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 (5G) 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 [11]. 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 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. Some research efforts have been devoted to combining SVC with wireless caching. For example, Wu et al. in [23] 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, the authors in [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, the authors in [30] provide the closed-form expression for EE and identify the conditions to obtain benefits from caching. Chen et al. in [8] and Zhang et al. in [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, the authors in [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 network of interest, whose locations follow homogeneous Poisson Point Processes (PPPs) $\Phi_M$ and $\Phi_S$ with densities $\lambda_M$ and $\lambda_S$, 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. 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 $N_2 = |\mathcal{N}_2|$. 
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

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

where $\alpha$ is the skewness parameter that characterizes the concentration of video requests [33]. 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

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

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

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 $N_1$ cache parts of the BL contents of the video files, while the cached parts are the same across the SBSs. For SBSs located in cluster $N_2$, 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 $N_1$ and $N_2$, respectively. They can be stacked into vectors $Q_1 = [Q_{1,1}, Q_{1,2}, ..., Q_{1,F}]$ and $Q_2 = [Q_{2,1}, Q_{2,2}, ..., Q_{2,F}]$.
- Scheme II (SVC-based random caching). The SBSs located in clusters $N_1$ and $N_2$ 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 $N_1$ and $N_2$, respectively. They can be stacked into vectors $T_1 = [T_{1,1}, T_{1,2}, ..., T_{1,F}]$ and $T_2 = [T_{2,1}, T_{2,2}, ..., T_{2,F}]$.

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

### 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,

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 $N_2$, which are closer to the user and are capable of providing stronger signal strength than those in cluster $N_2$, are responsible for caching and delivering BLs, while SBSs located in cluster $N_2$ 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.
the signal-to-interference ratio (SIR) of the designated user is expressed as

\[
\text{SIR}_{S, BL} = \frac{\left| \sum_{k \in N_{1,f}} h_{S,k} \sqrt{r_{S,k}} \right|^2}{\sum_{n \in \Phi_{S} \setminus N_{1}} \left| h_{S,n} \right|^2 P_{S} r_{S,n}^{\alpha_{S}} + \sum_{m \in \Phi_{M}} \left| h_{M,m} \right|^2 P_{M} r_{M,m}^{\alpha_{M}}},
\]

(3)

where \( N_{1,f} \) denotes the set of SBSs located in cluster \( N_{1} \) 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, \( \alpha_{S} \) and \( \alpha_{M} \) 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 \( N_{1} \), 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

\[
\text{SIR}_{S, EL} = \frac{\left| \sum_{k \in N_{2,f}} h_{S,k} \sqrt{r_{S,k}} \right|^2}{\sum_{n \in \Phi_{S} \setminus N_{2}} \left| h_{S,n} \right|^2 P_{S} r_{S,n}^{\alpha_{S}} + \sum_{m \in \Phi_{M}} \left| h_{M,m} \right|^2 P_{M} r_{M,m}^{\alpha_{M}}},
\]

(4)

where \( N_{2,f} \) denotes the set of SBSs located in cluster \( N_{2} \) that cache the 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

\[
\text{SIR}_{M} = \frac{\left| h_{M,m_0} \right|^2 P_{M} r_{M,m_0}^{\alpha_{M}}}{\sum_{n \in \Phi_{S}} \left| h_{S,n} \right|^2 P_{S} r_{S,n}^{\alpha_{S}} + \sum_{m \in \Phi_{M}\setminus m_0} \left| h_{M,m} \right|^2 P_{M} r_{M,m}^{\alpha_{M}}},
\]

(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

\[
P_{\text{Total}} = P_{\text{TR}} + P_{\text{CA}} + P_{\text{BH}} + P_{\text{Fix}},
\]

(6)

where \( P_{\text{TR}} \), \( P_{\text{CA}} \), \( P_{\text{BH}} \) and \( P_{\text{Fix}} \) 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], [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

\[
P_{\text{CA}} = c_{\text{ca}} N_{\text{ca}},
\]

(7)

where \( c_{\text{ca}} \) is the caching coefficient in W/bit and \( N_{\text{ca}} \) is the total number of data bits cached in the local storage of SBSs belonging to clusters \( N_{1} \) and \( N_{2} \).

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

\[
P_{\text{BH}} = c_{\text{bh}} N_{\text{bh}},
\]

(8)

where \( c_{\text{bh}} \) is the coefficient of backhaul power consumption in W/bit and \( N_{\text{bh}} \) 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, \( c_{\text{ca}} \) is typically less than \( c_{\text{bh}} \).

As for the fixed power consumption \( P_{\text{Fix}} \), 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

\[
P_{\text{Fix}} = (N_{1} + N_{2}) P_{S, \text{Fix}} + P_{M, \text{Fix}},
\]

(9)

where \( P_{S, \text{Fix}} \) and \( P_{M, \text{Fix}} \) 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

\[
R_{M}(\gamma) \triangleq W \mathbb{E} \{ \log_{2} (1 + \text{SIR}_{M}) | \text{SIR}_{M} \geq \gamma \},
\]

(10)

where \( W \) is the spectrum bandwidth and \( \gamma \) 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., \( P(SIR_M \geq \gamma) \), which can be provided by Lemma 1.

**Lemma 1.** When the user is served by the nearest MBS under the minimum QoS requirement \( \gamma \), the successful transmission probability is given in (12), where 
\[
G_\alpha(x) = \int_0^\infty \frac{1}{1+tx^2} dt.
\]

**Proof:** See Appendix A. ■

From (11), we can find that the derived expression for \( P(SIR_M \geq \gamma) \) 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 \( P(SIR_M \geq \gamma) \) can be simplified as
\[
P(SIR_M \geq \gamma) = (1 + \lambda_M^{-1} \gamma^2 \left( \frac{P_S}{P_M} \right)^{\frac{1}{2}} + \lambda_M \arccot(\gamma^{-\frac{1}{2}}))^{-1}.
\]

**Proof:** See Appendix B. ■

In Corollary 1, the simplified form of \( P(SIR_M \geq \gamma) \) 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., \( \frac{P_S}{P_M} \), \( P(SIR_M \geq \gamma) \) 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., \( \frac{\lambda_S}{\lambda_M} \), \( P(SIR_M \geq \gamma) \) deteriorates. This is because there are more SBSs, leading to severer cross-tier interference.

- As the minimum QoS requirement \( \gamma \) increases, the user can impose strict video quality requirement and 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
\[
R_{S,BL}(\gamma_{BL}, n_1) \triangleq \text{WE} \{ \log_2 (1 + \text{SIR}_{S,BL}) | \text{SIR}_{S,BL} \geq \gamma_{BL} \},
\]
\[
R_{S,EL}(\gamma_{EL}, n_2) \triangleq \text{WE} \{ \log_2 (1 + \text{SIR}_{S,EL}) | \text{SIR}_{S,EL} \geq \gamma_{EL} \},
\]
where \( \gamma_{BL} \) and \( \gamma_{EL} \) are the minimum QoS requirements for delivering BL and EL contents, respectively. \( n_1 = |N_1,f| \leq N_1 \) and \( n_2 = |N_2,f| \leq N_2 \) 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., \( P(SIR_{S,BL} \geq \gamma_{BL}) \) and \( P(SIR_{S,EL} \geq \gamma_{EL}) \). 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 (13) and (14), where \( c = \frac{\gamma_{EL}}{\sum_{k=1}^{n_2} x_{EL,k}^{\gamma_{EL}}} \), \( d = \frac{\gamma_{BL}}{\sum_{k=1}^{n_1} x_{BL,k}^{\gamma_{BL}}} \), \( x_{BL} = [x_{BL,1}, ..., x_{BL,n_1}] \) and \( x_{EL} = [x_{EL,1}, ..., x_{EL,n_2}] \).

**Proof:** See Appendix D. ■

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

**Corollary 2.** When \( \alpha_M = \alpha_S = 4 \), \( P(SIR_{S,BL} \geq \gamma_{BL}) \) and
\[ P(SIR_M \geq \gamma) = 2\pi \lambda_M \int_0^\infty x \exp(-\pi(x')^{2}G_{a,s}(0) - \lambda_M x^2(\gamma SIR + 1))dx, \tag{11} \]

\[ P(SIR_{S, BL} \geq \gamma_{BL}) = \int_0^a \prod_{k=1}^{n_1} \frac{2x_{BL,k}}{a^2} \exp(-\pi u\gamma_{BL}^2)dx_{BL}. \tag{12} \]

\[ P(SIR_{S, EL} \geq \gamma_{EL}) = \int_a^b \prod_{k=1}^{n_2} \frac{2x_{EL,k}}{b^2 - \alpha^2} \exp(-\pi v\gamma_{EL}^2)dx_{EL}, \tag{13} \]

where the expressions for \( P(SIR_{S, BL} \geq t|r_{S, BL} = x_{BL}) \) and \( P(SIR_{S, BL} \geq \gamma_{BL}|r_{S, BL} = x_{BL}) \) can be found in [57], and the expressions for \( P(SIR_{S, EL} \geq t|r_{S, EL} = x_{EL}) \) and \( P(SIR_{S, EL} \geq \gamma_{EL}|r_{S, EL} = x_{EL}) \) 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

\[ P_{Total,1} = P_{TR,1} + P_{CA,1} + P_{BH,1} + P_{Fix}, \tag{24} \]

where

\[ P_{TR,1} = \sum_{f=1}^{F} p_f(\zeta_s N_1 ||Q_{1,f}||_0 + g_{HDV}(f) N_2 ||Q_{2,f}||_0 P_{S}), \tag{25} \]

\[ P_{CA,1} = c_{ca} \sum_{f=1}^{F} (Q_{1,f} L_B N_1 + Q_{2,f} L_{EL} N_2), \tag{26} \]

and

\[ P_{BH,1} = c_{bh} \sum_{f=1}^{F} p_f(1 - Q_{1,f}) L_B + g_{HDV}(1 - Q_{2,f}) L_{EL}. \tag{27} \]
In (25), \( \zeta_S \) and \( \zeta_M \) 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 fractional caching and cooperative transmission schemes, the sum rate of the designated user is given by

\[
R_{\text{Sum},1} = \sum_{f=1}^{F} p_f \left( (1 - Q_{1,f}) R_M(\gamma_{BL}) + g_{\eta}(f)(1 - Q_{2,f}) R_M(\gamma_{EL}) \right)
\]

served by the nearest MBS to obtain BL and EL contents

\[
+ Q_{1,f} R_S(\gamma_{BL}, N_1) + g_{\eta}(f) Q_{2,f} R_S(\gamma_{EL}, N_2). \]

served by cooperative SBSs to obtain BL and EL contents

(28)

From (28), intuitions can be found that the performance of \( R_{\text{Sum},1} \) can predominate 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

\[
\max_{Q_1, Q_2} EE_1 = \frac{R_{\text{Sum},1}}{P_{\text{Total},1}} \tag{29a}
\]

\[
s.t. \quad \sum_{f=1}^{F} Q_{1,f} = M_B, \tag{29b}
\]

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

\[
0 \leq Q_{1,f} \leq 1, 0 \leq Q_{2,f} \leq 1, \forall f, \tag{29d}
\]

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 \( P_{\text{Total},1} \) 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 \( P_{\text{TR},1} \) can be approximated as

\[
P_{\text{TR},1} = \sum_{f=1}^{F} p_f (\zeta_S(N_1 f_0(Q_{1,f}) + g_{\eta}(f) N_2 f_0(Q_{2,f})) P_S + \zeta_M f_0(1 - Q_{1,f}) + g_{\eta}(f) f_0(1 - Q_{2,f})) P_M), \tag{30}
\]

where

\[
f_0(x) = \log(\frac{x}{\theta} + 1)/\log(\frac{1}{\theta} + 1). \tag{31}
\]

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) to (29d).

2) For \( f \in \{1, \ldots, F\} \), calculate \( \hat{Q}_{1,f} \), and then obtain

\[
\hat{Q}_{1,f}(t+1) = Q_{1,f}(t) + \epsilon(t) \frac{\partial EE_1}{\partial Q_{1,f}}|_{Q_{1,f}=Q_{1,f}(t), Q_{2,f}=Q_{2,f}(t)}.
\]

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 \( \sum_{f=1}^{F} \min\{|\hat{Q}_{1,f}(t+1) - u|^+, 1\} = MB \).

4) For \( f \in \{1, \ldots, F\} \), calculate \( \frac{\partial EE_1}{\partial Q_{2,f}} \), and then obtain

\[
\hat{Q}_{2,f}(t+1) = Q_{2,f}(t) + \epsilon(t) \frac{\partial EE_1}{\partial Q_{2,f}}|_{Q_{1,f}=Q_{1,f}(t), Q_{2,f}=Q_{2,f}(t)}.
\]

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 \( \sum_{f=1}^{F} \min\{|\hat{Q}_{2,f}(t+1) - v|^+, 1\} = ME \).

6) If convergence, the algorithm terminates. Otherwise, set \( t = t + 1 \) and \( \epsilon(t) = \frac{1}{t} \), then go back to Step 2.

In (31), \( \theta \) is a constant parameter to reflect the smoothness of \( f_0(x) \). A larger value for \( \theta \) 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], [13], [30], 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 \( \lim_{t \to \infty} \epsilon(t) = 0 \), \( \lim_{T \to \infty} \sum_{t=1}^{T} \epsilon(t) = \infty \) and \( \lim_{T \to \infty} \sum_{t=1}^{T} c(t) < \infty \). We set \( \epsilon(t) = \frac{1}{t} \) in the \( t \)-th iteration. Note that Steps 3) and 5) give the projections of \( Q_{1,f}(t+1) \) and \( Q_{2,f}(t+1) \) onto the optimization variable set so that constraints (29b) and (29c) are satisfied, where \( \left[ x \right]^+ \triangleq \max[x, 0] \).

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

\[
P_{\text{Total},2} = P_{\text{TR},2} + P_{\text{CA},2} + P_{\text{BH},2} + P_{\text{Fix}}, \tag{32}
\]

where

\[
P_{\text{TR},2} = \sum_{f=1}^{F} p_f (\zeta_S(N_1 T_{1,f} + g_{\eta}(f) N_2 T_{2,f}) P_S + \zeta_M (1 - T_{1,f}) + g_{\eta}(f) (1 - T_{2,f}) P_M), \tag{33}
\]

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

\[ P_{BH,2} = c_{bh} \sum_{f=1}^{F} p_{f}((1 - T_{1,f})^{N_{1}L_{B}} + g_{HDV}(1 - T_{2,f})^{N_{2}L_{E}}). \]  

(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 \( N_{1} \) and \( N_{2} \) remain to be determined. Under Scheme II, each SBS randomly selects BL and EL contents to cache in its local storage under probability distributions \( T_{1} \) and \( T_{2} \), 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 \( N_{1} \) which cache BL content of the \( f \)-th file, i.e., \( |N_{1,f}| \), follows the binomial distribution with parameters \( N_{1} \) and \( T_{1,f} \), while \( |N_{2,f}| \) follows the binomial distribution with parameters \( N_{2} \) and \( T_{2,f} \). Therefore, the sum rate of the designated user can be written as

\[ R_{sum,2} = \sum_{f=1}^{F} \sum_{n_{i}=1}^{N_{i}} \left( (1 - T_{1,f})^{N_{1}R_{M}(\gamma_{BL})} + g_{HDV}(1 - T_{2,f})^{N_{2}R_{M}(\gamma_{EL})} \right) \]

\[ \text{as served by the nearest MBS to obtain BL and EL contents} \]

\[ + \sum_{n_{i}=1}^{N_{1}} C_{N_{1}}^{n_{i}}(1 - T_{1,f})^{n_{i}}(1 - T_{1,f})^{N_{1}-n_{i}} R_{S,BL}(\gamma_{BL,n_{i}}) \]

\[ + \text{as served by cooperative SBSs to obtain BL contents} \]

\[ + \sum_{n_{2}=1}^{N_{2}} C_{N_{2}}^{n_{2}}(1 - T_{2,f})^{n_{2}}(1 - T_{2,f})^{N_{2}-n_{2}} R_{S,EL}(\gamma_{EL,n_{2}}) \]

\[ \text{as served by cooperative SBSs to obtain EL contents} \]  

(36)

From (36), it can be concluded that \( R_{sum,2} \) 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 (36) and (35), the EE maximization problem with BL and EL caching probability design can be formulated as

\[ \max_{T_{1}, T_{2}} \quad \frac{R_{sum,2}}{P_{Total,2}} \]  

(37a)

s.t. \[ \sum_{f=1}^{F} T_{1,f} = M_{B}, \]  

(37b)

\[ \sum_{f=1}^{F} T_{2,f} = M_{E}, \]  

(37c)

\[ 0 \leq T_{1,f} \leq 1, \quad 0 \leq T_{2,f} \leq 1, \forall f, \]  

(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.

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 \( N_{1} \) and \( N_{2} \), \( M_{B} \) 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 \( N_{1} \) and \( N_{2} \), the same fractions of the BL and EL contents are cached, respectively.
- Independent Content Placement (ICP): The SBSs belonging to clusters \( N_{1} \) and \( N_{2} \) randomly select \( M_{B} \) and

### Table I

VALUES OF SIMULATION PARAMETERS

| Parameter        | Value         |
|------------------|---------------|
| \( N_{1}, N_{2} \) | 4             |
| \( P_{S}, P_{M} \) | 23 dBm (default), 43 dBm |
| \( \lambda_{S}, \lambda_{M} \) | 1/(1000π), 1/(2500π) |
| \( a, b \)       | 50 m, 100 m   |
| \( \alpha_{S}, \alpha_{M} \) | 4          |
| \( M \)          | 500 Mbits (default) |
| \( F \)          | 20            |
| \( L_{B}, L_{E} \) | 100 Mbits, 200 Mbits |
| \( \alpha \)      | 1 (default)   |
| \( W \)          | 10 Mbits      |
| \( \gamma_{BL}, \gamma_{EL} \) | 10 dB (default), 5 dB (default) |
| \( \zeta_{S}, \zeta_{M} \) | 4.7         |
| \( c_{es} \)     | 6.25 \times 10^{-12} W/bit [30] |
| \( c_{bh} \)     | 5 \times 10^{-7} W/bit [30]   |
| \( P_{S, Fix} \)  | 6.8 W [37]   |
| \( P_{M, Fix} \)  | 130 W [37]    |
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. 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 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 strategies is evident.
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 $\theta$, 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. 5 shows the EE performance of different caching schemes under various values of $\gamma_{BL}$. 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 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 $\alpha$ is plotted. Note that larger $\alpha$ 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 $\alpha$ 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 ex-
pressions 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

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

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

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

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

$$I_{S_1} = \sum_{n \in \Phi_S \setminus \Phi_2} |h_{S,n}|^2 P_{S_1}^{-\alpha_S},$$

$$I_{M_1} = \sum_{m \in \Phi_M \setminus m_0} |h_{M,m}|^2 P_{M_1}^{-\alpha_M}.$$ Relying on stochastic geometry, $P(\text{SIR}_M \geq \gamma|r_{M,m_0} = x)$ can be calculated as

$$P(\text{SIR}_M \geq \gamma|r_{M,m_0} = x) = \frac{P(|h_{M,m_0}|^2 P_{M_1}^{-\alpha_M} \geq \gamma r_{M,m_0})}{P(r_{M,m_0} = x)}$$

where (a) follows the fact that $|h_{M,m_0}|^2 \sim \exp(1)$ and $\exp(\mu)$ denotes the exponential distribution with mean $\mu$. 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

$$\mathcal{L}_{I_{S_1}}(k_1) = E_{I_{S_1}}[\exp(- \sum_{n \in \Phi_S} k_1 I_{S_1,n})]$$

$$= E_{I_{S_1}}[\prod_{n \in \Phi_S} \exp(-k_1 |h_{S,n}|^2 P_{S_1}^{-\alpha_S})]$$

$$= E_{I_{S_1}}[\prod_{n \in \Phi_S} \frac{1}{1 + k_1 P_{S_1}^{-\alpha_S}}]$$

$$= \exp(-2\pi \lambda_S \int_0^\infty (1 - \frac{1}{1 + k_1 P_{S_1}^{-\alpha_S}}) dp \rho$$

$$= \exp(-\pi \lambda_S (k_1 P_{S_1}^{\frac{1}{2}}) \int_0^\infty \frac{1}{1 + \frac{1}{1 + k_1 P_{S_1}^{-\alpha_S}}} \rho \rho \exp(1) \] \tag{41}$$

We denote $G_\alpha(x) = \int_x^\infty \frac{1}{1 + \frac{1}{1 + \rho}} \exp(1) \] \tag{42}$

Therefore, we can obtain that

$$P(\text{SIR}_M \geq \gamma|r_{M,m_0} = x) = \exp(-\pi \lambda_M \gamma^2 \gamma^{-\frac{1}{\alpha_M}} G_{\alpha_M}(\gamma^{-\frac{1}{\alpha_M}})) \tag{43}$$

Finally, substituting (42) into (39), $P(\text{SIR}_M \geq \gamma)$ 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 $G_{\alpha_M}(\gamma^{-\frac{1}{\alpha_M}}) = \frac{\pi}{2} - \arctan(\gamma^{-\frac{1}{2}}) = \arccot(\gamma^{-\frac{1}{2}})$. Next, the simplified form of $P(\text{SIR}_M \geq \gamma)$ can be obtained as follows

$$P(\text{SIR}_M \geq \gamma) = \int_0^\infty f_{r_{M,m_0}}(x) \exp(-\pi x^2 \gamma^2 \frac{\pi}{2} \lambda_S P_{S_1}^{\frac{1}{2}} + \lambda_M \arccot(\gamma^{-\frac{1}{2}})) dx$$

$$= \pi \lambda_M \int_0^\infty \exp(-\pi x^2 (\lambda_M$$

$$+ \frac{1}{2} \frac{\pi}{2} \lambda_S P_{S_1}^{\frac{1}{2}} + \lambda_M \arccot(\gamma^{-\frac{1}{2}}))) dx^2.$$
\[= \lambda_M (\lambda_M + \gamma \left( \frac{\pi}{2} \lambda_S \left( \frac{P_S}{P_M} \right)^{\frac{1}{2}} + \lambda_M \arccot (\gamma \frac{1}{\lambda_M} \right))^{-1}
\]

\[(1 + \lambda_M \gamma \left( \frac{\pi}{2} \lambda_S \left( \frac{P_S}{P_M} \right)^{\frac{1}{2}} + \lambda_M \arccot (\gamma \frac{1}{\lambda_M} \right))^{-1}. \]

**APPENDIX C**

**PROOF OF THEOREM 1**

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

\[E[\log_2 (1 + y_M) | \Omega(y_M)] = \int_0^\infty f_{m_0}(x)dx \int_0^\infty \log_2 (1 + y_M) g(y_M | \Omega(y_M)) dy_M
\]

\[= \int_0^\infty f_{m_0}(x)dx \int_0^\infty \ln (1 + y_M) g(y_M | \Omega(y_M)) dy_M = \int_0^\infty f_{m_0}(x)dx \int_0^\infty \left( \int_0^\infty \frac{1}{1 + t} dt \right) g(y_M | \Omega(y_M)) dy_M,
\]

\[= \int_0^\infty \frac{1}{1 + t} \left[ \int_0^\infty f_{m_0}(x)dx \int_0^\infty \frac{P(\text{SIR}_M \geq \gamma | r_{M,m_0} = x)}{P(SIR_M \geq \gamma | r_{M,m_0} = x)} dt \right]
\]

\[= \log_2(1 + \gamma) + \frac{1}{ln 2} \int_0^\infty f_{m_0}(x)dx \int_0^\infty \frac{P(\text{SIR}_M \geq \gamma | r_{M,m_0} = x)}{P(SIR_M \geq \gamma | r_{M,m_0} = x)}(1 + t) dt,
\]

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

\[\int_0^\infty \left( \int_0^\infty \frac{1}{1 + t} dt \right) g(y_M | \Omega(y_M)) dy_M = \int_0^\infty \frac{1}{1 + t} dt \int_0^\infty g(y_M | \Omega(y_M)) dy_M = \int_0^\infty g(y_M | \Omega(y_M)) dy_M = P(y_M \geq t | \Omega(y_M)).
\]

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

**APPENDIX D**

**PROOF OF LEMMA 2**

Let vectors \( r_{S_{BL}} = [r_{SBL1}, r_{SBL2}, ..., r_{SBL,n_1}] \) and \( r_{S_{EL}} = [r_{S_{EL1}}, r_{S_{EL2}}, ..., r_{S_{EL,n_2}}] \) 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 \( r_{S_{BL}} \) and \( r_{S_{EL}} \) are given by (3)

\[ f(r_{SBL1}, ..., r_{SBL,n_1}) = \prod_{k=1}^{n_1} \frac{2r_{SBL,k}}{a^2}, \quad (45) \]

\[ f(r_{S_{EL1}}, ..., r_{S_{EL,n_2}}) = \prod_{k=1}^{n_2} \frac{2r_{S_{EL,k}}}{b^2 - a^2}, \quad (46) \]

**A. The derivation of \( P(SIR_{S_{BL}} \geq \gamma_{BL}) \)**

From the definition of successful transmission probability, \( P(SIR_{S_{BL}} \geq \gamma_{BL}) \) can be expressed as

\[ P(SIR_{S_{BL}} \geq \gamma_{BL}) = \int_0^a \cdots \int_0^a f(x_{SBL1}, ..., x_{SBL,n_1}) P(SIR_{SBL} \geq \gamma_{BL} | r_{SBL} = x_{BL}) dx_{BL}. \quad (47) \]

For notational simplicity, we let

\[ I_{S2} = \sum_{n \in \Phi_S \setminus \mathcal{N}_1} |h_{S,n}|^2 P_{SBL,n}^{-\alpha_S}, \]

\[ I_{M2} = \sum_{m \in \Phi_M} |h_{M,m}|^2 P_{M BL,m}^{-\alpha_M}. \]

Next, \( P(SIR_{S_{BL}} \geq \gamma_{BL} | r_{S_{BL}} = x_{BL}) \) is calculated as

\[ P(SIR_{S_{BL}} \geq \gamma_{BL} | r_{S_{BL}} = x_{BL}) = P\left( \sum_{k=1}^{n_1} h_{S,k} \sqrt{P_{SBL,k}} x_{BL,k}^2 \geq \gamma_{BL}(I_{S2} + I_{M2}) \right)
\]

\[ = E_{I_{S2}, I_{M2}}[\exp(-\frac{1}{\sum_{k=1}^{n_1} x_{BL,k}^{-\alpha_S}}) \gamma_{BL}(I_{S2} + I_{M2}) P_{SBL}^{-1}] \]

\[ = L_{I_{S2}} \left( \sum_{k=1}^{n_1} x_{BL,k}^{-\alpha_S} \right) L_{I_{M2}} \left( \frac{\gamma_{BL} P_{SBL}^{-1}}{\sum_{k=1}^{n_1} x_{BL,k}^{-\alpha_S}} \right), \quad (48) \]

where (d) follows the fact that \( \left| \sum_{k=1}^{n_1} h_{S,k} \sqrt{P_{SBL,k}} x_{BL,k}^2 \right| \sim P_{SBL}^{-1} \exp\left(\frac{1}{\sum_{k=1}^{n_1} x_{BL,k}^{-\alpha_S}}\right) \sim P_{SBL}^{-1} \left| \sum_{k=1}^{n_1} x_{BL,k}^{-\alpha_S} \right| \]

following the similar steps shown before, \( L_{I_{S2}}(k_2) \) and \( L_{I_{S2}}(k_2) \) are calculated as

\[ L_{I_{S2}}(k_2) = \exp(-\pi \lambda_S(k_2 P_S)^{\frac{1}{2}} G_{\alpha_S} a^2 (k_2 P_S)^{-\frac{2}{2}}), \quad (49) \]

\[ L_{I_{S2}}(k_2) = \exp(-\pi \lambda_M(k_2 P_M)^{\frac{1}{2}} G_{\alpha_M} (a^2 c^{-\alpha_S} - \pi \lambda_M(c P_M)^{\frac{1}{2}} G_{\alpha_M}(0)), \quad (50) \]

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

\[ P(SIR_{S_{BL}} \geq \gamma_{BL} | r_{S_{BL}} = x_{BL}) = \exp(-\pi \lambda_S \frac{ce^{-\frac{\alpha_S}{2}} G_{\alpha_S} (a^2 P_M)^{-\frac{2}{2}} - \pi \lambda_M(c P_M)^{\frac{1}{2}} G_{\alpha_M}(0)), \)

\[ (51) \]

Finally, substituting (51) and (45) into (47), we can obtain the expression for \( P(SIR_{S_{BL}} \geq \gamma_{BL}) \). Thus, the proof of \( P(SIR_{S_{BL}} \geq \gamma_{BL}) \) is completed.
B. The derivation of $P(SIR_{S,EL} \geq \gamma_{EL})$

Following the similar methods, $P(SIR_{S,EL} \geq \gamma_{EL})$ can be expressed as

$$P(SIR_{S,EL} \geq \gamma_{EL}) = \int_a^b \cdots \int_a^b f(x_{EL1}, \ldots, x_{ELn}) \, d\mathbf{x}_{EL}. \quad (52)$$

In order to simplify notations, we let

$$I_{S3} = \sum_{n \in N_1} |h_{S,n}|^2 P_{S} r_{ST}^{-\alpha_S},$$

$$I_{S4} = \sum_{n \in \Phi \setminus (N_1 \cup N_2)} |h_{S,n}|^2 P_{S} r_{ST}^{-\alpha_S}.$$ 

Afterwards, the expression for $P(SIR_{S,EL} \geq \gamma_{EL}|r_{S,EL} = \mathbf{x}_{EL})$ can be calculated as

$$P(SIR_{S,EL} \geq \gamma_{EL}|r_{S,EL} = \mathbf{x}_{EL}) = \mathcal{L}_{I_{S3}}(k_3) \mathcal{L}_{I_{S4}}(k_3) \mathcal{L}_{I_{S5}}(k_3), \quad (53)$$

where

$$k_3 = \frac{\gamma_{EL} P_S^{-1}}{\sum_{k=1}^{N_3} x_{EL,k}}.$$ 

$$\mathcal{L}_{I_{S3}}(k_3) = \exp(-\pi \lambda S(k_3 P_S)^{\frac{\alpha_S}{\sigma}}) \int_0^{\frac{\pi}{2}} \frac{1}{1 + t d t}, \quad (54)$$

$$\mathcal{L}_{I_{S4}}(k_3) = \exp(-\pi \lambda S(k_3 P_S)^{\frac{\alpha_S}{\sigma}}) G_{\alpha_S}(b^2(k_3 P_S)^{\frac{\alpha_S}{\sigma}}). \quad (56)$$

Thus, (53) can be rewritten as

$$P(SIR_{S,EL} \geq \gamma_{EL}|r_{S,EL} = \mathbf{x}_{EL}) = \exp(-\pi \lambda d^2 P_M \frac{M}{P_S} G_{\alpha_M}(0)$$

$$- \lambda S d^{\frac{\alpha_S}{\sigma}} \left( \int_0^{\frac{\pi}{2}} \frac{1}{1 + t d t} + G_{\alpha_S}(b^2(d^{\frac{\alpha_S}{\sigma}})) \right). \quad (57)$$

Substituting (57) and (46) into (52), we can obtain the theoretical expression for $P(SIR_{S,EL} \geq \gamma_{EL}).$
[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, Apr. 2016.

[32] Y. Huang, X. Zhang, J. Zhang, and J. Tang, “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. on Computer Commun. (INFOCOM 1999)*, vol. 1, New York, USA, Mar. 1999, pp. 126–134.

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

[36] D. P. Bertsekas, “Nonlinear programming,” *Journal of the Operational Research Society*, vol. 48, no. 3, pp. 334–334, 1997.

[37] S. S. Yong, 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.