Energy-Efficient Caching for Scalable Videos in Heterogeneous Networks

Xuewei Zhang, *Student Member, IEEE*, Tiejun Lv, *Senior Member, IEEE*, Wei Ni, *Senior Member, IEEE*, John M. Cioffi, *Fellow, IEEE*, Norman C. Beaulieu, *Fellow, IEEE*, and Y. Jay Guo, *Fellow, IEEE*

***Abstract*— By suppressing repeated content deliveries, wireless caching has the potential to substantially improve the energy efficiency (EE) of the fifth-generation communication networks. In this paper, we propose two novel energy-efficient caching schemes in heterogeneous networks, namely, scalable video coding (SVC)-based fractional caching and SVC-based random caching, which can provide on-demand video services with different perceptual qualities. We derive the expressions for successful transmission probabilities and ergodic service rates. Based on the derivations and the established power consumption models, the EE maximization problems are formulated for the two proposed caching schemes. By taking logarithmic approximations of the l_0-norm, the problems are efficiently solved by the standard gradient projection method. Numerical results validate the theoretical analysis and demonstrate the superiority of our proposed caching schemes, compared to three benchmark strategies.**

***Index Terms*— Energy efficiency (EE), heterogeneous networks, scalable video coding (SVC), standard gradient projection method, wireless caching.**

I. Introduction

EXPOSED to information explosion and data tsunami, we have witnessed explosive traffic surges for socializing, working and entertainment. It is forecasted that the total amount of data traffic is expected to achieve at 100 exabytes in 2023, and over 75% of this traffic is expected to be generated from bandwidth-demanding multimedia video services [1]. Notably, there are a large number of repeated deliveries for popular video files [2], which would cause huge resource wastes and aggravate traffic burdens over backhaul links. Therefore, innovative techniques are desired to address the repeated deliveries of bandwidth-demanding popular videos. In light of this, wireless caching has been proposed as the appealing candidate technique in the fifth generation (5G) communication networks [3]. Wireless caching can effectively relieve the severe traffic burden over backhaul links and reduce service delay [4]–[6]. Additionally, it also exhibits strong potential to reduce power consumption and improve energy efficiency (EE) of wireless systems [7].

Wireless caching allows base stations (BSs) to prefetch video files from the core network through capacity-limited backhaul links during off-peak hours. These videos can be placed in the local storage of the BSs [8], and delivered to users when requested. Thereby, it can relieve the requests of backhaul bandwidth during peak-hours. Depending on different content placement strategies, caching can be typically classified into two categories, including uncoded caching [4], [5] and coded caching [9]–[12]. Uncoded caching aims at storing complete video files in each of the BS, which is very suitable for popular videos. In order to effectively leverage the cumulative cache size of nearby BSs, coded caching enables each BS to store different fragments or proportions of the encoded contents, and a recipient of the video can construct the content file based on a set of pre-defined decoding rules. Recently, the network coding-based caching has been accepted as a promising content placement scheme, such as the maximum distance separable (MDS) coding-based caching schemes [10]–[13]. Random caching schemes have also gained a lot of interest [14], [15], where video files or their combinations are randomly placed under certain probability distributions to yield optimal successful transmission probability. Note that random caching can be regarded as a special case of uncoded caching, since complete video files are cached; however, the caching probabilities have yet to be determined. Till now, wireless caching has been extensively studied in cloud radio access networks (C-RAN) [4], [16], heterogeneous networks [11], [14], device-to-device (D2D) communications [17], [18], small cell networks (SCNs) [13], [19], [20] and networks combined with D2D and SCNs [21], etc. Consensus has been reached that wireless caching is able to reduce power consumption, service delay and backhaul-link traffic burden, as well as improve the request hit ratio.

Perceptual requirements of video subscribers can have strong impact on the design of wireless caching system. For example, people generally request standard viewing quality for news reports and sports games, and high viewing quality

Manuscript received December 11, 2017; revised April 17, 2018; accepted April 18, 2018. Date of publication June 7, 2018; date of current version October 30, 2018. This work was supported by the National Natural Science Foundation of China under Grant 61671072. *(Corresponding author: Tiejun Lv.)*

X. Zhang, T. Lv, and N. C. Beaulieu are with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: zhangxw@bupt.edu.cn; lvtiejun@bupt.edu.cn; nborm@bupt.edu.cn).

W. Ni is with Data61, Commonwealth Scientific and Industrial Research, Sydney, NSW 2122, Australia (e-mail: wei.ni@data61.csiro.au).

J. M. Cioffi is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: cioffi@stanford.edu).

Y. J. Guo is with the Global Big Data Technologies Centre, University of Technology Sydney, Ultimo, NSW 2007, Australia (e-mail: jay.guo@uts.edu.au).

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/JSAC.2018.2844998

for movies and TV series. However, this issue has not been captured in the aforementioned existing studies. Scalable video coding (SVC), as part of the H.265 standard [22], is able to flexibly remove part of the video bit streams to adjust to various user requirements and network states while guaranteeing acceptable video quality. To elaborate a little further, in SVC, each video file is divided into a base layer (BL) and multiple enhancement layers (ELs). Videos with only BL can provide fundamental viewing quality, while EL contents can complement to the BL to provide superior video quality. It is worth noting that EL contents cannot be decoded without the corresponding BL [23]. Some research efforts have been devoted to combining SVC with wireless caching. For example, Wu and Zhang [24] analyze the successful transmission rate and backhaul load in cellular networks by caching and transmitting scalable videos, and confirm the effectiveness of taking different viewing quality requirements into consideration. Additionally, Zhan and Wen [25] propose an SVC-based layer placement strategy, which can significantly reduce the average content download time. These works focus on analyzing the successful transmission rate, backhaul load and average content download time. In the near future, more diverse and comprehensive performance metrics, such as EE, are advocated so that more performance gain can be obtained through the combination of wireless caching and SVC.

In a different yet relevant context of 5G, EE is an important design criterion [26], [27], where the energy saving should not be at the cost of the quality of service (QoS) [7], [28]. The optimal EE design is critical, and can make preferable balance between total power consumption and spectrum efficiency (SE) so as to improve resource utilization [29]. It is obvious that this issue also merits consideration in cache-enabled networks. Specifically, Liu and Yang [30] provide the closed-form expression for EE and identify the conditions to obtain benefits from caching. Chen *et al.* [8] and Zhang *et al.* [31] also derive the expressions for EE by exploiting stochastic geometry. Furthermore, some researches focus on the design of optimal content placement policy to enhance EE. To be more specific, Gabry *et al.* [11] minimize the total power consumption to improve EE by designing the placement scheme of the coded packets under the MDS coding-based caching strategy. These works have been focusing on the optimization of EE, especially the minimization of total power consumption to improve EE. They fail to apply to the more practical scenario, where different viewing requirements need to be taken into consideration. It is of great importance to provide scalable video services when concentrating on EE optimizations in cache-enabled networks.

To fill this void, in this paper, we investigate the energy-efficient caching schemes to yield optimal EE performance in cache-enabled heterogeneous networks, where video files to be requested are encoded by SVC and are divided into a BL and an EL. The BL and EL contents are locally cached and cooperatively transmitted by two clusters of small-cell BSs (SBSs). We propose two caching schemes, namely, SVC-based fractional caching and SVC-based random caching, to improve the EE of the cache-enabled networks. A comprehensive performance analysis is carried out to characterize the EE of cache-enabled networks supporting SVC. To derive the closed-form expressions for EE, we establish the power consumption models and theoretically analyze the ergodic service rates. Afterwards, the optimization problems are formulated to maximize EE of the two proposed caching schemes, which are then effectively solved by the standard gradient projection method after taking the approximations of l_0-norm.

The main contributions of this paper are summarized as follows:

- We propose two new SVC-based caching schemes to support scalable video services. Both the content popularity and the quality preference of the videos are taken into account. In specific, standard definition video (SDV) and high definition video (HDV) can both be provided to the users, which are able to improve resource utilization and thereby enhance EE.
- Relying on stochastic geometry, when the designated user is served by the nearest MBS and cooperative SBSs, the closed-form expressions for successful transmission probabilities and ergodic service rates are derived, from which some useful insights are shed.
- It is more challenging to deal with the proposed EE optimization problems. After taking logarithmic approximations of the l_0-norm, we transform the objective functions into the continuous and differential ones. Finally, according to their specific forms, the EE optimization problems are efficiently solved by the proposed standard gradient projection method.

The rest of this paper is organized as follows. Section II presents the system model, including the network model, SVC-based caching schemes, channel model and power consumption model of the cache-enabled heterogeneous networks. Closed-form expressions for successful transmission probabilities and ergodic service rates are derived in Section III. Under the two proposed SVC-based caching schemes, the EE optimization problems are formulated and solved in Section IV and Section V, respectively. Extensive simulation results are presented in Section VI, and this paper concludes in Section VII.

II. System Model

In this section, we present the system model of the cache-enabled networks supporting SVC, including the network model, SVC-based caching schemes, channel model and power consumption model.

A. Network Model

Fig. 1 illustrates the heterogeneous network of interest, including an MBS tier and an SBS tier. Multiple MBSs and SBSs are independently and identically distributed in the observed network, whose locations follow homogeneous Poisson Point Processes (PPPs) Φ_{M} and Φ_{S} with densities λ_M and λ_S, respectively. Without loss of generality, our analysis is carried out for a designated user, which is located at the center of the network [8], [14], [31], [32]. All the BSs and the user are assumed to be equipped with a single antenna.

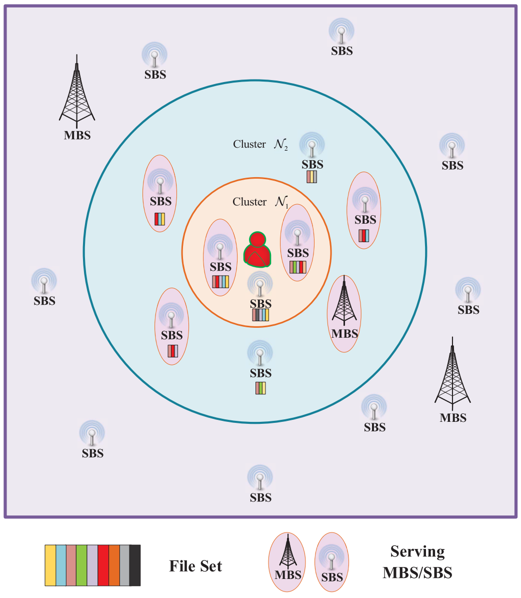


Fig. 1. In the proposed system model, the designated user is located at the center of the observed network and requests the video file colored in red. The serving SBSs are those who cache the BL and EL contents colored in red and are located in the circle and annular areas. If the user cannot obtain complete video layers, the nearest MBS will be triggered to retrieve the remaining video contents.

To implement different caching and transmission assignments, the SBSs are grouped into two clusters. In specific, taking the user's position as the center, SBSs within the circle with radius a form cluster $\mathcal{N}_1$ and the number of SBSs in this cluster is $N_1 = |\mathcal{N}_1|$. The SBSs located in the annulus with radii a and b $(a < b)$ form cluster $\mathcal{N}_2$, and the number of SBSs in this cluster is $N_2 = |\mathcal{N}_2|$.

Employing SVC, each video file is encoded into a BL and an EL.[1] The BL provides fundamental video quality, while users with both BL and EL contents can acquire superior video quality. Video files with only BL contents are defined as SDVs, and videos with both BL and EL contents are defined as HDVs [24]. The SBSs in cluster $\mathcal{N}_1$ are assigned to cache BL contents, while ELs are cached in the local storage of SBSs belonging to cluster $\mathcal{N}_2$.[2] When the designated user requests the f-th video file, the SBSs located in cluster $\mathcal{N}_1$ are responsible for transmitting the BL content of this file; and the SBSs located in cluster $\mathcal{N}_2$ will deliver the EL content if HDV is required. Some SBSs in clusters $\mathcal{N}_1$ and $\mathcal{N}_2$ that do not possess the required BL and EL contents remain silent until the next transmission process begins. When cooperative SBSs fail to provide the complete video layers, the nearest MBS is activated to deliver the remaining video contents from the core networks via backhaul links.

B. SVC-Based Caching Schemes

Assume that there are a total of F video files requested by the user. The sizes of each BL and EL are L_B and L_E, respectively.[3] The total cache size at each SBS is denoted by M. According to what mentioned above, the number of locally cached BL and EL contents can be given by $M_B = \lfloor (M/L_B) \rfloor$ and $M_E = \lfloor (M/L_E) \rfloor$. All videos are arranged in the descending order of popularity, where more popular videos are ranked with smaller indices. We assume that the probability of video requests follows the Zipf's law, as given by [33]

$$p_f = \frac{f^{-\alpha}}{\sum_{n=1}^{F} n^{-\alpha}}, \quad f = 1, 2, \ldots, F, \tag{1}$$

where α is the skewness parameter that characterizes the concentration of video requests [34]. The viewing quality preference for SDV and HDV is also considered. According to [24], the SDV perceptual preference of the f-th video file can be modeled as

$$g_{\text{SDV}}(f) = \frac{f-1}{F-1}, \quad f = 1, 2, \ldots, F. \tag{2}$$

The HDV perceptual preference can be accordingly denoted as $g_{\text{HDV}}(f) = 1 - g_{\text{SDV}}(f)$.

Based on the above content transmission protocol and user request model, we propose two caching schemes, which are described as follows:

- Scheme I (SVC-based fractional caching). The SBSs located in cluster $\mathcal{N}_1$ cache parts of the BL contents of the video files, where the cached parts are the same across the SBSs. For SBSs located in cluster $\mathcal{N}_2$, some parts of the EL contents are cached in their local storage, where the cached parts are also the same across these SBSs. Let $Q_{1,f}$ and $Q_{2,f}$ denote the caching fractions of the BL and EL contents of the f-th video file in SBSs belonging to clusters $\mathcal{N}_1$ and $\mathcal{N}_2$, respectively. They can be stacked into vectors $\mathbf{Q}_1 = [Q_{1,1}, Q_{1,2}, \ldots, Q_{1,F}]$ and $\mathbf{Q}_2 = [Q_{2,1}, Q_{2,2}, \ldots, Q_{2,F}]$.
- Scheme II (SVC-based random caching). The SBSs located in clusters $\mathcal{N}_1$ and $\mathcal{N}_2$ randomly cache the complete BL and EL contents under certain probability distributions, respectively. Let $T_{1,f}$ and $T_{2,f}$ denote the probabilities for caching BL and EL contents of the f-th video file in SBSs belonging to clusters $\mathcal{N}_1$ and $\mathcal{N}_2$, respectively. They can be stacked into vectors $\mathbf{T}_1 = [T_{1,1}, T_{1,2}, \ldots, T_{1,F}]$ and $\mathbf{T}_2 = [T_{2,1}, T_{2,2}, \ldots, T_{2,F}]$.

In the following sections, we will derive the optimal caching distributions to yield optimal EE of the cache-enabled network.

[1] In this paper, we consider two-layer video caching and transmission for illustration convenience. When the video files are divided into more layers, there can be more clusters of SBSs to cooperatively serve the user, while the performance analysis and the problem formulation can follow the exactly same steps proposed in the paper. To this end, the proposed two-layer video caching and transmission model is instrumental, and can be readily extended to the multi-layer cases.

[2] BL content is the most fundamental part of a scalable video, since the decoding of EL largely depends on the received BL. Owing to the paramount importance of BLs, the SBSs located in cluster $\mathcal{N}_1$, which are closer to the user and are capable of providing stronger signal strength than those in cluster $\mathcal{N}_2$, are responsible for caching and delivering BLs, while SBSs located in cluster $\mathcal{N}_2$ are assigned to provide ELs.

[3] Without loss of generality, equal video sizes are assumed in this paper. Of course, we can also consider different content sizes in the proposed schemes, which has little effect on the performance analysis.

C. Channel Model

Interference typically dominates over noises in modern cellular networks. This paper is therefore focusing on an interference-limited case, where the background additive white Gaussian noise is comparatively negligible at the user, as compared to the co-channel interference [13], [14], [24]. When the cached BL content of the f-th video file is received, the signal-to-interference ratio (SIR) of the designated user is expressed as

$$\mathrm{SIR}_{S,BL} = \frac{\left|\sum_{k\in\mathcal{N}_{1,f}} h_{S,k}\sqrt{P_S}r_{S,k}^{-\frac{\alpha_S}{2}}\right|^2}{\sum_{n\in\Phi_\mathrm{S}\backslash\mathcal{N}_1}|h_{S,n}|^2P_Sr_{S,n}^{-\alpha_S}+\sum_{m\in\Phi_\mathrm{M}}|h_{M,m}|^2P_Mr_{M,m}^{-\alpha_M}}, \tag{3}$$

where $\mathcal{N}_{1,f}$ denotes the set of SBSs located in cluster $\mathcal{N}_1$ that cache the BL content of the f-th video file. $h_{S,k}$ and $h_{M,m}$ are the channel gains from the k-th SBS and the m-th MBS, following the complex Gaussian distribution with zero mean and unit variance, i.e., $\mathcal{CN} \sim (0,1)$. P_S and P_M are the transmit powers of SBS and MBS, respectively. $r_{S,k}$ is the distance between the k-th SBS and the user. Likewise, $r_{M,m}$ is the distance between the m-th MBS and the user. Additionally, α_S and α_M are the path loss exponents from the SBSs and MBSs to the user. In (3), the first term in the denominator of the right hand side accounts for the interference from other SBSs, except those located in cluster $\mathcal{N}_1$, and the second term represents the cross-tier interference from all the MBSs. When the cached EL content of the f-th video file is delivered to the user, the received SIR is given by

$$\mathrm{SIR}_{S,EL} = \frac{\left|\sum_{k\in\mathcal{N}_{2,f}} h_{S,k}\sqrt{P_S}r_{S,k}^{-\frac{\alpha_S}{2}}\right|^2}{\sum_{n\in\Phi_\mathrm{S}\backslash\mathcal{N}_2}|h_{S,n}|^2P_Sr_{S,n}^{-\alpha_S}+\sum_{m\in\Phi_\mathrm{M}}|h_{M,m}|^2P_Mr_{M,m}^{-\alpha_M}}, \tag{4}$$

where $\mathcal{N}_{2,f}$ denotes the set of SBSs located in cluster $\mathcal{N}_2$ that cache EL content of the f-th video file. As mentioned, the required video layers may not be completely provided by cooperative SBSs. In this case, the remaining video contents can proceed to be delivered by the nearest MBS. As a result, the received SIR is denoted as

$$\mathrm{SIR}_{M} = \frac{\left|h_{M,m_0}\right|^2 P_M r_{M,m_0}^{-\alpha_M}}{\sum_{n\in\Phi_\mathrm{S}}|h_{S,n}|^2P_Sr_{S,n}^{-\alpha_S}+\sum_{m\in\Phi_\mathrm{M}\backslash m_0}|h_{M,m}|^2P_Mr_{M,m}^{-\alpha_M}}, \tag{5}$$

where m_0 refers to the nearest MBS that provides the user with the required video content.

D. Power Consumption Model

The total power consumption of the cache-enabled network can be modeled as

$$P_\mathrm{Total} = P_\mathrm{TR} + P_\mathrm{CA} + P_\mathrm{BH} + P_\mathrm{Fix}, \tag{6}$$

where P_TR, P_CA, P_BH and P_Fix are the power consumptions for data transmission, content caching, backhaul delivery and other fixed budgets, respectively. Details of these kinds of power consumptions are illustrated as follows.

After placing parts of the video layers in the local storage of SBSs, the caching power consumption exists. As described in [11] and [35], the caching power consumption is proportional to the number of the data bits stored in the BSs. Therefore, the caching power consumption is calculated as

$$P_\mathrm{CA} = c_{ca}N_{ca}, \tag{7}$$

where c_{ca} is the caching coefficient in $\mathrm{W/bit}$ and N_{ca} is the total number of data bits cached in the local storage of SBSs belonging to clusters $\mathcal{N}_1$ and $\mathcal{N}_2$.

Due to the limited storage of SBSs, all required video layers cannot be locally cached. Parts of them need to be retrieved from the nearest MBS via backhaul links. This leads to the backhaul power consumption. The backhaul consumption is proportional to the total number of the data bits transmitted via backhaul links, which is given by

$$P_\mathrm{BH} = c_{bh}N_{bh}, \tag{8}$$

where c_{bh} is the coefficient of backhaul power consumption in $\mathrm{W/bit}$ and N_{bh} is the total number of data bits delivered by backhaul links. In practical implementations, for video files with the same sizes, caching them typically consumes less power consumption than delivering them via microwave backhaul links. In this sense, c_{ca} is typically less than c_{bh} [30].

As for the fixed power consumption P_Fix, it is mainly caused by site-cooling, controlling and running circuit components and the oscillator. In the proposed network model, the fixed power consumption is calculated as

$$P_\mathrm{Fix} = (N_1 + N_2)P_\mathrm{S,Fix} + P_\mathrm{M,Fix}, \tag{9}$$

where $P_\mathrm{S,Fix}$ and $P_\mathrm{M,Fix}$ are the fixed power consumption constants for the SBSs and MBSs, respectively.

III. The Ergodic Service Rate

In this section, we derive the expressions for successful transmission probabilities and ergodic service rates when the user is served by the nearest MBS and cooperative SBSs, which provides the corner stone to derive the expressions for EE.

A. The Ergodic Service Rate From the Nearest MBS

Given limited storage capacity of SBSs, they can hardly cache all of the BL and EL contents locally. When the complete content layers of the required videos cannot be delivered to the user's side, the nearest MBS proceeds to provide the remaining video contents. In the following, we derive the ergodic service rate in the case where the user is served by its nearest MBS.

Definition 1: When the nearest MBS provides the user with the required video layers, the ergodic service rate is written as

$$R_M(\gamma) \triangleq W\mathbb{E}\left\{\log_2(1+\mathrm{SIR}_M)|\mathrm{SIR}_M \geq \gamma\right\}, \tag{10}$$

where W is the spectrum bandwidth and γ is set as the minimum QoS requirement. $\mathbb{E}\{\cdot\}$ takes the expectation with respect to small-scale fading, and the locations of PPP-distributed MBSs and SBSs.

Note that in (10), the minimum QoS requirement is guaranteed with the given $R_M(\gamma)$. To derive the expression for $R_M(\gamma)$, we have to first derive the successful transmission probability, i.e., $P(\mathrm{SIR}_M \geq \gamma)$, which can be provided by Lemma 1.

Lemma 1: When the user is served by the nearest MBS under the minimum QoS requirement γ, the successful transmission probability is given in (11), as shown at the top of the next page, where $G_\alpha(x) = \int_x^\infty \frac{1}{1+t^{\frac{\alpha}{2}}}\mathrm{d}t$.

Proof: See Appendix A. ■

From (11), we can find that the derived expression for $P(\mathrm{SIR}_M \geq \gamma)$ can be complicated. Fortunately, when $\alpha_M = \alpha_S = 4$, (11) can be expressed in a much simpler form, as shown in the following corollary.

Corollary 1: When $\alpha_M = \alpha_S = 4$, the expression for $P(\mathrm{SIR}_M \geq \gamma)$ can be simplified as

$$P(\mathrm{SIR}_M \geq \gamma) = (1+\lambda_M^{-1}\gamma^{\frac{1}{2}}(\frac{\pi}{2}\lambda_S(\frac{P_S}{P_M})^{\frac{1}{2}} + \lambda_M \operatorname{arccot}(\gamma^{-\frac{1}{2}})))^{-1}. \tag{14}$$

Proof: See Appendix B. ■

In Corollary 1, the simplified form of $P(\mathrm{SIR}_M \geq \gamma)$ is obtained, from which some useful insights are shed, as follows.

- With the increasing ratio between the transmit powers of SBS and MBS, i.e., $\frac{P_S}{P_M}$, $P(\mathrm{SIR}_M \geq \gamma)$ degrades. This is due to the fact that SBSs with larger signal strength can produce stronger cross-tier interference while the useful signal strength remains the same.
- With the increasing ratio between the PPP densities of SBS and MBS, i.e., $\frac{\lambda_S}{\lambda_M}$, $P(\mathrm{SIR}_M \geq \gamma)$ deteriorates. This is because there are more SBSs, leading to severer cross-tier interference.
- As the minimum QoS requirement γ increases, the user can impose strict video quality requirement and $P(\mathrm{SIR}_M \geq \gamma)$ decreases. The reason is because the co-exiting intra-tier and inter-tier interference restricts further improvement of the successful transmission probability.

Based on the expression for $P(\mathrm{SIR}_M \geq \gamma)$ in Lemma 1, $R_M(\gamma)$ can be calculated by employing the following theorem.

Theorem 1: When the user is served by the nearest MBS, the ergodic service rate is expressed as

$$R_M(\gamma) = W\log_2(1+\gamma) + \frac{2\pi\lambda_M W}{\ln 2}\int_0^\infty x e^{-\lambda_M \pi x^2}\mathrm{d}x \times \int_\gamma^\infty \frac{P(\mathrm{SIR}_M \geq t | r_{M,m_0} = x)}{P(\mathrm{SIR}_M \geq \gamma | r_{M,m_0} = x)(1+t)}\mathrm{d}t, \tag{15}$$

where the expressions for $P(\mathrm{SIR}_M \geq t | r_{M,m_0} = x)$ and $P(\mathrm{SIR}_M \geq \gamma | r_{M,m_0} = x)$ can be found in (43), as given in Appendix A.

Proof: See Appendix C. ■

Remark 1: From Theorem 1, it is obvious that a many of factors have impact on the ergodic service rate, such as the minimum QoS requirement, the PPP densities of SBSs and MBSs, and the system bandwidth. Among them, the minimum QoS requirement plays the most important role in the ergodic service rate, since the two terms shown in (15) are both related to the minimum QoS requirement. As the QoS requirement grows, the first term of (15) increases. In the meanwhile, the successful transmission probability decreases, leading to a reduction of the second term of (15). When the increase of the first term can compensate for the loss of the second term, the ergodic service rate is enhanced. Otherwise, the performance deteriorates. Thus, it is not always the case that high QoS requirement would lead to good ergodic service rate.

B. The Ergodic Service Rates From Cooperative SBSs

We proceed to derive the expressions for ergodic service rates when the required BL and EL contents are delivered by cooperative SBSs, which are denoted by $R_{S,BL}$ and $R_{S,EL}$, respectively. The definitions of $R_{S,BL}$ and $R_{S,EL}$ are given in the following.

Definition 2: When cooperative SBSs provide the user with the required BL and EL contents, the ergodic service rates are defined as

$$R_{S,BL}(\gamma_{BL}, n_1) \triangleq W\mathbb{E}\{\log_2(1+\mathrm{SIR}_{S,BL}) | \mathrm{SIR}_{S,BL} \geq \gamma_{BL}\}, \tag{16}$$

$$R_{S,EL}(\gamma_{EL}, n_2) \triangleq W\mathbb{E}\{\log_2(1+\mathrm{SIR}_{S,EL}) | \mathrm{SIR}_{S,EL} \geq \gamma_{EL}\}, \tag{17}$$

where γ_{BL} and γ_{EL} are the minimum QoS requirements for delivering BL and EL contents, respectively. $n_1 = |\mathcal{N}_{1,f}| \leq N_1$ and $n_2 = |\mathcal{N}_{2,f}| \leq N_2$ are the numbers of the serving SBSs that cache the required BL and EL contents.

Remark 2: According to Definitions 1 and 2, it can be seen that the ergodic service rates are obtained under the constraint of minimum QoS requirements. Thus, in the EE optimization problems formulated later, the QoS constraints are inherently satisfied in the devised EE expressions, and we only focus on the cache size restrictions.

In order to obtain the expressions devised in Definition 2, we need to acquire the successful transmission probabilities when the designated user receives BL and EL contents from cooperative SBSs, i.e., $P(\mathrm{SIR}_{S,BL} \geq \gamma_{BL})$ and $P(\mathrm{SIR}_{S,EL} \geq \gamma_{EL})$. The successful transmission probabilities are given in the following lemma.

Lemma 2: When there are n_1 and n_2 SBSs to provide BL and EL contents, respectively, the successful transmission probabilities are shown in (12) and (13), as shown at the top of the next page, where $c = \frac{\gamma_{BL}}{\sum_{k=1}^{n_1} x_{BL,k}^{-\alpha_S}}$, $d = \frac{\gamma_{EL}}{\sum_{k=1}^{n_2} x_{EL,k}^{-\alpha_S}}$, $\mathbf{x}_{BL} = [x_{BL,1}, \ldots, x_{BL,n_1}]$ and $\mathbf{x}_{EL} = [x_{EL,1}, \ldots, x_{EL,n_2}]$.

Proof: See Appendix D. ■

Considering the special case of $\alpha_M = \alpha_S = 4$, we are able to achieve the simplified forms of $P(\mathrm{SIR}_{S,BL} \geq \gamma_{BL})$ and $P(\mathrm{SIR}_{S,EL} \geq \gamma_{EL})$, as dictated in the following corollary.

Corollary 2: When $\alpha_M = \alpha_S = 4$, $P(\mathrm{SIR}_{S,BL} \geq \gamma_{BL})$ and $P(\mathrm{SIR}_{S,EL} \geq \gamma_{EL})$ can be simplified as

$$P(\mathrm{SIR}_{S,BL} \geq \gamma_{BL}) = \int_0^a \cdots \int_0^a \prod_{k=1}^{n_1} \frac{2x_{BL,k}}{a^2} \exp(-\pi u c^{\frac{1}{2}})\mathrm{d}\mathbf{x}_{BL}, \tag{18}$$

$$P(\mathrm{SIR}_M \geq \gamma) = 2\pi\lambda_M \int_0^\infty x \exp(-\pi(\lambda_S(\gamma \frac{P_S}{P_M} x^{\alpha_M})^{\frac{2}{\alpha_S}} G_{\alpha_S}(0) - \lambda_M x^2(\gamma^{\frac{2}{\alpha_M}} G_{\alpha_M}(\gamma^{-\frac{2}{\alpha_M}}) + 1)))\mathrm{d}x, \quad (11)$$

$$P(\mathrm{SIR}_{S,BL} \geq \gamma_{BL}) = \int_0^a \int_0^a \cdots \int_0^a \prod_{k=1}^{n_1} \frac{2x_{BL,k}}{a^2} \exp(-\pi(\lambda_S c^{\frac{2}{\alpha_S}} G_{\alpha_S}(a^2 c^{-\frac{2}{\alpha_S}}) - \lambda_M (c\frac{P_M}{P_S})^{\frac{2}{\alpha_M}} G_{\alpha_M}(0)))\mathrm{d}\mathbf{x}_{BL}, \quad (12)$$

$$P(\mathrm{SIR}_{S,EL} \geq \gamma_{EL}) = \int_a^b \int_a^b \cdots \int_a^b \prod_{k=1}^{n_2} \frac{2x_{EL,k}}{b^2 - a^2}$$
$$\exp(-\pi(\lambda_M(d\frac{P_M}{P_S})^{\frac{2}{\alpha_M}} G_{\alpha_M}(0) - \lambda_S d^{\frac{2}{\alpha_S}}(\int_0^{a^2 d^{-\frac{2}{\alpha_S}}} \frac{1}{1+t^{\frac{\alpha_S}{2}}}\mathrm{d}t + G_{\alpha_S}(b^2 d^{-\frac{2}{\alpha_S}}))))\mathrm{d}\mathbf{x}_{EL}, \quad (13)$$

$$P(\mathrm{SIR}_{S,EL} \geq \gamma_{EL}) = \int_a^b \cdots \int_a^b \prod_{k=1}^{n_2} \frac{2x_{EL,k}}{b^2 - a^2} \exp(-\pi v d^{\frac{1}{2}})\mathrm{d}\mathbf{x}_{EL}, \quad (19)$$

where

$$u = \lambda_S \mathrm{arccot}(a^2 c^{-\frac{1}{2}}) + \frac{\pi}{2}\lambda_M(\frac{P_M}{P_S})^{\frac{1}{2}}, \quad (20)$$

$$v = \lambda_S(\arctan(a^2 d^{-\frac{1}{2}}) + \mathrm{arccot}(b^2 d^{-\frac{1}{2}})) + \frac{\pi}{2}\lambda_M(\frac{P_M}{P_S})^{\frac{1}{2}}. \quad (21)$$

Proof: Corollary 2 can be proved by following the steps shown in Appendix B, and therefore suppressed. ■

Based on Lemma 2, $R_{S,BL}(\gamma_{BL}, n_1)$ and $R_{S,EL}(\gamma_{EL}, n_2)$ can be derived, as presented in Theorem 2.

Theorem 2: Let vectors $\mathbf{r}_{S_{BL}} = [r_{BL,1}, r_{BL,2}, \ldots, r_{BL,n_1}]$ and $\mathbf{r}_{S_{EL}} = [r_{EL,1}, r_{EL,2}, \ldots, r_{EL,n_2}]$ denote the positions of the serving SBSs belonging to clusters $\mathcal{N}_1$ and $\mathcal{N}_2$, respectively. When the designated user is served by cooperative SBSs to obtain BL and EL contents, the ergodic service rates can be calculated as

$$R_{S,BL}(\gamma_{BL}, n_1) = W\log_2(1+\gamma_{BL}) + \frac{W}{\ln 2}\int_0^a \int_0^a \cdots \int_0^a \prod_{k=1}^{n_1} \frac{2x_{BL,k}}{a^2}\mathrm{d}\mathbf{x}_{BL} \times \int_{\gamma_{BL}}^\infty \frac{P(\mathrm{SIR}_{S,BL} \geq t|\mathbf{r}_{S_{BL}} = \mathbf{x}_{BL})}{P(\mathrm{SIR}_{S,BL} \geq \gamma_{BL}|\mathbf{r}_{S_{BL}} = \mathbf{x}_{BL})(1+t)}\mathrm{d}t, \quad (22)$$

$$R_{S,EL}(\gamma_{EL}, n_2) = W\log_2(1+\gamma_{EL}) + \frac{W}{\ln 2}\int_a^b \int_a^b \cdots \int_a^b \prod_{k=1}^{n_2} \frac{2x_{EL,k}}{(b^2-a^2)}\mathrm{d}\mathbf{x}_{EL} \times \int_{\gamma_{EL}}^\infty \frac{P(\mathrm{SIR}_{S,EL} \geq t|\mathbf{r}_{S_{EL}} = \mathbf{x}_{EL})}{P(\mathrm{SIR}_{S,EL} \geq \gamma_{EL}|\mathbf{r}_{S_{EL}} = \mathbf{x}_{EL})(1+t)}\mathrm{d}t, \quad (23)$$

where the expressions for $P(\mathrm{SIR}_{S,BL} \geq t|\mathbf{r}_{S_{BL}} = \mathbf{x}_{BL})$ and $P(\mathrm{SIR}_{S,BL} \geq \gamma_{BL}|\mathbf{r}_{S_{BL}} = \mathbf{x}_{BL})$ can be found in (51), and the expressions for $P(\mathrm{SIR}_{S,EL} \geq t|\mathbf{r}_{S_{EL}} = \mathbf{x}_{EL})$ and $P(\mathrm{SIR}_{S,EL} \geq \gamma_{EL}|\mathbf{r}_{S_{EL}} = \mathbf{x}_{EL})$ can be found in (57).

Proof: This theorem can be proved by following the steps shown in Appendix C, and therefore omitted for brevity. ■

From Theorem 2, some useful insights can also be observed, which are similar to those shown in Remark 1. For avoiding repetition, they are omitted here. From the above derivations, we have successfully obtained the expressions for ergodic service rates, which are the key intermediate steps to gain the EE expressions.

IV. The EE Optimization Problem for Scheme I

In this section, we derive the expressions for total power consumption and the sum rate, based on which the EE maximization problem is formulated. The proposed optimization problem is approximated, and then efficiently solved by the proposed standard gradient projection method.

A. EE Optimization Problem Formulation

Under Scheme I, the total power consumption of the network of interest can be quantified as

$$P_{\mathrm{Total},1} = P_{\mathrm{TR},1} + P_{\mathrm{CA},1} + P_{\mathrm{BH},1} + P_{\mathrm{Fix}}, \quad (24)$$

where

$$P_{\mathrm{TR},1} = \sum_{f=1}^F p_f(\zeta_S(N_1 \|Q_{1,f}\|_0 + g_{\mathrm{HDV}}(f) N_2 \|Q_{2,f}\|_0)P_S + \zeta_M(\|1-Q_{1,f}\|_0 + g_{\mathrm{HDV}}(f)\|1-Q_{2,f}\|_0)P_M), \quad (25)$$

$$P_{\mathrm{CA},1} = c_{ca}\sum_{f=1}^F (Q_{1,f} L_B N_1 + Q_{2,f} L_E N_2), \quad (26)$$

and

$$P_{\mathrm{BH},1} = c_{bh}\sum_{f=1}^F p_f((1-Q_{1,f})L_B + g_{\mathrm{HDV}}(1-Q_{2,f})L_E). \quad (27)$$

In (25), ζ_S and ζ_M are the power efficiency coefficients of the power amplifiers of the SBSs and MBSs, respectively, and $\|\cdot\|_0$ stands for l_0-norm.

Based on the proposed SVC-based factional caching and cooperative transmission schemes, the sum rate of the designated user is given by

$$R_{\mathrm{Sum},1} = \sum_{f=1}^F p_f \underbrace{((1-Q_{1,f})R_M(\gamma_{BL}) + g_{\mathrm{HDV}}(f)(1-Q_{2,f})R_M(\gamma_{EL})}_{\text{served by the nearest MBS to obtain BL and EL contents}} + \underbrace{Q_{1,f}R_{S,BL}(\gamma_{BL}, N_1) + g_{\mathrm{HDV}}(f)Q_{2,f}R_{S,EL}(\gamma_{EL}, N_2))}_{\text{served by cooperative SBSs to obtain BL and EL contents}}. \quad (28)$$

From (28), intuitions can be found that the performance of $R_{\mathrm{Sum},1}$ can predominately depend on the caching fractions, the number of serving SBSs and the minimum QoS requirements. Moreover, if HDV service is required by the user, the SBSs located in cluster $\mathcal{N}_2$ are activated to deliver the EL contents. This can improve the sum rate performance. According to the expressions for total power consumption and sum rate presented above, the EE maximization problem with BL and EL caching fraction design can be formulated as

$$\max_{\mathbf{Q}_1,\mathbf{Q}_2} EE_1 = \frac{R_{\mathrm{Sum},1}}{P_{\mathrm{Total},1}} \tag{29a}$$

$$\text{s.t.} \sum_{f=1}^{F} Q_{1,f} = M_B, \tag{29b}$$

$$\sum_{f=1}^{F} Q_{2,f} = M_E, \tag{29c}$$

$$0 \leq Q_{1,f} \leq 1, \quad 0 \leq Q_{2,f} \leq 1, \ \forall f, \tag{29d}$$

where (29b) and (29c) are the cache size restrictions of each SBS when caching BL and EL contents, respectively; (29d) indicates the feasible solution regions of the optimization variables $Q_{1,f}$ and $Q_{2,f}$.

B. The Proposed Algorithm

In (29), constraints (29b) to (29d) form a convex variable set, while the objective function has a complicated expression. Moreover, the l_0-norm in $P_{\mathrm{Total},1}$ makes the problem more challenging to deal with. Note that the l_0-norm can be approximated by a logarithmic function, an exponential function or an arctangent function [4]. Without loss of generality, the logarithmic function is adopted in this paper as the smooth function. Then, the l_0-norm in $P_{\mathrm{TR},1}$ can be approximated as

$$\begin{aligned}\hat{P}_{\mathrm{TR},1} = \sum_{f=1}^{F} p_f(&\zeta_S(N_1 f_\theta(Q_{1,f}) + g_{\mathrm{HDV}}(f) N_2 f_\theta(Q_{2,f}))P_S \\ &+ \zeta_M(f_\theta(1-Q_{1,f}) + g_{\mathrm{HDV}}(f) f_\theta(1-Q_{2,f}))P_M),\end{aligned} \tag{30}$$

where

$$f_\theta(x) = \log(\frac{x}{\theta}+1)/\log(\frac{1}{\theta}+1). \tag{31}$$

In (31), θ is a constant parameter to reflect the smoothness of $f_\theta(x)$. A larger value for θ leads to a smoother function but less accurate approximation. After the approximations, the objective function (29a) is a continuous and differentiable function with respect to $Q_{1,f}$ and $Q_{2,f}$. Due to the specific form of the transformed EE optimization problem, the standard gradient projection algorithm is employed [14], [15], [36], and the suboptimal solution of (29) can be obtained, as summarized in Algorithm 1. In the proposed algorithm, $\epsilon(t)$ is the iteration step size, which satisfies $\lim\limits_{t\to 0}\epsilon(t) = 0$, $\lim\limits_{T\to\infty}\sum_{t=1}^{T}\epsilon(t) = \infty$ and $\lim\limits_{T\to\infty}\sum_{t=1}^{T}\epsilon^2(t) < \infty$. We set $\epsilon(t) = \frac{1}{t}$ in the t-th iteration. Note that Steps 3) and 5) give the projections of $\hat{Q}_{1,f}(t+1)$ and $\hat{Q}_{2,f}(t+1)$ onto the optimization variable set so that constraints (29b) and (29c) are satisfied, where $[x]^+ \triangleq \max[\mathrm{x}, 0]$.

Algorithm 1 The Proposed Standard Gradient Projection Method for Solving the EE Optimization Problem (29)

```
1) Initialization: Set $t = 1$, $\epsilon(1) = 1$, and find $Q_{1,f}$ and $Q_{2,f}$ that are feasible for constraints (29b)-(29d).
2) For $f \in \{1,\ldots,F\}$, calculate $\frac{\partial EE_1}{Q_{1,f}}$, and then obtain
   $\hat{Q}_{1,f}(t+1) = Q_{1,f}(t) + \epsilon(t)\frac{\partial EE_1}{Q_{1,f}} |_{Q_{1,f}=Q_{1,f}(t),Q_{2,f}=Q_{2,f}(t)}$.
3) For $f \in \{1,\ldots,F\}$, calculate $Q_{1,f}(t+1) = \min\{[\hat{Q}_{1,f}(t+1) - u']^+, 1\}$, where $u'$ satisfies $\sum_{f=1}^{F}\min\{[\hat{Q}_{1,f}(t+1) - u']^+, 1\} = M_B$.
4) For $f \in \{1,\ldots,F\}$, calculate $\frac{\partial EE_1}{Q_{2,f}}$, and then obtain
   $\hat{Q}_{2,f}(t+1) = Q_{2,f}(t) + \epsilon(t)\frac{\partial EE_1}{Q_{2,f}} |_{Q_{1,f}=Q_{1,f}(t),Q_{2,f}=Q_{2,f}(t)}$.
5) For $f \in \{1,\ldots,F\}$, calculate $Q_{2,f}(t+1) = \min\{[\hat{Q}_{2,f}(t+1) - v']^+, 1\}$, where $v'$ satisfies $\sum_{f=1}^{F}\min\{[\hat{Q}_{2,f}(t+1) - v']^+, 1\} = M_E$.
6) If convergence, the algorithm terminates. Otherwise, set $t = t+1$ and $\epsilon(t) = \frac{1}{t}$, then go back to Step 2).
```

V. The EE Optimization Problem for Scheme II

For Scheme II, according to the power consumption model established previously, the total power consumption is given by

$$P_{\mathrm{Total},2} = P_{\mathrm{TR},2} + P_{\mathrm{CA},2} + P_{\mathrm{BH},2} + P_{\mathrm{Fix}}, \tag{32}$$

where

$$\begin{aligned}P_{\mathrm{TR},2} = \sum_{f=1}^{F} p_f(&\zeta_S(N_1 T_{1,f} + g_{\mathrm{HDV}}(f) N_2 T_{2,f})P_S \\ &+ \zeta_M((1-T_{1,f}) + g_{\mathrm{HDV}}(f)(1-T_{2,f}))P_M),\end{aligned} \tag{33}$$

$$P_{\mathrm{CA},2} = c_{ca}\sum_{f=1}^{F}(T_{1,f}L_B N_1 + T_{2,f}L_B N_2), \tag{34}$$

and

$$P_{\mathrm{BH},2} = c_{bh}\sum_{f=1}^{F} p_f((1-T_{1,f})^{N_1}L_B + g_{\mathrm{HDV}}(1-T_{2,f})^{N_2}L_E). \tag{35}$$

To derive the expression for EE, the sum rate expression needs to be first derived. To this end, the numbers of the serving SBSs in clusters $\mathcal{N}_1$ and $\mathcal{N}_2$ remain to be determined. Under Scheme II, each SBS randomly selects BL and EL contents to cache in its local storage under probability distributions $\mathbf{T}_1$ and $\mathbf{T}_2$, respectively. As a result, the number of the serving SBSs follows the binomial distribution. To be more specific, the number of serving SBSs in cluster $\mathcal{N}_1$ which cache BL content of the f-th file, i.e., $|\mathcal{N}_{1,f}|$, follows the binomial distribution with parameters N_1 and $T_{1,f}$, while $|\mathcal{N}_{2,f}|$ follows the binomial distribution with parameters N_2

and $T_{2,f}$. Therefore, the sum rate of the designated user can be written as

$$\begin{aligned}
&R_{\text{Sum},2}\\
&=\sum_{f=1}^{F} p_f\\
&\underbrace{((1-T_{1,f})^{N_1} R_M(\gamma_{BL}) + g_{\text{HDV}}(f)(1-T_{2,f})^{N_2} R_M(\gamma_{EL})}_{\text{served by the nearest MBS to obtain BL and EL contents}}\\
&+\sum_{n_1=1}^{N_1} \underbrace{C_{N_1}^{n_1}(T_{1,f})^{n_1}(1-T_{1,f})^{N_1-n_1} R_{S,BL}(\gamma_{BL},n_1)}_{\text{served by cooperative SBSs to obtain BL contents}} +\\
&\underbrace{g_{\text{HDV}}(f)\sum_{n_2=1}^{N_2} C_{N_2}^{n_2}(T_{2,f})^{n_2}(1-T_{2,f})^{N_2-n_2} R_{S,EL}(\gamma_{EL},n_2)).}_{\text{served by cooperative SBSs to obtain EL contents}}
\end{aligned} \tag{36}$$

From (36), it can be concluded that $R_{\text{sum},2}$ depends on the caching probabilities, the number of serving SBSs, and the minimum QoS requirements. Particularly, when there are more serving SBSs, the requested video layers are more likely to be obtained locally. This can lead to less power consumption and relieve traffic congestion in backhaul. According to (32) and (36), the EE maximization problem with BL and EL caching probability design can be formulated as

$$\max_{\mathbf{T}_1,\mathbf{T}_2} \ EE_2 = \frac{R_{\text{Sum},2}}{P_{\text{Total},2}} \tag{37a}$$

$$\text{s.t.} \ \sum_{f=1}^{F} T_{1,f} = M_B, \tag{37b}$$

$$\sum_{f=1}^{F} T_{2,f} = M_E, \tag{37c}$$

$$0 \leq T_{1,f} \leq 1, \quad 0 \leq T_{2,f} \leq 1, \ \forall f, \tag{37d}$$

where (37b) and (37c) are the cache size constraints of each SBS when caching the BL and EL contents, respectively. (37d) specifies the feasible solution regions of the caching probabilities $T_{1,f}$ and $T_{2,f}$. It is obvious that (37) has the same form as (29). Therefore, in the same manner, this problem can also be effectively solved by using the standard gradient projection method, and more details can refer to Algorithm 1.

Remark 3: For Scheme II, the SBSs are capable of caching the complete video layers, while Scheme I aims at caching parts of the BL and EL contents. In Scheme II, video layers of the popular files are more likely to be completely cached and deliveries from the backhaul links are prevented when these videos are requested. In contrast, for Scheme I, each SBS caches parts of the popular video layers, and the remaining contents need to be retrieved from backhauls, consuming more power than caching. As a consequence, it can be concluded that randomly caching the complete video layers provides better EE than fractionally caching parts of them. The superiority of Scheme II is demonstrated in terms of EE, especially in the presence of high backhaul power consumption. These conclusions can be collaborated by extensive simulations as will be shown in Section VI.

TABLE I
VALUES OF SIMULATION PARAMETERS

Parameter	Value
N_1, N_2	4
P_S, P_M	23 dBm (default), 43 dBm
λ_S, λ_M	$1/(100^2\pi)$, $1/(250^2\pi)$
a, b	50 m, 100 m
α_S, α_M	4
M	500 Mbits (default)
F	20
L_B, L_E	100 Mbits, 200 Mbits
α	1 (default)
W	10 Mbits
γ_{BL}, γ_{EL}	10 dB (default), 5 dB (default)
ζ_S, ζ_M	4.7
c_{ca}	6.25×10^{-12} W/bit [30]
c_{bh}	5×10^{-7} W/bit [30]
$P_{\text{S,Fix}}$	6.8 W [37]
$P_{\text{M,Fix}}$	130 W [37]

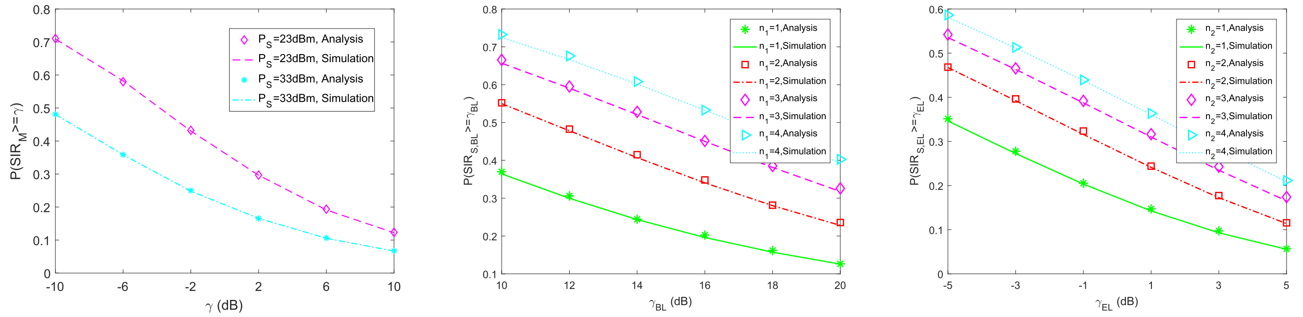
VI. SIMULATION RESULTS

In this section, we show the simulation results of the derived expressions for successful transmission probabilities and ergodic service rates, as well as the EE performance under the proposed SVC-based caching schemes. The simulation parameters are listed in Table I. For comparison purpose, three benchmarks are simulated, which are described as follows:

- Most Popular Content Placement (MPCP): For SBSs located in clusters $\mathcal{N}_1$ and $\mathcal{N}_2$, M_B BL contents and M_E EL contents of the most popular video files are cached in their local storage, respectively.
- Uniform Content Placement (UCP): Regardless of the video popularity, for SBSs located in clusters $\mathcal{N}_1$ and $\mathcal{N}_2$, the same fractions of the BL and EL contents are cached, respectively.
- Independent Content Placement (ICP): The SBSs belonging to clusters $\mathcal{N}_1$ and $\mathcal{N}_2$ randomly select M_B and M_E different BL and EL contents to cache in their local storage.

In Fig. 2, we show the successful transmission probabilities derived from theoretical analysis and Monte Carlo simulations. The plots of theoretical analysis and Monte Carlo simulations are overlapped, confirming the accuracy of successful transmission probabilities given in Lemmas 1 and 2. A common trend is also revealed that higher QoS requirements lead to lower successful transmission probabilities. From Fig. 2 (a), it can be seen that a larger value of P_S results in lower $P(\text{SIR}_M \geq \gamma)$. The reason for this is because, with the increase of P_S, the co-existing SBSs can produce stronger interference towards the serving MBS, and hence reduce $P(\text{SIR}_M \geq \gamma)$. Moreover, Figs. 2 (b) and (c) lead to the conclusion that with more cooperative SBSs, the successful transmission probability can be improved.

In Fig. 3, we show the ergodic service rates, i.e., $R_M(\gamma)$, $R_{S,BL}(\gamma_{BL}, n_1)$ and $R_{S,EL}(\gamma_{EL}, n_2)$, with varying numbers

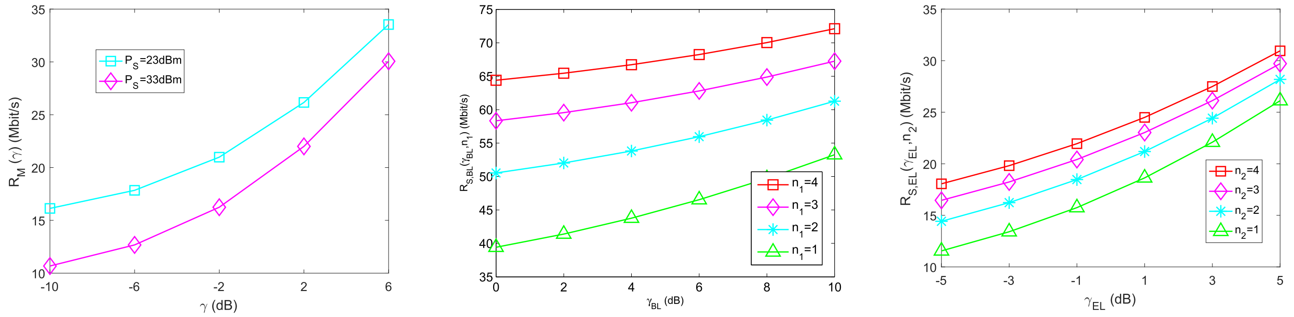


(a) $P(\mathrm{SIR}_M \geq \gamma)$ versus varying γ under different P_S.

(b) $P(\mathrm{SIR}_{S,BL} \geq \gamma_{BL})$ versus varying γ_{BL} under different n_1.

(c) $P(\mathrm{SIR}_{S,EL} \geq \gamma_{EL})$ versus varying γ_{EL} under different n_2.

Fig. 2. The successful transmission probabilities when the user is served by the nearest MBS and cooperative SBSs.



(a) $R_M(\gamma)$ versus varying γ_{BL} under different P_S.

(b) $R_{S,BL}(\gamma_{BL}, n_1)$ versus varying γ_{BL} under different n_1.

(c) $R_{S,EL}(\gamma_{EL}, n_2)$ versus varying γ_{EL} under different n_2.

Fig. 3. The ergodic service rates when the user is served by the nearest MBS and cooperative SBSs.

of cooperative SBSs and QoS requirements. Under our parameter settings, the ergodic service rates increase as the QoS requirements grow, since the improvement of the first term in (15), (22) and (23) can make up for the loss of the second term. However, higher P_S can lead to lower $R_M(\gamma)$ due to a reduction of the successful transmission probability, which coincides with the conclusion drawn from Fig. 2 (a). Moreover, in Figs. 3 (b) and 3 (c), the ergodic service rates are improved as n_1 and n_2 grow, since there are more cooperative SBSs to enhance the successful transmission probabilities. Therefore, it can be concluded that, in large scale heterogeneous networks with multiple SBSs, the performance of ergodic service rates can be further enhanced.

Fig. 4 presents the relationship between the EE performance and the transmit power of the SBSs under different caching schemes. The superiority of our proposed SVC-based caching schemes is validated. With the increase of P_S, the EE grows and the growth slows down. When P_S grows further, though the sum rate increases, the improvement of the sum rate cannot scale up with the increase of the total power consumption. This results in the degraded EE performance. When we obtain the optimal caching fractions under Scheme I, the optimal caching fractions are used to validate the accuracy of the l_0-norm approximation. As shown in Fig. 4, the approximation is accurate, and the performance loss resulting from the approximation is marginal. It can be concluded that when selecting the proper smooth parameter θ, the l_0-norm of the caching fractions can be accurately estimated. For the ICP scheme, the file selection process is performed under each channel realization. When there are sufficient channel realizations, the EE of ICP is comparable to that of UCP.

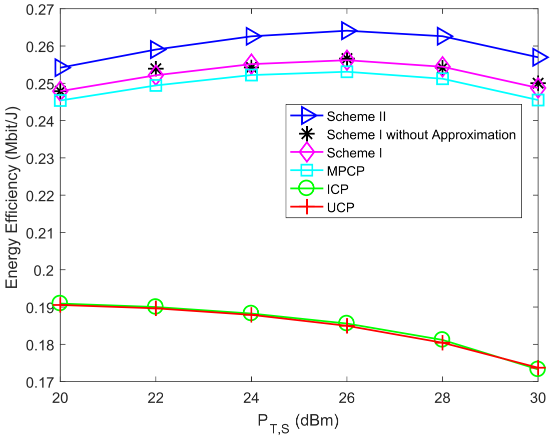


Fig. 4. The EE performance versus P_S under different caching schemes.

Fig. 5 shows the EE performance of different caching schemes under various values of γ_{BL}. It is obvious that the proposed caching schemes are superior to the three benchmarks, especially achieving much more performance gain than the UCP and ICP schemes. Furthermore, it can be seen that Scheme II can achieve higher EE than Scheme I. This indicates that randomly caching complete video layers provides better

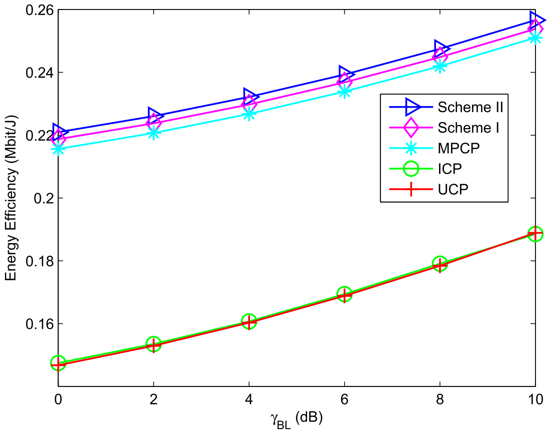


Fig. 5. The EE performance versus γ_{BL} under different caching schemes.

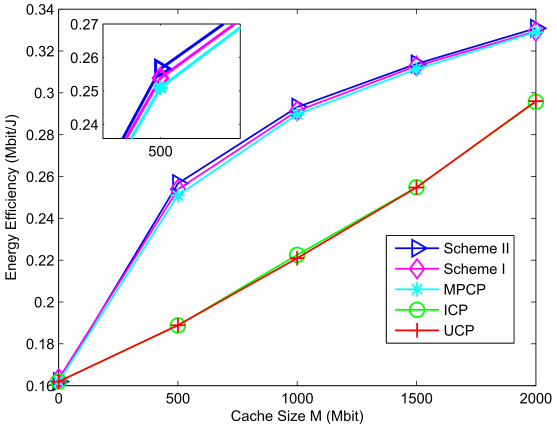


Fig. 6. The EE performance versus cache size under different caching schemes.

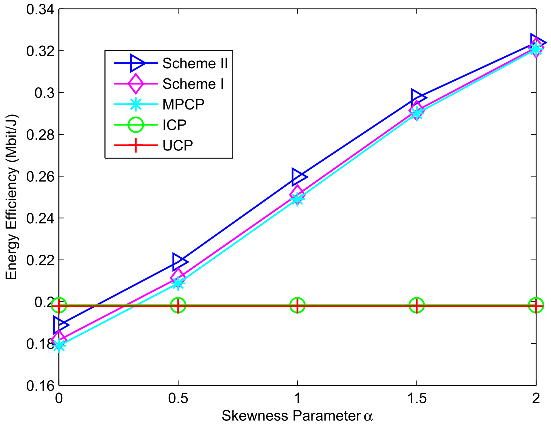


Fig. 7. The EE performance versus skewness parameter under different caching schemes.

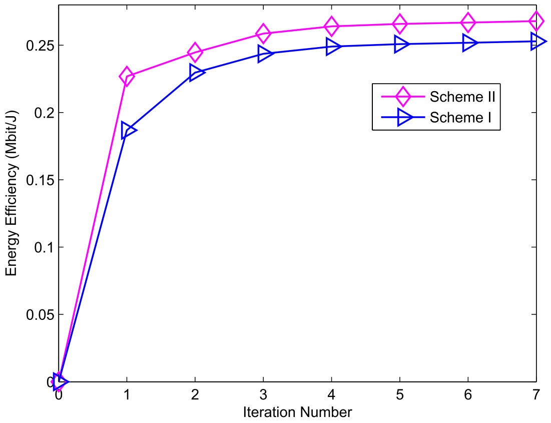


Fig. 8. The convergence property of the proposed algorithm.

EE than storing only parts of them, even under the optimal caching fractions.

In Fig. 6, we present the EE performance of different caching schemes with varying cache size M. When a larger cache size is equipped at each SBS, more video contents can be locally cached. Therefore, the demand for backhaul links can be relieved substantially, which in turn can significantly reduce service delay and backhaul power consumption. In practice, for video files with the same sizes, caching them can consume less power consumption than retrieving them from microwave backhaul links. As a result, larger cache sizes can lead to higher EE. Particularly, when $M = 0$, all caching schemes have the same EE. This is equivalent to the case with no caching. The proposed caching schemes provide better EE in small cache size region, and reach the maximum performance gap at about $M = 600$ Mbits. However, when the cache size grows further, the performance gaps gradually diminish, since all video files will be cached when the cache size of each SBS is large enough.

In Fig. 7, the relationship between the EE performance and skewness parameter α is plotted. Note that larger α means that fewer video files can meet the majority of user requests. Therefore, video files with smaller indices are much more likely to be stored in the local cache of SBSs. When α is small, the UCP and ICP schemes provide better EE. In this case, video popularities are uniformly distributed, and the UCP and ICP schemes are able to increase file diversity, and then increase request hit ratio.

In Fig. 8, we show the convergence property of the proposed algorithm. It can be seen that the proposed algorithm is able to converge after a small number of iterations, which validates the effectiveness of the standard gradient projection method in practical implementations.

VII. Conclusion

This paper proposed two energy-efficient SVC-based caching schemes to boost EE in cache-enabled heterogeneous networks. Based on the proposed caching schemes, we established the power consumption model, and derived the expressions for successful transmission probabilities and ergodic service rates. We further formulated two EE maximization problems, which are subject to the cache size constraint of each SBS. After taking approximations of the l_0-norm, the EE optimization problems can be efficiently solved. Numerical results confirmed the accuracy of our analysis as well as the superiority of the proposed caching schemes to three benchmarks.

Appendix A
Proof of Lemma 1

In the proposed model, the designated user is located at the origin of the observed network and the distance between the nearest MBS and the user is denoted by r_{M,m_0}, whose probability density function (PDF) is shown as [8]

$$f_{m_0}(r_{M,m_0}) = 2\pi\lambda_M r_{M,m_0} e^{-\lambda_M \pi r^2_{M,m_0}}. \tag{38}$$

Then, $P(\mathrm{SIR}_M \geq \gamma)$ is calculated as

$$P(\mathrm{SIR}_M \geq \gamma) = \int_0^\infty f_{m_0}(x) P(\mathrm{SIR}_M \geq \gamma | r_{M,m_0} = x)\mathrm{d}x. \tag{39}$$

For simplicity of notations, the interference from all SBSs and other non-serving MBSs are denoted by

$$\begin{aligned} I_{S_1} &= \sum_{n\in\Phi_\mathrm{S}\setminus\mathcal{N}_2} \left|h_{S,n}\right|^2 P_S r_{S,n}^{-\alpha_S}, \\ I_{M_1} &= \sum_{m\in\Phi_\mathrm{M}\setminus m_0} \left|h_{M,m}\right|^2 P_M r_{M,m}^{-\alpha_M}. \end{aligned}$$

Relying on stochastic geometry, $P(\mathrm{SIR}_M \geq \gamma | r_{M,m_0} = x)$ can be calculated as

$$\begin{aligned} P(\mathrm{SIR}_M &\geq \gamma | r_{M,m_0} = x) \\ &= P\Big(\frac{\left|h_{M,m_0}\right|^2 P_M r_{M,m_0}^{-\alpha_M}}{I_{S_1} + I_{M_1}} \geq \gamma | r_{M,m_0} = x\Big) \\ &= P(\left|h_{M,m_0}\right|^2 \geq \gamma P_M^{-1} x^{\alpha_M}(I_{S_1} + I_{M_1})) \\ &\overset{(a)}{=} \mathbb{E}_{I_{S_1}, I_{M_1}}[\exp(-(I_{S_1} + I_{M_1})\gamma P_M^{-1} x^{\alpha_M})] \\ &= \mathcal{L}_{I_{S_1}}(\gamma P_M^{-1} x^{\alpha_M}) \mathcal{L}_{I_{M_1}}(\gamma P_M^{-1} x^{\alpha_M}), \end{aligned} \tag{40}$$

where (a) follows the fact that $\left|h_{M,m_0}\right|^2 \sim \exp(1)$ and $\exp(\mu)$ denotes the exponential distribution with mean μ. Additionally, $\mathcal{L}_{I_{S_1}}(\gamma P_M^{-1} x^{\alpha_M})$ and $\mathcal{L}_{I_{M_1}}(\gamma P_M^{-1} x^{\alpha_M})$ are the Laplace transforms of interference I_{S_1} and I_{M_1}, respectively. Let $k_1 = \gamma P_M^{-1} x^{\alpha_M}$. Then, $\mathcal{L}_{I_{S_1}}(k_1)$ can be obtained as follows

$$\begin{aligned} \mathcal{L}_{I_{S_1}}(k_1) &= \mathbb{E}_{I_{S_1}}[\exp(-\sum_{n\in\Phi_\mathrm{S}} k_1 I_{S_1,n})] \\ &= \mathbb{E}_{I_{S_1}}[\prod_{n\in\Phi_S} \exp(-k_1 \left|h_{S,n}\right|^2 P_S r_{S,n}^{-\alpha_S})] \\ &= \mathbb{E}_{I_{S_1}}[\prod_{n\in\Phi_S} \frac{1}{1 + k_1 P_S r_{S,n}^{-\alpha_S}}] \\ &= \exp(-2\pi\lambda_s \int_0^\infty (1 - \frac{1}{1 + k_1 P_S \rho^{-\alpha_S}})\rho \mathrm{d}\rho \\ &= \exp(-\pi\lambda_s (k_1 P_S)^{\frac{2}{\alpha_S}} \int_0^\infty \frac{1}{1 + t^{\frac{\alpha_S}{2}}}\mathrm{d}t). \end{aligned} \tag{41}$$

We denote $G_\alpha(x) = \int_x^\infty \frac{1}{1+t^{\frac{\alpha}{2}}}\mathrm{d}t$, then $\mathcal{L}_{I_{S_1}}(k_1) = \exp(-\pi\lambda_s (k_1 P_S)^{\frac{2}{\alpha_S}} G_{\alpha_S}(0))$. In the similar manner, $\mathcal{L}_{I_{M_1}}(k_1)$ can be calculated as

$$\begin{aligned} \mathcal{L}_{I_{M_1}}(k_1) &= \mathbb{E}_{I_{M_1}}[\exp(-\sum_{m\in\Phi_\mathrm{M}\setminus m_0} k_1 I_{M_1,m})] \\ &= \mathbb{E}_{I_{M_1}}[\prod_{m\in\Phi_\mathrm{M}\setminus m_0} \exp(-k_1 \left|h_{M,m}\right|^2 P_M r_{M,m}^{-\alpha_M})] \\ &= \mathbb{E}_{I_{M_1}}[\prod_{m\in\Phi_\mathrm{M}\setminus m_0} \frac{1}{1 + k_1 P_M r_{M,m}^{-\alpha_M})} \\ &= \exp(-2\pi\lambda_M \int_x^\infty (1 - \frac{1}{1 + k_1 P_M \rho^{-\alpha_M}})\rho \mathrm{d}\rho \\ &= \exp(-\pi\lambda_M (k_1 P_M)^{\frac{2}{\alpha_M}} G_{\alpha_M}(x^2 (k_1 P_M)^{-\frac{2}{\alpha_M}})). \end{aligned} \tag{42}$$

Therefore, we can obtain that

$$\begin{aligned} &P(\mathrm{SIR}_M \geq \gamma | r_{M,m_0} = x) \\ &\quad = \exp(-\pi\lambda_M x^2 \gamma^{\frac{2}{\alpha_M}} G_{\alpha_M}(\gamma^{-\frac{2}{\alpha_M}})) \\ &\quad\quad \times \exp(-\pi\lambda_S (\gamma \frac{P_S}{P_M} x^{\alpha_M})^{\frac{2}{\alpha_S}} G_{\alpha_S}(0)). \end{aligned} \tag{43}$$

Finally, substituting (43) and (38) into (39), $P(\mathrm{SIR}_M \geq \gamma)$ is obtained. □

Appendix B
Proof of Corollary 1

When $\alpha_M = \alpha_S = 4$, it is intuitional to obtain that $G_{\alpha_S}(0) = \frac{\pi}{2}$ and $G_{\alpha_M}(\gamma^{-\frac{2}{\alpha_M}}) = \frac{\pi}{2} - \arctan(\gamma^{-\frac{2}{\alpha_M}}) = \mathrm{arccot}(\gamma^{-\frac{2}{\alpha_M}})$. Next, the simplified form of $P(\mathrm{SIR}_M \geq \gamma)$ can be obtained as follows

$$\begin{aligned} &P(\mathrm{SIR}_M \geq \gamma) \\ &\quad = \int_0^\infty f_{m_0}(x) \exp(-\pi x^2 \gamma^{\frac{1}{2}} (\frac{\pi}{2}\lambda_S (\frac{P_S}{P_M})^{\frac{1}{2}} \\ &\quad\quad + \lambda_M \mathrm{arccot}(\gamma^{-\frac{1}{2}})))\mathrm{d}x \\ &\quad = \pi\lambda_M \int_0^\infty \exp(-\pi x^2 (\lambda_M \\ &\quad\quad + \gamma^{\frac{1}{2}} (\frac{\pi}{2}\lambda_S (\frac{P_S}{P_M})^{\frac{1}{2}} + \lambda_M \mathrm{arccot}(\gamma^{-\frac{1}{2}}))))\mathrm{d}x^2 \\ &\quad = \lambda_M (\lambda_M + \gamma^{\frac{1}{2}} (\frac{\pi}{2}\lambda_S (\frac{P_S}{P_M})^{\frac{1}{2}} + \lambda_M \mathrm{arccot}(\gamma^{-\frac{1}{2}})))^{-1} \\ &\quad = (1 + \lambda_M^{-1} \gamma^{\frac{1}{2}} (\frac{\pi}{2}\lambda_S (\frac{P_S}{P_M})^{\frac{1}{2}} + \lambda_M \mathrm{arccot}(\gamma^{-\frac{1}{2}})))^{-1}. \end{aligned}$$

□

Appendix C
Proof of Theorem 1

For notational simplicity, let $y_M = \mathrm{SIR}_M$ and $\Omega(y_M)$ denotes the successful transmission event $\mathrm{SIR}_M \geq \gamma$. The conditional PDF of y_M is then denoted as $g(y_M|\Omega(y_M))$. Therefore, we have the following derivations

$$\begin{aligned} &\mathbb{E}[\log_2(1 + y_M)|\Omega(y_M)] \\ &= \int_0^\infty f_{m_0}(x)\mathrm{d}x \int_0^\infty \log_2(1 + y_M) g(y_M|\Omega(y_M))\mathrm{d}y_M \\ &= \frac{1}{\ln 2} \int_0^\infty f_{m_0}(x)\mathrm{d}x \int_0^\infty \ln(1 + y_M) g(y_M|\Omega(y_M))\mathrm{d}y_M \\ &= \frac{1}{\ln 2} \int_0^\infty f_{m_0}(x)\mathrm{d}x \int_0^\infty (\int_0^{y_M} \frac{1}{1+t} dt) g(y_M|\Omega(y_M)\mathrm{d}y_M \\ &\overset{(b)}{=} \frac{1}{\ln 2} \int_0^\infty f_{m_0}(x)\mathrm{d}x \int_0^\infty \frac{1}{1 + t}\mathrm{d}t \int_t^\infty g(y_M|\Omega(y_M))\mathrm{d}y_M \\ &\overset{(c)}{=} \frac{1}{\ln 2} \int_0^\infty f_{m_0}(x)\mathrm{d}x \int_0^\infty \frac{1}{1 + t} P(y_M \geq t|\Omega(y_M))\mathrm{d}t \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\ln 2}\int_0^\infty f_{m_0}(x)\mathrm{d}x \\
&\quad \int_0^\infty \frac{1}{1+t}\frac{P(\mathrm{SIR}_M \ge t, \mathrm{SIR}_M \ge \gamma | r_{M,m_0}=x)}{P(\mathrm{SIR}_M \ge \gamma | r_{M,m_0}=x)}\mathrm{d}t \\
&= \frac{1}{\ln 2}\int_0^\infty f_{m_0}(x)\mathrm{d}x \\
&\quad \int_0^\infty \frac{1}{1+t}\frac{P(\mathrm{SIR}_M \ge \mathrm{max(t,}\gamma)|\mathrm{r_{M,m_0}=x})}{P(\mathrm{SIR}_M \ge \gamma | r_{M,m_0}=x)}\mathrm{d}t \\
&= \frac{1}{\ln 2}\int_0^\infty f_{m_0}(x)\mathrm{d}x \\
&\quad \left[\int_0^\gamma \frac{1}{1+t}\mathrm{d}t + \int_\gamma^\infty \frac{P(\mathrm{SIR}_M \ge t | r_{M,m_0}=x)}{P(\mathrm{SIR}_M \ge \gamma | r_{M,m_0}=x)(1+t)}\mathrm{d}t\right] \\
&= \log_2(1+\gamma) + \frac{1}{\ln 2} \\
&\quad \int_0^\infty f_{m_0}(x)\mathrm{d}x \int_\gamma^\infty \frac{P(\mathrm{SIR}_M \ge t | r_{M,m_0}=x)}{P(\mathrm{SIR}_M \ge \gamma | r_{M,m_0}=x)(1+t)}\mathrm{d}t,
\end{aligned} \tag{44}
$$

where (b) and (c) follow the facts that

$$
\begin{aligned}
&\int_0^\infty (\int_0^{y_M} \frac{1}{1+t}dt) g(y_M|\Omega(y_M)\mathrm{d}y_M \\
&\qquad = \int_0^\infty \frac{1}{1+t}\mathrm{d}t \int_t^\infty g(y_M|\Omega(y_M))\mathrm{d}y_M, \\
&\int_t^\infty g(y_M|\Omega(y_M))\mathrm{d}y_M = P(y_M \ge t|\Omega(y_M)).
\end{aligned}
$$

Finally, substituting (44) and (38) into (10), the expression for $R_M(\gamma)$ shown in (15) is obtained. □

Appendix D
Proof of Lemma 2

Let vectors $\mathbf{r}_{S_{BL}} = [r_{BL,1}, r_{BL,2}, \ldots, r_{BL,n_1}]$ and $\mathbf{r}_{S_{EL}} = [r_{EL,1}, r_{EL,2}, \ldots, r_{EL,n_2}]$ represent the positions of the serving SBSs located in two clusters. The serving SBSs in clusters $\mathcal{N}_1$ and $\mathcal{N}_2$ are independently and uniformly distributed within the circle with radius a and annulus with radii a and b. Therefore, the joint PDF of $\mathbf{r}_{S_{BL}}$ and $\mathbf{r}_{S_{EL}}$ are given by [8]

$$
f(r_{BL,1}, \ldots, r_{BL,n_1}) = \prod_{k=1}^{n_1} \frac{2r_{BL,k}}{a^2}, \tag{45}
$$

$$
f(r_{EL,1}, \ldots, r_{EL,n_2}) = \prod_{k=1}^{n_2} \frac{2r_{EL,k}}{b^2-a^2}. \tag{46}
$$

A. The Derivation of $P(\mathrm{SIR}_{S,BL} \ge \gamma_{BL})$

From the definition of successful transmission probability, $P(\mathrm{SIR}_{S,BL} \ge \gamma_{BL})$ can be expressed as

$$
\begin{aligned}
P(\mathrm{SIR}_{S,BL} \ge \gamma_{BL}) = \int_0^a \ldots \int_0^a f(x_{BL,1}, \ldots, x_{BL,n_1}) \\
P(\mathrm{SIR}_{S,BL} \ge \gamma_{BL} | \mathbf{r}_{S_{BL}} = \mathbf{x}_{BL})\mathrm{d}\mathbf{x}_{BL}.
\end{aligned} \tag{47}
$$

For notational simplicity, we let

$$
\begin{aligned}
I_{S_2} &= \sum_{n\in\Phi_\mathrm{S}\setminus\mathcal{N}_1} |h_{S,n}|^2 P_S r_{S,n}^{-\alpha_S}, \\
I_{M_2} &= \sum_{m\in\Phi_\mathrm{M}} |h_{M,m}|^2 P_M r_{M,m}^{-\alpha_M}.
\end{aligned}
$$

Next, $P(\mathrm{SIR}_{S,BL} \ge \gamma_{BL}|\mathbf{r}_{S_{BL}} = \mathbf{x}_{BL})$ is calculated as

$$
\begin{aligned}
&P(\mathrm{SIR}_{S,BL} \ge \gamma_{BL}|\mathbf{r}_{S_{BL}} = \mathbf{x}_{BL}) \\
&\quad = P(\left|\sum_{k=1}^{n_1} h_{S,k}\sqrt{P_S} x_{BL,k}^{-\frac{\alpha_S}{2}}\right|^2 \ge \gamma_{BL}(I_{S_2}+I_{M_2})) \\
&\quad \overset{(d)}{=} \mathbb{E}_{I_{S_2},I_{M_2}}[\exp(-\frac{1}{\sum_{k=1}^{n_1} x_{BL,k}^{-\alpha_S}}\gamma_{BL}(I_{S_2}+I_{M_2})P_S^{-1})] \\
&\quad = \mathcal{L}_{I_{S_2}}(\frac{\gamma_{BL}P_S^{-1}}{\sum_{k=1}^{n_1} x_{BL,k}^{-\alpha_S}})\mathcal{L}_{I_{M_2}}(\frac{\gamma_{BL}P_S^{-1}}{\sum_{k=1}^{n_1} x_{BL,k}^{-\alpha_S}}),
\end{aligned} \tag{48}
$$

where (d) follows the fact that $\left|\sum_{k=1}^{n_1} h_{S,k}\sqrt{P_S} r_{BL,k}^{-\frac{\alpha_S}{2}}\right|^2 \sim P_S^{-1}\exp(\frac{1}{\sum_{k=1}^{n_1} x_{BL,k}^{-\alpha_S}})$. Let $k_2 = \frac{\gamma_{BL}P_S^{-1}}{\sum_{k=1}^{n_1} x_{BL,k}^{-\alpha_S}}$. Following the similar steps shown before, $\mathcal{L}_{I_{S_2}}(k_2)$ and $\mathcal{L}_{I_{M_2}}(k_2)$ are calculated as

$$
\mathcal{L}_{I_{S_2}}(k_2) = \exp(-\pi\lambda_S(k_2P_S)^{\frac{2}{\alpha_S}}G_{\alpha_S}(a^2(k_2P_S)^{-\frac{2}{\alpha_S}})), \tag{49}
$$

$$
\mathcal{L}_{I_{M_2}}(k_2) = \exp(-\pi\lambda_M(k_2P_M)^{\frac{2}{\alpha_M}}G_{\alpha_M}(0)), \tag{50}
$$

respectively. Substituting formulas (49) and (50) into (48), we can obtain that

$$
\begin{aligned}
&P(\mathrm{SIR}_{S,BL} \ge \gamma_{BL}|\mathbf{r}_{S_{BL}} = \mathbf{x}_{BL}) \\
&\quad = \exp(-\pi\lambda_S c^{\frac{2}{\alpha_S}}G_{\alpha_S}(a^2c^{-\frac{2}{\alpha_S}}) - \pi\lambda_M(c\frac{P_M}{P_S})^{\frac{2}{\alpha_M}}G_{\alpha_M}(0)).
\end{aligned} \tag{51}
$$

Finally, substituting (51) and (46) into (47), we can obtain the expression for $P(\mathrm{SIR}_{S,BL} \ge \gamma_{BL})$. Thus, the proof of $P(\mathrm{SIR}_{S,BL} \ge \gamma_{BL})$ is completed.

B. The Derivation of $P(\mathrm{SIR}_{S,EL} \ge \gamma_{EL})$

Following the similar methods, $P(\mathrm{SIR}_{S,EL} \ge \gamma_{EL})$ can be expressed as

$$
\begin{aligned}
P(\mathrm{SIR}_{S,EL} \ge \gamma_{EL}) = \int_a^b \ldots \int_a^b f(x_{EL,1}, \ldots, x_{EL,n_2}) \\
P(\mathrm{SIR}_{S,EL} \ge \gamma_{EL}|\mathbf{r}_{S_{EL}} = \mathbf{x}_{EL})\mathrm{d}\mathbf{x}_{EL}.
\end{aligned} \tag{52}
$$

In order to simplify notations, we let

$$
\begin{aligned}
I_{S_3} &= \sum_{n\in\mathcal{N}_1} |h_{S,n}|^2 P_S r_{S,n}^{-\alpha_S}, \\
I_{S_4} &= \sum_{n\in\Phi_\mathrm{S}\setminus\{\mathcal{N}_1\cup\mathcal{N}_2\}} |h_{S,n}|^2 P_S r_{S,n}^{-\alpha_S}.
\end{aligned}
$$

Afterwards, the expression for $P(\mathrm{SIR}_{S,EL} \ge \gamma_{EL}|\mathbf{r}_{S_{EL}} = \mathbf{x}_{EL})$ can be calculated as

$$
\begin{aligned}
&P(\mathrm{SIR}_{S,EL} \ge \gamma_{EL}|\mathbf{r}_{S_{EL}} = \mathbf{x}_{EL}) \\
&\qquad = \mathcal{L}_{I_{S_3}}(k_3)\mathcal{L}_{I_{S_4}}(k_3)\mathcal{L}_{I_{M_2}}(k_3),
\end{aligned} \tag{53}
$$

where

$$
k_3 = \frac{\gamma_{EL}P_S^{-1}}{\sum_{k=1}^{n_2} x_{EL,k}^{-\alpha_S}}, \tag{54}
$$

$$\mathcal{L}_{I_{S_3}}(k_3) = \exp(-\pi\lambda_S(k_3P_S)^{\frac{2}{\alpha_S}}\int_0^{a^2(k_3P_S)^{-\frac{2}{\alpha_S}}}\frac{1}{1+t^{\frac{\alpha_S}{2}}}\mathrm{d}t), \tag{55}$$

$$\mathcal{L}_{I_{S_4}}(k_3) = \exp(-\pi\lambda_S(k_3P_S)^{\frac{2}{\alpha_S}}G_{\alpha_S}(b^2(k_3P_S)^{-\frac{2}{\alpha_S}}). \tag{56}$$

Thus, (53) can be rewritten as

$$\begin{aligned}&P(\mathrm{SIR}_{S,EL} \geq \gamma_{EL}|\mathbf{r}_{S_{EL}} = \mathbf{x}_{EL})\\&\quad= \exp(-\pi(\lambda_M(d\frac{P_M}{P_S})^{\frac{2}{\alpha_M}}G_{\alpha_M}(0)\\&\quad-\lambda_S d^{\frac{2}{\alpha_S}}(\int_0^{a^2d^{-\frac{2}{\alpha_S}}}\frac{1}{1+t^{\frac{\alpha_S}{2}}}\mathrm{d}t)+G_{\alpha_S}(b^2d^{-\frac{2}{\alpha_S}}))).\end{aligned} \tag{57}$$

Substituting (57) and (46) into (52), we can obtain the theoretical expression for $P(\mathrm{SIR}_{S,EL} \geq \gamma_{EL})$. □

References

[1] Ericsson. (Nov. 2017). *Ericsson Mobility Report*. [Online]. Available: http://www.ericsson.com/en/mobility-report/reports/november-2017

[2] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. M. 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.

[3] K. Poularakis, G. Iosifidis, V. Sourlas, and L. Tassiulas, "Exploiting caching and multicast for 5G wireless networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2995–3007, Apr. 2016.

[4] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6118–6131, Sep. 2016.

[5] H. Zhou, M. Tao, E. Chen, and W. Yu, "Content-centric multicast beamforming in cache-enabled cloud radio access networks," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, San Diego, CA, USA, Dec. 2015, pp. 1–6.

[6] D. Liu and C. Yang, "Cache-enabled heterogeneous cellular networks: Comparison and tradeoffs," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[7] Q. Wu, G. Y. Li, W. Chen, D. W. K. Ng, and R. Schober, "An overview of sustainable green 5G networks," *IEEE Wireless Commun.*, vol. 24, no. 4, pp. 72–80, Aug. 2017.

[8] Z. Chen, J. Lee, T. Q. S. Quek, and M. Kountouris, "Cooperative caching and transmission design in cluster-centric small cell networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 3401–3415, May 2017.

[9] F. Xu, M. Tao, and K. Liu, "Fundamental tradeoff between storage and latency in cache-aided wireless interference networks," *IEEE Trans. Inf. Theory*, vol. 63, no. 11, pp. 7464–7491, Nov. 2017.

[10] J. Liao, K. K. Wong, M. R. A. Khandaker, and Z. Zheng, "Optimizing cache placement for heterogeneous small cell networks," *IEEE Commun. Lett.*, vol. 21, no. 1, pp. 120–123, Sep. 2016.

[11] F. Gabry, V. Bioglio, and I. Land, "On energy-efficient edge caching in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3288–3298, Dec. 2016.

[12] D. Ko, B. Hong, J. H. Lee, and W. Choi, "Optimal file storing with cache memory in amorphous femto helper aided networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Paris, France, May 2017, pp. 1–6.

[13] X. Xu and M. Tao, "Modeling, analysis, and optimization of coded caching in small-cell networks," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3415–3428, Aug. 2017.

[14] W. Wen, Y. Cui, F.-C. Zheng, S. Jin, and Y. Jiang, "Random caching based cooperative transmission in heterogeneous wireless networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Paris, France, May 2017, pp. 1–6.

[15] Y. Cui and D. Jiang, "Analysis and optimization of caching and multicasting in large-scale cache-enabled heterogeneous wireless networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 1, pp. 250–264, Jan. 2017.

[16] Z. Zhao, M. Peng, Z. Ding, W. Wang, and H. V. Poor, "Cluster content caching: An energy-efficient approach to improve quality of service in cloud radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1207–1221, May 2016.

[17] B. Chen, C. Yang, and A. F. Molisch, "Cache-enabled device-to-device communications: Offloading gain and energy cost," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4519–4536, Jul. 2017.

[18] L. Zhang, M. Xiao, G. Wu, and S. Li, "Efficient scheduling and power allocation for D2D-assisted wireless caching networks," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2438–2452, Jun. 2015.

[19] S. Tamoor-ul-Hassan, M. Bennis, P. H. J. Nardelli, and M. Latva-Aho, "Caching in wireless small cell networks: A storage-bandwidth tradeoff," *IEEE Commun. Lett.*, vol. 20, no. 6, pp. 1175–1178, Jun. 2016.

[20] Z. Hu, Z. Zheng, T. Wang, and L. Song, "Caching as a service: Small-cell caching mechanism design for service providers," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 6992–7004, Oct. 2016.

[21] C. Yang, Y. Yao, Z. Chen, and B. Xia, "Analysis on cache-enabled wireless heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 131–145, Jan. 2016.

[22] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.

[23] P. Ostovari, J. Wu, A. Khreishah, and N. B. Shroff, "Scalable video streaming with helper nodes using random linear network coding," *IEEE/ACM Trans. Netw.*, vol. 24, no. 3, pp. 1574–1587, Jun. 2016.

[24] L. Wu and W. Zhang, "Caching-based scalable video transmission over cellular networks," *IEEE Commun. Lett.*, vol. 20, no. 6, pp. 1156–1159, Jun. 2016.

[25] C. Zhan and Z. Wen, "Content cache placement for scalable video in heterogeneous wireless network," *IEEE Commun. Lett.*, vol. 21, no. 12, pp. 2714–2717, Dec. 2017.

[26] M. Peng, K. Zhang, J. Jiang, J. Wang, and W. Wang, "Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks," *IEEE Trans. Veh. Technol.*, vol. 64, no. 11, pp. 5275–5287, Nov. 2015.

[27] G. Zhao, S. Chen, L. Zhao, and L. Hanzo, "Joint energy-spectral-efficiency optimization of comp and bs deployment in dense large-scale cellular networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4832–4847, Jul. 2017.

[28] S. Zhang, Q. Wu, S. Xu, and G. Y. Li, "Fundamental green tradeoffs: Progresses, challenges, and impacts on 5G networks," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 1, pp. 33–56, 1st Quart., 2017.

[29] M. Wang, H. Gao, and T. Lv, "Energy-efficient user association and power control in the heterogeneous network," *IEEE Access*, vol. 5, pp. 5059–5068, Mar. 2017.

[30] D. Liu and C. Yang, "Energy efficiency of downlink networks with caching at base stations," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 907–922, Apr. 2016.

[31] J. Zhang, X. Zhang, and W. Wang, "Cache-enabled software defined heterogeneous networks for green and flexible 5G networks," *IEEE Access*, vol. 4, pp. 3591–3604, 2016.

[32] Y. Huang, X. Zhang, J. Zhang, J. Tang, Z. Su, and W. Wang, "Energy-efficient design in heterogeneous cellular networks based on large-scale user behavior constraints," *IEEE Trans. Wireless Commun.*, vol. 13, no. 9, pp. 4746–4757, Sep. 2014.

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

[34] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, vol. 1, New York, NY, USA, Mar. 1999, pp. 126–134.

[35] Y. Xu, Y. Li, Z. Wang, T. Lin, G. Zhang, and S. Ci, "Coordinated caching model for minimizing energy consumption in radio access network," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Sydney, NSW, Australia, Jun. 2014, pp. 2406–2411.

[36] D. P. Bertsekas, "Nonlinear programming," *J. Oper. Res. Soc.*, vol. 48, no. 3, p. 334, 1997.

[37] Y. S. Soh, T. Q. S. Quek, M. Kountouris, and H. Shin, "Energy efficient heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 5, pp. 840–850, May 2013.

Xuewei Zhang (S'18) received the B.E. degree in communication engineering from Tianjin Polytechnic University, Tianjin, China, in 2015. She is currently pursuing the Ph.D. degree with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China. Her research interests include multicast beamforming, wireless caching, and resource allocation.

Tiejun Lv (M'08–SM'12) received the M.S. and Ph.D. degrees in electronic engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 1997 and 2000, respectively. From 2001 to 2003, he was a Post-Doctoral Fellow with Tsinghua University, Beijing, China. In 2005, he joined the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, as a Full Professor. From 2008 to 2009, he was a Visiting Professor with the Department of Electrical Engineering, Stanford University, Stanford, CA, USA. He has authored over 60 published IEEE journal papers and 170 conference papers on the physical layer of wireless mobile communications. His current research interests include signal processing, communications theory, and networking. He received the Program for New Century Excellent Talents in University Award from the Ministry of Education, China, in 2006, and the Nature Science Award from the Ministry of Education, China, in 2015, for the hierarchical cooperative communication theory and technologies.

Wei Ni received the B.E. and Ph.D. degrees in electronic engineering from Fudan University, Shanghai, China, in 2000 and 2005, respectively. He was a Post-Doctoral Research Fellow with Shanghai Jiaotong University, from 2005 to 2008, a Research Scientist and the Deputy Project Manager with the Bell Labs R&I Center, Alcatel/Alcatel-Lucent, from 2005 to 2008, and a Senior Researcher with Devices Research and Development, Nokia, from 2008 to 2009. He is currently a Team Leader, a Senior Scientist, and the Project Leader with CSIRO, Sydney, Australia. He also holds honorary positions with the University of New South Wales, Macquarie University, and the University of Technology Sydney. His research interests include optimization, game theory, graph theory, as well as their applications to network and security.

Dr. Ni served as a Program Committee Member for the CHINACOM 2014 and a TPC Member for the IEEE ICC 2014, the IEEE ICCC 2015, the IEEE EICE 2014, and the IEEE WCNC 2010. He has been serving as the Vice Chair for the IEEE NSW VTS Chapter since 2018. He served as an Editor for the *Hindawi Journal of Engineering* from 2012 to 2015, the Secretary for the IEEE NSW VTS Chapter from 2015 to 2018, the Track Chair for the VTC-Spring 2017, the Track Co-Chair for the IEEE VTC-Spring 2016, and the Publication Chair for the BodyNet 2015. He also served as the Student Travel Grant Chair for the WPMC 2014.

John M. Cioffi (F'96) received the B.S.E.E. degree from the University of Illinois at Urbana–Champaign, Champaign, IL, USA, in 1978, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 1984. He was with the Bell Laboratories, Murray Hill, NJ, USA, from 1978 to 1984. He was with IBM Research, Armonk, NY, USA, from 1984 to 1986. He founded the Amati Communications Corporation in 1991 (purchased by TI in 1997), where he was an Officer and the Director from 1991 to 1997. Since 1986, he has been a Professor of electrical engineering with Stanford University, where he is currently an Emeritus. He is currently on the Board of Directors of ASSIA, Redwood City, CA, USA, (the Chairman and the CEO), Alto Beam, Tinoq, Collinear, and the Marconi Foundation, while serving several advisory boards. He has authored over 600 papers and holds over 100 patents, of which many are heavily licensed including key necessary patents for the international standards in ADSL, VDSL, vectored VDSL, G.fast, DSM, LTE, massive multi in multi out, and various Wi-Fi methodologies. His specific interests are in high-performance digital transmission. He was a member of the U.S. National and U.K. Royal Academies of Engineering in 2001 and 2009. He was a recipient of the IEEE AG Bell Medal in 2010, the Kirchmayer Medal in 2014, the Millennium Medal in 2000, the Internet Hall of Fame Award in 2014, the Economist Magazine 2010 Innovations Award, the International Marconi Fellowship in 2006, the IEEE Kobayashi Award in 2001, the Armstrong Award in 2013, the BBWF Lifetime Achievement Award in 2014, the IEE J. J. Thomson Medal in 2000, the 1991 and 2007 IEEE Communications Magazine Best Paper Award, and numerous conference best paper awards.

Norman C. Beaulieu is currently a Beijing University of Posts and Telecommunications Thousand-Talents Scholar. He is an Academician of the Royal Society of Canada and an Academician of The Canadian Academy of Engineering. He is a fellow of the Institution of Engineering and Technology, U.K., a fellow of The Engineering Institute of Canada (EIC), and a Nicola Copernicus Fellow of Italy. He was a recipient of the IEEE Edwin Howard Armstrong Award and the IEEE Reginald Aubrey Fessenden Award, named for the inventors of frequency modulation and amplitude modulation, respectively. He was a recipient of the Royal Society of Canada Thomas W. Eadie Medal. He received the Médaille K. Y. Lo Medal from the EIC, and was the subject of a *TIME Magazine* feature article. He received the Canada E. W. R. Steacy Memorial Fellowship from the esteemed Natural Sciences and Engineering Research Council, the unique Special University Prize in Applied Science from The University of British Columbia, and the J. Gordin Kaplan Award for Research from the University of Alberta. He holds the third highest Web of Science ISI h-index in the world in the combined areas of communication theory and information theory. He received the title "State Especially Recruited Foreign Expert" certified upon him by the Minister of Human Resources and Social Insurance and the Vice Minister of the Organization Department, Y. Weimin.

Y. Jay Guo (F'14) received the bachelor's and master's degrees from Xidian University, China, in 1982 and 1984, respectively, and the Ph.D. degree from Xian Jiaotong University, China, in 1987. He has authored over 300 research papers and holds 22 patents in antennas and wireless systems. His research interests include antennas, millimeter-wave and terahertz communications, and sensing systems as well as big data. He is a member of the College of Experts of Australian Research Council. He is a fellow of the Australian Academy of Engineering and Technology and a fellow of the IET. He received a number of most prestigious Australian national awards and was named one of the most influential engineers in Australia in 2014 and 2015.

He held various senior leadership positions with Fujitsu, Siemens, and NEC, U.K. In 2014, he was the Director of CSIRO for over nine years, directing a number of ICT research portfolios. He is currently a Distinguished Professor and the Founding Director of the Global Big Data Technologies Centre, University of Technology Sydney, Australia.

Dr. Guo chaired numerous international conferences. He was the International Advisory Committee Chair of the IEEE VTC2017, the General Chair of the ISAP2015, the iWAT2014, and the WPMC2014, and the TPC Chair of the 2010 IEEE WCNC and the 2012 and 2007 IEEE ISCIT. He served as a Guest Editor for the special issues on "Antennas for Satellite Communications" and "Antennas and Propagation Aspects of 60–90 GHz Wireless Communications," of the IEEE TRANSACTIONS ON ANTENNAS AND PROPAGATION, the Special Issue on "Communications Challenges and Dynamics for Unmanned Autonomous Vehicles," of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, and the Special Issue on "5G for Mission Critical Machine Communications," of the *IEEE Network Magazine*.