# Performance Analysis and Optimization for Layer-Based Scalable Video Caching in 6G Networks

Junchao Ma, *Member, IEEE*, Lingjia Liu, *Senior Member, IEEE*, Bodong Shang, *Member, IEEE*, Shashank Jere, and Pingzhi Fan, *Fellow, IEEE*

***Abstract*— Scalable video caching is a promising technique to alleviate backbone traffic in sixth generation (6G) networks, and to serve users with video quality that adapts to varying channel conditions. In this paper, we develop a layer-based scalable video caching technique with non-orthogonal transmission by taking advantage of the layer feature in the scalable video. In addition, the impact of different serving base station selection algorithms is investigated. Our results indicate that both the caching placement design and transmission scheme design dominate the caching performance. To evaluate the interplay of these two policies, a tractable metric of Caching Aided Data Rate (CADR) is characterized and maximized by jointly optimizing the aforementioned two policies. Together with extensive Monte Carlo simulations, numerical results are also evaluated in this paper, demonstrating that the proposed Layer-based video Caching scheme with Non-Orthogonal Transmission (LCNOT) can achieve higher CADR performance than other baseline schemes.**

***Index Terms*— Layer-based scalable video caching, stochastic geometry, non-orthogonal transmission, caching aided data rate (CADR).**

## I. Introduction

WIRELESS networks have witnessed an explosive increase in mobile data traffic for years, and as predicted by Cisco, this traffic will reach 77.5 exabyte per month in 2022, 6 times higher than that in 2017. Around 79% of this mobile data comes from mobile video [1]. Guaranteeing and improving customers' Quality of Experience (QoE) of the received video content is critical for maximizing operators' revenue, thereby receiving significant attention in industry and academia [2], [3]. However, the increase in mobile video traffic incurs heavy pressure on the network, especially on the backhaul link between the remote core network and the nearby base stations (BSs). This pressure becomes a bottleneck in increasing revenue and improving the QoE of users [4], [5]. One potential solution to mitigate the backhaul pressure is to deploy caching capacity at local BSs to proactively store videos of interest before users request them. If users' requests are responded to by local cache, backhaul transmissions can be avoided and the latency of retrieving a requested video can be reduced accordingly [6], [7], [8], [9]. Due to the fact that the storage capacity of each BS is quite limited, only a small quantity of videos can be cached locally. Utilization of the limited caching capacity to respond to as many requests as possible, i.e., the analysis of caching placement design, is very important in the caching analysis [8], [10], [11]. Since caching is meaningless unless the desired content is successfully delivered to the targeted user, cached content delivery in the content transmission phase is another important issue in video caching analysis [8], [12]. And caching placement policies and content transmission schemes are coupled and should be jointly studied to maximize caching performance. Although significant research has been conducted in video caching and transmission, it is either the case that many problems still persist, or that video is treated as generic mobile data in such research. Therefore, exclusively analyzing video caching and transmission demands research more extensive that at present.

In this paper, video content is considered to have its own characteristics, rather than treating it generically [13]. To adapt to the conditions of different receivers (display size, video quality requirement, channel condition, etc.), a single video content contains multiple versions with different bit rates [14], [15]. Users may adaptively retrieve the appropriate version according to their particular conditions. For example, when a BS multicasts a video to multiple users simultaneously, users with a relatively better channel condition can get a better version of the requested video to maximize their received video quality, while users with a comparatively worse channel condition may receive the requested video with only the basic quality [16] to guarantee that they avail the minimum video watching experience. One promising coding technique to satisfy the aforementioned requirements is Scalable Video Coding (SVC) [17], [18], [19]. Via SVC, a video can be encoded into $L$ layers, including one base layer (BL) which comprises basic and essential information of the video, and $L-1$ enhancement layers (ELs) which contains

Manuscript received 10 October 2020; revised 25 March 2022 and 24 July 2022; accepted 21 October 2022; approved by IEEE/ACM Transactions on Networking Editor E. Uysal. Date of publication 21 November 2022; date of current version 18 August 2023. The work of Lingjia Liu was supported in part by the U.S. National Science Foundation (NSF) under Grant CNS-1811720. The work of Pingzhi Fan was supported by the NSFC Project under Grant 62020106001. *(Corresponding author: Junchao Ma.)*

Junchao Ma is with the School of Electrical and Information Engineering, Jiangsu University of Technology, Changzhou 213001, China (e-mail: junchao_ma@foxmail.com).

Lingjia Liu and Shashank Jere are with the Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA 24061 USA.

Bodong Shang is with the Department of Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, PA 15213 USA.

Pingzhi Fan is with the Key Laboratory of Information Coding and Transmission, Southwest Jiaotong University, Chengdu 610031, China.

Digital Object Identifier 10.1109/TNET.2022.3222931

1558-2566 © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on May 19,2024 at 20:39:23 UTC from IEEE Xplore. Restrictions apply.

the enhancement information and improves the received video quality. The BL of data can be encoded exclusively but the decoding of EL should be combined with lower layers of data. Layer $l$ cannot be decoded unless the previous $l-1$ layers are successfully received [20].

In this article, we analyze scalable video caching and transmission in sixth generation (6G) networks while fully considering SVC characteristics [21]. Specifically, in the caching placement phase, we apply a layer-based caching scheme, in which each layer of one video content is cached independently. Compared with the traditional content-based caching scheme in which all the layers must be stored once the video content is cached [22], the introduced layer-based caching scheme offers more caching flexibility and is more efficient in transmissions. This is because in some cases the cached high layer data cannot be retrieved due to poor channel conditions, rendering the caching meaningless. In the caching placement design, we apply the probabilistic caching which is widely used in the literature [8], [10]. Therefore, a layer of every content is cached or not is represented by a particular probability.

Specifically, before the transmission begins, the user needs to select a serving BS to retrieve its desired content. In this paper, we consider two serving transmitter selection scenarios. The first one is the Nearest Transmitter Selection (NTS) in which the user selects the nearest BS in order to maximize the transmission quality by attempting to experience minimal interference or near minimal interference from nearby BSs. However, since the caching status is not taken into consideration, the selected nearest BS may be absent from caching the desired video. In practice, this scenario is suitable for a distributed network where users have no idea about the caching placement of nearby BSs. All the requests should be forwarded to the nearest BS to enjoy the best channel condition. The other one is the Nearest Cached Transmitter Selection (NCTS) scenario. In this scenario, the nearest transmitter that caches the BL of the desired video is selected as the serving BS. As a result, the user can definitely get served by its serving BS, but the selected BS may be far away from the receiver and the user may suffer severe interference in this scenario. In practice, the NCTS scenario can be applied in a network in which a central gateway exists to maintain the caching status of BSs. Thus, as long as a request is received, the gateway can transmit the request to the nearest BS having the requested content.

In the content delivery, we adopt the power domain Non-Orthogonal Transmission (NOT) scheme, in which the BS multiplexes the layers of data to be transmitted with a part of transmit power [16], [23]. Through appropriate power allocation policy, the user can in turn decode part or all of the data layers by applying Successive Interference Cancellation (SIC) decoding method. Based on the number of layers of the requested video collected within the delay constraint, the user could experience a particular video quality with a certain data rate [24]. In this paper, the motivation that we apply the non-orthogonal transmission is as follows. First of all, by carefully allocating powers to the transmitted video layers, different importance and protections can be provided to these layers. Specifically, BLs can receive more power to improve its decoding probability, thus getting more protection in the transmission. Secondly, through non-orthogonal transmissions and the SIC decoding method, users can adaptively receive the number of layers most fitting their suffered channels. Therefore, a user with better channel can decode more than one layers of the transmitted video, while another user can only decode the BL at the same time due to poor channel condition. Thus, the scalability and layer feature of the SVC can be reflected in the transmission.

In the analysis, we apply Cache Aided Data Rate (CADR) as the performance metric to model users' satisfactions, since it can quantify the impact of caching placement design and transmission scheme in tandem, and also can reflect the unequal error protection (UEP) in the caching and transmission as well. Using stochastic geometry tools, we first characterize CADR performance as a function of caching probabilities and power allocation coefficients under different transmitter selection scenarios. Through joint optimization of the two parameters, the maximum CADR performance can be achieved. Additionally, the CADR performance of different transmitter selection algorithms and some other benchmark schemes are compared and investigated via extensive simulations. The simulation results demonstrate that the introduced Layer-based Caching with Non-Orthogonal Transmission (LCNOT) scheme outperforms the benchmark schemes with regard to CADR performance. The contributions of this paper can be listed as follows:

- Firstly, to take fully advantage of layer features of scalable video caching and transmission, in this paper, we introduce a LCNOT scheme including layer-based caching placement policy and non-orthogonal transmission. In the introduced LCNOT scheme, users can adaptively retrieve its requested content with a particular data rate depending on the caching placement probabilities and power allocation coefficients. Also, the impact of different transmitter selection scenarios is considered. To the best of authors' understanding, no previous literature studied this scheme before, and the analytical results in this paper can provide beneficial insights and inspirations for further research on this topic.
- Secondly, CADR is proposed to evaluate the caching performance since it can simultaneously quantify the impact of caching placement probabilities and the power allocation coefficients. The CADR metric is characterized using stochastic geometry, and maximized by formulating an optimization problem with regard to caching placement parameters and caching transmission parameters. To efficiently solve the optimization problem, an iteration-based solution is given and thereby the sub-optimal caching placement probabilities and power allocation coefficient are achieved accordingly.

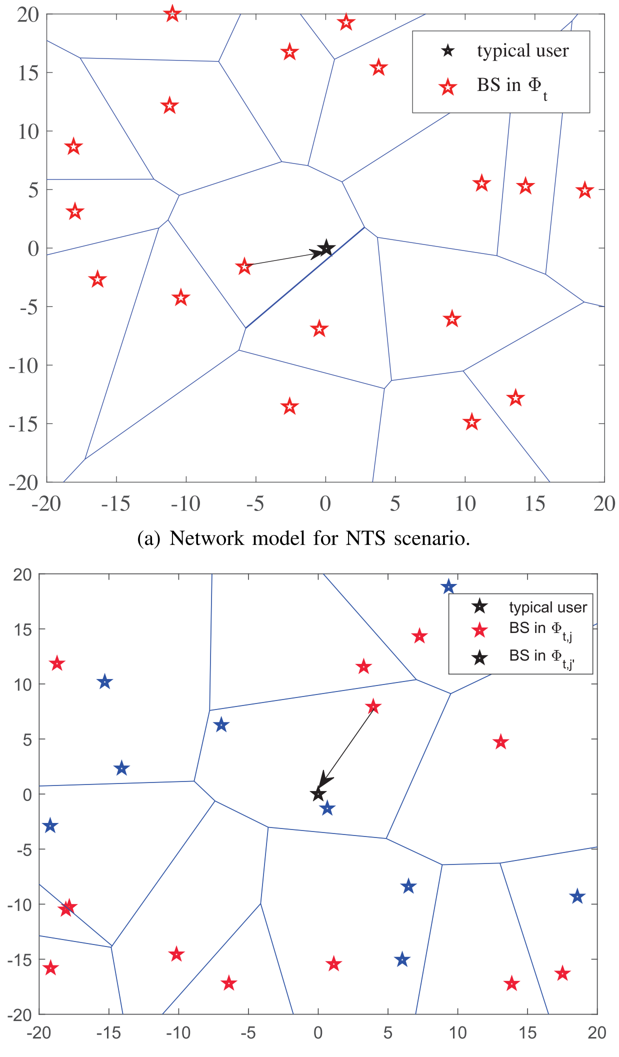
The remainder of this article is organized as follows. Literature review on SVC transmission and caching is presented in Section II. Then, we depict the system model in Section III which includes the network model, layer-based caching placement and NOT. In Section IV, we give the detailed characterization of CADR formulation and optimization under NTS and NCTS scenarios, respectively. Afterwards,

Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on May 19,2024 at 20:39:23 UTC from IEEE Xplore. Restrictions apply.

in Section V, extensive simulations are carried out to highlight the advancement of our introduced LCNOT scheme. Finally, Section VI concludes this article.

## II. Related Works

Recently, SVC transmission related research has attracted a lot of attention. To maximize the overall video quality received by users in multiple groups, [25] optimizes the resource allocation strategy and scalable multicast scheduling policy. Authors in [16] apply non-orthogonal transmission to improve the received quality of the scalable video compared to orthogonal transmission. As one of the first works of scalable video caching, [26] analyzes the scalable video caching structure, and studies the impact of different caching placement policies on backhaul offloading improvement as well as video transmission delay reduction. But no optimization is established and the traditional transmission scheme is adopted in this paper. Similarly with scalable video caching, Dynamic Adaptive Streaming over HTTP (DASH) based caching is studied in [27], in which users can dynamically choose video quality based on their requirements. However, the applied caching scheme is content-based and different versions of the same video are treated as different video contents, as a result of which one user may redundantly store some videos. Based on the aforementioned paper, in [28] the authors study video caching and transmission using DASH based caching and scalable video caching, considering each user to have a requirement for the quality of their desired video. By optimization of the caching policy exclusively, this paper aims to maximize the probability that the desired video with preferred quality requirement is successfully retrieved. The analysis in this paper ignores an important feature in SVC that a user can tolerate receiving a different quality of their desired video content if the requested version fails in its delivery. In addition, different energy-efficient scalable video caching schemes are designed in [11] to maximize the average delivery probability when a designated user requests content with a particular quality requirement. However, the proposed scheme is designed only for the designated user, without applicability to other users in the network. In our previous work, the SVC caching and non-orthogonal transmission are studied in which content-based caching placement is considered, ignoring the layer feature and scalability in the caching placement phase [22]. Poularakis et. al. in [29] optimize scalable video caching under the collaboration of multiple operators to minimize the delivery delay of users' requested content with a preferred video quality. Like [28], this paper also neglects the fact that users can tolerate a different video quality. Also, the transmission failure is underestimated in this paper. In [30], Hou et.al. consider the layer-based caching in the machine-type communication caching network, but the scalability in the transmission is not considered and the optimization of the caching and transmission processes is ignored to improve the users' content retrieval performance. [31] is quite relevant to our analysis because both works aim to optimize the caching placement policy at the caching placement phase and the power allocation at the transmission phase. But there are major differences between the two works. Firstly, the transmission considered in our paper is non-orthogonal transmission with multicast feature, while in [31] the traditional unicast feature is applied. Also, the power allocation in our paper is related to allocating different powers to video layers, while in [31] it is applied to adjust the portion of power consumed in retrieving content from backhaul and that used to transmit content from BS to users.



(a) Network model for NTS scenario.

(b) Network model for NCTS scenario.

Fig. 1. Applied network models where BSs are deployed with intensity $\lambda_t$.

## III. System Model

In this article, a wireless network drawn in Fig. 1 is considered in which BSs and users locate following Poisson Point Processes (PPPs) $\Phi_t$ and $\Phi_u$ with intensities $\lambda_t$ and $\lambda_u$, respectively. In a particular time slot, each user randomly requests a video content from the content library according to its popularity vector $\mathbf{p} = \{p(1), \cdots, p(j), \cdots, p(J)\}$, where $j$, $p(j)$, and $J$ denote the content index, the popularity of content $f_j$, and the size of the content library, respectively. Without loss of generality, it should be satisfied that $\sum_{j=1}^{J} p(j) = 1$ and $p(1) \geq p(2) \geq \cdots \geq p(J)$ [8].

Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on May 19,2024 at 20:39:23 UTC from IEEE Xplore. Restrictions apply.

In the caching placement phase, every BS adopts a probabilistic caching policy, and can cache $M_S$ bits at most [32]. For probabilistic caching placement, the caching decision of a particular layer $l$ of content $f_j$ can be modeled by a probability $q_l(j) \in [0,1]$, thus the total caching placement policy for layer $l$ is $\mathbf{q_l} = \{q_l(1), q_l(2), \cdots, q_l(J)\}$. Limited by the caching capacity of each BS, the caching probabilities should meet the requirement that

$$\sum_{j=1}^{J} \sum_{l=1}^{L} q_l\left(j\right) S_l \leq M_S, \tag{1}$$

where $S_l = tR_l$ is the size of the $l$-th layer of data, $t$ is the duration of one time slot, and $R_l$ is the data rate of layer $l$. Here it should be noted that a normalized time slot duration is considered which means $S_l \triangleq R_l$, and all the mentioned assumptions are valid in different time slots. The determination of caching decision for each layer is also assumed to be independent in the introduced caching placement policy.

In accordance with Slivnyak's theorem, we investigate the video retrieval of a typical user located at the origin in this paper, with other random users having the same performance stochastically. When a typical user requests a content $f_j$, it needs to select a serving BS before getting served [28]. In this article, two serving BS selection scenarios named NTS and NCTS are considered. As shown in Fig. 1(a), the nearest BS is selected as the serving BS regardless of the caching status in the NTS scenario. The user suffers near minimal interference from nearby BSs, but the caching status of the BS cannot be guaranteed since the requested content may not be cached by the serving BS. On the other contrary, in the NCTS scenario shown in Fig. 1(b), the nearest BS which caches the BL of the requested video $f_j$ is selected as the serving BS. This guarantees the user can find its desired content from the serving BS, but severe interference may be experienced when the serving BS is far away from the receiver.

In the transmission phase, an interference limited network is considered and the impact of noise is ignored [28]. The transmitted signal is assumed to suffer path loss and Rayleigh fading. Therefore, the received signal of a typical user is

$$y_0 = \sqrt{Pr^{-\alpha}} h x_0 + \sum_{k \in \Phi_{\text{in}}} \sqrt{P r_k^{-\alpha}} h_k x_k, \tag{2}$$

where $P$ is the transmit power of BSs, $r$ ($r_k$) is the distance between the serving BS ($k$-th interfering BS) and the typical user, and $\alpha > 2$ is the path loss exponent. $h(h_k) \sim \mathcal{CN}(0,1)$ expresses the Rayleigh fading parameter, and $\Phi_{\text{in}}$ denotes the set of interfering BSs.

Considering the layer features of SVC, in the transmission phase we adopt a power domain non-orthogonal video transmission scheme to transmit the desired video. If the first $l$ ($l = 1, 2, \cdots, L$) layers are cached and transmitted, at the transmitter side, $l_1$-th ($l_1 = 1, 2, \cdots, l$) layer of data is allocated with $b_{l_1} \in [0,1]$ part of the total transmit power, and thus in (2) $x_0 = \sum_{l_1=1}^{l} \sqrt{b_{l_1}} x_{l_1}$. The receiver applies the SIC technique to decode the requested video layer by layer from the received multiplexed signal. The process terminates until all the transmitted layers are decoded or an arbitrary layer is failed to be decoded. In practical, to balance the decoding complexity and scalability at the receiver side, we can adaptively set the number of content layers $L \leq 3$. Denote the received Signal-to-Interference Ratio (SIR) for the $l$-th layer of data by $\text{SIR}_l$ which is expressed in (3), and the $l$-th layer can be decoded if $\text{SIR}_l$ exceeds the decoding threshold $\theta_l$.

$$\text{SIR}_l = \frac{P b_l r^{-\alpha} |h|^2}{\sum_{i=l+1}^{L} P b_i r^{-\alpha} |h|^2 + I_R}, \tag{3}$$

where $I_R = \sum_{k \in \Phi_{\text{in}}} P r_k^{-\alpha} |h_k|^2$ and $\sum_{i=L+1}^{L} P b_i r^{-\alpha} |h|^2 \triangleq 0$. Here we assume the capacity-achieving channel coding method is adopted, as a result of which the decoding threshold $\theta_l$ and the experienced $l$-th layer data rate $R_l$ should satisfy $R_l = W \log(1+\theta_l)$, where $W$ is the allocated bandwidth [16].

*Theorem 1: According to the caching placement policy, when the first $l$ layers are cached and transmitted, the user can decode the first $l_1$ layers ($l_1 = 1, 2, \cdots, l$) of the requested video with the probability that*

$$P_r\left(D_{l_1,l}\right) = \begin{cases} P_r\left(|h_0|^2 > \Theta_1\right), & \text{if } C_1, \\ \cdots, & \\ P_r\left(|h_0|^2 > \Theta_m\right), & \text{if } C_m, \\ \cdots, & \\ P_r\left(|h_0|^2 > \Theta_{l_1}\right), & \text{if } C_{l_1}, \end{cases} \tag{4}$$

*where $m = 1, 2, \cdots, l_1$, $l_1 = 1, 2, \cdots, l$, $l = 1, 2, \cdots, L$, and*

$$\Theta_m = \frac{\theta_m I_R}{P r^{-\alpha}\left(b_m - \theta_m \sum_{i=m+1}^{l} b_l\right)}. \tag{5}$$

*$C_m$ means the condition satisfying the following equations for any $n = 1, \cdots, m-1, m+1, \cdots, l$ that*

$$\begin{cases} b_n > \dfrac{\theta_n}{\theta_m} b_m + \theta_n \displaystyle\sum_{i=n+1}^{m} b_i, & \text{if } n < m, \\ b_m < \dfrac{\theta_m}{\theta_n} b_n + \theta_m \displaystyle\sum_{i=m+1}^{n} b_i, & \text{if } n > m. \end{cases} \tag{6}$$

*It should be noted that $b = 1$ when $l = 1$, which indicates that all the power should be allocated to BL of $f_j$ if only BL data is cached by the serving BS. Denote $E_{l_1,l}$ as the event that the user only decodes the first $l_1$ layers when the first $l$ layers of the requested video are cached and transmitted, then the event $E_{l_1,l}$ happens with the probability that*

$$P_r\left(E_{l_1,l}\right) = \begin{cases} P_r\left(|h_0|^2 > \Theta_1\right) - P_r\left(|h_0|^2 > \Theta_{l_1+1}\right), & \text{if } M_1, \\ \cdots, & \\ P_r\left(|h_0|^2 > \Theta_{l_1}\right) - P_r\left(|h_0|^2 > \Theta_{l_1+1}\right), & \text{if } M_{l_1}, \\ 0, & \text{else}. \end{cases} \tag{7}$$

*with the assumption that $P_r\left(|h_0|^2 > \Theta_{l+1}\right) \triangleq 0$, and $M_m$ ($m = 1, \cdots, l_1$) means the conditions that satisfy*

$$\begin{cases} C_m, \\ b_m > \dfrac{\theta_m}{\theta_{l_1+1}} b_{l_1+1} + \theta_m \displaystyle\sum_{i=m+1}^{l_1+1} b_i. \end{cases} \tag{8}$$

Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on May 19,2024 at 20:39:23 UTC from IEEE Xplore. Restrictions apply.

*Proof:* The proof of Theorem 1 is shown in Appendix A. ■

Based on these probabilities derived above, we recall the responding process of the typical user's request. Once the user initiates a request towards video $f_j$, the request is forwarded to its serving BS based on the applied serving BS selection algorithm. The serving BS checks its caching status to see if $f_j$ is cached. If more than one layer is cached by the BS ($l \geq 2$), non-orthogonal transmission is applied and the user receives different layers of data depending on the suffered channel condition. If only the base layer data is cached ($l = 1$), all the power is allocated to the BL data, and the user may receive only the BL data or nothing at all. Otherwise, if BL cache is missed, the requester has to retrieve the content from its server via backhaul transmission which is beyond the scope of this paper. Therefore, the typical user adaptively enjoys different qualities of its requested video from BS caching with different data rates depending on the adopted caching placement performance which is expressed by caching placement policy, and the transmission performance which is determined by the power allocation policy. Accordingly, the average enjoyed data rate from BS caching when the typical user requests a video, i.e., CADR, can be formulated as

$$\begin{aligned} R = \sum_{j=1}^{J} p\left(j\right) \mathrm{P_r}\left[q_1\left(j\right)\right] \Bigg[ \sum_{l=2}^{L} \prod_{i=2}^{l} q_i\left(j\right)\left(1 - q_{i+1}\left(j\right)\right) \\ \times \sum_{l_1=1}^{l} \mathrm{P_r}\left(E_{l_1,l}\right) \sum_{k=1}^{l_1} R_k + \left(1 - q_2\left(j\right)\right) \mathrm{P_r}\left(E_{1,1}\right) R_1 \Bigg], \end{aligned} \tag{9}$$

where $\mathrm{P_r}[q_1\left(j\right)]$ represents the probability that the user finds the BL of $f_j$ from its serving BS, and its value is impacted by applied caching probability and serving BS selection scenario. From (9), it is obvious that the value of $R$ highly depends on the parameters $J, p, L, R_l, \theta, \mathbf{q}$, and $\mathbf{b}$, etc. Most of them are known by the network and are fixed during the system running, while the caching placement policy $\mathbf{q} = \{\mathbf{q_1}, \cdots, \mathbf{q_L}\}$ and the power allocation coefficients $\mathbf{b} = \{b_1, \cdots, b_L\}$ can be changed. Thus, to maximize the CADR performance, we establish the following optimization problem and jointly optimize the two parameters such that

$$\mathcal{P} \max_{\mathbf{q},\mathbf{b}} R \tag{10}$$

$$\text{s.t.} \begin{cases} 0 \leq q_l(j), b_l \leq 1, \\ \sum\limits_{l=1}^{L} b_l \leq 1, \\ \sum\limits_{j=1}^{J} \sum\limits_{l=1}^{L} q_l\left(j\right) \leq M_S, \quad \forall l \in [1, L] \\ b_{l_1} > \theta_{l_1} \sum\limits_{i=l_1+1}^{l} b_i, \qquad \forall l_1 \in [1, l-1] \end{cases} \tag{11}$$

The first constraint in (11) means the caching placement policy and power allocation scheme are both probability based. The second and third constraints are respectively the power allocation and caching placement parameter constraints, and the final constraint comes from (22).

## IV. Problem Solution

In this section, respectively under the NTS and NCTS scenarios, we first give the detailed derivation of CADR performance, and then effectively solve the problem $\mathcal{P}$ and get the sub-optimal caching placement probability and power allocation coefficient.

### A. NTS

As shown in Fig. 1(a), in the NTS scenario, when the typical user requests a content $f_j$, the request is forwarded to the nearest BS in $\Phi_t$. If the nearest BS contains the BL data of $f_j$, then it is selected as the serving BS and serves the typical user. Otherwise, if BL data caching is missed at the nearest BS, the request has to be responded by the remote server. Therefore the probability that the serving BS caches at least the BL of content $f_j$ is $\mathrm{P_r}\left[q_1\left(j\right)\right] = q_j$. According to [33], [34], the distance between the serving BS $X_0$ and the typical user, denoted by $r_0$, follows

$$f_{r_0}(r) = 2\pi\lambda_t r e^{-\pi\lambda_t r^2}. \tag{12}$$

Meanwhile, the interference at the typical user comes from BSs in $\Phi_t$ except $X_0$, thus $I_R$ can be represented as

$$I_R^{\mathrm{N}} = \sum_{k \in \Phi_t \backslash X_0} P|h_k|^2 r_k^{-\alpha}. \tag{13}$$

Here the superscript N means the variable $I_R$ is in the NTS scenario.

*Theorem 2: When the first $l$ layers of the requested content is cached and transmitted, the probability that the requester only decodes the first $l_1$ layers, which is derived in (7) for general case under the NTS scenario, is expressed as*

$$\begin{aligned} &P_r\left(E_{l_1,l}^N\right) = P_r\left(D_{l_1,l}^N\right) - P_r\left(D_{l_1+1,l}^N\right) \\ &= \begin{cases} \dfrac{1}{1 + s_1^N\left(\theta_1, \alpha, \mathbf{b}\right)} - \dfrac{1}{1 + s_1^N\left(\theta_{l_1+1}, \alpha, \mathbf{b}\right)}, & \textit{if } M_1, \\ \cdots, \\ \dfrac{1}{1 + s_1^N\left(\theta_{l_1}, \alpha, \mathbf{b}\right)} - \dfrac{1}{1 + s_1^N\left(\theta_{l_1+1}, \alpha, \mathbf{b}\right)}, & \textit{if } M_{l_1}, \\ 0, & \textit{else} \end{cases} \end{aligned} \tag{14}$$

*where* $\frac{1}{1+s_1^N(\theta_{l+1},\alpha,\mathbf{b})} \triangleq 0$, *and*

$$s_1^N\left(\theta_m, \alpha, \mathbf{b}\right) = \theta_m^{\frac{2}{\alpha}} \tau_m^{-\frac{2}{\alpha}} \int_{\theta_m^{-\frac{2}{\alpha}} \tau_m^{\frac{2}{\alpha}}}^{+\infty} \frac{1}{1 + u^{\frac{\alpha}{2}}} du,$$

$$\tau_m = b_m - \theta_m \sum_{i=m+1}^{l} b_i. \tag{15}$$

*Proof:* The detailed derivation is in Appendix B. ■

Substituting (14) into (9), we can get the expression of CADR under the NTS scenario. To have some insightful results and guidance in the following simulations, here we consider a special case that $L = 2$ [11], [16]. In the special case, one video constitutes one BL and one EL, which respectively corresponds to standard definition (SD) and high definition

Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on May 19,2024 at 20:39:23 UTC from IEEE Xplore. Restrictions apply.

(HD) versions of a video. Accordingly, the power allocation parameter $b = \{b_1, b_2\}$ is degraded to $b_1 = b$ and $b_2 = 1 - b$. The CADR performance under this special case is

$$\begin{aligned} R^{\mathrm{N}}_{L=2} = \sum_{j=1}^{J} p(j)\, q_1(j) \{ q_2(j) \left[ \mathrm{P_r}\left(E^{\mathrm{N}}_{1,2}\right) R_B \right. \\ \left. + \mathrm{P_r}\left(E^{\mathrm{N}}_{2,2}\right)(R_B + R_E) \right] \\ + \left[1 - q_2(j)\right] \mathrm{P_r}\left(E^{\mathrm{N}}_{1,1}\right) R_B \} , \end{aligned}$$

where $\mathrm{P_r}(E^{\mathrm{N}}_{1,2})$, $\mathrm{P_r}(E^{\mathrm{N}}_{2,2})$, and $\mathrm{P_r}(E^{\mathrm{N}}_{1,1})$ are given in (14), and $R_B \triangleq R_1$ and $R_E \triangleq R_2$ for illustration purpose. Therefore, the optimization problem $\mathcal{P}$ in (10) under the NTS case can be formulated as

$$\mathcal{P}^N_{N=2} \max_{\mathbf{q},b} R^{\mathrm{N}}_{L=2} \tag{16}$$

$$\text{s.t.} \begin{cases} \sum_{j=1}^{J} q_B(j) R_B + q_E(j) R_E \le M_S, \\ \dfrac{\theta_1}{\theta_1 + 1} < b \le 1, \end{cases} \tag{17}$$

This is a non-convex optimization since the objective function is non-convex with respect to power allocation parameter $b$ and caching placement probability $\mathbf{q}$. Achieving the global optimal solutions $\mathbf{q}^* = \{q^*_B(1), \cdots, q^*_B(J), q^*_E(1), \cdots, q^*_E(J)\}$ and $b^*$ needs more efforts and delicate analysis, and we will take it as our future work. Instead, to effectively calculate the sub-optimal solutions, we apply an iteration-based algorithm in which the original problem is decomposed into multiple sub-problems and are solved separately. As shown in Algorithm 1, we initially set $b^{(0)} = 0.5$ and $\mathbf{q}^{(0)} = \{0, \cdots, 0\}_{1\times 2J}$, and derive the local optimal $b$ and $\mathbf{q}$ iteratively. Specifically, in each iteration $k > 0$, given $b^{(k-1)}$, the optimization of (16) becomes a standard convex problem, and can be solved easily using Karush-Kuhn-Tucker (KKT) condition or *fmincon* function in MATLAB. Recall the problem in (16) with local optimal $\mathbf{q}^{(\mathbf{k})}$ and it is still non-convex due to the complex function in (15). Thus, we can get its local optimal solution $b^{(k)}$ by using one dimensional exhaustive algorithm. Concretely, we search all the potential $b$ from $\frac{\theta_1}{1+\theta_1}$ to 1 with step $\epsilon_b$ and select the $b$ that achieves the highest CADR performance as the local optimal $b^{(k)}$. The complexity of the searching process is $\mathcal{O}\left(\frac{1}{\epsilon_b(1+\theta_1)}\right)$. The progress goes to the next iteration $(k+1)$ with $b^{(k)}$ and $\mathbf{q}^{(\mathbf{k})}$, and terminates when the CADR performance becomes stable after $K$ iterations. The total complexity of the algorithm should be $\mathcal{O}\left(\frac{1}{K\epsilon_b(1+\theta_1)}\right)$. Typically, the algorithm converges after $K = 2$ or 3 iterations.

**Algorithm 1** Iteration-Based Algorithm for $\mathcal{P}^{\mathrm{N}}_{L=2}$

```
Input:
$\left(\mathbf{q}^{(0)} = \left\{\mathbf{q_1}^{(0)}, \mathbf{q_2}^{(0)}\right\}, b^{(0)}\right)$;   // initial caching placement policy and power allocation policy;
$k = 0$;
Calculate $\epsilon = R^{\mathrm{N}(0)}_{L=2}$;
Output:
while $\epsilon > 10^{-3}$ do
    $k = k + 1$;
    Solve $\mathcal{P}^{\mathrm{N}}_{L=2}$ with $b^{(k-1)}$ and get the caching placement parameters $\mathbf{q_1}^{(k)}$ accordingly;
    Solve $\mathcal{P}^{\mathrm{N}}_{L=2}$ again with $\mathbf{q_1}^{(k)}$ and $\mathbf{q_2}^{(k)}$ and calculate the $b^{(k)}$;
    Calculate $R^{(k)}_{L=2}$ with $\left(\mathbf{q_1}^{(k)}, \mathbf{q_2}^{(k)}, b^{(k)}\right)$;
    $\epsilon = R^{\mathrm{N}(k)}_{L=2} - R^{\mathrm{N}(k-1)}_{L=2}$;
end while
$K = k$;  // $K$ denotes the required number of iterations;
```

## B. NCTS

According to the caching status of video content $f_j$ and the thinning theorem of PPP, all the BSs in $\Phi_t$ can be partitioned into two parts: BSs with the content $f_j$ and BSs without the content $f_j$, following PPP $\Phi_{t,j}$ and $\Phi_{t,j'}$ with intensities $\lambda_{t,j} = \lambda_t q_1(j)$ and $\lambda_{t,j'} = \lambda_t \left[1 - q_1(j)\right]$, respectively. Here the mentioned term BS with content $f_j$ refers to the BS that caches the BL data of $f_j$. Under the NCTS scenario, the typical user's request $f_j$ will be forwarded to the nearest BS in $\Phi_{t,j}$ which may not be the nearest one in $\Phi_t$. In this case, the user has the probability of $\mathrm{P_r}\left[q_1(j)\right] = 1$ to find its requested video from the serving BS, and the distance from the user to the selected BS has the distribution that

$$f_{r_0}(r) = 2\pi\lambda_t q_1(j) r e^{-\pi\lambda_t q_1(j) r^2}. \tag{18}$$

Also the interfering BSs affecting the typical user comprise two parts: BSs except $X_0$ in $\Phi_{t,j}$ and BSs in $\Phi_{t,j'}$, thus,

$$I^{\mathrm{NC}}_R = \sum_{k\in\Phi_{t,j}\setminus X_0} P|h_k|^2 r_k^{-\alpha} + \sum_{k\in\Phi_{t,j'}} P|h_k|^2 r_k^{-\alpha}. \tag{19}$$

*Theorem 3: Substituting $I^{NC}_R$ into (4), a typical user can successfully receive the lowest $l_1$ layers of its requested content with a probability that*

$$P_r\left(D^{NC}_{l_1,l}\right) = \begin{cases} F(\theta_1), & \text{if } C_1 \\ \cdots, \\ F(\theta_{l_1}), & \text{if } C_{l_1} \end{cases} \tag{20}$$

*where*

$$\begin{aligned} & F(\theta_m) \\ & = \frac{q_1(j)}{q_1(j) + q_1(j)\, s^{NC}_1(\theta_m, \alpha, \mathbf{b}) + (1 - q_1(j))\, s^{NC}_2(\theta_m, \alpha, \mathbf{b})}, \end{aligned}$$

*and $s^{NC}_1(\theta_m, \alpha, \mathbf{b})$ and $s^{NC}_2(\theta_m, \alpha, \mathbf{b})$ are shown in (27). Further, the user only decodes the first $l_1$ layers with probability*

$$\begin{aligned} P_r\left(E^{NC}_{l_1,l}\right) &= P_r\left(D^{NC}_{l_1,l}\right) - P_r\left(D^{NC}_{l_1+1,l}\right) \\ &= \begin{cases} F(\theta_1) - F(\theta_{l_1+1}), & \text{if } M_1, \\ \cdots, \\ F(\theta_{l_1}) - F(\theta_{l_1+1}), & \text{if } M_{l_1}, \\ 0, & \text{else}. \end{cases} \end{aligned} \tag{21}$$

*Proof:* The detailed proof of Theorem 3 is shown in Appendix C. ■

The following derivation of $R^{\mathrm{NC}}$ and the corresponding optimization is similar to that in the NTS case. Here we omit the repeated description for brevity.

Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on May 19,2024 at 20:39:23 UTC from IEEE Xplore. Restrictions apply.

TABLE I
SIMULATION PARAMETER SETTINGS

| Parameter | Value |
|---|---|
| System bandwidth, $W$ | 10 MHz |
| Path loss exponent, $\alpha$ | 4 |
| EL data rate, $R_E$ | 2 Mbit/s |
| BL data rate, $R_B$ | 1.5 Mbit/s |
| BS density, $\lambda_u$ | $10^{-1}$BSs per $m^2$ |
| Content library, $J$ | 20 contents |
| Skewness parameter, $\gamma$ | 0.5 |
| Caching capacity, $M_S$ | 20 Mbit |

## V. PERFORMANCE EVALUATIONS

In this section, we carry out a series of simulations and numerical studies to evaluate the performance of the introduced LCNOT scheme. The parameters setting in simulations are given in TABLE I, unless otherwise stated. In order to study the impact of content popularity, we need to specify the distribution of the content popularity vector $\mathbf{p}$. Complying with the widely applied distribution in the literature [11], in this section, we assume the content popularity follows Zipf distribution with skewness parameter $\gamma$. That is, the popularity of content $f_j$ can be expressed as $p(j) = \frac{j^{-\gamma}}{\sum_{i=1}^{J} i^{-\gamma}}$.

To prove the superiority of our proposed scheme, we also investigate the following schemes in this section, and both NTS and NCTS scenarios are considered for these benchmark schemes. It is noted that the parameters in these schemes are also optimized.

- *Layer-based caching with orthogonal transmission (LCOT):* In the caching placement phase, layer-based caching is considered and orthogonal transmission is applied in the content delivery [16].
- *Content-based caching with non-orthogonal transmission (CCNOT):* Content-based caching placement policy is adopted which means if one video content $f_j$ is determined to be cached, all the layers are cached accordingly. Also, NOT is adopted in the content delivery phase.
- *Content-based caching with orthogonal transmission (CCOT):* Content-based caching placement and orthogonal transmission are respectively applied in the caching placement phase and content delivery phase.

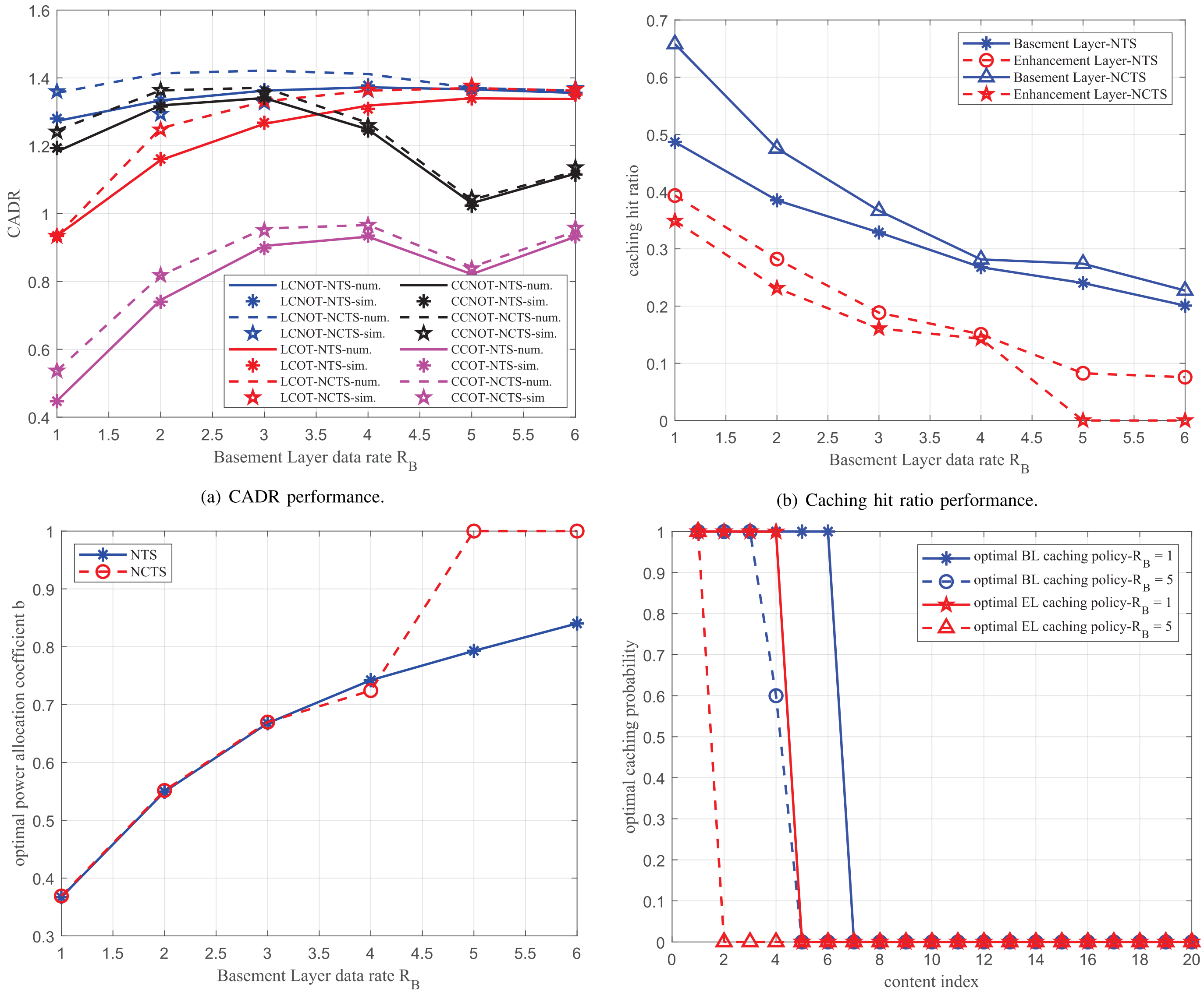
### A. Influence of BL Data Rate $R_B$

In this part, the impact of BL data rate on CADR performance is studied. From (16), we can find increasing $R_B$ can directly improve the value of CADR. On the other hand, due to the relationship that $\theta_B = 2^{R_B/W} - 1$, the increase of $R_B$ leads to a higher decoding threshold $\theta_B$, which further decreases the probabilities of successfully decoding BL of data and EL of data. As a comprehensive result, depending on which parameter dominates the final CADR performance, it may be not a monotonous trend for the CADR performance with $R_B$ increases from 1 Mbit/s to 6 Mbit/s, as shown in Fig. 2(a). From this figure, we have the following observations. For any compared schemes, the NCTS scenario performs better than the NTS scenario. Among all the potential schemes, the introduced LCNOT works the best while CCOT scheme works the worst. When $R_B$ is small, CCNOT outperforms the LCOT, with CADR approaching that of the introduced LCNOT. But as $R_B$ increases, the LCOT performs better and approaches the LCNOT gradually. This is because the threshold to successfully decode the base layers of the retrieved video increases with $R_B$. Since the successful transmission of base layer is a prerequisite of decoding enhance layers of the retrieved content, the optimal solutions tend to allocate more power for the base layer to improve its decoding probability. With almost all power allocated to the base layer of the transmitted video, the non-orthogonal transmission degrades to the orthogonal transmission, and the LCOT approaches LCNOT accordingly. This indicates that in the small $R_B$ regime, the caching placement policy design dominates the CADR performance and the layer-based caching policy is superior to the content-based caching policy, but the adopted caching transmission has very limited influence on the CADR. On the contrary, in the high $R_B$ regime, the caching transmission has a dominated impact on the CADR while the applied caching placement policy does not impact CADR performance too much. Therefore, we can focus on the caching placement policy design in the small $R_B$ regime, while concentrating the optimization of power allocaton when $R_B$ is large. What is more, we also give the simulation results of all the compared schemes, which match the theoretical results quite well, verifying the accuracy of the theoretical analysis.

In Fig. 2(b), the caching hit ratio (CHR) of BL and EL for the introduced LCNOT scheme under the NTS case and the NCTS case are respectively given. Different from the CHR definition in the literature that describes the average probability that one arbitrary user's requested content is cached by any local node, here we define CHR of one layer of content as the average probability that the particular layer of one user's requested content is cached by local nodes and is successfully transmitted and decoded. With $R_B$ increases, the required space to cache the BL of one content also increases, as a result of which, the caching capacity in terms of bits keeps unchanged but that in terms of number of layers and contents is reduced. This observation is also found in Fig. 2(c). Therefore, the CHR of EL and BL data is a decreasing function of $R_B$ with other parameters fixed. Since the successful decoding of BL data is a precondition to decode EL of data, the CHR of EL is always less than that of BL data. In addition, from the comparison of NTS and NCTS we can observe the NCTS case cares more about the caching and transmission of BL data than the NTS case and the LCNOT degrades to LCOT scheme in the $R_B > 3$ regime in the NCTS scenario.

In Fig. 2(c), the optimal power allocation coefficient $b$ for the LCNOT scheme under the NTS case and the NCTS case are respectively given as a function of $R_B$. With $R_B$ increases, the $\theta_B$ increased accordingly and the BS should offer more power to the BL to ensure its successful decoding, as a result of which the optimal $b$ is an increasing function of $R_B$ for both NTS and NCTS cases. When $R_B > 3$, all the power are allocated to the BL in the NCTS case. Also, in Fig. 2(d), we study the optimal caching policies for BL and EL of data under NTS scenario. The caching

Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on May 19,2024 at 20:39:23 UTC from IEEE Xplore. Restrictions apply.



(a) CADR performance.

(b) Caching hit ratio performance.

(c) Optimal resource allocation.

(d) Optimal caching placement policy.

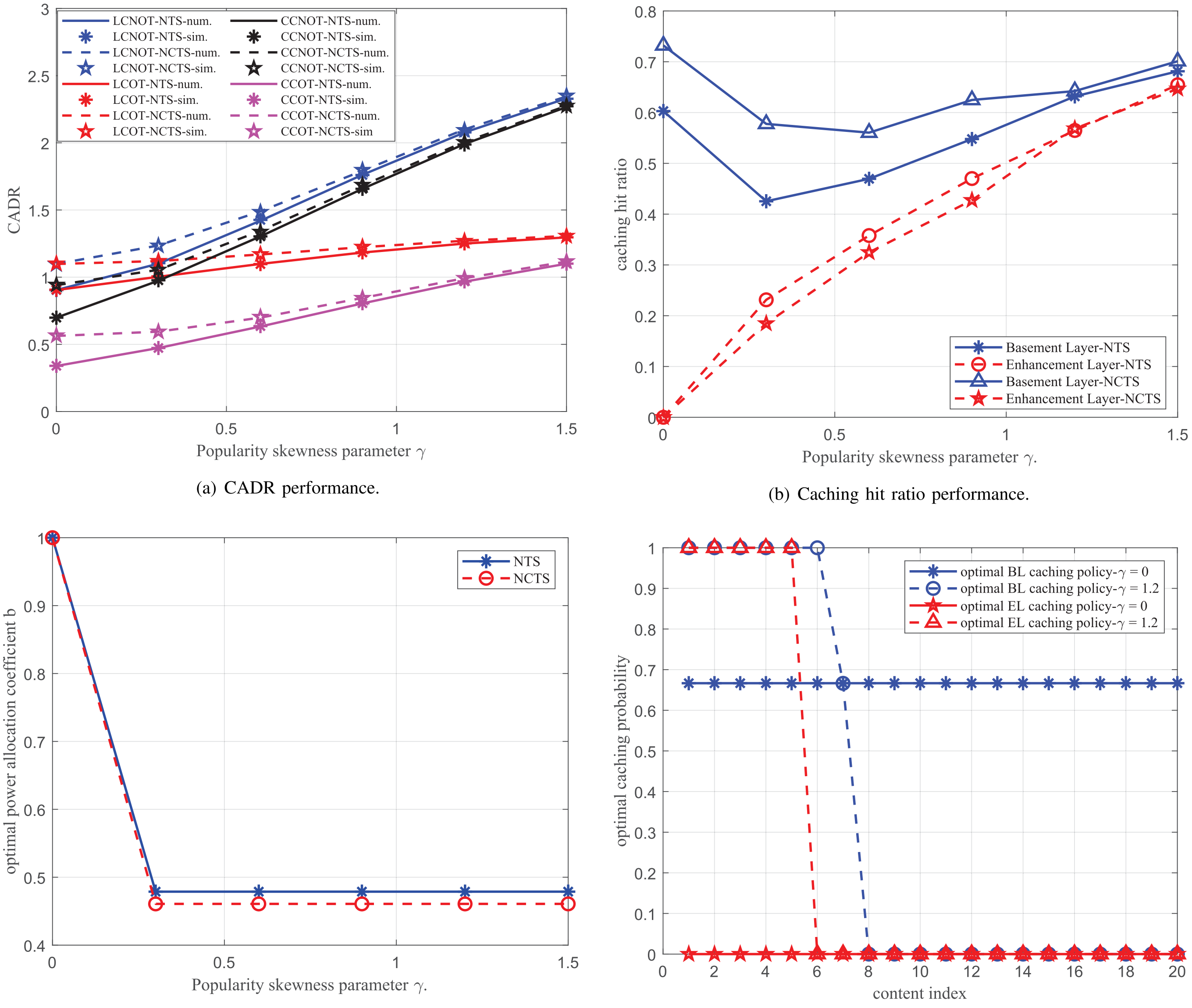
Fig. 2. CADR comparison with different BL data rates $R_B$.

probability is a decreasing function of content index $f_j$. This is intuitive since the popularity of a content with smaller index is higher, and more requests target such content. Allocating larger caching probabilities to more popular contents can achieve higher CADR performance. Another interesting observation is that the probability based caching degrades to the popularity-based caching that only the most popular BL and EL of data are cached until the caching space is fully utilized. It should be noted that since NTS case and NCTS case have similar caching probabilities, here we only present the optimal caching probabilities in the NTS case fo the sake of illustration.

### B. Influence of Popularity Skewness Parameter $\gamma$

In this part, we propose to analyze the impact of content popularity skewness parameter $\gamma$ on the CADR performance. Specifically, with a higher $\gamma$, users' requests are more skewed to the most popular videos. By caching the most popular contents, users' requests are more likely to be hit by node caching, as a result of which the CADR performance can be improved. Therefore, as shown in Fig. 3(a), the enjoyed CADR performance of all the schemes is an increasing function of $\gamma$. In LCOT, all the caching capacity is allocated to BL data because no EL will be involved in the transmission, and every BS can cache the BL of almost all the contents, thus the change of $\gamma$ does not impact its performance too much. Also, the gap between NOT based schemes (LCNOT and CCNOT) and OT based schemes (LCNOT and CCOT) increases but the gap between layer-based caching policy and content-based policy with the same transmission scheme becomes narrow. This indicates that caching transmission plays a more important role with $\gamma$ increases. Some other observations consistent with Fig. 2(a) can be seen such as the match of theoretical results and simulation results, and the outperformance of NCTS scenario against NTS scenario.

Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on May 19,2024 at 20:39:23 UTC from IEEE Xplore. Restrictions apply.



(a) CADR performance.

(b) Caching hit ratio performance.

(c) Optimal resource allocation.

(d) Optimal caching placement policy.

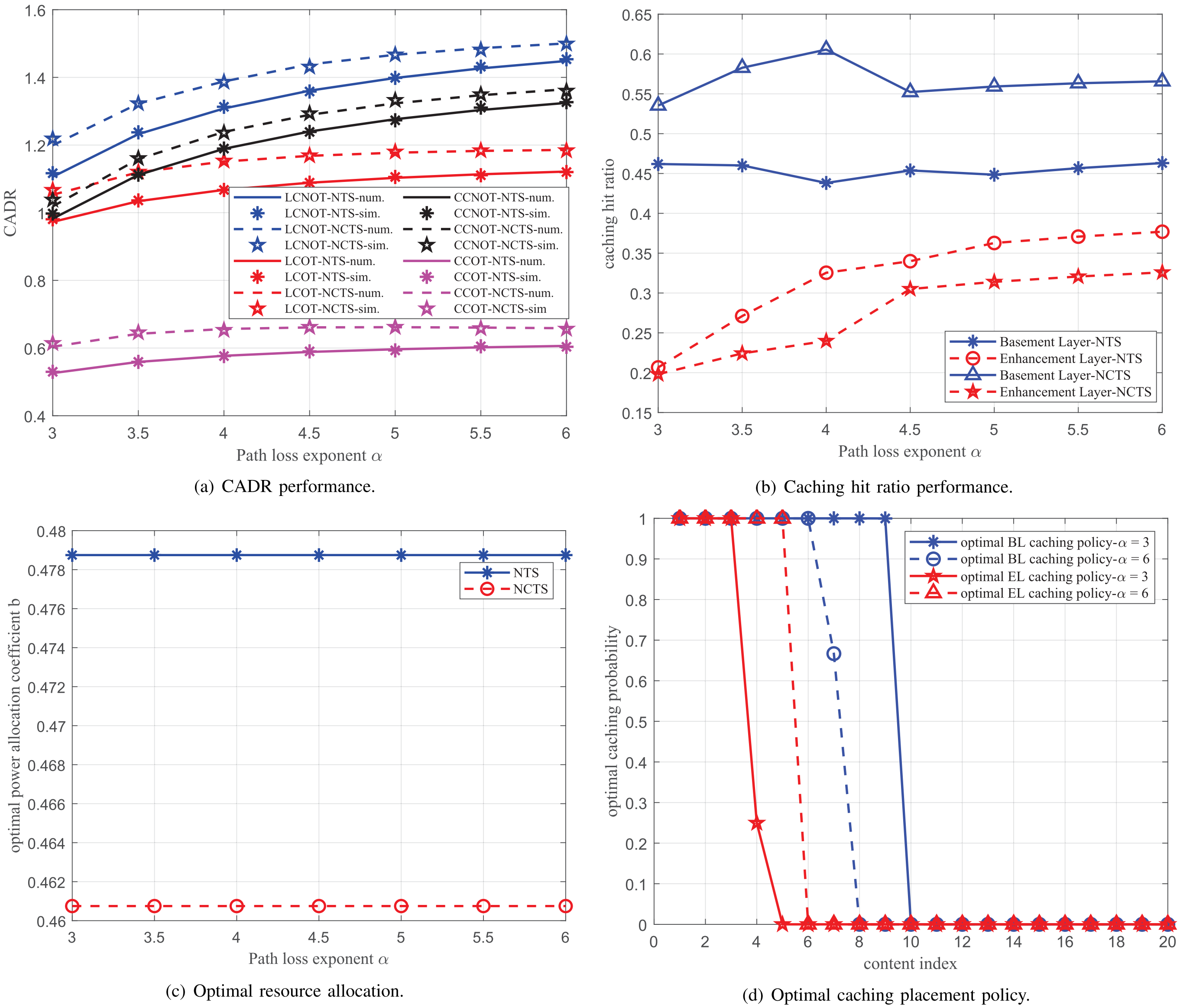
Fig. 3. CADR comparison with different popularity skewness parameters $\gamma$.

From Fig. 3(b) which plots the CHRs of BL and EL of data we have the following observations. The CHR of BL of data decreases when $\gamma$ changes from 0 to 0.3, and then increases gradually. With $\gamma$ increases, while the CHR of EL of data always increases with $\gamma$. This can be explained as follows. When $\gamma = 0$, all the contents have homogeneous popularity and users have the same probability to request any content. To handle this, the BS prefers to cache all the BL of every content uniformly. This also can be observed in Fig. 3(d) (see when $\gamma = 0$). Therefore, the CHR of BL reaches its maximum in this case. When $\gamma > 0$, the popularity of different contents differ and the gap goes shape with $\gamma$. The BS allocates more caching capacity to EL to achieve higher CADR performance (see the optimal caching when $\gamma = 1.2$ in Fig. 3(d)). As a consequence, the CHR of BL decreases and that of EL increases. With users' requests concentrating to the most popular contents, the CHR of BL increases gradually, but never exceeds its maximum value when $\gamma = 0$.

### C. Influence of Path Loss Exponent $\alpha$

Besides the influence of content related parameters (data rate $R_B$), caching placement related parameter (popularity skewness parameter $\gamma$), as shown in Fig. 4(a), in this part the impact of transmission related parameter, i.e., path loss exponent $\alpha$ on the CADR is studied. With $\alpha$ increases, the transmitted signals suffer less path loss and from Fig. 4(a) we can see the CADR of all the schemes is an increasing function of $\alpha$. As for the comparison of different schemes, the CCOT scheme performs the worst while the introduced LCNOT scheme performs the best. When $\alpha$ is small, the LCOT and CCNOT schemes perform quite similar. But with $\alpha$ increases, LCOT falls behind gradually. This indicates that

Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on May 19,2024 at 20:39:23 UTC from IEEE Xplore. Restrictions apply.



(a) CADR performance.

(b) Caching hit ratio performance.

(c) Optimal resource allocation.

(d) Optimal caching placement policy.

Fig. 4. CADR comparison with different path loss exponents $\alpha$.

the OT based scheme is not sensitive to the change of $\alpha$ (see LCOT and CCOT). In orthogonal transmission, BL of data gets all the transmit power and almost all transmission are successful, thus the CADR performs relative stable with $\alpha$. In comparison, the power is partitioned for BL and EL in the non-orthogonal transmission, not all the transmission, especially the transmission of EL, is successful. Thus, the increase of $\alpha$ improves the CADR performance a lot in these NOT schemes.

In Fig. 4(b) the trends of CHR of BL and EL are respectively plotted for the proposed LCNOT scheme. From the figure we can observe that the CHR of BL performs stable and higher than that of EL. This is because the transmission of BL data has higher chance to be successfully decoded but the successful transmission probability improves for the EL data with $\alpha$. In this case, the increase of $\alpha$ mainly impacts the transmission of EL data rather than the BL. It is worthy to note that with the increase of $\alpha$ the BS prefers to allocate more caching space to the EL data, as shown in Fig.4(d); but the transmit power allocation is not impacted as shown in Fig. 4(c). Therefore, under the NCTS, the BS allocates more caching space to EL from BL when $\alpha$ increases from $4$ to $4.5$, and the CHR of BL decreases suddenly while that of EL increases accordingly. Together with Fig. 2(c) and Fig. 3(c), we can find an interesting observation that in the proposed LCNOT with iteration based solution, the optimal $b$ does not change with caching placement related parameters and transmission environment as long as the caching probabilities are not all zeros.

## VI. Conclusion

In this article, we investigated the introduced LCNOT scheme for scalable video caching in cache-enabled wireless networks in terms of CADR. Two different serving BS selection algorithms named NTS and NCTS were considered.

Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on May 19,2024 at 20:39:23 UTC from IEEE Xplore. Restrictions apply.

Using stochastic geometry, the transmission performance in terms of the successful transmission probability of each layer when each video can be decoded into generalized layers was first derived. Then the CADR was characterized and maximized by jointly optimizing the caching placement policy and the caching transmission scheme. The optimization problem was effectively solved by using an iteration-based algorithm. Finally, numerical results were provided and were verified by Monte Carlo simulations. Results showed that the introduced LCNOT scheme outperforms the benchmark schemes, and NCTS performs better than NTS by adopting optimal caching placement probabilities and power allocation policy.

## Appendix A
## Proof of Theorem 1

A user can decode the first $l_1$ layers of its requested video when the first $l$ layers are cached are cached and transmitted with the probability that

$$\begin{aligned}\mathrm{P_r}\left(D_{l_1,l}\right) &= \mathrm{P_r}\left(\mathrm{SIR}_1 > \theta_1, \cdots, \mathrm{SIR}_m > \theta_m, \cdots, \mathrm{SIR}_{l_1} > \theta_{l_1}\right) \\ &= \mathrm{P_r}\left(\frac{b_1 P r_0^{-\alpha}|h_0|^2}{\sum_{i=2}^{l} b_i P r_0^{-\alpha}|h_0|^2 + I_R} > \theta_1, \cdots,\right. \\ &\times \left.\frac{b_{l_1} P r_0^{-\alpha}|h_0|^2}{\sum_{i=l_1+1}^{l} b_i P r_0^{-\alpha}|h_0|^2 + I_R} > \theta_{l_1}\right) \\ &\overset{(a)}{=} \mathrm{P_r}\left(|h_0|^2 > \Theta_1, \cdots, |h_0|^2\right. \\ &\quad \times \left. > \Theta_m, \cdots, |h_0|^2 > \Theta_{l_1}\right),\end{aligned}$$

where $\Theta_m$ is formulated in (5).

$$b_{l_1} > \theta_{l_1} \sum_{i=l_1+1}^{l} b_i, \tag{22}$$

in operation (a) should be satisfied for $l_1 = \{1, \cdots, l\}$ and $\sum_{i=l_1+1}^{l} b_i := 0$. $\frac{\theta_m}{b_m - \theta_m \sum_{i=m+1}^{l} b_i} > \frac{\theta_n}{b_n - \theta_n \sum_{i=n+1}^{l} b_i}$ holds for any $n \in \{1, \cdots, m-1, m+1, \cdots, l_1\}$ in $C_m$ that $\Theta_m > \max\{\theta_1, \cdots, \theta_{m-1}, \theta_{m+1}, \cdots, \theta_{l_1}\}$, which further can be transformed to (6). Accordingly, in this case the user only decodes the first $l_1$ layers with the probability that

$$\begin{aligned}\mathrm{P_r}\left(E_{l_1,l}\right) &= \mathrm{P_r}\left(\mathrm{SIR}_1 > \theta_1, \cdots, \mathrm{SIR}_{l_1} > \theta_{l_1}, \mathrm{SIR}_{l_1+1} < \theta_{l_1+1}\right) \\ &= \mathrm{P_r}\left(\mathrm{SIR}_1 > \theta_1, \cdots, \mathrm{SIR}_{l_1} > \theta_{l_1}\right) \\ &\quad - \mathrm{P_r}\left(\mathrm{SIR}_1 > \theta_1, \cdots, \mathrm{SIR}_{l_1+1} > \theta_{l_1+1}\right) \\ &= \mathrm{P_r}\left(D_{l_1,l}\right) - \mathrm{P_r}\left(D_{l_1+1,l}\right) \\ &= \begin{cases} \mathrm{P_r}\left(|h_0|^2 > \Theta_1\right) - \mathrm{P_r}\left(|h_0|^2 > \Theta_{l_1+1}\right), & \text{if } M_1, \\ \cdots, & \\ \mathrm{P_r}\left(|h_0|^2 > \Theta_{l_1}\right) - \mathrm{P_r}\left(|h_0|^2 > \Theta_{l_1+1}\right), & \text{if } M_{l_1}, \\ 0, & \text{else.} \end{cases}\end{aligned} \tag{23}$$

In the derivation, $\mathrm{P_r}\left(|h_0|^2 > \Theta_{l+1}\right) = \mathrm{P_r}\left(D_{l+1,l}\right) \triangleq 0$ is applied.

## Appendix B
## Proof of Theorem 2

In NTS scenario, $\mathrm{P_r}\left(|h_0|^2 > \Theta_m\right)$ derived in (4) is updated to

$$\begin{aligned}&\mathrm{P_r}\left(|h_0|^2 > \Theta_m^{\mathrm{N}}\right) \\ &= \mathrm{P_r}\left(|h_0|^2 > \frac{\theta_m I_R^{\mathrm{N}}}{P r_0^{-\alpha} \tau_m}\right) = E_{r_0}\left[\mathcal{L}_{I_R^{\mathrm{N}}}\left(\frac{\theta_m}{P r_0^{-\alpha} \tau_m}\right)\right] \\ &= E_{r_0}\left\{\exp\left[-2\pi\lambda_t \int_{r_0}^{\infty} \frac{1}{1 + \tau_m \theta_m^{-1}\left(\frac{r}{r_0}\right)^{\alpha}} r dr\right]\right\} \\ &= E_{r_0}\left[\exp\left(-\pi\lambda_t r_0^2 \theta_m^{\frac{2}{\alpha}} \tau_m^{-\frac{2}{\alpha}} \int_{\theta_m^{-\frac{2}{\alpha}} \tau_m^{\frac{2}{\alpha}}}^{\infty} \frac{1}{1 + u^{\frac{\alpha}{2}}} du\right)\right] \\ &= E_{r_0}\left\{\exp\left[-\pi\lambda_t r_0^2 s_1^{\mathrm{N}}\left(\theta_m, \alpha, \mathbf{b}\right)\right]\right\} \\ &= \int_0^{\infty} 2\pi\lambda_t r_0 \exp\left\{-\pi\lambda_t\left[1 + s_1^{\mathrm{N}}\left(\theta_m, \alpha, \mathbf{b}\right)\right] r_0^2\right\} dr_0 \\ &= \frac{1}{1 + s_1^{\mathrm{N}}\left(\theta_m, \alpha, \mathbf{b}\right)},\end{aligned} \tag{24}$$

where $\mathcal{L}_{T_R^{\mathrm{N}}}(s)$ is the Laplace transform of random variable $I_R^{\mathrm{N}}$ at $s$ with fixed $r_0$ and the detailed derivation of (24) can be found in [22]. Further, we can get the expressions of $\mathrm{P_r}\left(D_{l_1,l}^{\mathrm{N}}\right) - \mathrm{P_r}\left(D_{l_1+1,l}^{\mathrm{N}}\right)$ by substituting (24) into (7), and the probability that the typical user can only decode the first $l_1$ layers of its requested content in NTS case $\mathrm{P_r}\left(E_{l_1,l}^{\mathrm{N}}\right)$ can be derived as shown in (14). Theorem 2 is proved accordingly.

## Appendix C
## Proof of Theorem 3

Similar with NTS case, in the NCTS case

$$\begin{aligned}\mathrm{P_r}\left(|h_0|^2 > \Theta_m^{\mathrm{NC}}\right) &= \mathrm{P_r}\left(|h_0|^2 > \frac{\theta_m I_R^{\mathrm{NC}}}{P r_0^{-\alpha} \tau_m}\right) \\ &= E_{r_0}\left[\mathcal{L}_{I_R^{\mathrm{NC}}}\left(\frac{\theta_m}{P r_0^{-\alpha} \tau_m}\right)\right],\end{aligned} \tag{25}$$

where $I_R^{\mathrm{NC}}$ is shown in (19), and

$$\begin{aligned}&\mathcal{L}_{I_R^{\mathrm{NC}}}\left(\frac{\theta_m}{P r_0^{-\alpha} \tau_m}\right) \\ &= \exp\left[-2\pi\lambda_{t,j} \int_{r_0}^{\infty} \frac{r}{1 + \theta_m^{-1} \tau_m \left(\frac{r}{r_0}\right)^{\alpha}} dr\right] \\ &\quad \times \exp\left[-2\pi\lambda_{t,j'} \int_{0}^{\infty} \frac{r}{1 + \theta_m^{-1} \tau_m \left(\frac{r}{r_0}\right)^{\alpha}} dr\right] \\ &= \exp\left[-\pi\lambda_{t,j} r_0^2 \theta_m^{\frac{2}{\alpha}} \tau_m^{-\frac{2}{\alpha}} \int_{\theta_m^{-\frac{2}{\alpha}} \tau_m^{\frac{2}{\alpha}}}^{\infty} \frac{1}{1 + u^{\frac{\alpha}{2}}} du\right] \\ &\quad \times \exp\left[-\pi\lambda_{t,j'} r_0^2 \theta_m^{\frac{2}{\alpha}} \tau_m^{-\frac{2}{\alpha}} \int_0^{\infty} \frac{1}{1 + u^{\frac{\alpha}{2}}} du\right] \\ &= \exp\left[-\pi\lambda_{t,j} r_0^2 s_1^{\mathrm{NC}}\left(\theta_m, \alpha, \mathbf{b}\right)\right] \exp\left[-\pi\lambda_{t,j} r_0^2 s_2^{\mathrm{NC}}\left(\theta_m, \alpha, \mathbf{b}\right)\right] \\ &= \exp\left\{-\pi r_0^2\left[\lambda_{t,j} s_1^{\mathrm{NC}}\left(\theta_m, \alpha, \mathbf{b}\right) + \lambda_{t,j'} s_2^{\mathrm{NC}}\left(\theta_m, \alpha, \mathbf{b}\right)\right]\right\},\end{aligned} \tag{26}$$

Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on May 19,2024 at 20:39:23 UTC from IEEE Xplore. Restrictions apply.

where

$$\begin{cases} s_1^{\mathrm{NC}}\left(\theta_m, a, \mathbf{b}\right) = \theta_m^{\frac{2}{\alpha}} \tau_m^{-\frac{2}{\alpha}} \int_{\theta_m^{-\frac{2}{\alpha}} \tau_m^{\frac{2}{\alpha}}}^{\infty} \frac{1}{1+u^{\frac{\alpha}{2}}} du, \\ s_2^{\mathrm{NC}}\left(\theta_m, a, \mathbf{b}\right) = \theta_m^{\frac{2}{\alpha}} \tau_m^{-\frac{2}{\alpha}} \int_{0}^{\infty} \frac{1}{1+u^{\frac{\alpha}{2}}} du. \end{cases} \tag{27}$$

The derivation of (26) is detailed in [22]. Subsisting (26) and (18) back to (25), the closed form expression of $\mathrm{P_r}\left(|h_0|^2 > \Theta_m^{\mathrm{NC}}\right)$ can be achieved, and if we substitute it back to (7), then $\mathrm{P_r}\left(D_{l_1,l}^{\mathrm{NC}}\right)$ and $\mathrm{P_r}\left(E_{l_1,l}^{\mathrm{NC}}\right)$ can be derived as shown in (20) and (21), respectively. We omit the detailed description here for simplicity, and Theorem 3 is proved accordingly.

## References

[1] Cisco Visual Networking Index, *Global Mobile Data Traffic Forecast Update, 2016–2021*, Cisco White Paper, San Jose, CA, USA, 2017.

[2] R. Trestian, I.-S. Comsa, and M. F. Tuysuz, “Seamless multimedia delivery within a heterogeneous wireless networks environment: Are we there yet?” *IEEE Commun. Surveys Tuts.*, vol. 20, no. 2, pp. 945–977, 2nd Quart., 2018.

[3] Z. Zhang, Y. Yang, M. Hua, C. Li, Y. Huang, and L. Yang, “Proactive caching for vehicular multi-view 3D video streaming via deep reinforcement learning,” *IEEE Trans. Wireless Commun.*, vol. 18, no. 5, pp. 2693–2706, May 2019.

[4] T. V. Doan, L. Pajevic, V. Bajpai, and J. Ott, “Tracing the path to YouTube: A quantification of path lengths and latencies toward content caches,” *IEEE Commun. Mag.*, vol. 57, no. 1, pp. 80–86, Jan. 2019.

[5] M. Jaber, M. A. Imran, R. Tafazolli, and A. Tukmanov, “5G backhaul challenges and emerging research directions: A survey,” *IEEE Access*, vol. 4, pp. 1743–1766, 2016.

[6] L. Li, G. Zhao, and R. S. Blum, “A survey of caching techniques in cellular networks: Research issues and challenges in content placement and delivery strategies,” *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 1710–1732, 3rd Quart., 2018.

[7] S. Y. Lien, S. C. Hung, H. Hsu, and D. J. Deng, “Energy-optimal edge content cache and dissemination: Designs for practical network deployment,” *IEEE Commun. Mag.*, vol. 56, no. 5, pp. 88–93, May 2018.

[8] S. H. Chae and W. Choi, “Caching placement in stochastic wireless caching helper networks: Channel selection diversity via caching,” *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 6626–6637, Jun. 2016.

[9] J. Ma, L. Liu, B. Shang, and P. Fan, “Cache-aided cooperative device-to-device (D2D) networks: A stochastic geometry view,” *IEEE Trans. Commun.*, vol. 67, no. 11, pp. 7444–7455, Nov. 2019.

[10] Z. Chen, N. Pappas, and M. Kountouris, “Probabilistic caching in wireless D2D networks: Cache hit optimal versus throughput optimal,” *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 584–587, Mar. 2017.

[11] X. Zhang, T. Lv, W. Ni, J. M. Cioffi, N. C. Beaulieu, and Y. J. Guo, “Energy-efficient caching for scalable videos in heterogeneous networks,” *IEEE J. Sel. Areas Commun.*, vol. 36, no. 8, pp. 1802–1815, Aug. 2018.

[12] W. Wen, Y. Cui, F.-C. Zheng, S. Jin, and Y. Jiang, “Random caching based cooperative transmission in heterogeneous wireless networks,” *IEEE Trans. Commun.*, vol. 66, no. 7, pp. 2809–2825, Jul. 2018.

[13] A. Akhtar et al., “Low latency scalable point cloud communication in VANETs using V2I communication,” in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–7.

[14] H. Zhou, Y. Ji, X. Wang, and B. Zhao, “Joint resource allocation and user association for SVC multicast over heterogeneous cellular networks,” *IEEE Trans. Wireless Commun.*, vol. 14, no. 7, pp. 3673–3684, Jul. 2015.

[15] A. Tassi, I. Chatzigeorgiou, and D. Vukobratović, “Resource-allocation frameworks for network-coded layered multimedia multicast services,” *IEEE J. Sel. Areas Commun.*, vol. 33, no. 2, pp. 141–155, Feb. 2015.

[16] Z. Zhang, Z. Ma, M. Xiao, X. Lei, Z. Ding, and P. Fan, “Fundamental tradeoffs of non-orthogonal multicast, multicast, and unicast in ultra-dense networks,” *IEEE Trans. Commun.*, vol. 66, no. 8, pp. 3555–3570, Aug. 2018.

[17] J. M. Boyce, Y. Yan, J. Chen, and A. K. Ramasubramonian, “Overview of SHVC: Scalable extensions of the high efficiency video coding standard,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 1, pp. 20–34, Jan. 2016.

[18] T. Biatek, W. Hamidouche, J.-F. Travers, and O. Deforges, “Optimal bitrate allocation in the scalable HEVC extension for the deployment of UHD services,” *IEEE Trans. Broadcast.*, vol. 62, no. 4, pp. 826–841, Dec. 2016.

[19] L. Wu, Y. Zhong, W. Zhang, and M. Haenggi, “Scalable transmission over heterogeneous networks: A stochastic geometry analysis,” *IEEE Trans. Veh. Technol.*, vol. 66, no. 2, pp. 1845–1859, Feb. 2017.

[20] 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.

[21] C. Zhan and G. Yao, “SVC video delivery in cache-enabled wireless HetNet,” *IEEE Syst. J.*, vol. 12, no. 4, pp. 3885–3888, Dec. 2018.

[22] J. Ma, L. Liu, H. Song, R. Shafin, B. Shang, and P. Fan, “Scalable video transmission in cache-aided device-to-device networks,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 4247–4261, Jun. 2020.

[23] Z. Zhang, Z. Ma, M. Xiao, G. Liu, and P. Fan, “Modeling and analysis of non-orthogonal MBMS transmission in heterogeneous networks,” *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2221–2237, Oct. 2017.

[24] Y.-H. Hung, C.-Y. Wang, and R.-H. Hwang, “Optimizing social welfare of live video streaming services in mobile edge computing,” *IEEE Trans. Mobile Comput.*, vol. 19, no. 4, pp. 922–934, Apr. 2020.

[25] H. Zhu, Y. Cao, T. Jiang, and Q. Zhang, “Scalable NOMA multicast for SVC streams in cellular networks,” *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6339–6352, Dec. 2018.

[26] 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.

[27] M. Choi, J. Kim, and J. Moon, “Wireless video caching and dynamic streaming under differentiated quality requirements,” *IEEE J. Sel. Areas Commun.*, vol. 36, no. 6, pp. 1245–1257, Jun. 2018.

[28] D. Jiang and Y. Cui, “Analysis and optimization of caching and multicasting for multi-quality videos in large-scale wireless networks,” *IEEE Trans. Commun.*, vol. 67, no. 7, pp. 4913–4927, Jul. 2019.

[29] K. Poularakis, G. Iosifidis, A. Argyriou, I. Koutsopoulos, and L. Tassiulas, “Distributed caching algorithms in the realm of layered video streaming,” *IEEE Trans. Mobile Comput.*, vol. 18, no. 4, pp. 757–770, Apr. 2019.

[30] Y. Hou, N. Hu, Q. Cui, and X. Tao, “Performance analysis of scalable video transmission in machine-type-communication caching network,” *Int. J. Distrib. Sensor Netw.*, vol. 15, no. 1, p. 1550147718815851, 2019, doi: 10.1177/1550147718815851.

[31] D. Zhu, H. Lu, Z. Gu, Y. Lu, and F. Guo, “Joint power allocation and caching for SVC videos in heterogeneous networks,” in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–7.

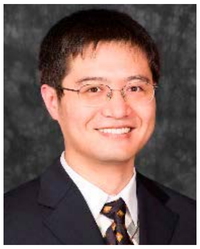
[32] D. Liu and C. Yang, “Caching policy toward maximal success probability and area spectral efficiency of cache-enabled HetNets,” *IEEE Trans. Commun.*, vol. 65, no. 6, pp. 2699–2714, Jun. 2017.

[33] J. G. Andrews, A. K. Gupta, and H. S. Dhillon, “A primer on cellular network analysis using stochastic geometry,” 2016, *arXiv:1604.03183*.

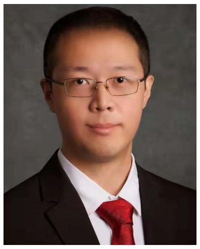
[34] M. Haenggi, “On distances in uniformly random networks,” *IEEE Trans. Inf. Theory*, vol. 51, no. 10, pp. 3584–3586, Oct. 2005.

**Junchao Ma** (Member, IEEE) received the Ph.D. degree from the Department of Information Science and Technology, Southwest Jiaotong University, in 2020. From 2017 to 2019, he was a Visiting Scholar with the Bradley Department of Electrical and Computer Engineering, Virginia Tech, under the fund of the China Scholarship Council (CSC). He is currently a Lecturer with the School of Electrical and Information Engineering, Jiangsu University of Technology. His current research interests include the Industrial Internet of Things (IIoT), the Internet of Vehicles, and video caching, computing, and communication.

Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on May 19,2024 at 20:39:23 UTC from IEEE Xplore. Restrictions apply.



**Lingjia Liu** (Senior Member, IEEE) received the B.S. degree in electronic engineering from Shanghai Jiao Tong University and the Ph.D. degree in electrical and computer engineering from Texas A&M University. He is currently an Associate Professor with the Bradley Department of Electrical and Computer Engineering, Virginia Tech. He is also an Associate Director for Affiliate Relations at the Wireless@Virginia Tech. Prior to that, he was an Associate Professor with the EECS Department, University of Kansas (KU). From 2008 and 2011, he was with the Standards & Mobility Innovation Laboratory, Samsung Research America, where he received the Global Samsung Best Paper Award in 2008 and 2010. He was leading Samsung's efforts on multiuser MIMO, CoMP, and HetNets in 3GPP LTE/LTE-advanced standards. His research interests include emerging technologies for 5G cellular networks, including machine learning for wireless networks, massive MIMO, massive machine-type communications, and mm-wave communications. He received the Air Force Summer Faculty Fellowship from 2013 to 2017, the Miller Scholarship at KU in 2014, the Miller Professional Development Award for Distinguished Research at KU in 2015, the 2016 IEEE GLOBECOM Best Paper Award, the 2018 IEEE ISQED Best Paper Award, the 2018 IEEE TCGCC Best Conference Paper Award, and the 2018 IEEE TAOS Best Paper Award.



**Bodong Shang** (Member, IEEE) received the B.Eng. degree from the School of Information Science and Technology, Northwest University, Xi'an, China, in 2015, the M.S. degree from the School of Telecommunications Engineering, Xidian University, Xi'an, in 2018, and the Ph.D. degree from the Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA, USA, in 2021. He is currently a Post-Doctoral Research Associate with the Department of Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, PA, USA. He held an internship position at Nokia Bell Labs, USA. His current research interests include vehicle-to-everything, non-terrestrial networks, reconfigurable intelligent surfaces, and edge computing. He was a recipient of the Chinese Government Award for Outstanding Self-Financed Students Abroad in 2021. He received the National Scholarship for graduate students in 2016 and 2017.

**Shashank Jere** received the B.S. degree in electrical and electronic engineering from Nanyang Technological University, Singapore, in 2014, and the M.S. degree in electrical engineering from the University of California at Los Angeles in 2016. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the Virginia Tech. From 2016 to 2019, he worked as a Platform and Product Development Engineer with Qualcomm Technologies Inc., San Diego, CA, USA. His current research interests include the broad area of wireless communications, neural networks, and machine learning.

**Pingzhi Fan** (Fellow, IEEE) received the M.Sc. degree in computer science from Southwest Jiaotong University, China, in 1987, and the Ph.D. degree in electronic engineering from Hull University, U.K., in 1994. Since 1997, he has been a Visiting Professor with Leeds University, U.K., and a Guest Professor with Shanghai Jiaotong University, since 1999. He is currently a Distinguished Professor and the Director of the Institute of Mobile Communications, Southwest Jiaotong University. He has over 290 research papers published in various international journals, and eight books (including edited), and is the inventor of 23 granted patents. His research interests include vehicular communications, wireless networks for big data, and signal design and coding. He served as a Board Member for the IEEE Region 10, IET (IEE) Council, and IET AsiaPacific Region. He is a Fellow of IET, CIE, and CIC. He was a recipient of the U.K. ORS Award in 1992, the NSFC Outstanding Young Scientist Award in 1998, and the IEEE VTS Jack Neubauer Memorial Award in 2018. He served as a General Chair or a TPC Chair for a number of international conferences, including VTC 2016 Spring, IWSDA 2017, and ITW 2018. He is the Founding Chair of the IEEE VTS BJ Chapter and IEEE ComSoc CD Chapter, and the IEEE Chengdu Section. He is an IEEE VTS Distinguished Lecturer from 2015 to 2019.

Authorized licensed use limited to: to IEEExplore provided by University Libraries | Virginia Tech. Downloaded on May 19,2024 at 20:39:23 UTC from IEEE Xplore. Restrictions apply.