FG can be redundantly stored in multiple SBSs. The scenario of FGs with overlapping subsets of files will be considered later, which will be compared with the nonoverlapping scenario. We denote the nth FG, $n \in \{1, \cdots, N\}$ by $\mathcal{G}_n$. The probability Q_n that an MU requests a file in FG $\mathcal{G}_n$, is thus given by

$$Q_n = \sum_{m,\ \mathrm{for} \mathcal{F}_\mathrm{m} \in \mathcal{G}_\mathrm{n}} q_m. \tag{2}$$

III. Probabilistic Small-Cell Caching Strategy

In this section, we introduce the PSC strategy, and formulate the probability that MUs can successfully download contents from the storage of the SBSs, which is an important performance metric of small-cell caching.

Generally, caching consists of two phases: a contents placement phase and a contents delivery phase [27]. In the contents placement phase, popular contents are transmitted and cached in the storage units of network devices that are close to MUs. In the contents delivery phase, the popular cached contents can be promptly retrieved for serving the MUs.

A. Contents Placement Phase

In the content placement phase of PSC, each SBS independently caches FG $\mathcal{G}_n$ with a specific caching probability, denoted by S_n. Hence, from the perspective of the entire network, the fraction of the SBSs that caches $\mathcal{G}_n$ equals to S_n. Since the distribution of SBSs in the network is modeled as an HPPP with the intensity of λ_s, according to the thinning theorem of HPPP [28], we can view the distribution of SBSs that cache $\mathcal{G}_n$ as a thinned HPPP with the intensity of $S_n \lambda_s$.

We assume that at a particular time instant, an MU can only request one file, and hence, the distribution of MUs who request the files in $\mathcal{G}_n$ can also be modeled as a thinned HPPP with the intensity $Q_n \lambda_u$. We treat the SBSs that cache $\mathcal{G}_n$ together with the MUs that request the files in $\mathcal{G}_n$ as the nth tier of the network, shortened as Tier-n.

B. Contents Delivery Phase

During the contents delivery phase, an MU that requests a file in $\mathcal{G}_n$ will associate with the nearest SBS that caches $\mathcal{G}_n$, and then attempts to download the file from it. We assume that only when the received signal-to-interference-and-noise-ratio (SINR) at the MU is above a prescribed threshold, can the requested file be successfully downloaded.

If the MU cannot download the requested file from the cached SBS, the requested file would be transmitted to the MU from a remote content provider, which means the data should flow across the Internet, the cellular core network, and the backhaul network, as illustrated in Fig. 1.

C. Probability of Successful Download

Recent surveys show that 96% of the operators consider backhaul as one of the most important challenges to small-cell deployments, and this issue is exacerbated in ultradense networks [29], [30]. If an MU can successfully download a requested file from storages of SBSs, the usage of the backhaul network will be greatly reduced and the transmission latency of a requested file will be significantly shortened. Therefore, we assume that a successful download of a requested file from storages of SBSs is always beneficial to the network performance. Accordingly, we focus on our attention on this SDP as the performance metric for small-cell caching in the following.

According to Slyvnyak's theorem for HPPP [28], an existing point in the process does not change the statistical distribution of other points of the HPPP. Therefore, the probability that an MU in Tier-n can successfully download the contents from SBSs can be obtained by analyzing the probability that a *typical* MU in Tier-n, say located at the origin, can successfully download the contents from its associated SBS in Tier-n.

When the MU considered requests a file in $\mathcal{G}_n$, its received SINR from its nearest SBS in Tier-n can be formulated as

$$\gamma_n(z) = \frac{P h_{x_0} z^{-\alpha}}{\sum_{x_j \in \Phi \setminus \{x_0\}} P h_{x_j} \|x_j\|^{-\alpha} + \sigma^2} \tag{3}$$

where σ^2 denotes the Gaussian noise power, z is the distance between the typical MU and its nearest SBS in Tier-n, x_j represents the locations of the interfering SBSs, Φ denotes the set of simultaneously active SBSs, and x_0 is the location of the serving BS at a distance of z. Additionally, $\|x_j\|$ denotes the distance between x_j and the typical MU, while h_{x_0} and h_{x_j} denote the corresponding channel gains.

Since the intercell interference is the dominant factor determining the signal quality in the operational cellular networks, especially when unity frequency reuse has been adopted for improving the spectrum efficiency, the minimum received SINR is used as the metric of successful reception. Let δ be the minimum SINR required for successful transmissions and $\mathcal{D}_n$ be the event that the typical Tier-n MU successfully receives the requested file from the associated Tier-n SBS. Then, the probability of $\mathcal{D}_n$ can be formulated as

$$\Pr(\mathcal{D}_n) = \Pr[\gamma_n(z) \geq \delta]. \tag{4}$$

Considering the request probabilities of $\mathcal{G}_n$ and based on the result of $\Pr(\mathcal{D}_n)$, we obtain the average probability that the MUs can successfully download contents from the storage of the SBSs, denoted by $\Pr(\mathcal{D})$, as

$$\Pr(\mathcal{D}) = \sum_{n=1}^{N} Q_n \cdot \Pr(\mathcal{D}_n). \tag{5}$$

In essence, $\Pr(\mathcal{D})$ quantifies the weighted sum of the SDP, where the weights are the request probabilities reflecting the importance of the files.

IV. Performance Analysis of Small-Cell Caching

In this section, we derive the SDP $\Pr(\mathcal{D})$ for the pair of network architectures. Some special cases are also considered with an aim to obtain more insights into the design of PSC.

A. Always-On Architecture

Our main result on the probability $\Pr(\mathcal{D})$ for the always-on architecture is summarized in Theorem 1.

Theorem 1: In the always-on architecture, the probability $\Pr(\mathcal{D})$ is given by

$$\begin{aligned}\Pr(\mathcal{D}) &= \sum_{n=1}^{N} Q_n \Pr(\mathcal{D}_n) \\ &= \sum_{n=1}^{N} Q_n \int_0^{\infty} \pi S_n \lambda_s \exp\left(-\frac{z^{\alpha}\delta\sigma^2}{P}\right) \\ &\exp\left(-\pi\lambda_s z^2 \left((1-S_n)C(\delta,\alpha) + S_n A(\delta,\alpha) + S_n\right)\right) dz^2\end{aligned} \tag{6}$$

where $A(\delta,\alpha) \triangleq \delta \frac{2}{\alpha-2}\, {}_2F_1(1, 1-\frac{2}{\alpha}; 2-\frac{2}{\alpha}; -\delta)$, and $C(\delta,\alpha) \triangleq \frac{2}{\alpha}\delta^{\frac{2}{\alpha}} B(\frac{2}{\alpha}, 1-\frac{2}{\alpha})$. Furthermore, ${}_2F_1(\cdot)$ denotes the hypergeometric function, and $B(\cdot)$ represents the beta function [31].

Proof: See Appendix A. ■

From (6), we conclude that the probability $\Pr(\mathcal{D})$ increases as the transmission power P grows, because $\exp(-\frac{z^{\alpha}\delta\sigma^2}{P})$ increases with P. Since it remains a challenge to obtain deeper insights from (6), which is not a closed-form expression, two special cases are examined in the sequel to gain deeper insight on the performance behavior of $\Pr(\mathcal{D})$.

1) Path-Loss Exponent $\alpha = 4$: According to 3GPP measurement [32], the typical value of the path-loss exponent for SBSs in practical environments is around 4. Substituting this typical value of $\alpha = 4$ into (6), we have

$$\begin{aligned}\Pr(\mathcal{D})\mid_{\alpha=4} = \sum_{n=1}^{N} Q_n \pi S_n \sqrt{\frac{\pi}{4\delta}\frac{P\lambda_s^2}{\sigma^2}}\,\mathrm{erfc}x\Bigg(\frac{\pi}{2}\cdot \\ \sqrt{\frac{P\lambda_s^2}{\delta\sigma^2}}\left(S_n + \frac{\pi}{2}\sqrt{\delta}(1-S_n) + S_n\sqrt{\delta}\arctan\sqrt{\delta}\right)\Bigg)\end{aligned} \tag{7}$$

where $\mathrm{erfc}x(x) \triangleq \exp(x^2)\mathrm{erfc}(x)$ is the scaled complementary error function [33].

Regarding the relationship between $\Pr(\mathcal{D})$ and λ_s, we propose Corollary 1.

Corollary 1: In the always-on architecture, for the special case of $\alpha = 4$, $\Pr(\mathcal{D})$ monotonically increases with the increase of λ_s.

Proof: See Appendix B. ■

From the results obtained in (6) that $\Pr(D)$ increases as P grows, and based on Corollary 1, we conclude that when $\alpha = 4$, the SDP $\Pr(\mathcal{D})$ can be improved by either increasing the SBSs' transmission power P or the SBSs' deployment intensity λ_s. Furthermore, since (7) can be viewed as a function of the variable $P\lambda_s^2$, the effect of increasing P to kP on $\Pr(\mathcal{D})$ is equivalent to increasing λ_s to $\sqrt{k}\lambda_s$, where k is a positive constant.

Moreover, according to the property of the function $\mathrm{erfc}x(x)$, i.e., $\lim_{x\to\infty}\mathrm{erfc}x(x) = \frac{1}{\sqrt{\pi}x}$, we have

$$\begin{aligned}&\lim_{P\to\infty}\Pr(\mathcal{D})\mid_{\alpha=4} = \lim_{\lambda_s\to\infty}\Pr(\mathcal{D})\mid_{\alpha=4} \\ &= \sum_{n=1}^{N}\frac{Q_n S_n}{\frac{\pi}{2}\sqrt{\delta} + (\sqrt{\delta}\arctan\sqrt{\delta} + 1 - \frac{\pi}{2}\sqrt{\delta})S_n}.\end{aligned} \tag{8}$$

From (8), we have Remark 1.

Remark 1: In the always-on architecture, given σ^2 and δ, the value of $\Pr(\mathcal{D})$ monotonically grows with the increase of P and λ_s, and it converges to a constant, when P or λ_s is sufficiently large.

2) Neglecting Noise, i.e., $\sigma^2 = 0$: In an interference-limited network, where the noise level is much lower than the interference, the impact of the noise can be neglected. In such cases, we assume that $\sigma^2 = 0$, and it follows that $\Pr(\mathcal{D})$ in (6) can be rewritten as

$$\Pr(\mathcal{D})\mid_{\sigma^2\to 0} = \sum_{n=1}^{N}\frac{Q_n S_n}{S_n A(\delta,\alpha) + (1-S_n)C(\delta,\alpha) + S_n}. \tag{9}$$

From (9), we have Remark 2.

Remark 2: In the always-on architecture operating in an interference-limited network, the probability of successful download depends only on the request probabilities and caching probabilities of the FGs, i.e., Q_n and S_n.

Note that in the scenario, where the different FGs may have an overlapping subset of files, the probability $\Pr(\mathcal{D})$ still has the same formulation as (6). However, all the subscripts n in (6) should be changed to m, because we should consider both the request probability and the caching probability of each file $\mathcal{F}_m$, i.e., S_m and Q_m, instead of each FG $\mathcal{G}_n$. Therefore, in this scenario, the specific SBSs that cache $\mathcal{F}_m$ and the MUs that request $\mathcal{F}_m$ are viewed as Tier-m. Since all the derivations are the same, our main results summarized in Theorem 1 as well as the aforementioned corollary and remarks, are still valid in conjunction with the subscript m. Hence we omit the analysis for this scenario with overlapping subsets of files for brevity.

B. Dynamic On–Off Architecture

As mentioned, in the dynamic on–off architecture an SBS is only active, when it has to provide services for the associated MUs. Specifically, an SBS in Tier-n is only active, when there is at least one MU in Tier-n located in its Voronoi cell. Hence, the probability that an SBS in Tier-n is active, which is denoted by $\Pr(\mathcal{A}_n)$, should be considered for the dynamic on–off architecture.

Our main result on the probability $\Pr(\mathcal{D})$ for the dynamic on–off architecture is summarized in Theorem 2.

Theorem 2: In the dynamic on–off architecture, the probability $\Pr(\mathcal{D})$ is given by

$$\begin{aligned}\Pr(\mathcal{D}) &= \sum_{n=1}^{N} Q_n \Pr(\mathcal{D}_n)\\ &= \sum_{n=1}^{N} Q_n \int_0^\infty \pi S_n \lambda_s \exp\left(-\frac{z^\alpha \delta \sigma^2}{P}\right) \exp\Bigg(-\pi \lambda_s z^2 \Bigg(\sum_{i=1, i\neq n}^{N}\\ &\quad \Pr(\mathcal{A}_i) S_i C(\delta,\alpha) + \Pr(\mathcal{A}_n) S_n A(\delta,\alpha) + S_n\Bigg)\Bigg) dz^2 \qquad (10)\end{aligned}$$

where $\Pr(\mathcal{A}_n)$ denotes the probability that an SBS in Tier-n is in the active mode, and

$$\Pr(\mathcal{A}_n) \approx 1 - \left(1 + \frac{Q_n \lambda_u}{3.5 S_n \lambda_s}\right)^{-3.5}. \qquad (11)$$

Proof: See Appendix C. ■

Compared to $\Pr(\mathcal{D})$ in the always-on architecture, $\Pr(\mathcal{D})$ in the dynamic on–off architecture also depends on the intensity of the MUs λ_u. The reason behind this is that the number of active SBSs in the network depends on the number of MUs in the network.

From (10), we have Remark 3.

Remark 3: In the dynamic on–off architecture, given σ^2 and δ, the value of $\Pr(\mathcal{D})$ monotonically increases with the increase of the transmission power P.

1) Neglecting Noise, i.e., $\sigma^2 = 0$: In an interference-limited network, substituting $\sigma^2 = 0$ into (10), we have

$$\Pr(\mathcal{D})\,|_{\sigma^2 \to 0} = \sum_{n=1}^{N} \frac{Q_n S_n}{\Pr(\mathcal{A}_n) S_n A(\delta,\alpha) + \sum_{i=1, i\neq n}^{N} \Pr(\mathcal{A}_i) S_i C(\delta,\alpha) + S_n}. \qquad (12)$$

From (12), we have Remark 4.

Remark 4: In the dynamic on–off architecture operating in an interference-limited network, the probability of successful download $\Pr(\mathcal{D})$ is independent of P, and depends only on Q_n, S_n as well as on the MU-to-SBS intensity ratio λ_u/λ_s.

When considering the scenario of FGs with overlapping subsets of files, the average probability $\Pr(\mathcal{D})$ cannot be formulated as the sum of $\Pr(\mathcal{D}_n)$ as in (5). Furthermore, we cannot formulate $\Pr(\mathcal{D})$ as $\Pr(\mathcal{D}) = \sum_{m=1}^{M} \Pr(\mathcal{D}_m)$, which we propose for the overlapping scenario in the always-on architecture. This is because in the dynamic on–off architecture the active probability of an SBS depends on the specific FG that it caches. Therefore, the analysis of $\Pr(\mathcal{D})$ in the dynamic on–off architecture considering the scenario with overlapping subsets of files requires further investigations as part of our future research.

V. Optimization of the Caching Probability

A larger $\Pr(\mathcal{D})$ always benefits the network because of 1) the backhaul saving and 2) the low-latency transmission of local contents from SBSs [2]. Based on such facts, in this section, we concentrate on maximizing $\Pr(\mathcal{D})$ by optimally designing the caching probabilities of the contents in the system, denoted by $\{S_n^{\text{Opt}} : n = 1, \ldots, N\}$.

Note that there is a paucity of literature on applying optimization theory relying on an objective function obtained from stochastic geometry analysis, especially, when aiming for deriving a closed-form expression of the optimal solution. In order to facilitate this optimization procedure, we ensure the mathematical tractability of the objective function by using a simple user association strategy and neglect the deleterious effects of noise.

A. Always-On Architecture

From (9), we can formulate the optimization problem of maximizing $\Pr(\mathcal{D})$ as

$$\begin{aligned}\max_{\{S_n\}}\ \Pr(\mathcal{D}) &= \max_{\{S_n\}} \sum_{n=1}^{N} \frac{Q_n S_n}{(1-S_n) C(\delta,\alpha) + S_n A(\delta,\alpha) + S_n}\\ \text{s.t.} &\sum_{n=1}^{N} S_n = 1\\ & S_n \geq 0,\ n = 1, \ldots, N.\end{aligned} \qquad (13)$$

The solution of Problem (13) is presented in Theorem 3.

Theorem 3: In the always-on architecture, the optimal caching scheme, which is denoted by the file caching PMF $\{S_n^{opt}\}$, that maximizes the average probability of successful download, is given by

$$S_n^{opt} = \left\lceil \frac{\sqrt{\frac{Q_n}{\xi}} - C(\delta,\alpha)}{A(\delta,\alpha) - C(\delta,\alpha) + 1} \right\rceil^{+}, \; n = 1, \ldots, N \qquad (14)$$

where $\sqrt{\xi} = \frac{\sum_{n=1}^{N^*} \sqrt{Q_n}}{(N^*-1)C(\delta,\alpha)+A(\delta,\alpha)+1}$, $\lceil \Omega \rceil^{+} \triangleq \max\{\Omega, 0\}$, and N^*, $1 \le N^* \le N$ satisfies the constraint that $S_n \ge 0 \; \forall n$.

Proof: It can be shown that the optimization Problem (13) is concave and can be solved by invoking the Karush−Kuhn−Tucker conditions [34]. The conclusion then follows. ■

From (14), when the request probability obeys $Q_n > \xi C^2(\delta,\alpha)$, $\mathcal{G}_n$ is cached with a caching probability of S_n^{opt}, otherwise, it is not cached. This optimal strategy implies that ideally the SBSs should cache the specific files with high request probabilities, while those files with low request probabilities should not be cached at all due to the limited storage of SBSs in the network. Moreover, we can see that from (14) the optimal caching probability of an FG is a linear function of the square root of its request probability.

Regarding the scenario of FGs associated with overlapping subsets of files, as we mentioned before, $\Pr(\mathcal{D})$ in this scenario has the same formulation as that in the nonoverlapping scenario. Therefore, the optimal caching probability of $\mathcal{F}_m$ in the scenario of FGs having overlapping subsets of files can be formulated as

$$S_m^{\text{Opt}} = \min\left\{ \left\lceil \frac{\sqrt{\frac{Q_m}{\xi}} - C(\delta,\alpha)}{A(\delta,\alpha) - C(\delta,\alpha) + 1} \right\rceil^{+}, 1 \right\} \qquad (15)$$

where $\sqrt{\xi} = \frac{\sum_{m=1}^{M^*} \sqrt{Q_m}}{(M^*-V)C(\delta,\alpha)+V(A(\delta,\alpha)+1)}$, and $M^*(1 \le M^* \le M)$, satisfies the constraint that $0 \le S_m \le 1 \; \forall m$, and V denotes the number of files in each FG.

Compared with the nonoverlapping scenario, the presence of overlapping subsets among the FGs provides a higher grade of diversity in the system. However, based on our simulations to be discussed in the sequel, we find that the gain of maximum $\Pr(\mathcal{D})$ obtained as a benefit of this diversity is limited, while the algorithm associated with the optimal caching strategy of (15) is more complex than that of (14).

B. Dynamic On–Off Architecture

In this architecture, as shown in (11), the probability $\Pr(\mathcal{A}_n)$ that an SBS in Tier-n is in the active mode, is a function of the ratio $Q_n \lambda_u / S_n \lambda_s$. Since the intensity of SBSs is much higher than the intensity of the MUs in this architecture, i.e., we have $\lambda_s \gg \lambda_u$, the SBS activity probability $\Pr(\mathcal{A}_n)$ in (11) can be approximated as

$$\Pr(\mathcal{A}_n) \approx \frac{Q_n \lambda_u}{S_n \lambda_s}. \qquad (16)$$

Substituting (16) into (12) and (5), we can formulate the optimization problem of maximizing the successful downloading probability as

$$\max_{\{S_n, \varepsilon_n\}} \Pr(\mathcal{D}) =$$

$$\max_{\{S_n, \varepsilon_n\}} \sum_{n=1}^{N} \frac{Q_n S_n}{Q_n \frac{\lambda_u}{\lambda_s} A(\delta,\alpha)\cdot\varepsilon_n + \sum_{i:i\neq n} Q_i \frac{\lambda_u}{\lambda_s} C(\delta,\alpha)\cdot\varepsilon_i + S_n}$$

$$\text{s.t.} \; \sum_{n=1}^{N} S_n = 1$$

$$S_n \ge 0, \; n = 1, \ldots, N$$

$$\varepsilon_n = \begin{cases} 1, & \text{if} \quad S_n > 0 \\ 0, & \text{if} \quad S_n = 0. \end{cases} \qquad (17)$$

Different from the optimization problem in (13), the variable ε_n is introduced to indicate whether $\mathcal{G}_n$ is cached. Due to the existence of ε_n, which implies 2^N hypotheses of file caching states, Problem (17) is difficult to solve. Nevertheless, we manage to find the solution and summarize it in Theorem 4.

Theorem 4: The optimal caching scheme, i.e., the optimal file caching PMF $\{S_n^{Opt}\}$, that maximizes the average probability of successful download, is given by

$$S_n^{Opt} = \begin{cases} \zeta_K \sqrt{Q_n \xi_K C(\delta,\alpha) - Q_n^2 (C(\delta,\alpha) - A(\delta,\alpha))} \\ - \left(\xi_K \frac{\lambda_u}{\lambda_s} C(\delta,\alpha) - Q_n \frac{\lambda_u}{\lambda_s} (C(\delta,\alpha) - A(\delta,\alpha)) \right), & n \le K \\ 0, & K < n \le N. \end{cases} \qquad (18)$$

where

$$\xi_K \triangleq \sum_{i=1}^{K} Q_i$$

$$\zeta_K \triangleq \frac{1 + K\xi_K \frac{\lambda_u}{\lambda_s} C(\delta,\alpha) - \xi_K \frac{\lambda_u}{\lambda_s} (C(\delta,\alpha) - A(\delta,\alpha))}{\sum_{i=1}^{K} \sqrt{Q_i \xi_K C(\delta,\alpha) - Q_i^2 (C(\delta,\alpha) - A(\delta,\alpha))}}. \qquad (19)$$

Regarding K, we have

$$K = \arg\max_{k} \left\{ D_k : k = 1, 2, \ldots, \widehat{N} \right\} \qquad (20)$$

where

$$D_k \triangleq \xi_k - \frac{\frac{\lambda_u}{\lambda_s} \left(\sum_{n=1}^{k} \sqrt{Q_n \xi_k C(\delta,\alpha) - Q_n^2 (C(\delta,\alpha) - A(\delta,\alpha))} \right)^2}{1 + k\xi_k \frac{\lambda_u}{\lambda_s} C(\delta,\alpha) - \xi_k \frac{\lambda_u}{\lambda_s} (C(\delta,\alpha) - A(\delta,\alpha))} \qquad (21)$$

and

$$\widehat{N} = \begin{cases} N, & \text{if} \;\; \frac{\lambda_u}{\lambda_s} < a_N \\ N-1, & \text{if} \;\; a_N \le \frac{\lambda_u}{\lambda_s} < a_{N-1} \\ \cdots \\ 1, & \text{if} \;\; a_2 \le \frac{\lambda_u}{\lambda_s}. \end{cases} \qquad (22)$$

Algorithm 1: Optimal Caching Probabilities in the Dynamic On–Off Architecture.

1: Set $j = N$.
2: Compute $\xi_j = \sum_{i=1}^{j} Q_i$, and ϑ_j and a_j in (24) and (23).
3: Compare $\frac{\lambda_u}{\lambda_s}$ with a_j. If $\frac{\lambda_u}{\lambda_s} < a_j$, go to Step 4; otherwise, set $j = j - 1$ and go to Step 2.
4: Set $\widehat{N} = j$.
5: Compute $\xi_k = \sum_{i=1}^{k} Q_i$ and D_k in (21), $k = 1, \cdots, \widehat{N}$.
6: Set $K = \arg\max_k \{D_k\}$.
7: Compute ξ_K and ζ_K in (19), then compute S_n^{Opt} in (18).

Furthermore, the segmentation parameter a_j, $j = 2, \ldots, N$ is given by

$$a_j = \frac{\vartheta_j}{(\vartheta_j \xi_j - Q_j)(C(\delta,\alpha) - A(\delta,\alpha)) + (1 - j\vartheta_j)\xi_j C(\delta,\alpha)} \tag{23}$$

where

$$\vartheta_j \triangleq \frac{\sqrt{Q_j \xi_j C(\delta,\alpha) - Q_j^2 (C(\delta,\alpha) - A(\delta,\alpha))}}{\sum_{i=1}^{j} \sqrt{Q_i \xi_j C(\delta,\alpha) - Q_i^2 (C(\delta,\alpha) - A(\delta,\alpha))}}. \tag{24}$$

Proof: See Appendix D. ■

To get a better understanding of Theorem 4, we propose Algorithm 1 to implement Theorem 4.

From Theorem 4, we have the following remarks.

Remark 5: In the always-on architecture, the optimal number of FGs to be cached depends only on $\{Q_n : n = 1, \cdots, N\}$. By contrast, in the dynamic on–off architecture, the optimal number of FGs to be cached depends not only on $\{Q_n\}$ but on the MU-to-SBS intensity ratio λ_u/λ_s in the network as well.

Remark 6: According to (22), given λ_u, more FGs tend to be cached in the SBSs, when λ_s becomes higher. Moreover, when the intensity of SBSs is not sufficiently high to cache all the FGs, the SBSs should cache the specific files with relatively high request probabilities, which is consistent with the conclusion for the always-on architecture.

Remark 7: In (18), with a practical region of the SINR threshold and path-loss exponent from 3GPP, i.e., for $\delta \in [0.5, 3]$ and $\alpha \in (2, 4]$, we have $\xi_K C(\delta,\alpha) \gg Q_n(C(\delta,\alpha) - A(\delta,\alpha))$, and the optimal caching probability $S_n^{Opt} \approx \zeta_K \sqrt{Q_n \frac{\lambda_u}{\lambda_s} \xi_K C(\delta,\alpha)} - \xi_K \frac{\lambda_u}{\lambda_s} C(\delta,\alpha)$. From (14) and (18), it is interesting to observe that the optimal caching scheme in both the always-on architecture and in the dynamic on–off architecture follow a square root law, i.e., S_n^{Opt} is a linear function of $\sqrt{Q_n}$.

VI. Numerical and Simulation Results

In this section, we present both our numerical and Monte-Carlo simulation results of $\Pr(\mathcal{D})$ in various scenarios. In the Monte-Carlo simulations, the performance is averaged over 1000 network deployments, where in each deployment SBSs and MUs are randomly distributed in an area of 5 × 5 km according to an HPPP distribution. The intensity of MUs in the network is 200/km^2. The transmission power of the SBSs, the noise power, the path-loss exponent, and the SINR threshold are set to 30 dBm, –104 dBm, 4 and 0.25(-6 dB), respectively [32]. In the simulations of the always-on architecture, the deployment intensity of SBSs is set to 80/km^2, while in the simulations of the dynamic on–off architecture, the intensity is set to 400/km^2.

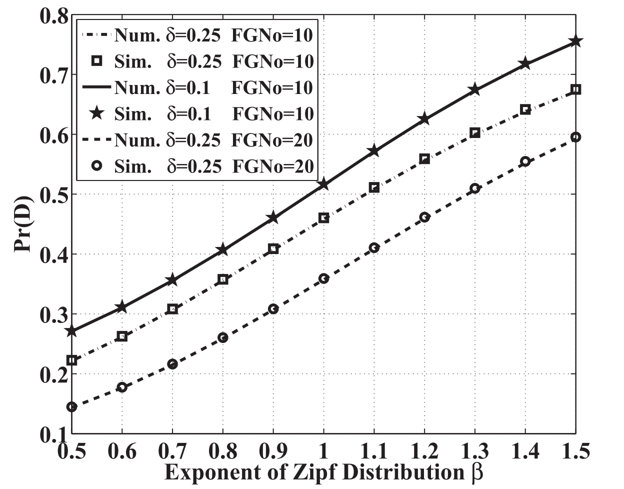


Fig. 2. Numerical and simulation results of $\Pr(\mathcal{D})$ of the O-PSC strategy in the always-on architecture.

Furthermore, we consider a file library consisting of $M = 100$ files, and we partition the file library into $N = 10$ FGs with a simple grouping strategy that the mth file belongs to $\mathcal{G}_n$ if $m \in [\frac{M}{N}(n-1)+1, \ldots, \frac{M}{N}n]\ \forall n \in \{1, \ldots, N\}$. Note that the specific choice of the file grouping strategy is beyond the scope of this paper and it does not affect our results, because it only changes the specific values of the request-PMF $\{Q_n\}$.

In addition, we consider the following two PSC strategies.

1) The request probability based PSC (RP-PSC) [12], where the caching probability of one FG equals to its request probability, i.e., $S_n = Q_n$. Intuitively, a particular FG is more popular than another, the RP-PSC strategy will designate more SBSs to cache it. This strategy is evaluated as a benchmark in our simulations.
2) The proposed optimized PSC (O-PSC) based on (14) in the always-on architecture and (18) in the dynamic on–off architecture, where $S_n = S_n^{\text{opt}}$.

A. Always-On Architecture

Fig. 2 compares the numerical and the simulation results concerning $\Pr(\mathcal{D})$ of the O-PSC strategy. First, it can be seen that the numerical results closely match the simulation results in all scenarios. In the following, we will focus on the analytical results only, due to the accuracy of our analytical results. Second, $\Pr(\mathcal{D})$ increases with the Zipf exponent β. With a larger β, the request probabilities of files are more unevenly distributed. In such cases, a few FGs dominate the requests and caching such popular FGs gives a large $\Pr(\mathcal{D})$. Third, $\Pr(\mathcal{D})$ will be lower, if the value of δ becomes higher. This is because when the SINR threshold is increased, the probability that the received SINR from the SBS storing the file exceeds this threshold is reduced. Finally, we can see that $\Pr(\mathcal{D})$ increases as the number of FGs

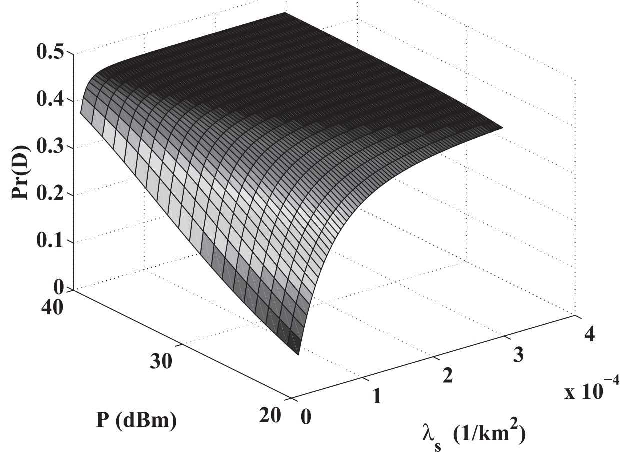


Fig. 3. $\Pr(\mathcal{D})$ of the O-PSC strategy with different P and λ_s in the always-on architecture.

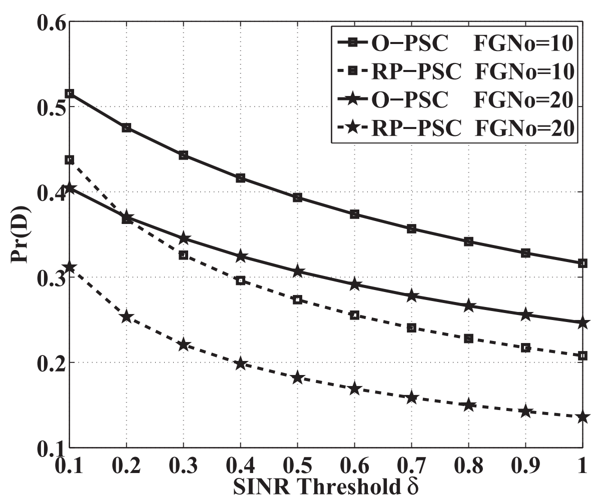


Fig. 4. Comparison of $\Pr(\mathcal{D})$ versus δ of the RP-PSC and O-PSC strategies in the always-on architecture.

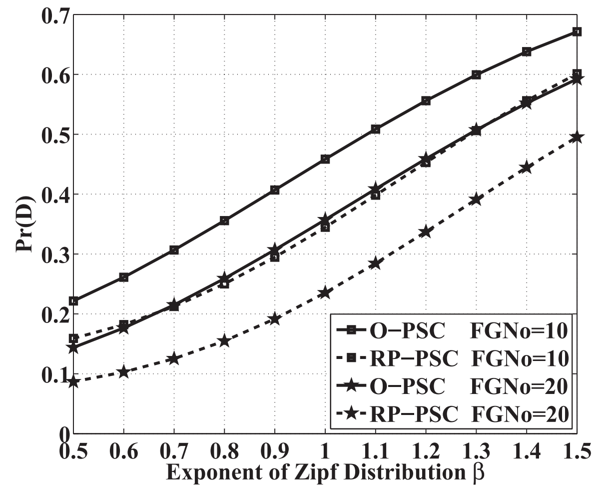


Fig. 5. Comparison of $\Pr(\mathcal{D})$ versus β of the RP-PSC and O-PSC strategies in the always-on architecture.

decreases. Since each SBS only caches one FG, decreasing the number of FGs implies that each SBS caches more files. Hence, this $\Pr(\mathcal{D})$ improvement comes from increasing the stored contents in each SBS.

Fig. 3 shows the SDP $\Pr(\mathcal{D})$ for the O-PSC strategy when the transmission power P of SBSs varies within 20–40 dBm and the deployment intensity λ_s of SBSs varies within 10–400/km^2. To highlight the asymptotic behavior of $\Pr(\mathcal{D})$ with the growth of P, we set the noise power to -50 dBm. We can see from the figure that $\Pr(\mathcal{D})$ increases monotonically with P or λ_s. The value of $\Pr(\mathcal{D})$ remains constant, when P or λ_s is sufficiently high. This result illustrates the limit of $\Pr(\mathcal{D})$ in the always-on architecture shown in (8).

In Fig. 4, we plot $\Pr(\mathcal{D})$ versus the SINR threshold δ to compare the performances of the RP-PSC and O-PSC strategies. We can see that the proposed O-PSC strategy exhibits a significantly better performance than the RP-PSC strategy. With the number of FGs $N = 10$, the performance gain in terms of $\Pr(\mathcal{D})$ provided by the O-PSC strategy ranges from 20% to 50%, when δ varies from 0.1 to 1. When δ is high, the probability that MUs can directly download the files from the storage of SBSs becomes small. In such cases, the advantage of optimizing the caching probabilities of the FGs is more obvious.

Even more significant $\Pr(\mathcal{D})$ improvement can be observed for the case of $N = 20$ than that for $N = 10$. A larger number of FGs means that less contents can be cached in each SBS, which implies a very limited storage capacity. In such cases, the benefit of optimizing the caching probabilities is more significant.

Fig. 5 compares $\Pr(\mathcal{D})$ in the context of RP-PSC and O-PSC strategies versus the Zipf exponents β. First, we can see that the proposed O-PSC strategy greatly outperforms the RP-PSC strategy in terms of $\Pr(\mathcal{D})$. With the number of FGs $N = 20$, the performance gain of $\Pr(\mathcal{D})$ ranges from 65% to 20% when β varies from 0.5 to 1.5. In other words, the $\Pr(\mathcal{D})$ improvement decreases, as β grows. The reason behind this trend is that for a large β, a small fraction of FGs dominate the file requests. Once the SBSs cache these very popular FGs, $\Pr(\mathcal{D})$ will become sufficiently high. Thus, the additional gain given by the optimization of caching probabilities becomes smaller. Furthermore, compared with the case $N = 10$, the $\Pr(\mathcal{D})$ improvement when $N = 20$ is more significant. The reason for this phenomenon has been explained above.

Fig. 6 compares $\Pr(\mathcal{D})$ in conjunction with the O-PSC strategies in the overlapping and nonoverlapping scenarios. Since the total number of files in our simulations is 100, in the figure, the curves of "FGNo = 10" and "FGNo = 20" are compared against the curves of "FilesPerGroup = 10" and "FilesPerGroup = 5," respectively. We can see that the performance of SDP in the scenario of FGs having overlapping subsets of files is better than that of the nonoverlapping subsets of files. The reason for this observation is that allowing overlapping amongst the different FGs provides a beneficial diversity of the FGs. Furthermore, we can see that when the SINR threshold is increased, the advantage of the overlapping scenario wanes. This is because when the SINR threshold is high, the O-PSC strategy tends to cache fewer popular files and the diversity of FGs becomes of limited benefit here.

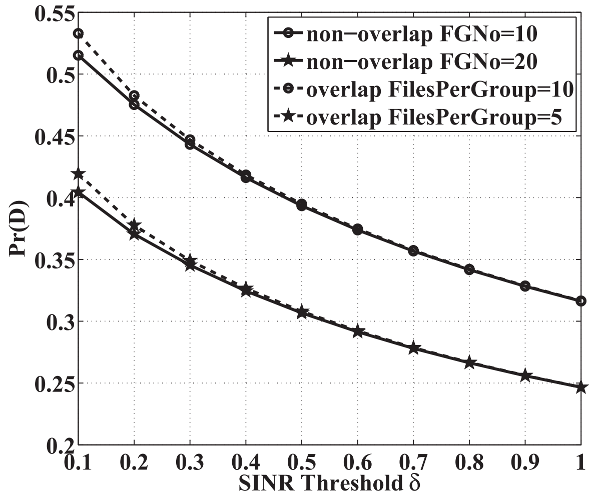


Fig. 6. Comparison of $\Pr(\mathcal{D})$ versus δ in the overlapping and nonoverlapping scenarios in the always-on architecture.

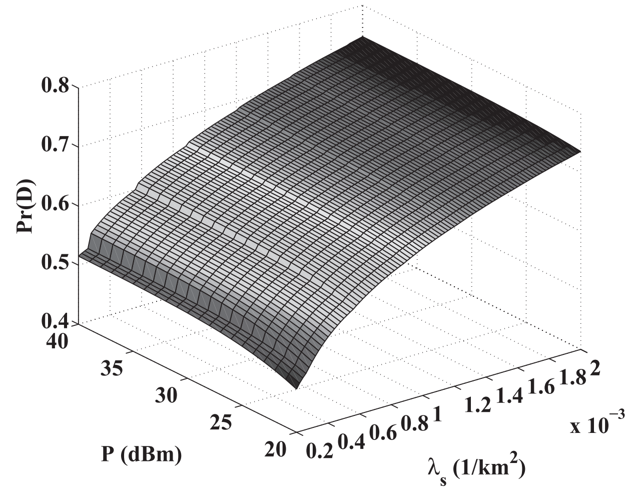


Fig. 8. $\Pr(\mathcal{D})$ with different P and λ_s in the dynamic on–off architecture.

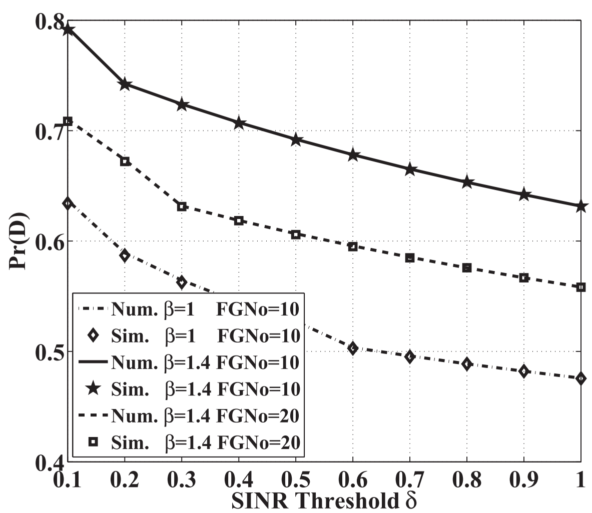


Fig. 7. Numerical and simulation results of $\Pr(\mathcal{D})$ of the O-PCP strategy in the dynamic on–off architecture.

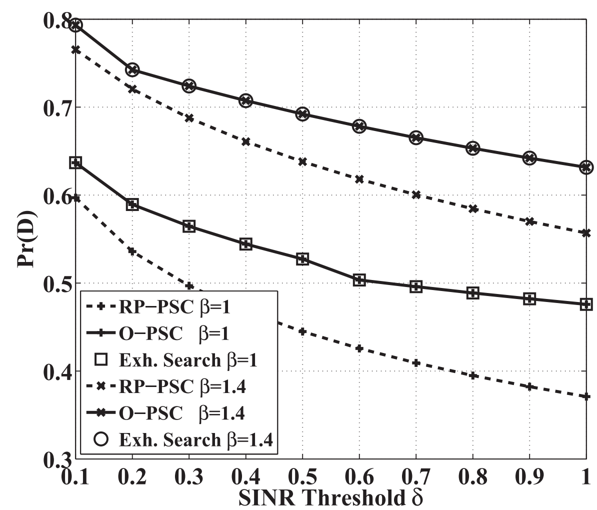


Fig. 9. Comparison of $\Pr(\mathcal{D})$ of the RP-PSC and O-PSC strategies versus δ in the dynamic on–off architecture.

B. Dynamic On–Off Architecture

Fig. 7 shows our comparison between the numerical and simulation results of $\Pr(\mathcal{D})$ for the O-PSC strategy. We can see from this figure that the numerical results closely match the simulation results in all scenarios. Similar phenomena can be observed as in the always-on architecture.

1) $\Pr(\mathcal{D})$ decreases upon increasing the SINR threshold δ.
2) $\Pr(\mathcal{D})$ increases with the Zipf exponent β.
3) $\Pr(\mathcal{D})$ increases when the number of FGs decreases.

The reasons behind these trends are the same as those discussed for the always-on architecture. Moreover, compared to Fig. 2, the value of $\Pr(\mathcal{D})$ in the dynamic on–off architecture of Fig. 7 is shown to be higher. The reason is that the dynamic on–off technique efficiently mitigates the potential avoidable interference in the network.

Fig. 8 shows the performance of $\Pr(\mathcal{D})$ for the O-PSC strategy in the dynamic on–off architecture, when the transmission power P of SBSs varies from 20 to 40 dBm and the SBS intensity λ_s varies from 200 to $2000/\text{km}^2$. We can see from this figure that $\Pr(\mathcal{D})$ increases monotonically, when either P or λ_s increases. Moreover, we can see that when P increases to a sufficiently high value, any further increase of P will no longer improve $\Pr(\mathcal{D})$. However, the increase of λ_s will always improve $\Pr(\mathcal{D})$, as seen in (12).

Fig. 9 compares $\Pr(\mathcal{D})$ of the RP-PSC and O-PSC strategies, when the SINR threshold δ varies. It can be seen from the figure that compared to the RP-PSC strategy, $\Pr(\mathcal{D})$ is obviously improved by the optimal caching PMF $\{S_n^{\text{Opt}}\}$ in the O-PSC strategy. With the Zipf exponent $\beta = 1$, the performance gain of $\Pr(\mathcal{D})$ ranges from 7% to 30%, when δ varies from 0.1 to 1. This observation is similar to that in the always-on architecture. That is, the $\Pr(\mathcal{D})$ improvement achieved by the O-PSC strategy is more pronounced, when the SINR threshold is higher. Furthermore, the $\Pr(\mathcal{D})$ improvement is higher when the Zipf exponent β is lower. The reason for this is explained above.

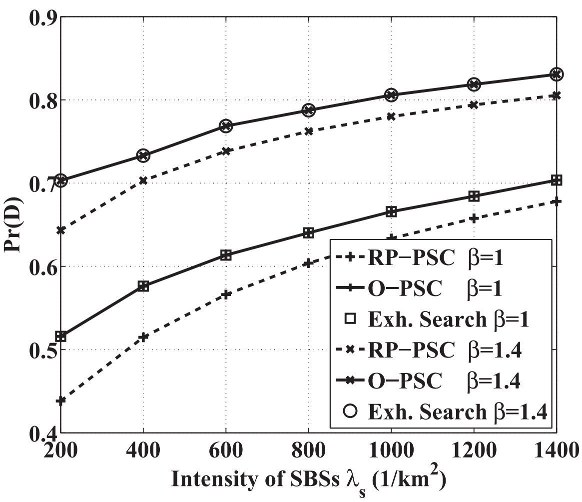


Fig. 10. Comparison of $\Pr(\mathcal{D})$ of the RP-PSC and O-PSC strategies versus λ_s in the dynamic on–off architecture.

Furthermore, in order to verify the optimality of the solution given by our algorithm, we plot the optimal solution obtained from the exhaustive search over all legitimate file caching states, denoted by "Exh. Search" in the figure. Observed from the figure that our solution exactly matches the optimal solution of "Exh. Search," which confirms our statement that the proposed solution achieves global optimality.

In Fig. 10, we portray $\Pr(\mathcal{D})$ of the RP-PSC and the O-PSC strategies versus the SBS intensity λ_s. First, it can be seen that compared with the RP-PSC strategy, the optimization of the caching probabilities in the O-PSC strategy improves $\Pr(\mathcal{D})$ in all scenarios. This $\Pr(\mathcal{D})$ improvement achieved by the O-PSC strategy wanes slightly when λ_s increases because when the SBS intensity is higher, each MU becomes capable of associating with multiple SBSs, and thus, the probability that MUs can successfully download contents from SBSs will be higher. In such a case, the $\Pr(\mathcal{D})$ improvement obtained by the optimization of the FG caching probabilities remains limited. In addition, we verify the optimality of our solution by comparing it to the optimal solution obtained from the exhaustive search.

VII. Conclusion

In this paper, based on stochastic geometry theory, we analyzed the performance of the PSC in a pair of network architectures. Specifically, we analyzed the probability $\Pr(\mathcal{D})$ that MUs can successfully download contents from the storage of SBSs. We concluded that increasing the SBSs' transmission power P or their deployment intensity λ_s is capable of increasing the SDP. However, in the always-on architecture, $\Pr(\mathcal{D})$ remains constant when P or λ_s is sufficiently high, while in the dynamic on–off architecture, $\Pr(\mathcal{D})$ always increases as λ_s grows. Furthermore, in order to maximize $\Pr(\mathcal{D})$, we optimized the caching probabilities of the FGs. Our results demonstrated that in the always-on architecture, the optimal subset of FGs depends on the contents request probabilities. In the dynamic on–off architecture, a piecewise defined function of MU-to-SBS intensity ratio λ_u/λ_s was introduced in order to find the optimal subset of FGs to be cached. Interestingly, a similar optimal caching probability law was found for both architectures, i.e., S_n^{Opt} is a linear function of $\sqrt{Q_n}$. Our simulation results showed that the proposed optimal caching probabilities of the FGs achieve a substantial gain in both architecture in terms of $\Pr(\mathcal{D})$ compared to the benchmark $S_n = Q_n$, because more caching resources are devoted to the more popular files in the proposed scheme.

Appendix A
Proof of Theorem 1

In Tier-n of the always-on architecture, where the intensity of the SBSs is $S_n\lambda_s$, the PDF of z, i.e., the distance between the typical MU and its nearest SBS, follows $f_Z(z) = 2\pi S_n\lambda_s z\exp(-\pi S_n\lambda_s z^2)$. From (3) and (4), we have

$$\begin{aligned}
&\Pr(\mathcal{D}_n) = \Pr\left(\gamma_n(z) \geq \delta\right)\\
&= \int_0^\infty \Pr\left[\frac{Ph_{x_0}z^{-\alpha}}{\sum_{x_j\in\Phi\backslash\{x_0\}} Ph_{x_j}\left\|x_j\right\|^{-\alpha} + \sigma^2} \geqslant \delta\right] f_Z(z)\,\mathrm{d}z\\
&\overset{(a)}{=} \int_0^\infty \mathbb{E}_I\left[\exp\left(-z^\alpha\delta I\right)\right]\exp\left(-\frac{z^\alpha\delta\sigma^2}{P}\right)\\
&\qquad 2\pi S_n\lambda_s z\exp(-\pi S_n\lambda_s z^2)\mathrm{d}z \qquad (25)
\end{aligned}$$

where (a) is obtained by $h_{x_0} \sim \exp(1)$ and $I \triangleq \sum_{x_j\in\Phi\backslash\{x_0\}} h_{x_j}\left\|x_j\right\|^{-\alpha}$ represents the interference.

The interference I consists of two independent parts: 1) I_1: the SBSs in other tiers, which are dispersed across the entire area of the network, and 2) I_2: the SBSs in the nth tier, whose distances from the typical MU are larger than z. Due to the independence of I_1 and I_2, we have $\mathbb{E}_I\left[\exp\left(-z^\alpha\delta I\right)\right] = \mathbb{E}_{I_1}\left[\exp\left(-z^\alpha\delta I_1\right)\right]\cdot \mathbb{E}_{I_2}\left[\exp\left(-z^\alpha\delta I_2\right)\right]$.

Since the distribution of the SBSs in Tier-i is viewed as an HPPP ϕ_i with $S_i\lambda_s$ and therefore, we have

$$\begin{aligned}
&\mathbb{E}_{I_1}\left[\exp\left(-z^\alpha\delta I_1\right)\right]\\
&= \mathbb{E}_{h_{x_j},x_j}\left[\prod_{x_j\in\sum_{i=1,i\neq n}^N \phi_i} \exp\left(-z^\alpha\delta h_{x_j}\left\|x_j\right\|^{-\alpha}\right)\right]\\
&\overset{(b)}{=} \mathbb{E}_{x_j}\left[\prod_{x_j\in\sum_{i=1,i\neq n}^N \phi_i} \frac{1}{1+z^\alpha\delta\left\|x_j\right\|^{-\alpha}}\right]\\
&\overset{(c)}{=} \exp\left(-\sum_{i=1,i\neq n}^N S_i\lambda_s\int_{\mathbb{R}^2}\left(1-\frac{1}{1+\delta z^\alpha\left\|x_j\right\|^{-\alpha}}\right)\mathrm{d}x_j\right)\\
&= \exp\left(-2\pi\sum_{i=1,i\neq n}^N S_i\lambda_s\frac{1}{\alpha}\delta^{\frac{2}{\alpha}}B\left(\frac{2}{\alpha},1-\frac{2}{\alpha}\right)z^2\right) \qquad (26)
\end{aligned}$$

where (b) uses $h_{x_j} \sim \exp(1)$, and (c) uses $\mathbb{E}\left[\prod_{v\in\Phi}\xi(v)\right] = \exp\left(-\lambda_\Phi\int\left(1-\xi(v)\right)\mathrm{d}v\right)$.

As for I_2, we have

$$\begin{aligned}
&\mathbb{E}_{I_2}\left[\exp\left(-z^{\alpha}\delta I_2\right)\right]\\
&= \exp\left(-S_n\lambda_s 2\pi \int_z^{\infty}\left(1-\frac{1}{1+z^{\alpha}\delta\left\|x_j\right\|^{-\alpha}}\right)\left\|x_j\right\|\mathrm{d}\left\|x_j\right\|\right)\\
&\overset{(d)}{=} \exp\left(-S_n\lambda_s\pi\delta^{\frac{2}{\alpha}}z^2\frac{2}{\alpha}\int_{\delta^{-1}}^{\infty}\frac{l^{\frac{2}{\alpha}-1}}{1+l}\mathrm{d}l\right)\\
&= \exp\left(-S_n\lambda_s\pi z^2\frac{2\delta}{\alpha-2}\,{}_2F_1\left(1,1-\frac{2}{\alpha};2-\frac{2}{\alpha};-\delta\right)\right)
\end{aligned} \tag{27}$$

where (d) uses $l \triangleq \delta^{-1}z^{-\alpha}\left\|x_j\right\|^{\alpha}$.

Our proof is completed by plugging (26) and (27) into (25). ■

APPENDIX B
PROOF OF COROLLARY 1

Since we have $\Pr(\mathcal{D}) = \sum_{n=1}^{N} Q_n \Pr(\mathcal{D}_n)$, to prove that $\Pr(\mathcal{D})$ increases with the increase of λ_s, we only have to prove that $\Pr(\mathcal{D}_n)$ increases monotonically upon increasing λ_s $\forall n$. Thus, in the following, we focus our attention on the proof that $\frac{\partial \Pr(\mathcal{D}_n)}{\partial \lambda_s} > 0$.

To simplify our discourse, we use $C_1 \triangleq \frac{\pi S_n}{2\sigma}\sqrt{\frac{\pi P}{\delta}}$, and $C_2 \triangleq \frac{\pi}{2\sigma}\sqrt{\frac{P}{\delta}}\left(S_n + \frac{\pi}{2}\sqrt{\delta}(1-S_n) + S_n\sqrt{\delta}\arctan\sqrt{\delta}\right)$. Obviously, we have $C_1 > 0$ and $C_2 > 0$. Then, $\Pr(\mathcal{D}_n)$ can be rewritten as

$$\Pr(\mathcal{D}_n) = C_1\lambda_s\exp(C_2^2\lambda_s^2)\mathrm{erfc}(C_2\lambda_s). \tag{28}$$

Hence, we have

$$\begin{aligned}
&\frac{\partial \Pr(\mathcal{D}_n)}{\partial \lambda_s} = C_1\lambda_s\exp(C_2^2\lambda_s^2)\left(1-\mathrm{erf}(C_2\lambda_s)\right)\\
&= \left(C_1\exp(C_2^2\lambda_s^2) + C_1\lambda_s\exp(C_2^2\lambda_s^2)2C_2^2\lambda_s\right)\mathrm{erfc}(C_2\lambda_s)\\
&\quad - C_1\lambda_s\exp(C_2^2\lambda_s^2)\frac{2}{\sqrt{\pi}}C_2\exp(-C_2^2\lambda_s^2)\\
&= C_1\exp(C_2^2\lambda_s^2)(1+2C_2^2\lambda_s^2)\mathrm{erfc}(C_2\lambda_s) - C_1C_2\lambda_s\frac{2}{\sqrt{\pi}}.
\end{aligned} \tag{29}$$

According to [35], the continued fraction expansion of the complementary error function is

$$\mathrm{erfc}(z) = \frac{z}{\sqrt{\pi}}\exp(-z^2)\frac{1}{z^2+\frac{a_1}{1+\frac{a_2}{z^2+\frac{a_3}{1+\cdots}}}}, a_m = \frac{m}{2}. \tag{30}$$

From (30), we have $\mathrm{erfc}(z) > \frac{z}{\sqrt{\pi}}\exp(-z^2)\frac{1}{z^2+\frac{1}{2}}$. Substituting $C_2\lambda_s$ for z, we have

$$\exp(C_2^2\lambda_s^2)\mathrm{erfc}(C_2\lambda_s) > \frac{C_2\lambda_s}{\sqrt{\pi}}\frac{1}{C_2^2\lambda_s^2+\frac{1}{2}}. \tag{31}$$

Substituting (31) into (29), we can prove that $\frac{\partial \Pr(\mathcal{D}_n)}{\partial \lambda_s} > 0$, which implies that $\Pr(\mathcal{D})$ increases monotonically upon increasing λ_s. ■

APPENDIX C
PROOF OF THEOREM 2

Similar to the derivation in Appendix A, in the dynamic on–off architecture, the intensity of SBSs in Tier-n is also $S_n\lambda_s$. Thus, in Tier-n the distance z between the typical MU and its nearest SBS follows the same PDF $f_Z(z)$ in the always-on architecture. It follows that we have a similar formulation for $\Pr(\mathcal{D}_n)$ in the dynamic on–off architecture, yielding

$$\begin{aligned}
\Pr(\mathcal{D}_n) = &\int_0^{\infty}\mathbb{E}_I\left[\exp\left(-z^{\alpha}\delta I\right)\right]\exp\left(-\frac{z^{\alpha}\delta\sigma^2}{P}\right)\\
&2\pi S_n\lambda_s z\exp(-\pi S_n\lambda_s z^2)\mathrm{d}z.
\end{aligned} \tag{32}$$

In the dynamic on–off architecture, the interference I only arrives from the SBSs in the active mode. According to [36], the activity probability $\Pr(\mathcal{A}_n)$ of the SBSs in Tier-n, can be formulated as

$$\Pr(\mathcal{A}_n) \approx 1 - \left(1+\frac{Q_n\lambda_u}{3.5S_n\lambda_s}\right)^{-3.5}.$$

As in Appendix A, we divide the interference into two parts: $I = I_1 + I_2$. The first part of interference I_1 is inflicted by the active SBSs in any Tier-i, $i \neq n$, which can be viewed as a homogeneous PPP with the intensity of $\Pr(\mathcal{A}_i)S_i\lambda_s$. Hence, we update (26) as follows:

$$\begin{aligned}
&\mathbb{E}_{I_1}\left[\exp\left(-z^{\alpha}\delta I_1\right)\right]\\
&= \exp\left(-2\pi\sum_{i=1:i\neq n}^{N}\Pr(\mathcal{A}_i)S_i\lambda_s\frac{1}{\alpha}\delta^{\frac{2}{\alpha}}B\left(\frac{2}{\alpha},1-\frac{2}{\alpha}\right)z^2\right).
\end{aligned} \tag{33}$$

The second part of the interference I_2 comes from the active SBSs in Tier-n located in the area outside the circle with radius z. We update (27) as follows:

$$\begin{aligned}
&\mathbb{E}_{I_2}\left[\exp\left(-z^{\alpha}\delta I_2\right)\right] = \exp\left(-\Pr(\mathcal{A}_n)\right.\\
&\left.S_n\lambda_s\pi z^2\frac{2\delta}{\alpha-2}\,{}_2F_1\left(1,1-\frac{2}{\alpha};2-\frac{2}{\alpha};-\delta\right)\right).
\end{aligned} \tag{34}$$

Integrating (33) and (34) into (32) completes the proof. ■

APPENDIX D
PROOF OF THEOREM 4

Note that in the following proof, we simplify the notation by introducing $a \triangleq \frac{\lambda_u}{\lambda_s}$, $C \triangleq C(\delta,\alpha)$, and $A \triangleq A(\delta,\alpha)$.

First, we investigate the optimization Problem (17) for a given indicator vector $\boldsymbol{\varepsilon}$. Let us denote by N^* the number of ones in $\boldsymbol{\varepsilon}$, and by $\{n_j\}$ the subscript of the ones in N^*. Then, we have a new optimization problem represented as

$$\begin{aligned}\max_{\{S_{n_j}\}} \quad & \sum_{j=1}^{N^*} \frac{Q_{n_j} S_{n_j}}{Q_{n_j} aA + \sum_{i:i\neq j} Q_{n_i} aC + S_{n_j}} \\ \text{s.t.} \quad & \sum_{j=1}^{N^*} S_{n_j} = 1 \\ & S_{n_j} > 0 \; \forall j = 1, \ldots N^*. \end{aligned} \tag{35}$$

If we neglect the constraint $S_{n_j} > 0$, the solution to Problem (35) is presented in Lemma 1.

Lemma 1: Neglecting the constraint $S_{n_j} > 0$, the optimal solution for Problem (35) is given by

$$S_{n_j}^{Opt} = \zeta \sqrt{Q_{n_j} C\xi - Q_{n_j}^2 (C - A)} - \left[\xi aC - Q_{n_j} a(C - A)\right] \tag{36}$$

where we have $\zeta \triangleq \frac{1+N^*\xi aC - \xi a(C-A)}{\sum_{i=1}^{N^*} \sqrt{Q_{n_i}\xi C - Q_{n_i}^2 (C-A)}}$ and $\xi \triangleq \sum_{j=1}^{N^*} Q_{n_j}$.

Proof: See Appendix E.

From (36), we propose Lemma 2.

Lemma 2: Given the request probabilities of two FGs cached, where $Q_{n_i} > Q_{n_j}$, according to (36), we have $S_{n_i}^{Opt} > S_{n_j}^{Opt}$.

Proof: See Appendix F.

Based on Lemma 2, we have $S_{n_{j^*}}^{\text{Opt}} = \min\{S_{n_j}^{\text{Opt}}\}$ where $n_{j^*} = \arg\min_{n_j}\{Q_{n_j}\}$. Hence, the constraint $S_{n_j} > 0$, $\forall j = 1, \ldots N^*$, is equivalent to $S_{n_{j^*}} > 0$. In order to ensure that $S_{n_{j^*}}^{\text{Opt}} > 0$, based on (36), we have

$$a < a_{n_{j^*}}, \; a_{n_{j^*}} \triangleq \frac{\vartheta_{n_{j^*}}}{(\vartheta_{n_{j^*}}\xi - Q_{n_{j^*}})(C-A) + (1 - N^*\vartheta_{n_{j^*}})\xi C} \tag{37}$$

where

$$\vartheta_{n_{j^*}} \triangleq \frac{\sqrt{Q_{n_{j^*}}\xi C + Q_{n_{j^*}}^2 (A - C)}}{\sum_{i=1}^{N^*} \sqrt{Q_{n_i}\xi C + Q_{n_i}^2 (A - C)}}. \tag{38}$$

Hence, (36) only becomes the optimal solution of Problem (35), when a meets the requirement (37).

Substituting the optimal solution in (36) into (35), we obtain the maximum value of $\Pr(\mathcal{D})$ for the given indicator vector $\boldsymbol{\varepsilon}$, yielding

$$D_{N^*} = \xi - \frac{a\left(\sum_{j=1}^{N^*} \sqrt{Q_{n_j} C\xi + Q_{n_j}^2 (A - C)}\right)^2}{1 + N^*\xi aC + \xi a(A - C)}. \tag{39}$$

Second, we extend the Problem (35) to Problem (17). Based on the analysis above, given the indicator vector $\boldsymbol{\varepsilon}_1$, when $a < a_{\boldsymbol{\varepsilon}_1}$ in (37), we can obtain the maximum $\Pr(\mathcal{D})$ denoted by $D_{\boldsymbol{\varepsilon}_1}$ in (39). For $\boldsymbol{\varepsilon}_2$, if we have $a_{\boldsymbol{\varepsilon}_2} > a_{\boldsymbol{\varepsilon}_1}$, then provided $a < a_{\boldsymbol{\varepsilon}_1}$ holds, we have $a < a_{\boldsymbol{\varepsilon}_2}$. Thus, $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$ are both reasonable for this optimization problem. Through the comparison of $D_{\boldsymbol{\varepsilon}_1}$ and $D_{\boldsymbol{\varepsilon}_2}$, we can find the right choice between $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$. Then obtain the optimal solution of $\{S_n\}$ in form of (36).

Using $\{Q_n\}$, we can obtain the segmentation parameters for a in (37). The smallest segmentation parameter is obtained when $\boldsymbol{\varepsilon}$ contains N ones, which is denoted by a_N. When $a < a_N$, i.e., λ_s is high enough, all FGs can be cached in SBSs. Then, with the increase of a, i.e., the decrease of λ_s, some FGs cannot be cached, where a reduced number of ones appear in $\boldsymbol{\varepsilon}$. Since we have $Q_1 > Q_2 > \cdots > Q_N$, the unpopular FGs will be discarded one by one. Accordingly, we can obtain both $\boldsymbol{\varepsilon}_i$ as well as the segmentation parameter a_i. As a result, a piecewise defined function regarding a is obtained like the number of ones in $\boldsymbol{\varepsilon}$ is shown in (20). ■

Appendix E
Proof of Lemma1

Neglecting the constraint $S_{n_j} > 0$, it becomes plausible that Problem (35) is a concave maximization problem. Adopting the Lagrange multiplier Λ, we have

$$\Lambda(\mathbf{S}, \lambda) = \sum_{j=1}^{N^*} \frac{Q_{n_j} S_{n_j}}{Q_{n_j} aA + \sum_{i=1:i\neq j}^{N^*} Q_{n_i} aC + S_{n_j}} + \lambda\left(\sum_{j=1}^{N^*} S_{n_j} - 1\right). \tag{40}$$

Using $\xi \triangleq \sum_{j=1}^{N^*} Q_{n_j}$ and $\frac{\partial \Lambda}{\partial S_{n_j}} = 0$, we have

$$\frac{Q_{n_j} aC\xi + Q_{n_j}^2 a(A - C)}{\left(aC\xi + a(A - C)Q_{n_j} + S_{n_j}\right)^2} + \lambda = 0 \; \forall n_j. \tag{41}$$

Since $\sum_{j=1}^{N^*} S_{n_j} = 1$, we have

$$S_{n_j}^{\text{Opt}} = \zeta \sqrt{Q_{n_j} aC\xi + Q_{n_j}^2 a(A - C)} - \left[\xi aC + Q_{n_j} a(A - C)\right] \tag{42}$$

where

$$\zeta \triangleq \frac{1 + N^*\xi aC + \xi a(A - C)}{\sum_{i=1}^{N^*} \sqrt{Q_{n_i}\xi aC + Q_{n_i}^2 a(A - C)}}. \tag{43}$$

■

Appendix F
Proof of Lemma 2

First, based on the optimal solution given in (36), we have

$$\frac{\partial S_{n_j}^{\text{Opt}}}{\partial Q_{n_j}} = \zeta \frac{\sqrt{a}}{2} \frac{C\xi + 2Q_{n_j}(A - C)}{\sqrt{Q_{n_j} C\xi + Q_{n_j}^2 (A - C)}} + a(C - A). \tag{44}$$

Since $C(\alpha, \delta) > A(\alpha, \delta) > 0$, we have $\frac{\partial S_{n_j}^{\text{Opt}}}{\partial Q_{n_j}} \geq 0$ when $Q_{n_j} \leq \frac{\xi}{2}\frac{C}{C-A}$, which means $S_{n_j}^{\text{Opt}}$ increases with the growth of Q_{n_j}, when Q_{n_j} is no bigger than $\frac{\xi}{2}\frac{C}{C-A}$.

1) Since $Q_{n_j} \leq \xi$, if $\frac{C}{C-A} \geq 2$, for all Q_{n_j}, $\frac{\partial S_{n_j}^{\text{Opt}}}{\partial Q_{n_j}} > 0$, and the proof is completed.

2) For $\frac{C}{C-A} < 2$, we consider the following case. Since $\frac{C}{C-A} > 1$, we have $\frac{\xi}{2}\frac{C}{C-A} > \frac{\xi}{2}$. Because $\sum_j Q_{n_j} = \xi$, among

the N^* FGs cached, there is only one FG associated with $Q_{n_j} > \frac{\xi}{2}\frac{C}{C-A}$. We denote the request probability of this popular file by Q_1 and its caching probability by S_1^{Opt}. Since the request probabilities of other cached FGs must be less than $\frac{\xi}{2}\frac{C}{C-A}$, and $\frac{\partial S_{n_j}^{\mathrm{Opt}}}{\partial Q_{n_j}} > 0$ when Q_{n_j} in this region, the highest caching probability among these less popular FGs occurs when only two FGs are cached. That is, the other FG with request probability $Q_2 = \xi - Q_1$. Denoted by S_2^{Opt} its caching probability. We have

$$S_1^{\mathrm{Opt}} - S_2^{\mathrm{Opt}} = \zeta\sqrt{a}\left(\sqrt{Q_1 C\xi + Q_1^2(A-C)} - \sqrt{Q_2 C\xi + Q_2^2(A-C)}\right) + (Q_1 - Q_2)a(C-A). \quad (45)$$

Since $Q_1 C\xi + Q_1^2(A-C) - Q_2 C\xi - Q_2^2(A-C) = (Q_1 - Q_2)\xi a A > 0$, we have $S_1^{\mathrm{Opt}} - S_2^{\mathrm{Opt}} > 0$. Thus, for the dominate FG, its caching probability also dominates.

Combining the two parts above, we complete the proof. ■

Acknowledgment

The authors would like to thank Dr. H. (Henry) Chen from the University of Sydney, Australia, for his helpful discussions and suggestions.

References

[1] CISCO, "Cisco visual networking index: Global mobile data traffic forecast update, 2014–2019 White Paper," Feb. 2014.

[2] D. Lopez-Perez, M. Ding, H. Claussen, and A. H. Jafari, "Towards 1 Gbps/UE in cellular systems: Understanding ultra-dense small cell deployments," Mar. 2015.

Q2 [3] J. Erman, A. Gerber, M. Hajiaghayi, D. Pei, S. Sen, and O. Spatscheck, "To cache or not to cache: The 3G case," *IEEE Internet Comput.*, vol. 15, no. 2, pp. 27–34, Mar. 2011.

[4] U. Niesen, D. Shah, and G. W. Wornell, "Caching in wireless networks," *IEEE Trans. Inf. Theory*, vol. 58, no. 10, pp. 6524–6540, Oct. 2012.

[5] J. Li, H. Chen, Y. Chen, Z. Lin, B. Vucetic, and L. Hanzo, "Pricing and resource allocation via game theory for a small-cell video caching system," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 8, pp. 2115–2129, Aug. 2016.

[6] 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, Feb. 2014.

[7] S. Woo, E. Jeong, S. Park, J. Lee, S. Ihm, and K. Park, "Comparison of caching strategies in modern cellular backhaul networks," in *Proc. 11th Annu. Int. Conf. Mobile Syst., Appl., Serv.*, New York, NY, USA, 2013, pp. 319–332. [Online]. Available: http://doi.acm.org/10.1145/2462456.2464442

[8] H. Ahlehagh and S. Dey, "Video-aware scheduling and caching in the radio access network," *IEEE/ACM Trans. Netw.*, vol. 22, no. 5, pp. 1444–1462, Oct. 2014.

[9] K. Shanmugam, N. Golrezaei, A. Dimakis, A. Molisch, and G. Caire, "Femtocaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

[10] N. Golrezaei, A. Molisch, A. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Commun. Mag.*, vol. 51, no. 4, pp. 142–149, Apr. 2013.

[11] 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. 2015.

[12] N. Golrezaei, P. Mansourifard, A. Molisch, and A. Dimakis, "Base-station assisted device-to-device communications for high-throughput wireless video networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, pp. 3665–3676, Jul. 2014.

[13] H. J. Kang and C. G. Kang, "Mobile device-to-device (D2D) content delivery networking: A design and optimization framework," *J. Commun. Netw.*, vol. 16, no. 5, pp. 568–577, Oct. 2014.

[14] C. Yang, Y. Yao, Z. Chen, and B. Xia, "Analysis on cache-enabled wireless heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 131–145, Jan. 2015.

[15] M. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inform. Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.

[16] K. Poularakis, G. Iosifidis, and L. Tassiulas, "Approximation algorithms for mobile data caching in small cell networks," *IEEE Trans. Commun.*, vol. 62, no. 10, pp. 3665–3677, Oct. 2014.

[17] J. Li, Y. Chen, Z. Lin, W. Chen, B. Vucetic, and L. Hanzo, "Distributed caching for data dissemination in the downlink of heterogeneous networks," *IEEE Trans. Commun.*, vol. 63, no. 10, pp. 3553–3568, Oct. 2015.

[18] E. Bastug, M. Bennis, and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs," *EURASIP J. Wireless Commun. Netw.*, vol. 2015, no. 1, p. 41, 2015. Q3

[19] G. Vettigli, M. Ji, A. Tulino, J. Llorca, and P. Festa, "An efficient coded multicasting scheme preserving the multiplicative caching gain," in *Proc. IEEE Conf. Comput. Commun. Workshops*, Apr. 2015, pp. 251–256.

[20] I. Ashraf, L. Ho, and H. Claussen, "Improving energy efficiency of femtocell base stations via user activity detection," in *Proc. IEEE Wireless Commun. Network. Conf.*, Apr. 2010, pp. 1–5.

[21] 3GPP, "Tentative 3GPP timeline for 5G," Mar. 2015.

[22] QUALCOMM, "1000x: More small cells. hyper-dense small cell deployments," Jun. 2014.

[23] C. Yang, J. Li, and M. Guizani, "Cooperation for spectral and energy efficiency in ultra-dense small cell networks," *IEEE Wireless Commun.*, vol. 23, no. 1, pp. 64–71, Feb. 2016.

[24] N. Saxena, A. Roy, and H. Kim, "Traffic-aware cloud ran: A key for green 5g networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 1010–1021, Apr. 2016.

[25] H. Dhillon, R. Ganti, F. Baccelli, and J. Andrews, "Modeling and analysis of K-tier downlink heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 550–560, Apr. 2012.

[26] M. Zinka, K. Suh, Y. Gu, and J. Kurose, "Characteristics of YouTube network traffic at a campus network—Measurements, models, and implications," *Comput. Netw.*, vol. 53, no. 4, pp. 501–514, Mar. 2009.

[27] M. Maddah-Ali and U. Niesen, "Decentralized coded caching attains order-optimal memory-rate tradeoff," *IEEE/ACM Trans. Netw.*, vol. 23, no. 4, pp. 1029–1040, Aug. 2014.

[28] D. Stoyan, W. Kendall, and J. Mecke, *Stochastic Geometry and its Applications*, 2nd ed. Hoboken, NJ, USA: Wiley, 1995.

[29] S. C. Forum, "Scf049: Backhaul technologies for small cells (release 4)," Feb. 2014. Q4

[30] C. Nicoll, "3G and 4G small cells create big challenges for MNOs," Mar. 2013.

[31] I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series, and Products*, 7th ed. Amsterdam, The Netherlands: Elsevier, 2007.

[32] 3GPP, "Further advancements for E-UTRA physical layer aspects," 3GPP, France, Tech. Rep. v.9.0.0, Mar. 2010.

[33] W. Cody, "Algorithm 715: SPECFUN—A portable FORTRAN package of special function routines and test drivers," *ACM Trans. Math. Softw.*, vol. 19, no. 1, pp. 22–30, Mar. 1993.

[34] H. Kuhn and A. Tucker, "Nonlinear programming," in *Proc. 2nd Berkeley Symp. Math. Statist. Probability*, 1951, pp. 481–492.

[35] A. Cuyt, V. Petersen, B. Verdonk, H. Waadeland, and W. Jones, *Handbook of Continued Fractions for Special Functions*. Berlin, Germany: Springer-Verlag, 2008.

[36] S. Lee and K. Huang, "Coverage and economy of cellular networks with many base stations," *IEEE Commun. Lett.*, vol. 16, no. 7, pp. 1038–1040, Jul. 2012.

Youjia Chen received the B.S. and M.S degrees in communication engineering from Nanjing University, Nanjing, China, in 2005 and 2008, respectively. She is currently working toward the Ph.D. degree in wireless engineering with the University of Sydney, Sydney, Australia.

From 2008 to 2009, she was with Alcatel Lucent Shanghai Bell. From August 2009 until now, she has been with the College of Photonic and Electrical Engineering, Fujian Normal University, China. Her research interests include resource management, load balancing, and caching strategy in heterogeneous cellular networks.

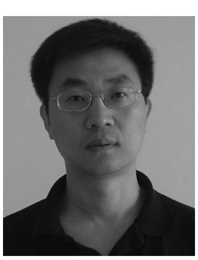
Ming Ding (M'12) received the B.S. and M.S. degrees (with first class Hons.) in electronics engineering and Ph.D. degree in signal and information processing from Shanghai Jiao Tong University (SJTU), Shanghai, China, in 2004, 2007, and 2011, respectively.

From September 2007 to September 2011, while at the same time working as a Researcher/Senior Researcher Sharp Laboratories of China (SLC), after achieving the Ph.D. degree, he continued working with SLC as a Senior Researcher/Principal Researcher until September 2014, when he joined National Information and Communications Technology Australia (NICTA). In July 2016, Commonwealth Scientific and Industrial Research Organization (CSIRO) and NICTA joined forces to create Data61, where he continued as a Senior Research Scientist in this new R&D center in Sydney, NSW, Australia. He has authored more than 30 papers in IEEE journals and conferences, all in recognized venues, and about 20 3GPP standardization contributions, as well as a Springer book entitled *Multi-point Cooperative Communication Systems: Theory and Applications* (Springer-Verlag, 2013). In addition, as the first inventor, he holds 15 CN, seven JP, three US, two KR patents, and has co-authored another 100+ patent applications on 4G/5G technologies.

Dr. Ding has been the Guest Editor/Cochair/TPC member of several IEEE top-tier journals/conferences, e.g., the IEEE Journal on Selected Areas in Communications, the IEEE Communications Magazine, the IEEE Globecom Workshops, etc. He received the Presidents Award from the SLC in 2012 for his inventions and publications and served as one of the key members in the 4G/5G standardization team when it was awarded in 2014 as the Sharp Company Best Team: LTE 2014 Standardization Patent Portfolio.

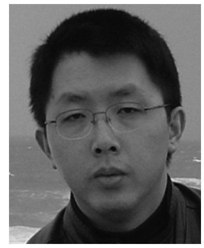
Jun Li (M'09–SM'16) received the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University, Shanghai, China in 2009.

From January 2009 to June 2009, he was a Research Scientist with the Department of Research and Innovation, Alcatel Lucent Shanghai Bell. From June 2009 to April 2012, he was a Postdoctoral Fellow with the School of Electrical Engineering and Telecommunications, University of New South Wales, Australia. From April 2012 to June 2015, he was a Research Fellow with the School of Electrical Engineering, University of Sydney, Australia. From June 2015 to the present, he has been a Professor with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, China. His research interests include network information theory, channel coding theory, wireless network coding, and cooperative communications.



Zihuai Lin (S'98–M'06–SM'10) received the Ph.D. degree in electrical engineering from Chalmers University of Technology, Gothenburg, Sweden, in 2006.

Prior to his Ph.D. degree, he has held positions at Ericsson Research, Stockholm, Sweden. Following having received the Ph.D. degree, he was a Research Associate Professor with Aalborg University, Aalborg, Denmark, and is currently with the School of Electrical and Information Engineering, University of Sydney, Sydney, Australia. His research interests include source/channel/network coding, coded modulation, MIMO, OFDMA, SC-FDMA, radio resource management, cooperative communications, small-cell networks, 5G cellular systems, etc.



Guoqiang Mao (S'98–M'02–SM'08) received the Ph.D. degree in telecommunications engineering from Edith Cowan University, Perth, Australia, in 2002.

Between 2002 and 2014, he was with the School of Electrical and Information Engineering, University of Sydney, Sydney, Australia. He joined the University of Technology Sydney in February 2014 as a Professor of wireless networking and the Director of Center for real-time information networks. The Center is among the largest university research centers in Australia in the field of wireless communications and networking. He has published about 200 papers in international conferences and journals, which have been cited more than 4000 times. His research interest includes intelligent transport systems, applied graph theory and its applications in telecommunications, Internet of Things, wireless sensor networks, wireless localization techniques, and network performance analysis.

Dr. Mao is an Editor of the IEEE Transactions on Wireless Communications (since 2014) and the IEEE Transactions on Vehicular Technology (since 2010). He received Top Editor award for outstanding contributions to the IEEE Transactions on Vehicular Technology in 2011, 2014, and 2015. He is a Cochair of the IEEE Intelligent Transport Systems Society Technical Committee on Communication Networks. He has served as a Chair, Cochair, and Technical Program Committee Member in a large number of international conferences.

Lajos Hanzo (F'08) received the M.S. degree in electronics and the Ph.D. degree from the Technical University of Budapest, Budapest, Hungary, in 1976 and 1983, respectively. He received the prestigious Doctor of Sciences research degree in wireless communications from the University of Southampton, U.K., in 2004.

In 2016, he was admitted to the Hungarian Academy of Science, Budapest, Hungary. During his 40-year career in telecommunications, he has held various research and academic posts in Hungary, Germany, and the U.K. Since 1986, he has been with the School of Electronics and Computer Science, University of Southampton, U.K., where he holds the Chair in telecommunications. He has successfully supervised 111 Ph.D. students, coauthored 20 John Wiley/IEEE Press books on mobile radio communications, totalling in excess of 10 000 pages, published 1600+ research contributions on IEEE Xplore, acted both as Technical Program Committee member and General Chair of IEEE conferences, presented keynote lectures, and received a number of distinctions. Currently he is directing a 60-strong academic research team, working on a range of research projects in the field of wireless multimedia communications sponsored by industry; the Engineering and Physical Sciences Research Council (EPSRC), U.K.; and the European Research Council.s Advanced Fellow Grant. He is an enthusiastic supporter of industrial and academic liaison, and he offers a range of industrial courses. He has 25 000+ citations and an H-index of 60. For further information on research in progress and associated publications, see http://www-mobile.ecs.soton.ac.uk. Dr. Hanzo is also a Governor of the IEEE Vehicular Technology Society. During 2008–2012, he was the Editor-in-Chief of the IEEE Press and a Chaired Professor with Tsinghua University, Beijing, China. In 2009, he received an honorary doctorate award by the Technical University of Budapest and in 2015, from the University of Edinburgh, Edinburgh, U.K., as well as the Royal Society.s Wolfson Research Merit Award. He is a Fellow of the Royal Academy of Engineering, The Institution of Engineering and Technology, and EURASIP.

QUERIES

Q1. Author: Please supply index terms/keywords for your paper. To download the IEEE Taxonomy go to http://www.ieee.org/documents/taxonomy_v101.pdf.

Q2. Author: Please update Ref. [2].

Q3. Author: Please provide page range in Ref. [18].

Q4. Author: Please provide complete bibliographic details in Refs. [29] and [30].