

A Novel Content Caching and Delivery Scheme for Millimeter Wave Device-to-Device Communications

Theshani Nuradha*, Tharaka Samarasinghe[†], Kasun T. Hemachandra*

*Department of Electronic and Telecommunication Engineering, University of Moratuwa, Moratuwa, Sri Lanka

[†]Department of Electrical and Electronic Engineering, University of Melbourne, Victoria, Australia

Email: theshanin@uom.lk, tharakas@uom.lk, kasunh@uom.lk

Abstract—A novel content caching strategy is proposed for a cache enabled device-to-device (D2D) network where the user devices are allowed to communicate using millimeter wave (mmWave) D2D links (> 6 GHz) as well as conventional sub 6 GHz cellular links. The proposed content placement strategy maximizes the successful content delivery probability of a line of sight D2D link. Furthermore, a heuristic algorithm is proposed for efficient content delivery. The overall scheme improves the successful traffic offloading gain of the network compared to conventional cache-hit maximizing content placement and delivery strategies. Significant energy efficiency improvements can also be achieved in ultra-dense networks.

Index Terms—content caching, device-to-device communications, millimeter wave, ultra-dense networks

I. INTRODUCTION

Small cells and device-to-device (D2D) communications are envisioned to be promising technologies for enhancing the quality of service (QoS), throughput and energy efficiency of next generation wireless networks [1], [2]. Due to scarcity in existing cellular spectrum, millimeter wave (mmWave) frequencies have been considered as an enabling technology for high speed D2D communications. In D2D aided cellular networks, the successful establishment of D2D connections depends on the availability of popular files in proximity devices. Therefore, content placement in user devices is of paramount importance for D2D based traffic offloading. This paper presents a novel content caching strategy for a cache enabled D2D network, where the user devices are allowed to communicate using mmWave D2D links as well as conventional sub 6 GHz cellular links.

Cache placement schemes based on cache hit probability maximization [3], and cache aided throughput maximization [4], can be found in the literature for conventional cellular networks. In these works, optimal caching probabilities are obtained such that the achieved content diversity leads to better network performance. When it comes to mmWave networks, a D2D aware caching policy that splits the most popular content into two content groups, and randomly distributes the content groups among the users is proposed in [5]. The partitioning of the most popular content is performed based on fairness considerations, such that the two content groups have equal self cache hit probability in a user device. The content placement does not consider the characteristics of mmWave

propagation and the effects of blockage when placing content. A cache hit probability maximization based optimal cache placement in a mmWave ad-hoc network is studied in [6]. The paper omits the effect of interference from other D2D links, which can be crucial factor on network performance.

In this paper, we consider a cloud radio access network (C-RAN) operating in the sub 6 GHz band, and supports D2D communications using mmWave spectrum. A content placement scheme is proposed for user device caching, considering the propagation characteristics of mmWave links and the interference from other D2D links, which makes it different to [5] and [6]. In addition, it also considers the popularity and application specific QoS constraints of different files, which makes it more applicable to next generation wireless networks, where QoS measures such as latency are considered to be key performance indicators. Thus, we maximize a different metric, referred to as the successful content reception probability, which is the probability of reception without violating the file specific QoS constraints. The main contributions of the paper are as follows:

- A content placement scheme is proposed for user devices by solving an optimization problem, that maximizes the successful content delivery probability within the line of sight (LoS) region of a D2D transmitter. Optimal caching probabilities of a multitude of heterogeneous files, that have their own rate constraints for successful reception, are obtained.
- The cache placement scheme is coupled with a user association scheme to further improve the offloading gain without violating QoS constraints.
- The performance of the network in terms of successful delivery of content and offloading gain is evaluated both analytically and through simulations, to clearly highlight the gains of the proposed content placement and the user association schemes with respect to energy efficiency as well.

The overall scheme improves the successful traffic offloading gain of the network compared to conventional cache-hit maximizing content placement and delivery strategies. Significant energy efficiency improvements can also be achieved in ultra-dense networks.

The paper organization is as follows. Section II presents the system model and the problem formulation. The solution to the optimization problem which leads to the content placement

This work is supported by the Senate Research Council, University of Moratuwa, Sri Lanka, under grant SRC/LT/2018/2.

strategy, and the user association scheme are presented in Section III. The performance of the proposed schemes are evaluated theoretically and numerically in Section IV and Section V, respectively. Section VI concludes the paper.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. Topological Model

We consider a C-RAN, where the remote radio heads (RRHs) are spatially distributed according to a homogeneous Poisson point process (HPPP) Φ_R with intensity λ_R . The RRHs are wirelessly connected to the edge cloud (fronthaul links) and the edge clouds are connected to the core network (backhaul links). The spatial distribution of the mobile user (MU) devices is modelled using an independent HPPP Φ_u with intensity λ_u . Content to a MU may either be delivered from a RRH through a cellular link that operates in the sub 6 GHz band (transmit power P_c , wavelength W_c , bandwidth B_c) or from another MU through a mmWave D2D link (transmit power P_d , wavelength W_d , bandwidth B_d). Content delivery through a D2D link will only be possible if the requested file is in the cache of another proximity MU. Each MU is capable of caching M_d files of equal size.

RRHs are equipped with omni-directional transmitting antennas while the MUs are equipped with omni-directional antennas for cellular communications and directional antennas for mmWave communications. Similar to [5], sectorized antenna pattern is adopted at the transmitters to approximate the antenna pattern for the mmWave links. It is assumed that the antennas of the transmitter and the receiver are perfectly aligned for desired links while the interfering transmitter antenna bore-sight is uniformly distributed over $[0, 2\pi]$. This means the probability distribution of the i.i.d. random antenna gain G' associated an interfering mmWave link is given by $\Pr\{G' = G_m^2\} = \left(\frac{\Delta\theta}{2\pi}\right)^2$, $\Pr\{G' = G_s^2\} = \left(\frac{2\pi - \Delta\theta}{2\pi}\right)^2$, and $\Pr\{G' = G_m G_s\} = \frac{2\Delta\theta(2\pi - \Delta\theta)}{(2\pi)^2}$, where G_m and G_s denote the main and side lobe gains, respectively, and $\Delta\theta$ denotes the angle of deviation from the antenna bore-sight.

We refer to the MUs that request content as active MUs. At a given time, the probability of an MU being active is ρ . The remaining MUs, which we refer to as inactive MUs, can serve as potential D2D transmitters. It is assumed that all RRHs are active at a given time. Without loss of generality, we consider a typical MU located at the origin for our analysis.

B. Channel Model

For the cellular links that operate below 6 GHz, the simple path loss model with a path loss exponent α_c is used to model the location dependent path loss. For mmWave links, the average LoS ball model is used [7], [8]. According to this model, a link is considered as a LoS link if the link is shorter than D_L . Otherwise, it is considered a non-line of sight (NLoS) link. The average size and the density of the blockages determine D_L [9]. For mmWave links, blockage effects induce different path loss exponents α_L and α_N for LoS and NLoS links, respectively, with $\alpha_L < \alpha_N$. We assume fast Rayleigh fading where the fading power is exponentially distributed with unit

mean [5]. The received signal-to-interference-plus-noise-ratio (SINR) when receiving file i from the D2D transmitter located at $x \in \Phi_d$ is given by

$$\text{SINR}_{d,i,x} = \frac{G_m^2 h_x r_x^{-\alpha_d}}{\hat{N} + \sum_{y \in \Phi_d \setminus \{x\}} h_y G' r_y^{-\alpha_d}}, \quad (1)$$

where Φ_d , h_x and r_x denote the point process of the active D2D transmitters, fading power and the distance between the MU and the D2D transmitter at x , respectively, $\alpha_d \in \{\alpha_L, \alpha_N\}$, $\hat{N} = \frac{16\pi^2 N_o F_N B_d}{P_d W_d^2}$, N_o is the noise power spectral density and F_N is the noise figure of the receiver. Similarly, an expression for the received SINR when receiving file i through a cellular link from the RRH $x \in \Phi_R$, which we denote by $\text{SINR}_{c,i,x}$, can be obtained by replacing subscript d in (1) with subscript c and by replacing both antenna gains G_m^2 and G' by $G_T G_R$, where G_T and G_R are the antenna gains of the transmitting RRH and the receiving MU, respectively.

C. Content Placement

We assume that MUs request content from a finite content library of N files of equal size, and the file requests follow a Zipf distribution of popularity exponent ϵ . Thus, the probability of requesting the i -th most popular file is given by

$$\beta_i = \frac{i^{-\epsilon}}{\sum_{j=1}^N j^{-\epsilon}}.$$

The rate and the delay constraints of the files may vary with the file type and the associated application. The rate constraint for the i -th most popular file is denoted by R_i , and this constraint necessitates an SINR greater than $T_i = 2^{\frac{R_i}{B}} - 1$, for a link having a bandwidth of $B \in \{B_d, B_c\}$.

Due to the limited storage capacity of the MUs, the propagation characteristics of mmWaves, and the rate requirements of different files, caching content at MUs should be done in an efficient manner. In this paper, we focus on designing a content placement scheme that offloads the traffic to the D2D devices without violating the rate (QoS) constraints, which are considered to be crucial in next generation networks. Moreover, considering the effects of blockage, it is preferred to have LoS D2D links. Thus, we focus on the successful LoS reception probability (SLP) for the i -th most popular file, which we define as

$$\text{SLP}_i = \Pr\{\text{SINR}_{d,i,\hat{x}} \geq T_i \cap r_{\hat{x}} \leq D_L\}, \quad (2)$$

where \hat{x} denotes the location of the closest D2D transmitter who has the i -th file in its cache. The SLP depicts the probability of receiving content from the nearest LoS D2D transmitter without violating the rate threshold.

Let the probability of the i -th file being stored in the cache of an MU be q_i . We define \mathbf{q} , an N -dimensional caching probability vector $\mathbf{q} = [q_1, \dots, q_N]$, which denotes the probabilities of an MU caching the N files in the content library. We focus on finding \mathbf{q} that maximizes the average successful LoS reception probability (ASLP), defined as

$$\text{ASLP}(\mathbf{q}) = \sum_{i=1}^N \beta_i \text{SLP}_i,$$

where the SLP_i values are averaged over the request probabilities, while not violating the MU storage constraints statistically (on average). Note that our objective function captures both popularity and the QoS requirements of different content, and also the effects of wireless propagation. The optimization problem can be formulated as

$$\begin{aligned} & \underset{\mathbf{q}}{\text{maximize}} \quad \text{ASLP}(\mathbf{q}) = \sum_{i=1}^N \beta_i \text{SLP}_i \\ & \text{subject to} \quad 0 \leq q_i \leq 1, \quad i = 1, \dots, N, \\ & \quad \quad \quad \sum_{i=1}^N q_i \leq M_d. \end{aligned} \quad (3)$$

III. SYSTEM DESIGN

A. Successful LoS Reception Probability

An expression for SLP_i can be obtained using fundamentals of stochastic geometry [10]. That is, from (2),

$$\begin{aligned} \text{SLP}_i &= \Pr \left\{ h_{\hat{x}} > T_i r_{\hat{x}}^{\alpha_L} \left(\hat{I} + \hat{N} \right) \mid r_{\hat{x}} \leq D_L \right\} \Pr \{ r_{\hat{x}} \leq D_L \} \\ &= \int_0^{D_L} L_{\hat{I}}(T_i x^{\alpha_L}) e^{-\hat{N} T_i x^{\alpha_L}} f_{r_{\hat{x}}}(x) dx, \end{aligned}$$

where $\hat{I} = \sum_{y \in \Phi_d \setminus \{\hat{x}\}} h_y G' r_y^{-\alpha_d}$ and $L_{\hat{I}}(S)$ is the Laplace transform of the interference from mmWave D2D links. The Laplace transform of the D2D interference is given by

$$\begin{aligned} L_{\hat{I}}(S) &= \mathbb{E} \left[\prod_{y \in \Phi_d \setminus \{\hat{x}\}} \mathbb{E}_{G', h_y} \left[e^{-\frac{G' h_y r_y^{-\alpha_d} S}{G_m^2}} \right] \right] \\ &\stackrel{(a)}{=} \exp \left(-\rho \lambda_u p_d \int_0^{2\pi} \int_0^\infty \left(1 - \mathbb{E}_{G', h_y} \left[e^{-\frac{G' h_y z^{-\alpha_d} S}{G_m^2}} \right] \right) z dz d\phi \right) \\ &\stackrel{(b)}{=} \exp \left[-2\pi \rho \lambda_u p_d \mathbb{E}_{G'} \left[\int_0^{D_L} \left(1 - \frac{G_m^2}{1 + G' z^{-\alpha_L} S} \right) z dz \right. \right. \\ &\quad \left. \left. + \int_{D_L}^\infty \left(1 - \frac{G_m^2}{1 + G' z^{-\alpha_N} S} \right) z dz \right] \right], \end{aligned}$$

where (a) follows from the probability generating functional (PGFL) of the PPP, p_d is the probability of receiving the requested content via a D2D link, which has to be separately calculated, as shown later Section IV, and the expectation in (b) can be evaluated by using the PDF of G' given in Section II, which would result in a product of three exponential functions.

It is not hard to see that the resulting expression, which has multiple integrals, makes it prohibitively hard for us to use it in a meaningful manner in the optimization problem. We have an N -dimensional non-convex constrained optimization problem, and even obtaining the optimum solution numerically is not trivial. Hence, we make few approximations to obtain a mathematically tractable expression for SLP_i , *i.e.*, we obtain a convex approximation of the objective function such that we can solve the optimization problem in closed form. We note that these approximations are made only to design the content placement policy, and all assumptions are relaxed in the remainder of the paper, which includes the performance

evaluation and the numerical evaluations in Sections IV and V, respectively.

Firstly, we neglect small-scale fading with regards to mmWave propagation since it causes only minor changes in received power when the transmitter is within the LoS region [11], [12]. Secondly, we consider the worst case of D2D interference where all the user requests are catered by D2D transmitters, and approximate the random interference using the average worst case interference. To overcome the singularity when computing the D2D interference averaged only over the large-scale fading, we use the bounded path loss model $g(r) = \min(1, r^{-\alpha_d})$ to model the path loss from the interferers, similar to [13]. With these approximations, and by using Campbell's theorem, the average interference can be written as

$$\begin{aligned} \bar{I} &= \mathbb{E} \left[\sum_{y \in \Phi_d \setminus \{x\}} G' \min(1, r_y^{-\alpha_d}) \right] \\ &= \frac{\lambda_u \rho}{4\pi} (G_m \Delta\theta + G_s (2\pi - \Delta\theta))^2 \\ &\quad \times \left(\frac{\alpha_L - 2D_L^{2-\alpha_L}}{\alpha_L - 2} + \frac{2D_L^{2-\alpha_N}}{\alpha_N - 2} \right). \end{aligned}$$

From (2), and by considering the maximum search discovery distance to initiate a D2D communication link to be D_R , we have

$$\text{SLP}_i = \Pr \left\{ r_{\hat{x}} \leq \min(\hat{D}_i, D_L, D_R) \right\},$$

where

$$\hat{D}_i = \left[\frac{G_m^2}{T_i \bar{I} + T_i \hat{N}} \right]^{\alpha_L}.$$

On the assumption that the transmitters having the i -th content stored in their cache form a PPP of intensity $\lambda_u (1 - \rho) q_i$, and by using the distribution of the distance to the nearest MU in a PPP, we have

$$\text{SLP}_i = 1 - \exp(-\pi \lambda_u (1 - \rho) q_i D_{i,c}^2), \quad (4)$$

where $D_{i,c} = \min\{\hat{D}_i, D_L, D_R\}$.

B. Optimum Content Placement

Once the approximations are applied, it is straightforward to see that the optimization problem becomes convex. The Lagrangian can be written as

$$\begin{aligned} \mathcal{L}(\mathbf{q}, \mu) &= -1 + \sum_{i=1}^N \beta_i \exp(-\pi \lambda_u (1 - \rho) q_i D_{i,c}^2) \\ &\quad + \mu \left(\sum_{i=1}^N q_i - M_d \right), \end{aligned}$$

where μ is the non-negative Lagrangian multiplier. By applying Karush-Kuhn-Tucker (KKT) conditions, we have

$$q_i(\mu) = -\frac{\ln[\mu / (\beta_i \pi \lambda_u (1 - \rho) D_{i,c}^2)]}{[\lambda_u (1 - \rho) D_{i,c}^2]},$$

and according to the first inequality constraint, the optimum q_i^* should satisfy $q_i^* = \min \{\max \{q_i(\mu^*), 0\}, 1\}$, and according to the second constraint, which is met with equality, gives us $\sum_{i=1}^N \min \{\max \{q_i(\mu^*), 0\}, 1\} = M_d$. This can be used to find μ^* through a simple root finding algorithm such as bisection search [3], [4].

C. User Association

In a hybrid (multi-tier) wireless network, it is important to associate users with the appropriate tiers to achieve traffic offloading while maintaining QoS constraints. In our system model, an MU may receive content via a LoS D2D link, a NLoS D2D link or a cellular link. Therefore, it is important to recognize the appropriate method of content delivery. From (4), one can see that an MU can successfully receive the i -th file from a D2D transmitter in the LoS region if the distance to the D2D transmitter is less than $D_{i,c}$. However, we have assumed the worst case D2D interference when obtaining $D_{i,c}$. This means, it may be possible to increase this threshold further without violating the rate constraints, which will facilitate more offloading. Since q is now defined, the probability of receiving the requested content from a D2D transmitter within the LoS region can be calculated as $\gamma = \sum_{i=1}^N \beta_i (1 - q_i^*) (1 - \exp(-\pi \lambda_u (1 - \rho) q_i^* D_L^2))$, where we have used the fact that the D2D mode initiates when the required content is not found in self-cache. We can use γ to scale the worst-case average interference to have a tighter approximation of the interference from the active D2D transmitters. Hence, assuming that the interference is dominated by the D2D transmitters in the LoS region, we can obtain an updated distance threshold value that satisfies the QoS requirement as

$$\hat{D}_{i,L} = \left[\frac{G_m^2}{T_i \gamma \bar{I} + T_i \hat{N}} \right]^{1/\alpha_L}.$$

By replacing subscript L with N , we can obtain a similar distance threshold $\hat{D}_{i,N}$ for a transmitter in the NLoS region.

The proposed user association scheme can be summarized as follows. For the i -th file, the MU first checks its own cache. If not found, it checks with D2D transmitters who are closer than

$$D_{i,u} = \min \left\{ \hat{D}_{i,L}, \max \left\{ \hat{D}_{i,N}, D_L \right\}, D_R \right\},$$

for a D2D connection. If both actions fail, the MU connects to the nearest RRH through a cellular link. The process is summarized in Algorithm 1.

The rationale behind the distance thresholds can be explained as follows. When $\hat{D}_{i,L} < D_L$, it is straightforward that the threshold is $\hat{D}_{i,L}$, as any transmitter outside this (both LoS and NLoS) will not satisfy the rate constraints. When $\hat{D}_{i,L} > D_L$, we particularly focus on transmitters between D_L and $\hat{D}_{i,L}$, who are NLoS according to the channel model. Whether these NLoS transmitters can transmit successfully or not will depend on the value of $\hat{D}_{i,N}$. To this end, if $\hat{D}_{i,N} < D_L$, none of the NLoS transmitters will be able to transmit successfully. Thus, we set the threshold as D_L .

Algorithm 1 User Association Scheme

```

1:  $f_i \leftarrow$  Requested file
2:  $A_L \leftarrow$  circular region with the radius  $\min \left\{ \hat{D}_{i,L}, D_R \right\}$ 
3:  $A_N \leftarrow$  circular region with the radius  $\min \left\{ \hat{D}_{i,N}, D_R \right\}$ 
4: if  $f_i$  in the device self cache then
5:   Get file from self cache
6: else if  $\hat{D}_{i,L} \leq D_L$  AND  $f_i$  in  $A_L$  then
7:   Get file from the closest LoS D2D transmitter
8: else if  $\hat{D}_{i,N} > D_L$  AND  $f_i$  in  $A_N$  then
9:   Get File from the closest NLoS D2D transmitter
10: else
11:   if  $f_i$  in Edge cloud then
12:     Get file from the edge cloud (fronthaul link)
13:   else
14:     Get file from the core network (backhaul link)
15:   end if
16: end if

```

However, when $D_L \leq \hat{D}_{i,N} \leq \hat{D}_{i,L}$, all NLoS transmitters between D_L and $\hat{D}_{i,N}$ will satisfy the rate constraints, thus we pick $\hat{D}_{i,N}$ as the threshold. Hence, overall, the distance threshold is given by $\max \left\{ \hat{D}_{i,N}, D_L \right\}$. The user association policy tries to make use of candidate NLoS transmitters as well, to further facilitate offloading. Note that we will have $\hat{D}_{i,L} > \hat{D}_{i,N}$ for all meaningful link lengths as $\alpha_N > \alpha_L$.

IV. PERFORMANCE ANALYSIS

Having placed content, and have decided on the user association policy, a performance analysis of the network presented in Section II will be carried out in this section. Note that the assumptions made in Section III are relaxed in this analysis since the assumptions were made only to obtain a mathematically tractable objective function.

An offloading event occurs when a content request is served by self cache or via a D2D link. To this end, the probability of finding the required content in the device cache itself is given by $p_s = \sum_{i=1}^N \beta_i q_i^*$. The probability of receiving the requested content via a D2D link is given by

$$p_d = \sum_{i=1}^N \beta_i (1 - q_i^*) (1 - \exp(-\pi \lambda_u (1 - \rho) q_i^* D_{i,u}^2)). \quad (5)$$

Thus, $p_d + p_s$ gives us the offloading probability.

We define the successful reception probability (SP) as the probability of an MU receiving content without violating the rate constraints. The SP through a D2D link is given by $\sum_{i=1}^N \beta_i (1 - q_i^*) \Pr \{ \text{SINR}_{d,i,\hat{x}} \geq T_i \cap r_{\hat{x}} \leq D_{i,u} \}$, and similarly, the SP through the cellular network is given by $\sum_{i=1}^N \beta_i (1 - q_i^*) e^{-\pi \lambda_u (1 - \rho) q_i^* D_{i,u}^2} \Pr \{ \text{SINR}_{c,i,x} \geq T_i \}$. The sum of these two probabilities and p_s gives us SP.

An expression for $\Pr \{ \text{SINR}_{c,i,x} \geq T_i \}$ can be obtained using the fundamentals of stochastic geometry, by following a

similar approach to the one shown in Section III. Considering the closest RRH to be located at $x \in \Phi_R$,

$$\begin{aligned} \Pr\{\text{SINR}_{c,i,x} \geq T_i\} &= \Pr\{h_x > T_i r_x^{\alpha_c} (\hat{I}_c + \hat{N}_c)\} \\ &= \int_0^\infty L_{\hat{I}_c}(T_i z^{\alpha_c}) e^{-\hat{N}_c T_i z^{\alpha_c}} f_{r_x}(z) dz, \end{aligned}$$

where $\hat{I}_c = \sum_{y \in \Phi_R \setminus \{x\}} h_y r_y^{-\alpha_c}$, $\hat{N}_c = \frac{16\pi^2 N_o F_N B_c}{G_T G_R P_c W_c^2}$, and f_{r_x} is the PDF of the distance to the nearest RRH, given by $f_{r_x}(z) = 2\pi\lambda_R z e^{-\pi\lambda_R z^2}$, for $z \geq 0$. Furthermore, we have

$$\begin{aligned} L_{\hat{I}_c}(S) &= \mathbb{E} \left[\prod_{y \in \Phi_R \setminus \{x\}} \mathbb{E}_{h_y} \left[e^{-h_y r_y^{-\alpha_c} S} \right] \right] \\ &= \exp \left(-2\pi\lambda_R \int_x^\infty \left(1 - \frac{1}{1 + v^{-\alpha_c} S} \right) v dv \right). \end{aligned}$$

An expression for $\Pr\{\text{SINR}_{d,i,\hat{x}} \geq T_i \cap r_{\hat{x}} \leq D_{i,u}\}$ can be obtained along similar lines, and by appropriately changing the distance limits in the integration. To this end, we get

$$\begin{aligned} \Pr\{\text{SINR}_{d,i,\hat{x}} \geq T_i \cap r_{\hat{x}} \leq D_{i,u}\} &= \int_0^{D1} L_{\hat{I}}(T_i z^{\alpha_L}) e^{-\hat{N} T_i z^{\alpha_L} / G_m^2} f_{r_{\hat{x}}}(z) dz \\ &+ \int_{D_L}^{D2} L_{\hat{I}}(T_i z^{\alpha_N}) e^{-\hat{N} T_i z^{\alpha_N} / G_m^2} f_{r_{\hat{x}}}(z) dz \end{aligned}$$

where $f_{r_{\hat{x}}}$ is the PDF of the distance to the nearest MU in a PPP of intensity $\lambda_u (1 - \rho) q_i^*$, $D1 = \min\{D_L, \hat{D}_{i,L}, D_R\}$ and $D2 = \max\{D_L, \min(\hat{D}_{i,N}, D_R)\}$. Moreover,

$$L_{\hat{I}}(S) = \exp \left(-2\pi\rho\lambda_u p_d \mathbb{E}_{G'} \left[\int_0^\infty \left(1 - \frac{G_m^2}{1 + G' S y^{-\alpha(y)}} \right) y dy \right] \right),$$

where $\alpha(y) = [1 - (1 - \alpha_L) \mathbf{1}_{\{y \leq D_L\}}] [1 - (1 - \alpha_N) \mathbf{1}_{\{y > D_L\}}]$, the expectation can be straightforwardly evaluated using the PDF of G' given in Section III, and p_d is given by (5).

V. NUMERICAL RESULTS AND DISCUSSIONS

In this section, the performance of the proposed scheme is evaluated using simulations. The simulation parameters are set to align with previous works [3], [5], [9], [12] and presented in Table I. We consider $R_i = R \forall i \in \{1, \dots, N\}$ for simplicity. We compare the performance of the proposed system (S-1) with the system proposed in [3] (S-2), where the content is placed to maximize the cache hit probability and the content delivery is based on D2D links within a radius of D_R , which is the maximum discovery distance of an MU.

Fig. 1 compares the SP performance of the overall systems, and S-1 outperforms S-2 in all considered scenarios. When D_L decreases (blockage density increases), the SP reduces in both systems. However, we can observe the performance gap between S-1 and S-2 increasing. Since S-1 has prioritized the LoS region for both content placement and delivery, it is more robust to changes in D_L . On the other hand, S-2, that has focused on D_R (which is generally larger than D_L), may encounter frequent unsuccessful D2D transmissions when

TABLE I
SIMULATION PARAMETERS

RRH density λ_R	10/km ²
User requesting probability ρ	0.5
Path loss exponent ($f_c = 1$ GHz) α_c	2.5
Path loss exponent ($f_c = 28$ GHz) α_L, α_N	2.1, 4
Transmit power of a RRH P_c	100 mW
Power consumption for backhaul P_b	1 W
Transmit power of a user device P_d	2 mW
Main lobe gain of the user antenna G_m	9dB
Side lobe gain of the user antenna G_s	-9dB
Noise power density N_o	-178 dB/Hz
Noise figure F_N	10 dB
Content Library size N	100
Edge cloud cache capacity M_e	50
Device cache capacity M_d	2
Bandwidth in cellular link B_c	20 MHz
Bandwidth in mmWave link B_d	1 GHz
Maximum search discovery distance of a device D_R	150 m

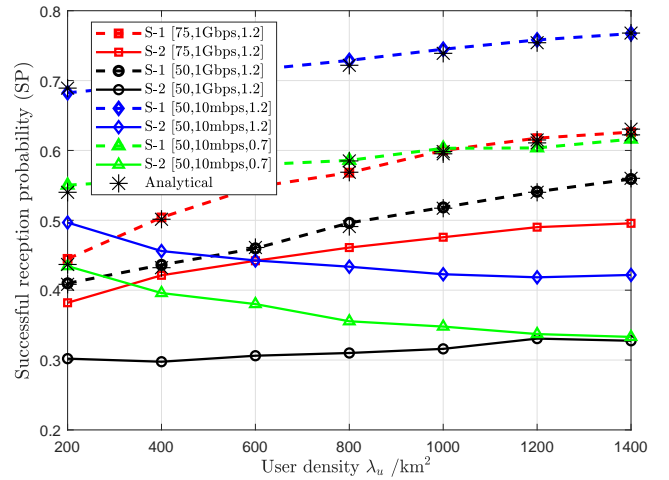


Fig. 1. The behavior of the overall SP of the system with λ_u for different $[D_L, R, \epsilon]$ combinations.

the blockage density increases. When both rate constraints and D_L reduce, the SP of S-2 decreases with λ_u . This is due to the increased interference from D2D links and the QoS requirements not being satisfied with S-2. However, since S-1 considers the effect of interference in both the content placement and delivery (user association), the SP increases with λ_u , making S-1 a promising approach for future ultra-dense networks. When ϵ is reduced while keeping the other parameters fixed, the overall success of both the systems reduce. The reduction in ϵ leads to the content requests spreading out over a large range of files, which in fact reduces the probability of successfully receiving a file over a D2D link. Since both systems have averaged the objective functions over all possible files, a similar trend is observed for all values of ϵ . Fig. 2 compares the offloading probability (OP_d) of the two systems. S-2 having a higher offloading probability is rather obvious since considering D_R leads to a larger offloading region with S-2. However, the figure also conveys that a considerable portion of the offloaded traffic in S-2 will not be successfully delivered, and hence, the successful offloading probability (SOP_d) of S-2 is lower than S-1. When D_L

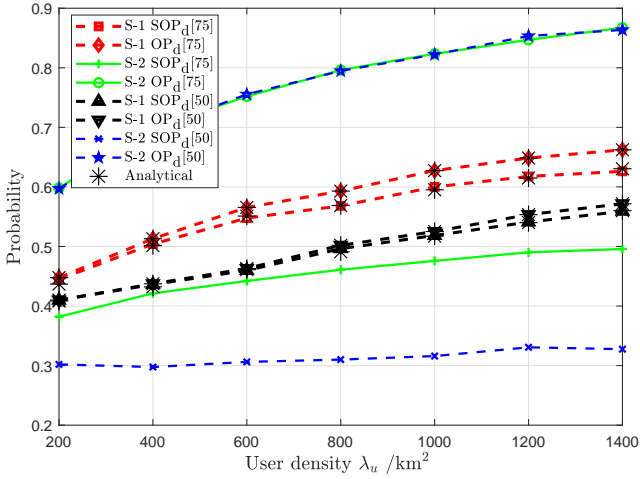


Fig. 2. Offloading and the successful offloading probabilities of the D2D network for 1 Gbps data rate, where $\epsilon = 1.2$.

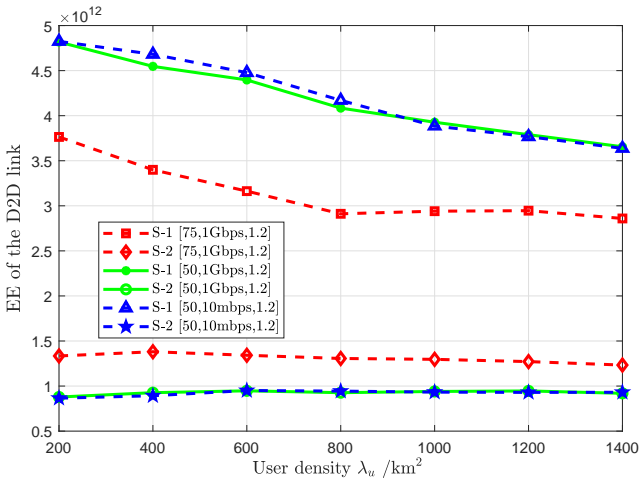


Fig. 3. The Energy Efficiency of the D2D link with λ_u for different $[D_L, R, \epsilon]$ combinations.

reduces, the unsuccessful offloading of S-2 increases, but in S-1, the gap between the two offloading probabilities remains low as it is more robust to changes in D_L , as described with respect to Fig. 1.

Energy efficiency (EE) can be computed by the ratio between the average throughput from successful transmissions and the average power consumption, per user request. Some insights on the EE of the two systems can be inferred from Fig. 2. When λ_u increases, the probability of initiating a D2D link increases in both schemes. This leads to the average throughput from successful transmissions increasing, and the average power consumption decreasing. However, the average throughput from successful transmissions of S-1 increases at a higher rate with λ_u compared to S-2. On the other hand, the average power consumption of S-2 decreases at a higher rate with λ_u compared to S-1 due to the higher offloading. Therefore, the overall EE of the two systems are almost similar. However, one can identify that the EE of D2D links is significantly higher for S-1, compared to S-2, which is illustrated in Fig. 3. This is due to the unsuccessful D2D trans-

missions in S-2 resulting in device energy wastage. When λ_u is varied from 200 to 1400 /km², for $D_L = 75\text{m}$ and $\epsilon = 1.2$, on average, a 1.3 fold improvement in terms of EE is observed with S-1 compared to S-2. This improves significantly when D_L is further reduced. For example, on average, a 3.5 fold improvement can be observed at $D_L = 50\text{m}$. This depicts the superior performance of S-1 in ultra-dense networks with high blockage.

VI. CONCLUSIONS

The performance of a cache enabled D2D network, where D2D communications occur exclusively in the mmWave band, has been studied using a stochastic geometric framework. As a result, a novel content caching scheme in user devices to maximize the successful content delivery probability of LoS D2D links has been introduced. The performance gains of the proposed content placement and delivery schemes have been highlighted through simulations. The numerical results have shown that the proposed scheme achieves higher successful content offloading, improved energy efficiency while satisfying QoS requirements of the users, and superior performance in ultra-dense networks with high blockage.

REFERENCES

- [1] F. Jameel, Z. Hamid, F. Jabeen, S. Zeadally, and M. A. Javed, "A survey of device-to-device communications: Research issues and challenges," *IEEE Commun. Surv. Tut.*, vol. 20, pp. 2133–2168, Apr. 2018.
- [2] R. I. Ansari, C. Chrysostomou, S. A. Hassan, M. Guizani, S. Mumtaz, J. Rodriguez, and J. J. P. C. Rodrigues, "5G D2D networks: Techniques, challenges, and future prospects," *IEEE Syst. J.*, vol. 12, pp. 3970–3984, Dec. 2018.
- [3] Z. Chen and M. Kountouris, "D2D caching vs. small cell caching: Where to cache content in a wireless network?," in *Proc. IEEE International Workshop on Signal Processing Advances in Wireless Communications*, pp. 1–6, Jul. 2016.
- [4] Z. Chen, N. Pappas, and M. Kountouris, "Probabilistic caching in wireless D2D networks: Cache hit optimal versus throughput optimal," *IEEE Commun. Letters*, vol. 21, pp. 584–587, Mar. 2017.
- [5] N. Giatzoglou, K. Ntontin, E. Kartsakli, A. Antonopoulos, and C. Verikoukis, "D2D-aware device caching in mmWave-cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 35, pp. 2025–2037, Sep. 2017.
- [6] S. Vuppala, T. X. Vu, S. Gautam, S. Chatzinotas, and B. Ottersten, "Cache-aided millimeter wave ad-hoc networks," in *Proc. IEEE Wireless Communications and Networking Conference*, pp. 1–6, Apr. 2018.
- [7] T. Bai and R. W. Heath, "Coverage and rate analysis for millimeter-wave cellular networks," *IEEE Trans. Wireless Commun.*, vol. 14, pp. 1100–1114, Feb. 2015.
- [8] Y. Zhu, L. Wang, K. Wong, and R. W. Heath, "Secure communications in millimeter wave ad hoc networks," *IEEE Trans. Wireless Commun.*, vol. 16, pp. 3205–3217, May 2017.
- [9] J. G. Andrews, T. Bai, M. N. Kulkarni, A. Alkhateeb, A. K. Gupta, and R. W. Heath, "Modeling and analyzing millimeter wave cellular systems," *IEEE Trans. Commun.*, vol. 65, pp. 403–430, Jan. 2017.
- [10] M. Haenggi, J. G. Andrews, F. Baccelli, A. Dousse, and M. Franceschetti, "Stochastic geometry and random graphs for the analysis and design of wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 27, pp. 1029–1046, Sep. 2009.
- [11] F. Wang, H. Wang, H. Feng, and X. Xu, "A hybrid communication model of millimeter wave and microwave in D2D network," in *Proc. IEEE Vehicular Technology Conference*, pp. 1–5, May 2016.
- [12] Y. Zhu, G. Zheng, L. Wang, K. Wong, and L. Zhao, "Performance analysis and optimization of cache-enabled small cell networks," in *Proc. IEEE Global Telecommunications Conference*, pp. 1–6, Dec. 2017.
- [13] N. Deng, M. Haenggi, and Y. Sun, "Millimeter-wave device-to-device networks with heterogeneous antenna arrays," *IEEE Trans. Commun.*, vol. 66, pp. 4271–4285, Sep. 2018.