Analysis and Optimization of Service Delay for Multiquality Videos in Multitier Heterogeneous Network With Random Caching

Xuewei Zhang, Tiejun Lv, Senior Member, IEEE, Yuan Ren, Wei Ni, Senior Member, IEEE, and Norman C. Beaulieu, Life Fellow, IEEE

Abstract—Aiming to minimize service delay, we propose a new random caching scheme in device-to-device (D2D) assisted heterogeneous network. To support diversified viewing qualities of multimedia video services, each video file is encoded into a base layer (BL) and multiple enhancement layers (ELs) by scalable video coding (SVC). A super layer, including the BL and several ELs, is transmitted to every user. We define and quantify the service delay of multiquality videos by deriving successful transmission probabilities when a user is served by a D2D helper, a small-cell base station (SBS), and a macrocell base station. We formulate a delay minimization problem subject to the limited cache sizes of D2D helpers and SBSs. The structure of the optimal solutions to the problem is revealed, and then an improved standard gradient projection method is designed to effectively obtain the solutions. Both theoretical analysis and Monte Carlo simulations validate the successful transmission probabilities. Compared with three benchmark caching policies, the proposed SVC-based random caching scheme is superior in terms of reducing the service delay.

Index Terms—Heterogeneous network, random caching, scalable video coding (SVC), service delay, super layer.

I. INTRODUCTION

The service modes of wireless communications are transferring from connection-oriented services [1], such as voice call and short message, to content-oriented services, such as on-demand multimedia video [2]. The amount of data traffic is experiencing a more significant surge than ever before. It is predicted that the total amount of data traffic will reach 100 exabytes by 2023, and multimedia video services will account for most of the 100 exabytes [3]. Under this circumstance, backhaul with finite bandwidth is expected to become increasingly restrictive when retrieving requested contents from the core network to wireless edges [4], i.e., coexisting base stations (BSs) and user equipment. The limited capacity of backhaul is one of the most restrictive factors, especially for time-sensitive and real-time video services. Aiming to relieve this pressing limitation of backhaul and mitigate service delays, wireless caching is proposed as a promising technique, and attracts strong attention in the fifth-generation (5G) communication networks and beyond [5].

With proactive caching enabled in wireless edges, video files requested by users can be prefetched to the local storage of wireless edges via backhauls [6]. The content placement is performed in light-traffic time periods. Cached contents can be delivered to the users, if requested. According to the types of cached contents, wireless caching can be typically classified into uncoded caching and coded caching. Earlier studies focused on the design of uncoded caching [5], [7], in which uncoded video files, especially those popular ones, are placed in the local caches of wireless edges. Later, wireless caching is extended to coded caching [8], [9], where complete videos are first encoded into different data packets and then these coded packets are locally stored by the proposed caching strategies.

Among many caching schemes, random caching, also known as probabilistic caching, is an important class of wireless caching [10]–[13], where complete video files or their combinations are prefetched to be cached under a certain caching distribution that can be optimized. In [10] and [11], by optimizing the successful transmission probabilities, the random caching distributions were determined. Zheng et al. [12] derived the content hit probability and its approximation for throughput analysis. By maximizing these two metrics, the caching probabilities were optimized. In our recent paper [13], we studied random caching in the heterogeneous network. The random caching probabilities were optimized to maximize the energy efficiency of the considered network.

With no assistance of BSs, device-to-device (D2D) communications allow users to establish direct links with their nearby neighbors. This helps reduce the overall transmission power of the system, and improve the system throughput [14], [15]. By integrating wireless caching into D2D communications, data traffic can be offloaded from small-cell BSs (SBSs) and macrocell BSs (MBSs), relieving traffic congestion and reducing service.
delay [16]–[19]. Chen et al. [16] evaluated the offloading gain and energy cost of D2D helpers, when the offloading opportunity was maximized. In [17], a machine learning model was proposed to capture the content popularity and request preference in D2D communications. Deng and Haenggi [18] focused on the energy cost of D2D helpers, and proposed two hybrid caching schemes to reduce the cost. To optimize the system throughput, Zhang et al. [19] took both D2D-link scheduling and resource allocation into account in single-hop D2D communications.

Given the limited backhaul capacity, ever-changing channel conditions, and varying user requirements, multiquality video services are in increasingly high demand, including multimedia services for standard definition videos (SDVs) and high definition videos (HDVs). To provide diversified perceptual viewing experiences to mobile users, scalable video coding (SVC), developed for advanced video coding [20], has attracted a lot of interest. With the aid of SVC, each video can be divided into a base layer (BL) and several enhancement layers (ELs) [21]. The BL contains the most basic information of the scalable video, and the file only containing the BL can be decoded as SDV, which has the lowest viewing quality. Successive ELs, together with the BL, can provide HDV. More layers provide better video quality, and the video with all divided layers can exhibit the most excellent viewing quality [22]. More technical details for the encoding and decoding process of SVC can refer to [20]. SVC has been applied to wireless caching in the literature. Zhan and Yao [23] maximized the total throughput of cache-enabled heterogeneous network by jointly optimizing SVC-based retrieving decision and data rate allocation. In [24], given the layered structure of video files, the data traffic delivered over backhaul was minimized. In our earlier work [4], we proposed an SVC-based layer placement scheme and maximized the average amount of offloaded traffic, so that most data traffic was retrieved from SBSs and the pressure was relieved on the MBS.

For large-scale video transmissions, the limited backhaul capacity is often the bottleneck of the system. Congestions in backhaul would lead to unacceptable latency. Hence, effective performance metric of service delay is crucial, and needs to be carefully designed [25]. Relying on queuing theory, Amer et al. [26] derived the average delay for unit request, and minimized the delay with the greedy algorithm. A weighted average delay for unit request was considered in [27], through which the bandwidth allocation and caching probability distribution were yielded. In [28], a learning-based caching scheme was devised in D2D-assisted network, with the objective of minimizing the average transmission delay. The delay was also minimized by jointly designing the caching and user association strategies in [29]. As mentioned earlier, mobile users can request different viewing qualities according to their preference or network states, whereas the study of SVC-supported wireless caching is still in a very earlier stage. On the other hand, provided SVC is in place, the unecessary video layers may not need to be delivered. This can significantly reduce the service delay. Therefore, delay analysis of SVC-based video retrievals is important.

This article presents a new random caching scheme in D2D-assisted three-tier heterogeneous network, consisting of D2D, SBS, and MBS tiers, to minimize the service delay. To provide diversified viewing qualities of video services, each video file is encoded by SVC. A super layer, containing the BL and several ELs, is delivered for providing multiquality multimedia video services. A user can be served by the nearest D2D helper or SBS, which caches the requested super layers. When requested contents cannot be locally provided, the nearest MBS is responsible for retrieving the contents from the core network via its backhaul at the additional cost of resource and latency.

In the proposed SVC-based random caching scheme, D2D helpers and SBSs randomly select super layers to cache, and the caching probabilities can adapt to the delay performance of the three-tier heterogeneous network. The key contributions can be summarized as follows.

1) Any requested videos are encoded by SVC, and super layers, each of which consists of a BL and several successive ELs, are cached randomly. By sending super layers to mobile users, diversified viewing qualities can be achieved. This can avoid SVC decoding at the users, hence reducing the service delay, as compared to the conventional separate transmissions of different layers.

2) We define the partial service delays when a user is served by a D2D helper, an SBS, or an MBS. We derive the corresponding successful transmission probabilities, from which the expressions for partial service delays can be attained, and so can the overall averaged service delay.

3) Subject to the limited cache sizes of D2D helpers and SBSs, the delay minimization problem is formulated. The structure of the optimal solution to the delay minimization problem is discussed. An improved standard gradient projection method is accordingly developed to provide suboptimal solutions for the random caching probabilities.

The rest of this article is arranged as follows. Section II presents the network model, channel model, and SVC-based random caching scheme. In Section III, we first define the service delay, and then derive the successful transmission probabilities to establish the expression for service delay. In Section IV, the delay minimization problem is formulated and solved. Numerical results are presented in Section V. Finally, concluding remarks are provided in Section VI.

II. SYSTEM MODEL

In this section, we provide the system model of the considered heterogeneous network, including network model, channel model and SVC-based random caching protocol.

A. Network Model

The three-tier heterogeneous network we consider consists of an MBS tier, an SBS tier, and a D2D tier. The SBSs and D2D helpers have limited cache sizes, and can store parts of requested video contents during off-peak hours. The MBSs, SBSs, and D2D helpers are single-antenna and randomly distributed, following independently and identically distributed Poisson point processes (PPPs) $\Phi_m$, $\Phi_s$, and $\Phi_d$ with density parameters $\lambda_m$, $\lambda_s$, and $\lambda_d$, respectively. Some of the users act as D2D helpers; and the rest can receive the requested contents from the D2D helpers, SBSs, or MBSs. Without loss of generality, we randomly choose one user that demands a video service as the typical user, and set the position of this user as the origin of the observed network [6], [10].

Taking the location of the typical user as the center, the D2D helpers

\[ \Phi_m, \Phi_s, \Phi_d \]

\[ \lambda_m, \lambda_s, \lambda_d \]

\[ \Phi_m, \Phi_s, \Phi_d \]

\[ \lambda_m, \lambda_s, \lambda_d \]

\[ \Phi_m, \Phi_s, \Phi_d \]

\[ \lambda_m, \lambda_s, \lambda_d \]

\[ \Phi_m, \Phi_s, \Phi_d \]

\[ \lambda_m, \lambda_s, \lambda_d \]

\[ \Phi_m, \Phi_s, \Phi_d \]

\[ \lambda_m, \lambda_s, \lambda_d \]

\[ \Phi_m, \Phi_s, \Phi_d \]

\[ \lambda_m, \lambda_s, \lambda_d \]

\[ \Phi_m, \Phi_s, \Phi_d \]

\[ \lambda_m, \lambda_s, \lambda_d \]

\[ \Phi_m, \Phi_s, \Phi_d \]

\[ \lambda_m, \lambda_s, \lambda_d \]

\[ \Phi_m, \Phi_s, \Phi_d \]

\[ \lambda_m, \lambda_s, \lambda_d \]

\[ \Phi_m, \Phi_s, \Phi_d \]

\[ \lambda_m, \lambda_s, \lambda_d \]

\[ \Phi_m, \Phi_s, \Phi_d \]

\[ \lambda_m, \lambda_s, \lambda_d \]

\[ \Phi_m, \Phi_s, \Phi_d \]

\[ \lambda_m, \lambda_s, \lambda_d \]

\[ \Phi_m, \Phi_s, \Phi_d \]

\[ \lambda_m, \lambda_s, \lambda_d \]

\[ \Phi_m, \Phi_s, \Phi_d \]

\[ \lambda_m, \lambda_s, \lambda_d \]

\[ \Phi_m, \Phi_s, \Phi_d \]

\[ \lambda_m, \lambda_s, \lambda_d \]

\[ \Phi_m, \Phi_s, \Phi_d \]

\[ \lambda_m, \lambda_s, \lambda_d \]

\[ \Phi_m, \Phi_s, \Phi_d \]

\[ \lambda_m, \lambda_s, \lambda_d \]

\[ \Phi_m, \Phi_s, \Phi_d \]

\[ \lambda_m, \lambda_s, \lambda_d \]

\[ \Phi_m, \Phi_s, \Phi_d \]

\[ \lambda_m, \lambda_s, \lambda_d \]

\[ \Phi_m, \Phi_s, \Phi_d \]

\[ \lambda_m, \lambda_s, \lambda_d \]
are distributed in the circular area with radius $r_c$ [16], [17].

Compared with ordinary D2D equipment, the adopted D2D helper is endowed with limited storage capacity. When the requested video contents can be found in the local caches of D2D helper, it can provide cached contents to its neighboring users by D2D communications. With the aid of cache-enabled D2D helpers, more multimedia data traffic can be locally provided, thus alleviating the heavy traffic burden of BSs and backhaul links. The BSs that are likely to serve the user are distributed in the cell with radius $r_d$. We assume $r_d \geq r_c$. Likewise, limited cache storage is also allocated to each BS. The D2D helpers operate in the overlay mode [30]. The spectrums allocated to the BSs and D2D helpers are orthogonal. When the typical user is receiving contents from the D2D helper, interference only comes from other nonserving helpers. The BSs and D2D helpers are also assigned with orthogonal spectrums. Therefore, the intertier interference can be avoided [31]. The D2D helpers are labeled in the ascending order of their distances from the user, and so are BSs and MBSs.

Line-of-sight propagation typically dominates small-scale wireless communications [32], [33]. Therefore, we assume that the nearest of the nodes storing the requested content is chosen as the serving one. Specifically, in the collaborative area, the nearest D2D helper that caches the requested video content, denoted by $d_0$, is selected as the serving helper. The detailed content request and user connection steps are illustrated as follows. At the beginning of each content transmission process, the MBS is able to collect user requests, and then the MBS broadcasts these video requests to D2D helpers. Afterward, each helper checks its content library to determine whether the requested contents are locally cached. If some helpers store the requested content, they will reply to the MBS, and the MBS will choose the nearest helper in user’s collaborative area as the serving one. Based on the steps shown above, the communication link between this serving helper and the content-demanding user is established, and then the requested contents can be delivered to the user via D2D communications. If all D2D helpers in the collaborative area fail to provide the requested video file, the nearest of the BSs storing the content, denoted by $s_0$, serves the user. In the worst case where the requested video cannot be retrieved from the local caches of D2D helpers and BSs, the nearest MBS, denoted by $m_0$, retrieves the content for the user via its capacity-constrained backhaul. For illustration convenience, the D2D helpers and BSs that cache the requested video files are referred to as potential serving D2D helpers and BSs, respectively. An example of the considered network model is provided in Fig. 1.

\[ \text{SIR}_d = \frac{|h_{d_0}^0|^2 (r_0^d)^{-\alpha_d}}{\sum_{k \in \Phi_d, k \neq d_0} |h_k^d|^2 (r_k^d)^{-\alpha_d}} \]  \hspace{1cm} (1)

where $h_{d_0}^0$ and $h_k^d$ are the small-scale fading from the D2D helper $d_0$ and other nonserving D2D helpers, obeying the complex Gaussian distribution with zero mean and unit variance, i.e., $CN(0,1)$; $r_0^d$ is the distance between $d_0$ and the user; $r_k^d$ is the distance between helper $k$ and the user; and $\alpha_d$ is the path loss exponent from the D2D helpers to the user, which satisfies $\alpha_d > 2$.

If the user fails to retrieve the contents from the local cache of the D2D helpers, the nearest potential serving MBS $s_0$ is assigned to transmit the videos. The received SIR achieved by SBS $s_0$ can be written as

\[ \text{SIR}_s = \frac{|h_{s_0}^0|^2 (r_0^s)^{-\alpha_s}}{\sum_{k \in \Phi_s, k \neq s_0} |h_k^s|^2 (r_k^s)^{-\alpha_s}} \]  \hspace{1cm} (2)

where $h_{s_0}^0$ and $h_k^s$ are the small-scale fading from the SBS $s_0$ and other nonserving SBSs, following $CN(0,1)$; $r_0^s$ is the distance between $s_0$ and the user; $r_k^s$ is the distance between SBS $k$ and the user; and $\alpha_s$ is the path loss exponent from the SBSs to the user, satisfying $\alpha_s > 2$.

When the requested video file needs to be retrieved from the core network through backhaul, the user is served by the nearest MBS $m_0$. The SIR delivered by $m_0$ can be written as

\[ \text{SIR}_m = \frac{|h_{m_0}^m|^2 (r_0^m)^{-\alpha_m}}{\sum_{k \in \Phi_m, k \neq m_0} |h_k^m|^2 (r_k^m)^{-\alpha_m}} \]  \hspace{1cm} (3)

where $h_{m_0}^m$ and $h_k^m$ are the small-scale fading from the nearest MBS $m_0$ and other farther MBSs, following $CN(0,1)$; $r_0^m$ is the distance between MBS $m_0$ and the user; $r_k^m$ is the distance between MBS $k$ and the user; and $\alpha_m > 2$ is the path loss exponent from the MBSs to the user.

\section{SVC-Based Caching Scheme}

We aim to provide multiquality multimedia video services to mobile users. The requested video files are preprocessed by SVC into multiple layers, and the number of divided layers is $L$. Among these layers, $l = 1$ indicates the BL, and $l = 2, \ldots, L$ indicate the ELs. The $L$-level video qualities can be provided to
the user. The video with quality level \( l \) consists of the content layers indexed from 1 to \( l \), a BL and \((l - 1)\) successive ELs. We define the video with quality level \( l \) as super layer \( l \). The user with the super layer can decode the requested video with the preferred viewing quality. In our recent works [4] and [35], the BSs were designed to cooperatively transmit different video layers to the user. At the user, successive interference cancellation (SIC) was employed to separate the individual signals from multiple serving SBSs, and then the requested video file was decoded. This would cause extra decoding latency and resource overhead. In this article, with the super layer, the user does not need to run SIC and SVC decoding, reducing computational complexity and service delay.

A video library is located in the core network, where there are \( F \) video files that can be requested by the user. By applying SVC, each video file is divided into \( L \) layers. The size of each layer is \( s_{f,l} \). The size of super layer \( l \) from the \( f \)th video is \( c_{f,l} = \sum_{k=1}^{L} s_{f,k} \).\(^4\) The videos are ranked in the descending popularity order. The Zipf law has been widely used to characterize the request probability [36]; however, this distribution is inadequate for wireless video requests. To fill this void, the Mandelbrot-Zipf (M-Zipf) law has been developed for cellular video requests [37], where the request probability of the \( f \)th video is given by

\[
p(f) = \frac{(f + q)^{-\alpha}}{\sum_{n=1}^{F} (n + q)^{-\alpha}}, \quad f = 1, 2, \ldots, F
\]

where \( \alpha \) is the skewness parameter to account for the degree of request concentration [36], and \( q \) is the plateau factor. With larger \( q \), there would be a smaller difference among the request probabilities of the most popular files. If \( q = 0 \), the M-Zipf distribution becomes the Zipf distribution. The values of both \( \alpha \) and \( q \) can be experimentally determined through real-world datasets, which is beyond the scope of this article.

Regarding the request probability of video quality, the preference for SDV of the \( f \)th video is \( g_{SDV}(f) = \frac{1}{F} \) [38], and therefore the preference for HDV is \( g_{HDV}(f) = 1 - g_{SDV}(f) \). When HDV is requested, all quality levels are assumed to have the same request popularity. To this end, the probability of the perceptual preference for super layer \( l \) from the \( f \)th video is given by

\[
p_{f,l} = \begin{cases} p(f) \cdot \frac{f-1}{F-1}, & l = 1 \\ p(f) \cdot \frac{f-l}{F-l}, & l = 2, \ldots, L. \end{cases}
\]

In the considered network, D2D helpers and SBSs are capable of caching and transmitting super layers to the typical user, where the caching probabilities are unknown. In this article, we design the random caching scheme for super layers in heterogeneous network. The \( f \)th video file with quality level \( l \) is randomly placed in the local caches of D2D helpers and SBSs with probabilities \( p_{f,l}^d \) and \( p_{f,l}^s \), respectively. The optimized random caching probabilities are collected by the MBS, and the MBS can determine whether to cache the super layers in D2D helpers and SBSs or not. The super layers can be locally placed according to these probabilities. With larger caching probability, the super layer is more likely to be cached and vice versa. The probabilities will be optimized later. For notational convenience, all values of \( p_{f,l}^d \) and \( p_{f,l}^s \) are collected in the matrices \( p^d \in \mathbb{R}^{F \times L} \) and \( p^s \in \mathbb{R}^{F \times L} \), respectively.

\(^4\)In this article, we consider the more general case, where the sizes of video files or content layers do not need to be the same. This means that the file size or the layer size has no effect on the design of caching protocol, performance analysis, and problem formulation.

### III. Service Delay Analysis

In this section, we first establish the expressions for partial service delays when the typical user is served by the nearest potential serving D2D helper, SBS, or MBS. To achieve the expressions for these delays, the successful transmission probabilities are derived, and in turn the overall service delay is obtained.

#### A. Performance Analysis of Service Delay

When referring to service delay, the transmission of the requested super layer is successful. The service delay is based on the successful transmission. A successful transmission refers to the case that the received SIR at the user exceeds a predefined quality-of-service (QoS) requirement, so that the minimum data rate for data transmission can be guaranteed.

The D2D helpers cache the super layer \( l \) of the \( f \)th video file with probability \( p_{f,l}^d \), and the potential serving D2D helpers caching the layer form a thinning PNN with density \( p_{f,l}^d \lambda_d \). According to the property of PNN, when there is no serving D2D helper in the collaborative area, the probability is \( \exp(-\lambda_d p_{f,l}^d \pi r_d^2) \). On the contrary, when there is at least one serving D2D helper in the area, this probability is

\[
\alpha_{f,l}^d = 1 - \exp(-\lambda_d p_{f,l}^d \pi r_d^2)
\]

which is also called the D2D association probability. Based on this, when the user is served by the nearest potential serving D2D helper, the partial service delay is given by

\[
D_{f,l}^d = \frac{c_{f,l}}{W_d \log_2(1 + \theta)} \left( 1 - \alpha_{f,l}^d \right) + \frac{c_{f,l}}{W_s \log_2(1 + \theta)}
\]

where \( P_{f,l}(\text{SIR} \geq \theta) \) is the successful transmission probability when the serving D2D helper delivers super layer \( l \) of the \( f \)th video, \( \theta \) is the minimum QoS requirement of the user, and \( W_d \) is the allocated system bandwidth for D2D tier.

Likewise, when there is no serving SBS in the cell, the probability is \( \exp(-\lambda_s p_{f,l}^s \pi r_d^2) \), and the SBS association probability is given by

\[
\alpha_{f,l}^s = 1 - \exp(-\lambda_s p_{f,l}^s \pi r_d^2)
\]

There are two cases that the user will be served by the SBS. The first case is that the user cannot find any D2D helper in the collaborative area with probability \( 1 - \alpha_{f,l}^d \). The second case is that the user can connect to the D2D helper but the received signal strength cannot meet the minimum QoS requirement, and the probability of this case is \( \alpha_{f,l}^d (1 - P_{f,l}(\text{SIR} \geq \theta)) \). Thus, the transmission delay when the user is served by the nearest potential serving SBS can be expressed as

\[
D_{f,l}^s = \left[ (1 - \alpha_{f,l}^d) + \alpha_{f,l}^d (1 - P_{f,l}(\text{SIR} \geq \theta)) \right] \times c_{f,l} \frac{P_{f,l}(\text{SIR} \geq \theta)}{W_s \log_2(1 + \theta)}
\]

When neither the potential serving D2D helpers nor SBSs can provide the requested super layer, the nearest MBS becomes the serving node. The requested super layer is first retrieved via backhaul from the core network, and then sent to the user.
The service delay is caused by backhaul retrieval and downlink transmission. To this end, the partial service delay is written as

\[ D_{f,t}^m = \left[1 - a_{f,t}^d P_{f,t}(\text{SIR}_d \geq \theta) - a_{f,t}^d P_{f,t}(\text{SIR}_a \geq \theta) \right] \]

\[ + a_{f,t}^d P_{f,t}(\text{SIR}_d \geq \theta) a_{f,t}^d P_{f,t}(\text{SIR}_a \geq \theta)] \]

\[ \times \left( \frac{C_{f,t}}{R_{\text{th}}} + P(\text{SIR}_m \geq \theta) \frac{C_{f,t}}{W_m \log_2(1+\theta)} \right) \]

\[ = \left(1 - a_{f,t}^d P_{f,t}(\text{SIR}_d \geq \theta) \right) \left(1 - a_{f,t}^d P_{f,t}(\text{SIR}_a \geq \theta) \right) \]

\[ \times C_{f,t} \left( \frac{1}{R_{\text{th}}} + P(\text{SIR}_m \geq \theta) \frac{C_{f,t}}{W_m \log_2(1+\theta)} \right) \] (10)

where \( P(\text{SIR}_m \geq \theta) \) is the successful transmission probability from the nearest MBS, \( W_m \) is the system bandwidth assigned for MBSs, and \( R_{\text{th}} \) is the backhaul data rate.

By taking the three types of service delay into account, the overall service delay for a successful transmission event is written as

\[ D(p^d, p^a) = \sum_{f=1}^{F} \sum_{l=1}^{L} P_{f,l} \left(D_{f,t}^d + D_{f,t}^s + D_{f,t}^m \right). \] (11)

From (7) to (11), we have the following findings.

1) The random caching probabilities of the D2D helpers and SBSs affect the association probabilities and the successful transmission probabilities, and thus have a strong impact on the service delay. In the following, given the limited cache sizes of D2D helpers and SBSs, the random caching vectors \( p^d \) and \( p^a \) are meticulously designed.

2) The three partial service delays are not parallel with each other, and there are some tradeoffs among them. When more required contents can be found in local caches, the user can be locally served by the D2D helper or the SBS. In this case, the MBS is less likely to serve the user, and the time-consuming backhaul delivery can be avoided. Thus, the overall service delay can be largely reduced.

### B. Successful Transmission Probabilities

In Section III-A, we present the delay expressions for three cases, in each of which the successful transmission probability is expected to be derived.

When the user is served by the nearest potential serving D2D helper, two cases are summarized as follows.

1) **Case 1**: In the collaborative area, the geographically nearest D2D helper is the serving D2D helper, and the probability of having the helper is \( p_{f,t}^d \).

2) **Case 2**: The geographically nearest helper is not the serving D2D helper, and the serving helper is one of the farther helpers in the observed collaborative area. The probability for this case is \( (1 - p_{f,t}^d) \).

Bearing the two cases in mind, the successful transmission probability of the D2D helper is written as

\[ P_{f,t}(\text{SIR}_d \geq \theta) = p_{f,t}^d P_{f,t}(\text{SIR}_d \geq \theta) + (1 - p_{f,t}^d) P_{f,t}^2(\text{SIR}_d \geq \theta) \] (12)

where \( P_{f,t}(\text{SIR}_d \geq \theta) \) and \( P_{f,t}^2(\text{SIR}_d \geq \theta) \) are the successful transmission probabilities with respect to Cases 1 and 2, respectively. Their expressions are provided in Theorem 1.

**Theorem 1**: In the collaborative area with radius \( r_c \), when the typical user is served by the nearest potential serving D2D helper, the successful transmission probabilities for Cases 1 and 2 are given, respectively, by

\[ P_{f,t}^1(\text{SIR}_d \geq \theta) = \begin{cases} \frac{p_{f,t}^d q_{f,t}^d(\theta \frac{\pi r_c}{\lambda})}{1 - \exp(-\lambda_d \pi r_c^2)}, & \text{if } 0 < p_{f,t}^d \leq 1 \\ 0, & \text{if } p_{f,t}^d = 0 \end{cases} \] (13)

\[ P_{f,t}^2(\text{SIR}_d \geq \theta) = \begin{cases} \frac{p_{f,t}^d q_{f,t}^d(0)}{1 - \exp(-\lambda_d \pi r_c^2)}, & \text{if } 0 < p_{f,t}^d \leq 1 \\ 0, & \text{if } p_{f,t}^d = 0 \end{cases} \] (14)

where

\[ q_{f,t}^d(x) = 1 - \exp(-\lambda_d \pi r_c^2) \left( p_{f,t}^d + \theta \frac{\pi r}{\lambda_d} G_{\alpha_d}(x) \right) \]

\[ G_{\alpha}(b) = \int_b^{\infty} \frac{1}{1 + x^2} \ dx. \] (15)

**Proof**: Refer to Appendix A.

**Remark 1**: The successful transmission probability is higher in Case 1 than in Case 2. The reason is provided as follows. In Case 1, the interference only comes from the D2D helpers that are farther away than \( d_0 \); however, in Case 2, the interference includes what is generated by the helpers closer than the serving node \( d_0 \). Therefore, the interference from Case 2 is severer, leading to a lower successful transmission probability. The same finding can be obtained when the user is served by the nearest potential serving SBS.

According to Theorem 1, the expression for \( P_{f,t}(\text{SIR}_d \geq \theta) \) can be written as

\[ P_{f,t}(\text{SIR}_d \geq \theta) = \frac{p_{f,t}^d}{1 - \exp(-\lambda_d \pi r_c^2)} \]

\[ \times \left( p_{f,t}^d \left( q_{f,t}^d(\theta \frac{\pi r_c}{\lambda}) - q_{f,t}^d(0) \right) + q_{f,t}^d(0) \right). \] (17)

Similarly, there are two cases in regards of the successful transmission when the typical user is served by the nearest potential serving SBS. The two cases are shown as follows.

1) **Case 3**: The geographically nearest SBS caches the requested video layer, and becomes the serving SBS. The probability is \( p_{f,t}^s \).

2) **Case 4**: The geographically nearest SBS does not cache the requested super layer. The serving node is one of the farther SBSs, with the probability \( (1 - p_{f,t}^s) \).

By taking the two cases into consideration, \( P_{f,t}(\text{SIR}_a \geq \theta) \) can be written as

\[ P_{f,t}(\text{SIR}_a \geq \theta) = p_{f,t}^s P_{f,t}^1(\text{SIR}_a \geq \theta) + (1 - p_{f,t}^s) P_{f,t}^4(\text{SIR}_a \geq \theta) \] (18)

where \( P_{f,t}^3(\text{SIR}_a \geq \theta) \) and \( P_{f,t}^4(\text{SIR}_a \geq \theta) \) are the successful transmission probabilities in Cases 3 and 4, respectively. The expressions for the probabilities can be found in the following theorem.

**Theorem 2**: If the typical user is served by the nearest potential serving SBS, the successful transmission probabilities
of Cases 3 and 4 are given by
\[ P^s_{f,l}(\text{SIR}_a \geq \theta) = \begin{cases} \frac{p^{f,l}_s q_{f,l}^s(x)}{1 - \exp(-\lambda_a p^{f,l}_s \pi r_d^2)}, & \text{if } 0 < p^{f,l}_s \leq 1 \\ 0, & \text{if } p^{f,l}_s = 0 \end{cases} \]
\[ P^s_{f,l}(\text{SIR}_a \geq \theta) = \begin{cases} \frac{p^{f,l}_s q_{f,l}^s(0)}{1 - \exp(-\lambda_a p^{f,l}_s \pi r_d^2)}, & \text{if } 0 < p^{f,l}_s \leq 1 \\ 0, & \text{if } p^{f,l}_s = 0 \end{cases} \]
\[ \text{where} \quad q_{f,l}^s(x) = \frac{1 - \exp\left(-\lambda_a \pi r_d^2 \left(p^{f,l}_s + \theta^{\frac{2}{m}} G_{\alpha_m}(x)\right)\right)}{p^{f,l}_s + \theta^{\frac{2}{m}} G_{\alpha_m}(x)}. \]

Proof: The proof can be obtained by referring to Appendix A.

Based on Theorem 2, the expression for \( P_{f,l}(\text{SIR}_a \geq \theta) \) can be written as
\[ P_{f,l}(\text{SIR}_a \geq \theta) = \frac{p^{f,l}_s}{1 - \exp(-\lambda_a p^{f,l}_s \pi r_d^2)} \times \left( p^{s}_f q_{f,l}^s(0) + q_{f,l}^s(x) \right). \]

When the nearest MBS is assigned to deliver the requested super layers, the successful transmission probability is shown in the following theorem.

Theorem 3: The successful transmission probability for the nearest MBS serving the typical user is given by
\[ P(\text{SIR}_m \geq \theta) = \left[ 1 + \theta^{\frac{2}{m}} G_{\alpha_m}(\theta^{-\frac{2}{m}}) \right]^{-1}. \]

Proof: See Appendix B.

Remark 2: Notice that, when \( a = 4 \), we can obtain that \( G_0(b) = \pi/2 - \arctan(b) = \arccot(b) \). Thus, if \( \alpha_d = \alpha_s = \alpha_m = 4 \), the closed-form expressions for (11), (17), (22), and (23) can be attained. Otherwise, they cannot be regarded as closed-form ones.

Remark 3: From (17) and (22), we observe that the PPP densities \( \lambda_d \) and \( \lambda_a \) have a strong impact on the successful transmission probabilities \( P_{f,l}(\text{SIR}_d \geq \theta) \) and \( P_{f,l}(\text{SIR}_a \geq \theta) \), respectively. From (23), it is seen that \( P(\text{SIR}_m \geq \theta) \) is independent of the PPP density \( \lambda_m \). We can infer that, the PPP density can affect the average number of serving nodes, and, in turn, the successful transmission probabilities \( P_{f,l}(\text{SIR}_d \geq \theta) \) and \( P_{f,l}(\text{SIR}_a \geq \theta) \). The probabilities also depend on the path-loss exponent and caching probability.

As a result, the expressions for successful transmission probabilities are obtained in (17), (22), and (23), and so are the partial service delays. By substituting (7)–(10) into (11), the expression for overall service delay is developed.

IV. PROBLEM FORMULATION AND THE PROPOSED ALGORITHM

In this section, we minimize the service delay subject to the finite cache sizes of D2D helpers and SBSs. Then, we exploit the structure of the optimal solutions, and develop the improved standard gradient projection method, from which the suboptimal caching probabilities can be achieved.

A. Problem Formulation

Given the expression for the overall service delay (11), the delay minimization problem is formulated as
\[ \min_{\mathbf{p}^d, \mathbf{p}^s} D(\mathbf{p}^d, \mathbf{p}^s) \]
subject to
\[ \sum_{f=1}^{F} \sum_{l=1}^{L} p^d_{f,l} c_{f,l} \leq M_d \]
and
\[ 0 \leq p^s_{f,l} \leq 1 \quad \forall f, \forall l \]
Constraints (24b) and (24c) indicate the cache size restrictions, where \( M_d \) and \( M_s \) are the cache sizes allocated to each D2D helper and SBS, respectively. Furthermore, inequalities (24d) and (24e) specify the feasible solution regions of the caching probabilities \( p^d_{f,l} \) and \( p^s_{f,l} \), respectively.

B. Proposed Algorithm

To solve the problem (24), we show the structure of the optimal solutions in the following theorem.

Theorem 4: The random caching probabilities \( p^d_{f,l} \) and \( p^s_{f,l} \) are maximized, if and only if the cache sizes of each D2D helper and SBS are fully utilized; or in other words, constraints (24b) and (24c) take equality.

Proof: The overall service delay includes three parts, i.e., the partial delays caused by D2D helpers, SBSs, and backhaul deliveries. In practice, the backhaul retrieval is much more time-consuming, since there are limited backhaul resources shared by many users. In other words, \( D^m_{f,l} \gg D^s_{f,l} \) and \( D^m_{f,l} \gg D^d_{f,l} \). Note that the values of \( D^m_{f,l} \) and \( D^s_{f,l} \) are comparable. To reduce the service delay, it is an effective manner to improve the content hit rate [39], which is the probability that requested contents can be satisfied by local caches. With the improved content hit rate, the user is more likely to acquire the requested super layers locally. Thus, the backhaul deliveries can be avoided and the service delay can be shortened.

By using the proposed caching and transmission schemes, the content hit rate of super layer \( l \) from the \( f \)th file is given by
\[ h_{l, f}^{\text{hit}} = 1 - \left( 1 - a^d_{f,l} P_{f,l}(\text{SIR}_d \geq \theta) \right) \left( 1 - a^s_{f,l} P_{f,l}(\text{SIR}_s \geq \theta) \right). \]

It can be seen that \( h_{l, f}^{\text{hit}} \) is a monotonically increasing function of \( p^d_{f,l} \) and \( p^s_{f,l} \). With larger caching probabilities, the content hit rate can be improved, and the service delay can be reduced to a lower level. Therefore, the minimum service delay can be attained when the cache sizes of D2D helpers and SBSs are fully utilized.

According to Theorem 4, the original optimization problem (24) can be reformulated as
\[ \min_{\mathbf{p}^d, \mathbf{p}^s} D(\mathbf{p}^d, \mathbf{p}^s) \]
subject to
\[ \sum_{f=1}^{F} \sum_{l=1}^{L} p^d_{f,l} c_{f,l} = M_d \]
Algorithm 1: The Standard Gradient Projection Method for Solving Problem (25).

1) Initialization: Set \( t = 1 \), and \( \epsilon(1) = 1 \). Input the accuracy threshold \( \Delta \), and find \( \tilde{p}_{d,f,l} \) and \( p_{s,f,l} \) feasible to (25b)–(25d).

2) Set \( \delta \ll \Delta \), and set the maximum number of iterations as \( T \).

3) while \( (t \leq T \) and \( \delta \leq \Delta \))

4) For all \( f \in \{1, \ldots, F\} \) and \( l \in \{1, \ldots, L\} \), calculate \( \frac{\partial D(p_{d,f,l}, p_{s,f,l})}{\partial p_{d,f,l}} \), and then obtain \( \tilde{p}_{d,f,l}(t+1) = p_{d,f,l}(t) - \epsilon(t) \left( \frac{\partial D(p_{d,f,l}, p_{s,f,l})}{\partial p_{d,f,l}} \right)_{p_{d,f,l}=p_{d,f,l}(t), p_{s,f,l}=p_{s,f,l}(t)} \). \( (26) \)

5) For all \( f \in \{1, \ldots, F\} \) and \( l \in \{1, \ldots, L\} \), calculate \( p_{s,f,l}(t+1) = \min \{ [\tilde{p}_{s,f,l}(t+1) - u_1]^+, 1 \} \). \( (27) \)

6) For all \( f \in \{1, \ldots, F\} \) and \( l \in \{1, \ldots, L\} \), evaluate \( \frac{\partial D(p_{d,f,l}, p_{s,f,l})}{\partial p_{s,f,l}} \), and then obtain \( \tilde{p}_{s,f,l}(t+1) = p_{s,f,l}(t) - \epsilon(t) \left( \frac{\partial D(p_{d,f,l}, p_{s,f,l})}{\partial p_{s,f,l}} \right)_{p_{d,f,l}=p_{d,f,l}(t), p_{s,f,l}=p_{s,f,l}(t)} \). \( (29) \)

7) For all \( f \in \{1, \ldots, F\} \) and \( l \in \{1, \ldots, L\} \), evaluate \( p_{s,f,l}(t+1) = \min \{ [\tilde{p}_{s,f,l}(t+1) - u_2]^+, 1 \} \). \( (30) \)

8) \( \Delta = |D(p_{d,f,l}, p_{s,f,l})^{(t+1)} - D(p_{d,f,l}, p_{s,f,l})^{(t)}| \)

9) \( t = t + 1 \) and update the stepsize \( \epsilon(t) \).

10) end

In Algorithm 1, Steps 4) and 6) calculate the partial derivatives of the service delay with respect to \( p_{d,f,l} \) and \( p_{s,f,l} \). Given a predefined step size \( \epsilon(t) \), \( \tilde{p}_{d,f,l}(t+1) \) and \( \tilde{p}_{s,f,l}(t+1) \) are updated according to formulas (26) and (29). They can be regarded as the optimal values of \( p_{d,f,l} \) and \( p_{s,f,l} \) achieved at the \( (t+1) \)th iteration without the cache size restrictions of each D2D helper and SBS. In order to not exceed the cache sizes, the values of \( \tilde{p}_{d,f,l}(t+1) \) and \( \tilde{p}_{s,f,l}(t+1) \) are projected onto the variable sets comprised of the cache size restrictions (25b) and (25c), as done in (27) and (30). \( [\cdot]^+ \) is employed to guarantee nonnegativity. Thus, \( \min \{ [\cdot]^+, 1 \} \) can maintain valid feasible value regions of \( p_{d,f,l} \) and \( p_{s,f,l} \), satisfying (25d). As the result of adopting Algorithm 1, suboptimal solutions for the random caching probabilities can be achieved while all constraints in (25) are satisfied. We proceed to discuss the property of the proposed algorithm. In the feasible variable region comprised by constraints (25b)–(25d), the optimal point is selected at the negative direction with the fastest descent at each iteration. Therefore, the obtained value of \( D(p_{d,f,l}, p_{s,f,l}) \) is less than or at least equal to the value sourced from the last iteration, and this algorithm will certainly converge. From the simulation results shown later, the proposed algorithm can converge to the suboptimal solution within a small number of iterations. At each iteration, the cost for calculating the partial derivatives dominates over the cost of the proposed algorithm, and the total required number of calculating the partial derivatives scales with \( FL \). Thus, when using \( O \) function for complexity analysis, according to [40] and [41], the overall computational complexity of the proposed algorithm is \( O(FL) \).

V. SIMULATION RESULTS

In order to show the correctness of the derived successful transmission probabilities, similar to [47], the numerical results of both analysis and Monte Carlo (MC) simulations for these expressions are displayed. The MC experiments for each point are performed more than 50,000 times. Next, we show the numerical results of the service delay. For comparison, three benchmark strategies are adopted as follows.

1) Most popular content placement (MPCP): The super layers from the most popular videos are locally cached in the D2D helpers and SBSs. This caching scheme is used in most uncoded caching, e.g., in [5] and [7].

2) Equal probability content placement (EPCP): In the local caches of D2D helpers and SBSs, the super layers of all video files are randomly stored with the same caching probability until all cache sizes are fully utilized. This scheme overlooks the file popularity distribution and perceptual quality preference.

3) Independent content placement (ICP): The D2D helpers and SBSs randomly select different super layers to cache, irrespective of the actual service requirements of users.

The values of simulation parameters are listed in Table I. We have given the references for most values of the simulation parameters, and the other values are set accordingly. In specific, the density of the SBS is set as ten times of the D2D helpers; and the range of the plateau factor should satisfy \( 0 \leq q \leq F \), and we set \( q = 5 \); and the cache sizes of the D2D helpers and SBSs are

\[ \sum_{f=1}^{F} \sum_{l=1}^{L} \tilde{p}_{d,f,l} c_{f,l} = M_d \]  
\[ \sum_{f=1}^{F} \sum_{l=1}^{L} \tilde{p}_{s,f,l} c_{f,l} = M_s \]  
\[ (24c), (24d), (24e). \]  

Note that the objective (25a) is differentiable; however, it is difficult to tell (25a) is convex due to its complex form. Constraints (25b)–(25d) are linear equalities and inequalities with regard to \( \tilde{p}_{d,f,l} \) and \( \tilde{p}_{s,f,l} \). They form a convex variable set. In light of this, the standard gradient projection method can be employed to solve this problem. The standard gradient projection method is an effective approach to solve the optimization problem with a differentiable objective function over a convex variable set [10]. The detailed procedure of solving this problem is summarized in Algorithm 1.
Fig. 2. Successful transmission probabilities when the typical user is served by the D2D helper, SBS, and MBS. (a) $P_{f,l} (\text{SIR}_d \geq \theta)$ versus $p_{d,l}^f$. (b) $P_{f,l} (\text{SIR}_s \geq \theta)$ versus $p_{s,l}^f$. (c) $P (\text{SIR}_m \geq \theta)$ versus $\theta$.

### TABLE I

| Parameters | Values |
|------------|--------|
| $\theta$   | 5 dB [42] |
| $F$        | 20 [43] |
| $L$        | 2 [44] |
| $s_{f,l}$  | 25 Mbits [39] |
| $q$        | 5 |
| $M_d, M_s$ | 200 Mbits, 500 Mbits |
| $\lambda_d, \lambda_s$ | 0.01 [16], 0.001 |
| $\alpha_d, \alpha_s, \alpha_m$ | 4 [46] |

Fig. 3. Derived service delay versus the D2D caching probability.

In Fig. 2, we plot the successful transmission probabilities. It is shown that the performance gap is negligible between our analysis and MC simulations. This validates (17), (22), and (23).

In Fig. 2(a) and (b), it is revealed that larger caching probabilities have stronger positive effects on successful transmission probabilities since the closer serving D2D helper and SBS can be found around the user, providing stronger received signal strength. From Fig. 2(c), we can conclude that a larger QoS requirement leads to a lower successful transmission probability. Furthermore, a smaller path loss exponent can adversely affect the successful transmission probability. This is because, though the received signal strength increases as the path loss exponent decreases, it fails to scale with the growing interference caused by other MBSs.

The service delay in (11) is shown in Fig. 3, varying with the caching probabilities. We can see that larger values of random caching probabilities lead to lower service delay. This is because larger caching probabilities indicate that more content requests can be satisfied in the local caches of the D2D helpers and SBSs, and the contents do not need to be retrieved via backhauls. In practical implementations, though larger caching probabilities are beneficial for reducing the service delay, the caching probabilities need to be carefully designed to meet the finite cache sizes.

In Fig. 4, the relationship between the service delay and the minimum QoS requirement $\theta$ is presented. We can see that the service delay increases as $\theta$ grows. The reason is as follows. When $\theta$ grows, the successful transmission probabilities are substantially decreased, which leads to a higher service delay. In addition, the proposed random caching scheme is superior to the three benchmark strategies in terms of service delay. This is because random caching is able to fully exploit the accumulated cache sizes of D2D helpers and SBSs, and more layers can be flexibly placed. As for the three benchmarks, the MPCP shows better delay performance than EPCP and ICP, since EPCP and ICP overlook the video popularity and the viewing quality preference.

In Fig. 5, we show the relationship between the service delay and the cache sizes $M_d$ and $M_s$. It is clear that a larger cache size results in a lower service delay. With a larger cache size, more requested super layers can be satisfied by the local caches of the D2D helpers and SBSs, avoiding time-consuming content retrievals from the core network via backhauls. Compared to Fig. 5(a) and (b), we infer that allocating more cache sizes to the D2D helpers leads to more reduced service delay. This is because the D2D helpers are geographically closer to the typical user, and obtaining the requested layers from them is more delay-effective.

The relationship between the service delay and the skewness parameter $\alpha$ is provided in Fig. 6. As $\alpha$ grows, the user requests increasingly focus on a small number of popular videos. The
Fig. 5. Service delay versus different cache sizes of the D2D helper and SBS. (a) Service delay versus the D2D cache size. (b) Service delay versus the SBS cache size.

Fig. 6. Service delay versus the skewness parameter.

Fig. 7. Service delay versus the backhaul transmission rate.

Fig. 8. Convergence property of the proposed algorithm.

In Fig. 7, we present the relationship between the service delay and the backhaul transmission rate $R_{bh}$. We see that a larger $R_{bh}$ results in a lower service delay, and the performance gap between the proposed caching scheme and MPLP decreases as $R_{bh}$ grows. This is due to the fact that, as $R_{bh}$ grows, the transmission rate of backhaul deliveries has an increasingly negligible impact on the service delay. When $R_{bh}$ is high enough, the backhaul capacity would be no longer a limiting factor in large-scale video distributions.

In Fig. 8, the convergence property of the proposed algorithm is shown. Obviously, Algorithm 1 can converge to the suboptimal solution after a small number of iterations, which validates the excellent convergence performance of the proposed algorithm. Additionally, we can observe that larger values of the minimum QoS requirement $\theta$ result in larger service delays, and in the meanwhile the conclusion derived from Fig. 4 is also clearly verified.

VI. CONCLUSION

In this article, we investigated the random caching scheme for delay minimization in D2D-assisted heterogeneous network. To provide diversified viewing qualities of multimedia video services, the super layers were transmitted to the user. We first analyzed the successful transmission probabilities, and then obtained the closed-form expression for the overall service delay. Based on this expression, we minimized the service delay efficiently by applying the improved standard gradient projection method. Numerical results validate our analysis of successful transmission probabilities, and the proposed random caching scheme was shown to be superior to the MPCP, EPCP, and ICP strategies.

APPENDIX A

PROOF OF THEOREM 1

We start with the calculation of $P^1_{f,l}(\text{SIR}_d \geq \theta)$. Keep in mind that, in Case 1, the intratier interference results from the D2D helpers that are farther away than $d_0$ from the typical user. Hence,
\[ P_{f,l}^1 (SIR_d \geq \theta) \] can be calculated as
\[ P_{f,l}^1 (SIR_d \geq \theta) = \int_0^r f_d(r)P_{f,l}^1 (SIR_d \geq \theta | r_0^d = r) \, dr \] (32)
where
\[ f_d(r) = \frac{2\pi \lambda_d \rho \exp \left( -\lambda_d \rho \frac{r^2}{\pi} \right)}{1 - \exp (-\lambda_d \rho \frac{r^2}{\pi})} \] (33)
is the probability density function (PDF) of the distance between the D2D helper \( d_0 \) and the user. For notational simplicity, the interference from other nonserving D2D helpers is \( I_d = \sum_{k \in \Phi_d \setminus \{d_0\}} |h_k|^2 (r_k^d)^{-\alpha_d} \). The conditional successful transmission probability \( P_{f,l}^1 (SIR_d \geq \theta | r_0^d = r) \) is calculated as
\[ P_{f,l}^1 (SIR_d \geq \theta | r_0^d = r) \]
\[ = \mathbb{P}_{f,l} \left( |h_0|^2 \geq I_d, r^{\alpha_d} \right) \]
\[ = \mathbb{E}_{I_d, h_0} \left[ \prod_{k \in \Phi_d \setminus \{d_0\}} \exp \left( -\theta r^{\alpha_d} \right) \right] \]
\[ = \mathbb{E}_{\Phi_d, h_0} \left[ \prod_{k \in \Phi_d \setminus \{d_0\}} \frac{1}{1 + \theta r^{\alpha_d} \left( r_k^d \right)^{-\alpha_d}} \right] \]
\[ \exp \left( -2\pi \lambda_d \int_0^r \frac{1}{1 + \theta r^{\alpha_d} \rho^{-\alpha_d}} \, \rho \, d\rho \right) \]
\[ = \exp \left( -\pi \lambda_d \theta \frac{r}{\pi} r^2 G_{\alpha_d} \left( \theta - \frac{r}{\pi} \right) \right) \] (34)
where \( G_{\alpha_d}(b) = \int_b^\infty \frac{1}{1 + \theta \frac{r}{\pi}} \, dr \). Since \( |h_0|^2 \) follows the exponential distribution with unit mean, its PDF is \( f_{|h_0|^2}(x) = \exp(-x) \), and then \( P_{f,l}^1 (|h_0|^2 \geq I_d, r^{\alpha_d} \theta) \) can be accordingly calculated. As \( I_d \) is another stochastic variable in \( P_{f,l}^1 (|h_0|^2 \geq I_d, r^{\alpha_d} \theta) \), the expectation of \( I_d \) is supposed to be considered, as shown in Step (a). \( |h_0|^2 \) also follows the exponential distribution with unit mean. Therefore, its expectation can be easily yielded based on its known PDF, as given in Step (b). Finally, Step (c) can be obtained by leveraging the probability generating function property of the PPP, and more details for this special property can refer to [48, Definition 4.3]. As a result, the expression for \( P_{f,l}(SIR_d \geq \theta) \) can be obtained by substituting (33) and (34) into (32), given as follows:
\[ P_{f,l}^1 (SIR_d \geq \theta) \]
\[ = \int_0^r \frac{2\pi \lambda_d \rho \exp \left( -\lambda_d \rho \frac{r^2}{\pi} \right)}{1 - \exp (-\lambda_d \rho \frac{r^2}{\pi})} \times \exp \left( -\lambda_d \pi r^2 \left( p_{f,l}^1 + \theta \frac{r}{\pi} G_{\alpha_d} \left( \theta - \frac{r}{\pi} \right) \right) \right) \, dr \]
\[ = \int_0^r \frac{\pi \lambda_d \rho \exp \left( -\lambda_d \rho \frac{r^2}{\pi} \right)}{1 - \exp (-\lambda_d \rho \frac{r^2}{\pi})} \times \exp \left( -\lambda_d \pi r^2 \left( p_{f,l}^1 + \theta \frac{r}{\pi} G_{\alpha_d} \left( \theta - \frac{r}{\pi} \right) \right) \right) \, dr \] (35)
From (35), we obtain that when \( p_{f,l}^1 = 0 \), the denominator of \( p_{f,l}^1 (SIR_d \geq \theta) \) equals to zero. In order to avoid this, it is stipulated that, when \( p_{f,l}^1 = 0 \), \( P_{f,l}^1 (SIR_d \geq \theta) = 0 \) holds. This is consistent with the practical situation, since the successful transmission probability from the D2D helper equals to zero when there is no potential serving helper. As a result, we can obtain the successful transmission probability in Case 1.

We proceed with Case 2. The expression for \( P_{f,l}^2 (SIR_d \geq \theta) \) can be obtained by
\[ P_{f,l}^2 (SIR_d \geq \theta) = \int_0^r f_d(r)P_{f,l}^2 (SIR_d \geq \theta | r_0^d = r) \, dr \] (36)
where the interferences come from the D2D helpers that are geographically closer than the D2D helper \( d_0 \) and the helpers that are farther away than \( d_0 \). The interferences generated by the closer and farther SBSs are denoted by \( I_{d_2} \) and \( I_{d_3} \), respectively. According to the above analysis, \( P_{f,l}^2 (SIR_d \geq \theta | r_0^d = r) \) can be obtained by
\[ P_{f,l}^2 (SIR_d \geq \theta | r_0^d = r) \]
\[ = \mathbb{P}_{f,l} \left( |h_0|^2 \geq I_{d_2} + I_{d_3}, r^{\alpha_d} \right) \]
\[ = \mathbb{E}_{I_{d_2}, I_{d_3}} \left[ \exp \left( -\theta r^{\alpha_d} \right) \right] \]
\[ = \mathbb{E}_{I_{d_2}, I_{d_3}} \left[ \exp \left( -\theta r^{\alpha_d} \right) \right] \]
\[ = \mathcal{L}_{I_{d_2}}(\theta r^{\alpha_d}) \mathcal{L}_{I_{d_3}}(\theta r^{\alpha_d}) \] (37)
where \( \mathcal{L}_{I_{d_2}}(\theta r^{\alpha_d}) \) and \( \mathcal{L}_{I_{d_3}}(\theta r^{\alpha_d}) \) are the Laplace transforms regarding \( I_{d_2} \) and \( I_{d_3} \), respectively. The Laplace transforms \( \mathcal{L}_{I_{d_2}}(\theta r^{\alpha_d}) \) and \( \mathcal{L}_{I_{d_3}}(\theta r^{\alpha_d}) \) are given by
\[ \mathcal{L}_{I_{d_2}}(\theta r^{\alpha_d}) = \exp \left( -\pi \lambda_d \theta \frac{r}{\pi} r^2 \left( G_{\alpha_d}(0) - G_{\alpha_d} \left( \theta - \frac{r}{\pi} \right) \right) \right) \] (38)
\[ \mathcal{L}_{I_{d_3}}(\theta r^{\alpha_d}) = \exp \left( -\pi \lambda_d \theta \frac{r}{\pi} r^2 G_{\alpha_d} \left( \theta - \frac{r}{\pi} \right) \right). \] (39)
Based on (38) and (39), \( P_{f,l}^2 (SIR_d \geq \theta) \) can be obtained by
\[ P_{f,l}^2 (SIR_d \geq \theta) \]
\[ = \int_0^r \frac{2\pi \lambda_d \rho \exp \left( -\lambda_d \rho \frac{r^2}{\pi} \right)}{1 - \exp (-\lambda_d \rho \frac{r^2}{\pi})} \times \exp \left( -\lambda_d \pi r^2 \left( p_{f,l}^1 + \theta \frac{r}{\pi} G_{\alpha_d} \left( \theta - \frac{r}{\pi} \right) \right) \right) \, dr \]
\[ = \int_0^r \frac{\pi \lambda_d \rho \exp \left( -\lambda_d \rho \frac{r^2}{\pi} \right)}{1 - \exp (-\lambda_d \rho \frac{r^2}{\pi})} \times \exp \left( -\lambda_d \pi r^2 \left( p_{f,l}^1 + \theta \frac{r}{\pi} G_{\alpha_d} \left( \theta - \frac{r}{\pi} \right) \right) \right) \, dr \]
\[ = \int_0^r \frac{\pi \lambda_d \rho \exp \left( -\lambda_d \rho \frac{r^2}{\pi} \right)}{1 - \exp (-\lambda_d \rho \frac{r^2}{\pi})} \times \exp \left( -\lambda_d \pi r^2 \left( p_{f,l}^1 + \theta \frac{r}{\pi} G_{\alpha_d} \left( \theta - \frac{r}{\pi} \right) \right) \right) \, dr \] (35)
On the other hand, when $P^3_{f,j} = 0$, we set $P^3_{f,j} (\text{SIR}_d \geq \theta) = 0$.

**APPENDIX B**

**PROOF OF THEOREM 3**

When the user is served by the nearest MBS, the successful transmission probability $P (\text{SIR}_m \geq \theta)$ is calculated as

$$P (\text{SIR}_m \geq \theta) = \int_0^\infty f_m (r) P (\text{SIR}_m \geq \theta | r = r) \, dr$$

where

$$f_m (r) = 2\pi \lambda_m r \exp \left( -\lambda_m r^2 \right).$$

We denote the interference from other non-serving MBSs as $I_m = \sum_{k \in \Phi \setminus m} h_k^2 r^2 \lambda_k$. $P (\text{SIR}_m \geq \theta | r = r)$ is calculated as

$$P (\text{SIR}_m \geq \theta | r = r) = \exp \left( -\frac{\alpha_m r^2}{\lambda_m} G_{\alpha_m} \left( \frac{\theta}{\alpha_m} \right) \right).$$

By substituting (42) into (40), we can obtain

$$P (\text{SIR}_m \geq \theta) = 2\pi \lambda_m \int_0^\infty r \exp \left( -\frac{\lambda_m r^2}{2} G_{\alpha_m} \left( \theta \frac{r^2}{\lambda_m} \right) \right) \, dr$$

$$= \pi \lambda_m \int_0^\infty \exp \left( -\frac{\lambda_m r^2}{2} \left( 1 + \frac{\alpha_m}{G_{\alpha_m} \left( \theta \frac{r^2}{\lambda_m} \right) \left( \frac{\alpha_m}{\theta} \right) \right) \right) \, dr$$

$$= \left[ 1 + \frac{\alpha_m}{G_{\alpha_m} \left( \theta \frac{\alpha_m}{\theta} \right) \left( \frac{\alpha_m}{\theta} \right) \right].$$

As a result, the expression for $P (\text{SIR}_m \geq \theta)$ is obtained.
Wei Ni (Senior Member, IEEE) received the B.E. and Ph.D. degrees in electronic engineering from Fudan University, Shanghai, China, in 2000 and 2005, respectively.

He is currently a Group Leader and Principal Research Scientist with the Commonwealth Scientific and Industrial Research Organization, Sydney, NSW, Australia, and an Adjunct Professor with the University of Technology, Sydney, Australia and an Honorary Professor with Macquarie University, Sydney, Australia. Prior to this, he was a Postdoctoral Research Fellow with Shanghai Jiaotong University from 2005 to 2008; Deputy Project Manager with the Bell Labs, Alcatel/Alcatel-Lucent from 2005 to 2008; and Senior Researcher with Devices R&D, Nokia from 2008 to 2009. His research interests include signal processing, stochastic optimization, as well as their applications to network efficiency and integrity.

Dr. Ni has been the Chair of IEEE Vehicular Technology Society (VTS) New South Wales (NSW) Chapter since 2020 and an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS since 2018. He was first the Secretary and then Vice-Chair of IEEE NSW VTS Chapter from 2015 to 2019, Track Chair for VTC-Spring 2017, Track Co-Chair for IEEE VTC-Spring 2016, Publication Chair for BodyNet 2015, and Student Travel Grant Chair for WPMC 2014.

Xuewei Zhang received the Ph.D. degree in information and communication information engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2020.

She is currently an Associate Professor with the School of Communications and Information Engineering, Xi’an University of Posts and Telecommunications, Xi’an, China. Her research interests include wireless edge caching, resource allocation, and heterogeneous networks.

Tiejun Lv (Senior Member, IEEE) received the M.S. and Ph.D. degrees in electronic engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 1997 and 2000, respectively.

From January 2001 to January 2003, he was a Postdoctoral Fellow with Tsinghua University, Beijing, China. In 2005, he was promoted as a Full Professor with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China. From September 2008 to March 2009, he was a Visiting Professor with the Department of Electrical Engineering, Stanford University, Stanford, CA, USA. He is the author of three books, more than 80 published IEEE journal papers, and 180 conference papers on the physical layer of wireless mobile communications. His research interests include signal processing, communications theory, and networking.

Dr. Lv was the recipient of the Program for New Century Excellent Talents in University Award from the Ministry of Education, China, in 2006. He was also the recipient of the Nature Science Award in the Ministry of Education of China for the hierarchical cooperative communication theory and technologies in 2015.

Norman C. Beaulieu (Life Fellow, IEEE) received the B.A.Sc. (honors), M.A.Sc., and Ph.D. degrees from the University of British Columbia, Vancouver, BC, Canada, in 1980, 1983, and 1986, respectively, all in electrical engineering.

He is an Academician of the Royal Society of Canada, and an Academician of the Canadian Academy of Engineering.

Prof. Beaulieu is a Fellow of the Institution of Engineering and Technology of the U.K., a Fellow of the Engineering Institute of Canada (EIC), and a Nicola Copernicus Fellow of Italy. He was the recipient of the esteemed Natural Sciences and Engineering Research Council of Canada E.W.R. Steacy Memorial Fellowship. He is the only person in the world to hold both the IEEE Edwin Howard Armstrong Award and the IEEE Reginald Aubrey Fessenden Award, named for the inventors of Frequency Modulation or FM, and Amplitude Modulation or AM, respectively. He is a Beijing University of Posts and Telecommunications Thousand-Talents Scholar. He holds the third highest Web of Science ISI h-index in the world in the combined areas of communication theory and information theory. He was the recipient of the title “State Especially Recruited Foreign Expert” certified upon him by Minister of Human Resources and Social Insurance, and Vice Minister of the Organization Department, Y. Weimin. He was also the recipient of the Royal Society of Canada Thomas W. Edie Medal, the Mdaillle K.Y. Lo Medal of the EIC, and was the subject of a TIME Magazine feature article. He was also awarded the unique Special University Prize in Applied Science of the University of British Columbia, and the J. Gordin Kaplan Award for Research of the University of Alberta.