Caching Policy and Cooperation Distance Design for Base Station Assisted Wireless D2D Caching Networks: Throughput and Energy Efficiency Optimization and Trade-Off

Ming-Chun Lee, *Student Member, IEEE*, and Andreas F. Molisch, *Fellow, IEEE* Ming Hsieh Department of Electrical Engineering
University of Southern California
Los Angeles, CA, USA
Email: mingchul@usc.edu, molisch@usc.edu

***Abstract*—This work investigates the optimal caching policy and cooperation distance design from both throughput and energy efficiency (EE) perspectives in base station (BS) assisted wireless device-to-device (D2D) caching networks. By jointly considering the effects of BS transmission, D2D-caching, and self-caching, and the impact of the cooperation distance, a clustering approach is proposed with specifically designed power control and resource reuse policies. The throughput and EE of two network structures are comprehensively analyzed, and designs aiming to optimize the throughout and EE respectively are proposed. We also characterize the trade-off between the throughput and EE and provide corresponding designs. Simulations considering practical parameters are conducted to verify the analyses and evaluate the proposed designs; they demonstrate superior performance compared to state-of-the art.**

I. Introduction

Demand for wireless video delivery services has dramatically increased in recent years and is expected to continue to grow [1], straining the capacity of wireless networks. One of the most promising approaches to resolve this challenge is caching at the wireless edge. Since the majority of video demand is generated by a few popular files, local caching of popular files can avoid unnecessary traffic as the demand can be satisfied without backhaul. In contrast to conventional approaches for throughput increase, such as network densification or use of more spectrum, wireless caching leverages the unique traffic characteristics of video content (i.e., asynchronous reuse) and cheap storage to improve the system capacity [2]–[4].

A. Literature Review

Wireless edge caching has been investigated in various scenarios. Femtocaching was first proposed based on low-cost helper nodes with no or limited backhaul [5] and has been generalized to heterogeneous networks in [6]–[9]. The combination of femtocaching and other techniques, such as multiple-input multiple-output techniques [10] and coded multicast [11]–[13], has also been widely explored. Self-caching is another approach that naturally leverages existing storage resource in devices [16]–[18]. With device-to-device (D2D) communications becoming widely available [14], wireless D2D caching networks, first suggested in [15], have been widely explored [15]–[28]. Specifically, in [16], the delay performance was optimized and studied by adjusting a sub-optimum caching policy and cooperation distance using empirical results in clustering D2D networks. In [17], the caching policy was optimized in pursuit of offloading the macrocell traffic to D2D connections. Caching policies to maximize hit-rate and throughput are investigated and compared in [18]. This paper concludes that the optimal-hit-rate policy generally does not optimize the throughput. An opportunistic cooperation approach to improve D2D transmission in caching networks by optimizing the cluster size and bandwidth allocation is investigated in [19] for a clustering network and simplified caching policy. The scaling laws of D2D caching networks were derived in [20]–[22]. By adopting clustering and assuming user locations on a grid, the papers derive the asymptotic scaling laws for both uncoded caching and coded multicast. In [23], the throughput-outage trade-off was investigated for a clustering network. Similar to [20]–[22], it adopts a simplified grid network for describing user distribution and focuses on characterizing asymptotic behavior.

In [24], an approach to optimize the content caching and delivery during the same time-frame was proposed. This is different from the common assumption that users have already cache files according to a policy before entering the network. In [25] and [26], designs for optimizing successful access probability were proposed subject to different constraints. EE performance was studied in [27], in which the transmission power consumption and battery life were jointly considered for designing a caching policy. However, this paper does not take the self-caching effect into consideration. The scheduling and power control policies for D2D caching networks were investigated in [28], and, based on the proposed policies, the caching policy and cooperative distance were empirically opti-

This work was supported in part by the National Science Foundation (NSF). Part of this work was presented at the 2018 IEEE International Conference on Communications [36].

mized. In consideration of user mobility, caching designs with uncertainty were investigated in [29], [30]. By considering individual user preference, recent research has started to design caching networks via using heterogeneous user preference modeling [31]–[35]. We note that, while there are many papers investigating self-caching and D2D caching, their interaction has not been well explored. Besides, even if self-caching could be influential to the system [16], [18], its impact was occasionally overlooked.

Caching policies in wireless D2D caching networks have been designed in pursuit of different objectives: cache hit rate, i.e., successful access probability or outage [17], [25], download delay [16], [26], throughput [21], [23], [28], and energy efficiency (EE) [24], [27]. However, different objectives generally conflict with one another and have their own disadvantages. When using cache hit rate, the designs aim to maximize the probability that a user can reach the desired file through D2D communications, while ignoring the potential help from the base station (BS). Also, an optimal hit rate does not actually mean that the system throughput is optimal as well [18]. Similarly, when considering network throughput, EE of users cannot be guaranteed. On the other hand, when focusing on optimizing EE, the network throughput might be sacrificed [36]. As a result, in order to improve the design of the system, it is necessary to comprehensively explore the trade-offs between different objectives.

The impact of cooperation distance on the D2D networks has been discussed in [16], [21], [23], [27]. Generally speaking, a larger cooperation distance can provide better caching cooperation between users, i.e., a user has a higher chance to obtain the desired content via D2D links, while on the downside it leads to the higher power consumption and lower frequency reuse gain [16], [23]. This trade-off motivates the interest in exploring the effect of cooperation distance. In [16], the effect of cooperation distance on the average delay was studied for both deterministic and random caching schemes. In [23], the optimal throughput–outage trade-off was characterized for different cooperation distances. Throughput–outage trade-offs of different caching schemes were also compared in [21]. It was demonstrated that, by well designing the cooperation distance, a simple decentralized caching policy can achieve a near-optimal throughput scaling. Targeting the optimal energy consumption, [27] provides an empirical understanding of cooperation distance design via using simulations. We note that although the optimization of cooperation has drawn some attentions, the optimization of cooperation distance has not been investigated in all aspects, especially concerning the trade-off between different objectives.

B. Main Contribution

From the previous discussion, it can be concluded that a comprehensive understanding of different optimal designs and their trade-offs is necessary. In addition, the investigation of designs that provide the best compromise between different objectives is still far from providing conclusive results. The insufficiency lies in several aspects: (i) lack of joint design of caching policy and cooperation distance; (ii) relying on numerical results and/or simplified models; (iii) disregard of the effect of self-caching and its interaction with D2D-caching; (iv) absence of analysis and design for trade-off between the fundamental throughput and energy efficiency aspects. Therefore our work aims to address these issues.

In this work, a BS-assisted D2D caching network is considered. We focus on optimizing caching policy and cooperation distance designs in terms of network throughput and EE, respectively. We also discuss the throughput–EE trade-off and the network designs to achieve this trade-off. To jointly consider effects of BS, D2D-caching, and self-caching, we consider the user being able to access the desired file through BS links, D2D links, or its own cache. To embody the effect of cooperation distance and to mitigate interference between D2D links, a cluster D2D network configuration [16], [23] is adopted with specified power control policy and frequency reuse approach. D2D communications are allowed only between users in the same cluster. Since different cooperation distances manifest different sizes of cluster, different cooperation gains, and power consumptions, this network configuration along with different objectives offers the flexibility in investigating different types of designs and their trade-offs.

Network throughput analyses of two network structures, which we call the ”random-push” and the ”prioritized-push” networks, are provided in this work. Although the prioritized-push network is more spectrally efficient and practical, it suffers from a complicated formulation that makes its exact optimization intractable. In contrast, the random-push throughput is easier to analyze. We thus first analyze the random-push network, and then building on the results, we provide tractable approximations for the throughput of the prioritized-push network. Since the throughput of the random-push network and the proposed approximation for the prioritized-push network are both concave functions when fixing a cooperation distance, the throughput-based design is thus converted to a standard concave program with a one-dimensional search, in which the solution can be effectively obtained by a simple quantization. To analyze the network EE, methodologies similar to the throughout analysis are exploited for firstly analyzing the network power consumption of the networks. Then the network EE can be obtained by combining analytical results of throughput and power consumption. Since the resulting expressions of EE are quasi-concave when fixing the cooperation distance, the EE-based optimization becomes solving a standard quasi-concave program, for which the optimal solutions are attainable.

To investigate the throughput–EE trade-off, the concept of pareto-optimiality in multi-objective optimization is exploited [37]. By introducing the weighted sum method [37], the optimal trade-off design problem is proposed and a solution approach is provided by exploiting results in [38]. We note that the trade-off design can be interpreted as the design providing a compromise between two distinct objectives via adopting different cooperation distances and caching policies. Simulations considering practical parameters and network configuraions are offered to validate our theoretical analysis and evaluate the proposed designs. The proposed designs can outperform designs that does not jointly consider effects of

BS communications, D2D communications, and self-caching in terms of the targeted objectives and provide better trade-off. The insights of the designs and the effects of critical parameters are also discussed.

Our main contributions are summarized as follows:[1]

- By jointly considering the effects of BS, D2D-caching, and self-caching, and the impact of the cooperation distance, we analyze network throughput and EE and propose the mathematically tractable approximate formulations in the clustering network considering both active and inactive users with specifically designed power control and resource reuse policies.
- By exploiting the throughput and EE formulations, we propose the corresponding caching policy and cooperation design problems. We also show that the proposed optimization problems can be effectively solved by converting to standard concave and quasi-concave programs along with a simply one-dimensional search.
- We characterize the trade-off between throughput and EE and formulate their trade-off design problem. To the best of our knowledge, this work is the first to address the trade-off between throughput and EE and provide the the corresponding trade-off design.
- By considering practical network parameters and configurations in simulations, we validate the proposed analyses and evaluate the proposed designs.

The remainder of the paper is organized as follows. In Sec. II, the adopted network configurations and system models are introduced. In Sec. III, throughput analysis and the corresponding optimization are provided. The EE formulation and its optimal design are proposed in Sec. IV. We analyze the throughput–EE trade-off and propose the trade-off design in Sec. V. Numerical results and corresponding discussions are provided in Sec. VI. Conclusions are provided at the end of this paper.

II. Content Caching and System Modeling of Base Station Assisted Wireless D2D Caching Networks

A. Network and System Models

This work considers a BS-assisted cache-enabled wireless D2D network and adopts the clustering presented in [16], [23]. To wit, a square cell with side length D is served by a BS and is split into several equal-sized square clusters with side length (henceforth called cluster size) d, where D2D communication is allowed between two devices within the same cluster. Then the number of clusters in a cell is $N = \frac{D^2}{d^2}$. With slight loss of practicality, a fractional number of clusters is allowed for mathematical tractability and simplicity. We consider two non-overlapping frequency bands for establishing BS communications and D2D communications, respectively. For communications between the BS and the devices, the time-frequency resources of the BS band are shared by all clusters via an orthogonal multiple access approach, such as FDMA. To guarantee the minimum video streaming quality, each BS link assigned to a user is allowed to obtain the same, fixed, amount of resources (bandwidth and power). Typically, the data rate achievable on such a BS link is significantly lower than on a D2D link. Since the amount of resources is limited, there exists a maximum number of users N_{BS} that can simultaneously use BS links. Adopting the clustering network structure provides the following benefits: (i) tractable closed-form expressions of critical metrics can be obtained;[2] (ii) the results are easily extensible to analyses of other aspects; and (iii) the resulting designs can serve as a benchmark/reference system for other systems, i.e., the performance achieved with a clustering approach constitutes an achievable lower bound.

D2D communications are considered only between users within the same cluster. Consequently, we call (with slight abuse of definition) d also the cooperation distance; in fact, "cluster size" and "cooperation distance" will be used interchangeably throughout this paper. Resources for D2D communications are spatially reused between the clusters. Such a reuse scheme evenly applies K colors to the clusters, and only the clusters with the same color can be active on the same time-frequency resource for D2D communications. Note that the adopted reuse scheme is analogous to the spatial reuse scheme in conventional cellular networks [39], and K is the reuse factor.

In this work, we adopt a simplified channel model in which only the path-loss effect is considered for mathematical tractability. The channel randomness, such as small-scale fading and shadowing, is ignored because link level optimization is not employed, the channel randomness can be averaged out by using frequency diversity and properties of Poisson point processes (PPPs), and the caching policy is designed and operated over a long time scale. The path-loss model is

$$20\log_{10}\frac{4\pi d_0}{\lambda_{\mathrm{c}}} + 10\alpha\log_{10}(\frac{d_{\mathrm{TR}}}{d_0}) \quad [\mathrm{dB}], \tag{1}$$

where d_{TR} is the distance between transmitter and receiver, λ_{c} is the wavelength of the carrier frequency, α is the pathloss exponent, d_0 is the break point distance. To restrict the interference between different clusters and maintain the received signal power with respect to the change of d, a power control policy is adopted such that[3]

$$E_D = \left[(\sqrt{K}-1)\frac{d}{d_0}\right]^{\alpha}\cdot(\frac{4\pi d_0}{\lambda_{\mathrm{c}}})^2\cdot\nu, \tag{2}$$

where E_D is the transmission power for D2D communications and ν, which is a choice of the designer, is the maximum allowable interference between two clusters using the same resource. Thus, by fixing ν to be sufficiently small, the interference can be effectively avoided. Besides, by this policy, the average received power of users in a cluster can be maintained even if the cluster size is adjusted for optimization purposes. This is mainly because E_{D} scales with d on the order of α. Note that this power control policy depends only on system

[1]Compared with our conference version [36], this paper extends the model to consider both active and inactive users and provides analyses of a more spectrally efficient and practical network structure.

[2]We note that a dynamic D2D scheduler, such as [28], provides better spectral efficiency; however it is very challenging to find the optimal caching policy for this case, and only some heuristic designs are known [28].

[3] Correcting our conference version [36], the multiplier $\sqrt{2}$ in the same equation of [36] is unnecessary. This revision generally does not have any impact on the results in this paper and in [36].

TABLE I: Summary of Notations

Notations	Descriptions
D; d; N; K	Cell size; cluster size (cooperation distance); number of clusters; reuse factor
λ_{c}; α; d_0; ν	Carrier frequency; path-loss exponent; breaking point distance; maximum allowable interference power
λ_{a}; λ_{i}; λ_{u}	Density of active users; of inactive users; of overall users
κ_{a}; κ_{i}; κ_{u}	(in a cluster) Average number of active users; of inactive users; of overall users
P_k^{a}; P_n^{i}; P_k^{u}	(in a cluster) Probability of number of active users to be k; number of inactive users to be n; number of overall users to be k
T_{B}; T_{D}; T_{S}	Throughput of using BS link; of using D2D link; of using self-caching
E_{B}; E_{D}	Power consumption of using BS link; of using D2D link as described in (2)
S; M;	Cache space of a user device; number of files in the library
b_m; a_m	Probability for file m to be cached in a device ; probability for an active user to request file m
P_{S}; $P_{\mathrm{B},k}$; $P_{\mathrm{D},k}$	Elementary access probabilities: refer the clear definitions to (3); to (4); to (5)

parameters, and no attempt is made to adapt it to the channel states/distances between TX and RX. Hence, given the system parameters, the transmission powers of all D2D links in the network are identical. Also note that, since interference is avoided when ν is sufficiently small, the interference between clusters will be ignored in the remainder of the paper.

In this work, users can obtain the desired content via their own caches, D2D communications, or BS communications with different transmission qualities and costs. We denote the throughput for a user that accesses the content via a BS link as T_{B}; via a D2D link as T_{D}; and from its own cache as T_S; and consider $T_{\mathrm{S}} \geq T_{\mathrm{D}} > T_{\mathrm{B}}$.[4] Note that we generally assume T_{B}, T_{D}, and T_{S} to be invariant with respect to the cluster size d, and these assumptions are reasonable when the power control policy in (2) is adopted, and the amount of BS resources assigned to each BS link are the same.[5] Furthermore, we assume that the throughput of the user is independent of the actual distance between the transmitter and receiver, which is practical when we have a fixed modulation-and-coding scheme. Similar to the throughput case, we denote the power consumption for a user to access the content via a BS link as E_{B}; the power consumption for a user to access the content using a D2D link is by definition E_{D} in (2); we consider only $E_{\mathrm{B}} > E_{\mathrm{D}}$.[6] Zero power consumption is assumed if the user can access the desired content from its own cache. For simplicity, we assume here that energy cost is purely determined by RF energy required for transmission; access to storage and coding/decoding is assumed to be negligible in comparison. We assume that the BS is equipped with an unlimited backhaul connected to repositories containing all contents in the library. Thus, the request from a user can always be satisfied (with a minimum video quality) if the BS link is available for that user.

We consider two different types of users in this work: active and inactive users. An active user is a user who places a request that needs to be satisfied and participates in the D2D cooperation (i.e., sends files to other users that request them); an inactive user is a user who does not place request of its own but still participates in the D2D cooperation. We consider both active and inactive users to be independently distributed according to homogeneous Poisson point processes (HPPPs) with user densities λ_{a} and λ_{i}, respectively. Hence the overall user distribution is an HPPP with density $\lambda_{\mathrm{u}} = \lambda_{\mathrm{a}} + \lambda_{\mathrm{i}}$.

The library consists of M files with all files having the same size. Each user is assumed to be able to cache S files in the device. A random caching policy [6] is employed by the users, and all users adopt the same caching policy. Denoting b_m as the probability for the user to cache file m, the caching policy is expressed as $\{b_m\}_1^M$, where $\sum_{m=1}^M b_m = S \leq M$. All users follow the identical request probability distribution. The request probability of a user for file m, i.e., the probability that a user wants file m is denoted as a_m with $0 \leq a_m \leq 1, \forall m$, and $\sum_{m=1}^M a_m = 1$. The notations used in this paper are summarized in Table I.

B. Elementary Access Probability Analysis

Here the elementary access probabilities of using different transmission approaches are analyzed. The results will serve as the foundation for further results in the subsequent sections.

Consider the caching policy $\{b_m\}_1^M$. The self-caching probability of a user is defined as the probability that the desired file of the user can be found in its own cache:

$$P_{\mathrm{S}} = \sum_{m=1}^{M} a_m b_m. \tag{3}$$

Then considering there are k users in a cluster, the probability that a user cannot find the desired content through self-caching or D2D communications is

$$P_{\mathrm{B},k} = \sum_{m=1}^{M} a_m (1-b_m)^k, \tag{4}$$

where $(1-b_m)^k$ is the probability that file m is not in the caches of users of the cluster, and therefore $a_m(1-b_m)^k$ is the probability that the user wants file m but file m is not in the caches of users of the cluster. Finally, when both BS and D2D links are available for a user, the probability that the user obtains the desired file via the D2D link is

$$P_{\mathrm{D},k} = 1 - P_{\mathrm{B},k} - P_{\mathrm{S}} = 1 - \sum_{m=1}^{M} a_m (1-b_m)^k - \sum_{m=1}^{M} a_m b_m. \tag{5}$$

[4]Here T_{B}, T_{D}, and T_{S} can be generalized to include the perspective of the user satisfaction by considering the effective or weighted throughput. Our results will hold as long as the inequality $T_{\mathrm{S}} \geq T_{\mathrm{D}} > T_{\mathrm{B}}$ holds.

[5]This assumption is in line with policies of network providers that do not charge video traffic to users as long as they opt for lowest possible quality.

[6]Similarly, here E_{B} and E_{D} could be generalized (by using the effective or weighted power consumption) to include the different impacts of power consumptions. For example, we can emphasize the importance of the power consumption of the users by rendering the user power consumption a larger weight. Our results will hold as long as the inequality $E_{\mathrm{B}} > E_{\mathrm{D}}$ holds.

III. Caching Policy and Cooperation Distance Design for Throughput Optimization

In this section, the caching policy and cooperation distance design is investigated for the goal of optimizing network throughput. We first analyze the network throughput considering two different network structures, i.e., the random-push and prioritized-push networks. Then the optimization approach is proposed. We note that although the prioritized-push network is more spectrally efficient, its throughput analysis is more challenging and builds on the analysis of the random-push network. Also, in the following analyses, we assume for simplicity that N_{BS} is sufficiently large to provide BS links to all users that need one. While this assumption might not be true in general, from the simulations, we can observe that outage occurs mostly when the cluster size is very small (the number of clusters is large), which is usually not the cooperation distance that we are interested in (see Fig. 5 in Sec. VI).

A. Throughput Analysis for Random-Push Networks

The random-push system operates as follows. For each cluster, the BS randomly chooses a user to serve without considering whether the user can obtain its desired content from its own cache. If the selected user can obtain the desired content from its own cache, the self-cache approach is used by the user; otherwise, the BS checks whether the desired content can be found through D2D links. If yes, the D2D communication is used; otherwise, the BS will serve the selected user by a BS link. The rest of the users then check whether they can obtain their desired contents from their own caches. If yes, their requests can be satisfied; otherwise, they wait to be selected by the BS in the future. This system is called *random-push* because the BS tends to push the content to the randomly selected user without considering whether the content has already been cached by this user. Note that since the resources of both the BS and D2D communications are shared in a cluster-based manner, we indicate that only a single user in a cluster is allowed to communicate at a time.

Now we analyze the throughput of the random-push system. Considering the HPPP, the numbers of active users k and inactive users n in a cluster are Poisson random variables with probability mass functions (pmfs) being

$$\begin{aligned} P_k^{\text{a}} &= \frac{(\kappa_{\text{a}})^k}{k!} e^{-\kappa_{\text{a}}}, k = 0, 1, 2, ..., \\ P_n^{\text{i}} &= \frac{(\kappa_{\text{i}})^n}{n!} e^{-\kappa_{\text{i}}}, n = 0, 1, 2, ..., \end{aligned} \tag{6}$$

respectively, where $\kappa_{\text{a}} = \lambda_{\text{a}} d^2$ and $\kappa_{\text{i}} = \lambda_{\text{i}} d^2$. Suppose the number of active and inactive users in the cluster are $k > 0$ and n, respectively. Using the derived access probabilities, the throughput of the user selected by the BS is

$$\begin{aligned} T_{\text{c,Ran},k,n} &= T_{\text{D}} P_{D,k+n} + T_B P_{\text{B},k+n} + T_{\text{S}} P_{\text{S}} \\ &= T_{\text{D}} + (T_{\text{B}} - T_{\text{D}}) \Big[\sum_{m=1}^{M} a_m (1 - b_m)^{(k+n)} \Big] \\ &\quad + (T_{\text{S}} - T_{\text{D}}) \sum_{m=1}^{M} a_m b_m. \end{aligned} \tag{7}$$

Hence the throughput of the cluster is shown in (8) on the top of the next page, where (a) is derived by using (6) and rearranging the summation; (b) is derived by using the similar approach as in (a). It follows that the throughput of the system is

$$T_{\text{s,Ran}} = N \cdot T_{\text{c,Ran}}. \tag{9}$$

Lemma 1-1: When given a fixed d, (9) is a non-decreasing concave function with respect to the feasible set $\mathcal{B} = \{0 \leq b_m \leq 1, \forall m\}$.

Proof. Consider $\mathcal{B} = \{0 \leq b_m \leq 1, \forall m\}$ and a given fixed d. Since $T_S \geq T_D \geq T_{\text{B}} > 0$, $\kappa_{\text{a}} > 0$, and $\kappa_{\text{i}} \geq 0$, it is simple to find that the first order partial derivative of $T_{\text{s,Ran}}$ is non-negative on $\mathcal{B}$. This leads to that $T_{\text{s,Ran}}$ is non-decreasing with respect to $b_m, \forall m$, over $\mathcal{B}$. To prove that $T_{\text{s,Ran}}$ is concave over $\mathcal{B}$, we note that the Hessian of $T_{\text{s,Ran}}$ is a diagonal matrix with diagonal entries being non-positive. Therefore the Hessian of $T_{\text{s,Ran}}$ is negative semidefinite over $\mathcal{B}$. ■

B. Throughput Analysis for Prioritized-Push Networks

Here we introduce the prioritized-push network, which is more practical and provides better spectral efficiency than the random-push network. The prioritized-push network operates as follows. For each cluster, every active user first checks whether their requests can be satisfied by their own caches. If yes, their requests are directly satisfied and they remain online for potential D2D cooperation; otherwise, they send the requests to the BS. The BS collects all the requests from the users who cannot be satisfied by self-caching and checks whether there exist users that can be satisfied by D2D links. If yes, the BS picks one to be served by the D2D communication. If not, the BS will randomly pick one user to be served by a BS link using a given amount of BS resources. Thus, there is at least one user being served either by D2D or BS in the cluster as long as there are active users and not all of them can be satisfied by self-caching. The same procedure is implemented for every cluster. It can be immediately understood that the prioritized-push network is more spectrally efficient than the random-push network which also serves one user per cluster, but which randomly picks one user to serve without checking whether there are other users that can use D2D communications.[7]

The throughput analysis for the prioritized-push network is more challenging. We thus, in the following, provide tractable approximations for them and use the approximations for conducting optimization. Suppose the number of active users in the cluster is $k > 0$. The probability of each active user to

[7]We note that, for the prioritized-push network, the number of served users by the D2D and BS links is proportional to the number of clusters, so that an increase in the cluster size automatically means a reduction in the number of non-self-served users (though the throughput still might increase, due to the higher throughput of D2D). Having said that, if we want to guarantee serving the same number of users by the D2D and BS links when the number of clusters are different, we can simply add some additional BS users who can then provides an additional throughput on the top of our adopted network structure; this does not affect the optimization of the caching policy.

$$
\begin{aligned}
T_{\text{c,Ran}} &= \sum_{n=0}^{\infty} P_n^{\text{i}} \sum_{k=1}^{\infty} P_k^{\text{a}} (T_{\text{c,Ran},k,n} + (k-1) T_{\text{S}} P_{\text{S}}) = \sum_{n=0}^{\infty} P_n^{\text{i}} \left[\left(\sum_{k=0}^{\infty} P_k^{\text{a}} T_{\text{c,Ran},k,n} + P_k^{\text{a}} (k-1) T_{\text{S}} P_{\text{S}} \right) - P_0^{\text{a}} (T_{\text{c,Ran},0,n} - T_{\text{S}} P_{\text{S}}) \right] \\
&\overset{(a)}{=} \sum_{n=0}^{\infty} P_n^{\text{i}} \Bigg[T_{\text{D}} + (T_{\text{B}} - T_{\text{D}}) \Bigg(\sum_{m=1}^{M} a_m (1-b_m)^n e^{-\kappa_{\text{a}} b_m} \underbrace{\sum_{k=0}^{\infty} \frac{(\kappa_{\text{a}}(1-b_m))^k}{k!} e^{-\kappa_{\text{a}}(1-b_m)}}_{=1} \Bigg) \\
&+ (\kappa_a - 1) T_{\text{S}} P_{\text{S}} + (T_{\text{S}} - T_{\text{D}}) \left(\sum_{m=1}^{M} a_m b_m \right) - e^{-\kappa_a} (T_{\text{c,Ran},0,n} - T_{\text{S}} P_{\text{S}}) \Bigg] \\
&\overset{(b)}{=} T_D (1 - e^{-\kappa_{\text{a}}}) + (T_{\text{B}} - T_{\text{D}}) \left[\sum_{m=1}^{M} a_m e^{-(\kappa_{\text{a}} + \kappa_{\text{i}}) b_m} \right] \\
&+ \left[(T_{\text{S}} - T_{\text{D}})(1 - e^{-\kappa_{\text{a}}}) + T_S (\kappa_{\text{a}} - 1 + e^{-\kappa_{\text{a}}}) \right] \left(\sum_{m=1}^{M} a_m b_m \right) - (T_{\text{B}} - T_{\text{D}}) e^{-\kappa_{\text{a}}} \left[\sum_{m=1}^{M} a_m e^{-\kappa_{\text{i}} b_m} \right].
\end{aligned} \tag{8}
$$

have its desired file *not* to be cached in the D2D network of a cluster is:

$$
\sum_{n=0}^{\infty} P_n^i \sum_{m=1}^{M} a_m (1-b_m)^{k+n} = \sum_{m=1}^{M} a_m (1-b_m)^k e^{-\kappa_i b_m}. \tag{10}
$$

Note that the derivation here follows the same approach as in (8). Using (10) and assuming that each user is independent,[8] the probability there is *no* potential D2D link in the cluster is:

$$
\left[\sum_{m=1}^{M} a_m (1-b_m)^k e^{-\kappa_i b_m} \right]^k. \tag{11}
$$

Then by ignoring the small probability that all users are served by either self-caching or BS, the sum throughput of the users $T_{\text{c,Pri}}$ in a cluster is approximated by (12) shown on the top of next page.[9] Note that the total throughput is simply $T_{\text{s,Pri}} = NT_{\text{c,Pri}}$. Eq. (12) is too complicated for conducting caching policy and cooperation optimizations. We thus propose further approximations for them. Obviously, the complication is due to the second term in (12). To approximate it, we distinguish between two cases: (i) $\kappa_a \leq \theta$ and (ii) $\kappa_a > \theta$, where $\theta \geq 1$ is a small number.[10] This distinction is because we want to use two different approximations for different cases, i.e., κ_a is small or large. When doing case 1,[11] since κ_a is small, i.e., there is a high probability to have a small number of active users, the most important terms of the summation are the first several terms. The following approximation with the parameter $\theta \geq 1$ is thus used:

$$
\begin{aligned}
&\sum_{k=1}^{\infty} P_k^a \left[\sum_{m=1}^{M} a_m (1-b_m)^k e^{-\kappa_i b_m} \right]^k \\
&\approx \sum_{k=1}^{\infty} P_k^a \left[\sum_{m=1}^{M} a_m (1-b_m)^k e^{-\kappa_i b_m} \right]^\theta
\end{aligned} \tag{13}
$$

Then observe that the inner summation is the convex combination of several points located on a convex curve, we have

$$
\begin{aligned}
&\sum_{k=1}^{\infty} P_k^a \left[\sum_{m=1}^{M} a_m (1-b_m)^k e^{-\kappa_i b_m} \right]^\theta \\
&\leq \sum_{k=1}^{\infty} P_k^a \left[\sum_{m=1}^{M} a_m (1-b_m)^{\theta k} e^{-\theta \kappa_i b_m} \right] \\
&\leq \sum_{k=1}^{\infty} P_k^a \left[\sum_{m=1}^{M} a_m (1-b_m)^k e^{-\theta \kappa_i b_m} \right],
\end{aligned} \tag{14}
$$

where the final inequality is because $(1 - b_m) \leq 1$. By using (14), (12) can be expressed as (15), which is a concave function (See Lemma 1-3 below), on the top of next page.

Considering $\kappa_a > \theta$, we approximate the outer exponent k using the mean value κ_a. We thus have the approximation shown in (16) on the top of next page, where the inequality is due to that x^{κ_a} is convex with respect to x when $x \geq 0$ and $\kappa_a \geq 1$ and that $E[g(x)] \geq g(E[x])$ when $g(.)$ is convex (Jensen's inequality). The resulting throughput is shown in (17) on the top of next page. To characterize (17), Lemma 1.2 is provided:

Lemma 1-2: Suppose $\kappa_a \geq 1$. $\left[\sum_{m=1}^{M} a_m e^{-(\kappa_a + \kappa_i) b_m} \right]^{\kappa_a}$ and $\left[\sum_{m=1}^{M} a_m e^{-\kappa_i b_m} \right]^{\kappa_a}$ are convex and non-increasing with respect to $\mathcal{B} = \{0 \leq b_m \leq 1, \forall m\}$.

Proof. See Appendix A. ■

Since (17) is still non-convex due to the difference of two convex functions, we further approximate it by dropping the

[8]This is generally not true because all users in the same cluster share the same D2D caching inventory.

[9]This approximation does not work effectively when adopting a caching policy tending to be selfish and in a system whose popularity distribution is highly concentrated, e.g., $\gamma = 1.3$ and $q = 0$. However, in practice, the optimal caching policy tends to be selfish only in the case that the popularity distribution is highly concentrated or when the density of active users are overwhelmingly large, which rarely happens in practice. We note that under the practically considered popularity distributions in the simulations, this approximation works well.

[10]The idea is similar to having the breaking-point in the path-loss model, and θ might be an empirically selected value.

[11]Actually case 1 is much less important than case 2 since the optimal design usually needs more users. The reason for considering case 1 is for the mathematical completeness.

$$\begin{aligned} T_{\text{c,Pri}} &\approx \sum_{k=1}^{\infty} P_k^a \left[T_{\text{D}} + (T_{\text{B}} - T_{\text{D}}) \left[\sum_{m=1}^{M} a_m (1-b_m)^k e^{-\kappa_i b_m} \right]^k + T_{\text{S}}(k-1) \sum_{m=1}^{M} a_m b_m \right] \\ &= T_{\text{D}}(1-e^{-\kappa_a}) + (T_{\text{B}} - T_{\text{D}}) \sum_{k=1}^{\infty} P_k^a \left[\sum_{m=1}^{M} a_m (1-b_m)^k e^{-\kappa_i b_m} \right]^k + T_{\text{S}} \left(\kappa_a - 1 + e^{-\kappa_a} \right) \sum_{m=1}^{M} a_m b_m. \end{aligned} \tag{12}$$

$$\begin{aligned} T_{\text{c,Pri-A1}} = T_{\text{D}}(1-e^{-\kappa_a}) &+ (T_{\text{B}} - T_{\text{D}}) \left[\sum_{m=1}^{M} a_m e^{-(\kappa_a + \theta\kappa_i) b_m} \right] \\ &- (T_{\text{B}} - T_{\text{D}}) e^{-\kappa_a} \left[\sum_{m=1}^{M} a_m e^{-\theta\kappa_i b_m} \right] + T_{\text{S}} \left(\kappa_a - 1 + e^{-\kappa_a} \right) \sum_{m=1}^{M} a_m b_m. \end{aligned} \tag{15}$$

$$\begin{aligned} &\sum_{k=1}^{\infty} P_k^a \left[\sum_{m=1}^{M} a_m (1-b_m)^k e^{-\kappa_i b_m} \right]^k \approx \sum_{k=0}^{\infty} P_k^a \left[\sum_{m=1}^{M} a_m (1-b_m)^k e^{-\kappa_i b_m} \right]^{\kappa_a} - e^{-\kappa_a} \left[\sum_{m=1}^{M} a_m e^{-\kappa_i b_m} \right]^{\kappa_a} \\ &\geq \left[\sum_{k=0}^{\infty} P_k^a \sum_{m=1}^{M} a_m (1-b_m)^k e^{-\kappa_i b_m} \right]^{\kappa_a} - e^{-\kappa_a} \left[\sum_{m=1}^{M} a_m e^{-\kappa_i b_m} \right]^{\kappa_a} = \left[\sum_{m=1}^{M} a_m e^{-(\kappa_a + \kappa_i) b_m} \right]^{\kappa_a} - e^{-\kappa_a} \left[\sum_{m=1}^{M} a_m e^{-\kappa_i b_m} \right]^{\kappa_a}. \end{aligned} \tag{16}$$

$$\begin{aligned} T_{\text{c,Pri-A2}} = T_{\text{D}}(1-e^{-\kappa_a}) &+ (T_{\text{B}} - T_{\text{D}}) \left[\sum_{m=1}^{M} a_m e^{-(\kappa_a + \kappa_i) b_m} \right]^{\kappa_a} \\ &- (T_{\text{B}} - T_{\text{D}}) e^{-\kappa_a} \left[\sum_{m=1}^{M} a_m e^{-\kappa_i b_m} \right]^{\kappa_a} + T_{\text{S}} \left(\kappa_a - 1 + e^{-\kappa_a} \right) \sum_{m=1}^{M} a_m b_m. \end{aligned} \tag{17}$$

third term in (17), resulting in a concave function (See Lemma 1-3 below):

$$\begin{aligned} &T_{\text{c,Pri-AC2}} = \\ &T_{\text{D}}(1-e^{-\kappa_a}) + (T_{\text{B}} - T_{\text{D}}) \left[\sum_{m=1}^{M} a_m e^{-(\kappa_a + \kappa_i) b_m} \right]^{\kappa_a} \\ &+ T_{\text{S}} \left(\kappa_a - 1 + e^{-\kappa_a} \right) \sum_{m=1}^{M} a_m b_m. \end{aligned} \tag{18}$$

We denote $T_{\text{c,Pri-AC}} = T_{\text{c,Pri-A1}}$ if $\kappa_a \leq \theta$; $T_{\text{c,Pri-AC}} = T_{\text{c,Pri-AC2}}$ otherwise. Then Lemma 1-3 characterize the properties of $T_{\text{c,Pri-AC}}$:

Lemma 1-3: When given a fixed d, $T_{\text{c,Pri-AC}}$ is a non-decreasing concave function with respect to the feasible set $\mathcal{B} = \{0 \leq b_m \leq 1, \forall m\}$.

Proof. The non-decreasing property and concavity of $T_{\text{c,Pri-A1}}$ can be proved by using the same approach in Lemma 1-1. We thus omit the proof. Regarding $T_{\text{c,Pri-AC2}}$, the proof is trivial by using Lemma 1-2 and observing that $T_B - T_D \leq 0$. ■

The simplification in (18) provides the tractability for optimizing caching policies, in which the throughput optimization problem becomes a standard concave program. To justify this simplification, we observe that the third term of (17) is due to the case that there is no active user in the cluster. Then because we generally consider the second part of the approximation to be useful when κ_a is large, this simplification could result in minor impact except for the point that κ_a is near the breaking-point θ. Thus, since the points where the simplification is not effective are not near the optimal cooperation point, the error is less important. Besides, when we consider directly solving the non-convex $T_{\text{c,Pri-A2}}$ using more advanced non-convex solution approaches, such as the concave-convex procedure [40], the performance does not improve.

C. Throughput-Based Caching Policy and Cooperation Distance Design

According to the analysis in Secs. III.A and III.B, we design the caching policy and cooperation distance by solving the following optimization problem:

$$\begin{aligned} \max_{d, b_m, \forall m=1,..,M} \quad & T_{\text{sys}} = N \cdot (T_{\text{c,Ran}} \quad \text{or} \quad T_{\text{c,Pri-AC}}) \\ \text{subject to} \quad & \textstyle\sum_{m=1}^{M} b_m \leq S, \quad 0 \leq b_m \leq 1, \forall m. \end{aligned} \tag{19}$$

To solve (19), we first observe that, if we can solve its sub-problem with any given d, the problem then becomes a simple one-dimensional problem with small range. Note that $d > 0$ is generally within 100 meters considering practical D2D communications, and, given the optimal solution is attainable when fixing d, the problem is solvable even by simple quantization without significant effort. We then provide the following proposition:

Proposition 1: When given a fixed d, (19) becomes a concave optimization problem and its optimal solution must be tight at the equality of the sum constraint, i.e., for the optimal solution $(b_m)^*, \forall k, m$, we have $\sum_{m=1}^{M}(b_m)^* = S$.

Proof. Follows from Lemmas 1-1 and 1-3. ■

By Proposition 1, the problem becomes a standard concave optimization problem, and any convex solver[12] can be used to solve the problem. The overall solving approach is summarized as following: cooperation distance d is firstly quantized to form sub-problems of (19). Then the optimal caching policies of the quantized sub-problems are attained by the convex solver. Finally, by comparing between throughput results of different sub-problems, we can obtain the optimal caching along with the optimal cooperation distance.

IV. Caching Policy and Cooperation DistanceDesign for Energy Efficiency Optimization

In this work, we define the EE (bits/Joule) as the ratio of the total average throughput (bits/s) and total average power consumption (Joule/s):

$$EE_{\text{sys}} = \frac{T_{\text{sys}}}{P_{\text{sys}}} = \frac{T_{\text{clu}}}{P_{\text{clu}}}, \tag{20}$$

where T_{sys} and P_{sys} are the average throughput and average power consumption of the system, respectively; T_{clu} and P_{clu} are the average throughput and average power consumption of a cluster of the system. In the following, the EE is first analyzed in random-push and prioritized-push networks. Then the design aiming to optimize the EE is proposed.

A. Energy Efficiency Analysis for Random-Push Networks

Recall that the average throughput of a cluster in the random-push network is derived in (8). Then by following the same approach, we can obtain the average power consumption of a cluster in (21) on the top of next page. By substituting (8) and (21) into (20), the EE of the random-push network is then derived.

Lemma 2-1: When given a fixed d, $\frac{T_{\text{c,Ran}}}{P_{\text{c,Ran}}}$ is a positive quasi-concave and non-decreasing function with respect to $\mathcal{B}$.

Proof. By noticing $E_{\text{B}} \geq E_{\text{D}}$ and following the same approach as in Lemma 1-1, $P_{\text{c,Ran}}$ can be proved to be a positive convex function and non-increasing with respect to $\mathcal{B}$. Then observe that $P_{\text{c,Ran}}$ is convex and non-increasing with respect to $\mathcal{B}$; $T_{\text{c,Ran}}$ is concave and non-decreasing with respect to $\mathcal{B}$; $P_{\text{c,Ran}}$ and $T_{\text{c,Ran}}$ are both positive. Thus, $\frac{T_{\text{c,Ran}}}{P_{\text{c,Ran}}}$ is a positive quasi-concave and non-decreasing function with respect to $\mathcal{B}$ [41]. ■

[12]General convex solvers need to find the Hessian matrix which requires a high computational cost as the dimension of the solution space is large. We hence note that the Lagrange multiplier based approach can be used to solve part of the problem, i.e., the part involving $T_{\text{c,Ran}}$ and $T_{\text{c,Pri-A1}}$, more effectively.

B. Energy Efficiency Analysis for Prioritized-Push Networks

By following similar ideas and derivations as for the throughput analysis in Sec. III.B, the power consumption of the prioritized-push network is approximated by expressions in (22) on the top of next page, in which the first expression is the power consumption counterpart of (12); the second is the counterpart of (15); the third is the counterpart of (17); and the forth is the counterpart of (18). Then again by substituting the approximations of the throughput and power consumption into (20), the approximation for EE in the prioritized-push network can be obtained. Denoting $P_{\text{c,Pri-AC}} = P_{\text{c,Pri-A1}}$ if $\kappa_a \leq \theta$; $P_{\text{c,Pri-AC}} = P_{\text{c,Pri-AC2}}$ otherwise, we have the following Lemma:

Lemma 2-2: When given a fixed d, $\frac{T_{\text{c,Pri-AC}}}{P_{\text{c,Pri-AC}}}$ is a positive quasi-concave and non-decreasing function with respect to $\mathcal{B}$.

Proof. By following the same approach as for the proof of Lemma 1-1, we can prove that $P_{\text{c,Pri-A1}}$ is positive convex and non-increasing with respect to $\mathcal{B}$. Also, by using Lemma 1-2, $P_{\text{c,Pri-AC2}}$ can be proved to be positive convex and non-increasing with respect to $\mathcal{B}$. Then by combining the above results, Lemma 2-2 is proved. ■

C. EE-Based Caching Policy and Cooperation Distance Design

By (20) and the EE analyses in Secs. IV.A and IV.B, the EE optimization problem is

$$\begin{array}{ll} \max\limits_{d,b_m,\forall m=1,..,M} & EE_{\text{sys}} = \left(\frac{T_{\text{c,Ran}}}{P_{\text{c,Ran}}} \quad \text{or} \quad \frac{T_{\text{c,Pri-AC}}}{P_{\text{c,Pri-AC}}}\right) \\ \text{subject to} & \sum_{m=1}^{M} b_m \leq S, \quad 0 \leq b_m \leq 1, \forall m. \end{array} \tag{23}$$

Here we use the same approach as in Sec. III.C in which we solve the sub-problems with quantized d, and then perform a one-dimensional search. Therefore we focus on solving the problem with fixed d. For this case we have the following proposition:

Proposition 2: For a fixed d, (23) is a standard quasi-concave problem and its optimal solution is tight at the equality of the sum constraint.

Proof. Again follows from Lemmas 2-1 and 2-2. ■

By Proposition 2, we know that (23) becomes a standard quasi-concave optimization with a convex feasible set when fixing d. Consequently, a standard solving procedure is used and briefly described as follows. By introducing a slack variable, the problem is equivalent to

$$\begin{array}{ll} \max\limits_{t,b_m,\forall m=1,..,M} & t \\ \text{subject to} & EE_{\text{sys}} \geq t \\ & \sum_{m=1}^{M} b_m \leq S, \quad 0 \leq b_m \leq 1, \forall m. \end{array} \tag{24}$$

Since P_{sys} is positive, for a fixed t, we have a convex feasibility problem:

$$\begin{array}{ll} \max\limits_{b_m,\forall m=1,..,M} & 0 \\ \text{subject to} & -T_{\text{sys}} + tP_{\text{sys}} \leq 0 \\ & \sum_{m=1}^{M} b_m \leq S, \quad 0 \leq b_m \leq 1, \forall m. \end{array} \tag{25}$$

Note that if (25) is feasible, t is achievable. Since (25) is solvable by standard convex solvers, by exploiting the

$$P_{\text{c,Ran}} = E_{\text{D}}(1-e^{-\kappa_{\text{a}}}) + (E_{\text{B}}-E_{\text{D}})\left[\sum_{m=1}^{M} a_m e^{-(\kappa_{\text{a}}+\kappa_{\text{i}})b_m}\right] - E_{\text{D}}(1-e^{-\kappa_{\text{a}}})\left(\sum_{m=1}^{M} a_m b_m\right) - (E_{\text{B}}-E_{\text{D}})e^{-\kappa_{\text{a}}}\left[\sum_{m=1}^{M} a_m e^{-\kappa_{\text{i}} b_m}\right]. \tag{21}$$

$$\begin{aligned}
P_{\text{c,Pri}} &\approx E_{\text{D}}(1-e^{-\kappa_a}) + (E_{\text{B}}-E_{\text{D}})\sum_{k=1}^{\infty} P_k^a \left[\sum_{m=1}^{M} a_m (1-b_m)^k e^{-\kappa_i b_m}\right]^k; \\
P_{\text{c,Pri-A1}} &= E_{\text{D}}(1-e^{-\kappa_a}) + (E_{\text{B}}-E_{\text{D}})\left[\sum_{m=1}^{M} a_m e^{-(\kappa_a+\theta\kappa_i)b_m}\right] - (E_{\text{B}}-E_{\text{D}})e^{-\kappa_a}\left[\sum_{m=1}^{M} a_m e^{-\theta\kappa_i b_m}\right]; \\
P_{\text{c,Pri-A2}} &= E_{\text{D}}(1-e^{-\kappa_a}) + (E_{\text{B}}-E_{\text{D}})\left[\sum_{m=1}^{M} a_m e^{-(\kappa_a+\kappa_i)b_m}\right]^{\kappa_a} - (E_{\text{B}}-E_{\text{D}})e^{-\kappa_a}\left[\sum_{m=1}^{M} a_m e^{-\kappa_i b_m}\right]^{\kappa_a}; \\
P_{\text{c,Pri-AC2}} &= E_{\text{D}}(1-e^{-\kappa_a}) + (E_{\text{B}}-E_{\text{D}})\left[\sum_{m=1}^{M} a_m e^{-(\kappa_a+\kappa_i)b_m}\right]^{\kappa_a}.
\end{aligned} \tag{22}$$

bisection or other adaptive approaches to adjust t, a solution arbitrarily close to the optimum can be obtained.[13]

V. Throughput–Energy Efficiency Trade-Off Analysis and Design

It can be observed that optimizing EE could be different from optimizing throughput, and there exists a trade-off between them. This section aims to characterize such trade-off and provide the trade-off design. To analyze the trade-off between throughput and EE, we need to consider a multi-objective optimization problem containing different objectives that could conflict with each other. That is to say, trade-offs between different objectives exist in the problem and a solution that dominates in all aspects generally does not exist. We thus introduce the pareto-optimality [37] in the throughput–EE domain in Proposition 3.

Proposition 3: A pareto-optimal solution is defined as the solution with d_{o} and $\{b_{m,\text{o}}\}_1^M$ such that there does not exist another feasible solution with d and $\{b_m\}_1^M$ satisfying the following conditions simultaneously:

$$\begin{aligned}
&T_{\text{sys}}(\{b_m\}_1^M, d) > T_{\text{sys}}(\{b_{m,\text{o}}\}_1^M, d_{\text{o}}); \\
&EE_{\text{sys}}(\{b_m\}_1^M, d) > EE_{\text{sys}}(\{b_{m,\text{o}}\}_1^M, d_{\text{o}}).
\end{aligned} \tag{26}$$

Since there could exist multiple pareto-optimal solutions, the collection of all such solutions is denoted as the pareto-optimal set Ω. A common approach to deal with multi-objective problems is to convert the problem into a single objective problem via the weighted sum method [37]. We then provide Proposition 4 that helps us in finding pareto-optimal solutions.

Proposition 4: The optimal solution of the following problem gives a solution in a pareto-optimal set Ω:

$$\begin{aligned}
\max_{d;b_m,\forall m=1,..,M} \quad & w_1 T_{\text{sys}} + w_2 EE_{\text{sys}} \\
\text{subject to} \quad & \textstyle\sum_{m=1}^{M} b_m \le S; 0 \le b_m \le 1, \forall m. \\
& d \in \text{feasible range}
\end{aligned} \tag{27}$$

[13]Again, part of the solution approach can be incorporated with the more efficient Lagrange multiplier based approach.

Note that (27) reduces to the throughput and EE optimization problem when considering $w_1 = 1$, $w_2 = 0$ and $w_1 = 0$, $w_2 = 1$, respectively.

Proof. We prove Proposition 4 by contradiction. Assume that the optimal solution d_{o} and $\{b_{m,\text{o}}\}$ of (27) does not give a pareto-optimal solution. Then there must exist a d and $\{b_m\}$ such that $T_{\text{sys}}(\{b_m\}, d) > T_{\text{sys}}(\{b_{m,\text{o}}\}, d_{\text{o}})$ and $EE_{\text{sys}}(\{b_m\}, d) > EE_{\text{sys}}(\{b_{m,\text{o}}\}, d_{\text{o}})$ are satisfied. It follows that

$$\begin{aligned}
&w_1 T_{\text{sys}}(\{b_m\}, d) + w_2 EE_{\text{sys}}(\{b_m\}, d) \\
&\quad > w_1 T_{\text{sys}}(\{b_{m,\text{o}}\}, d_{\text{o}}) + w_2 EE_{\text{sys}}(\{b_{m,\text{o}}\}, d_{\text{o}}).
\end{aligned} \tag{28}$$

This contradicts that d_{o} and $\{b_{m,\text{o}}\}$ are optimal for (27). ■

To solve (27), the analyses in Secs. III and IV are exploited. We again focus on solving the problem with a fixed cooperation distance d. Then by considering the approximations in the analyses and given a fixed d, we observe that the objective function of (27) has a special structure:

$$h(\mathbf{x}) + \frac{f(\mathbf{x})}{g(\mathbf{x})}, \mathbf{x} \in \mathcal{B}, \tag{29}$$

where $h(\mathbf{x})$ is concave, $f(\mathbf{x})$ is concave, and $g(\mathbf{x})$ is convex over $\mathcal{B}$, respectively. We note that this stucture can be clearly identified by denoting $h(\mathbf{x}) = T_{\text{sys}}$, $f(\mathbf{x}) = T_{\text{sys}}$, and $g(\mathbf{x}) = P_{\text{sys}}$, and by using Propositions 1 and 2. Note that in [38], (29) has been proven to be NP-complete and an efficient approach to find the ϵ-approximation of the global optimal solution of (29) was proposed. Thus, results and techniques in [38] can be exploited to solve (29).[14]

VI. Numerical Results

This section provides numerical results to validate our analyses and evaluate the proposed designs. For all simulations in this paper, we consider the following parameters and setup:

[14]Although only the minimization counterpart of (29) was explicitly investigated in [38]. According to [38] and our own investigations, concept, results, and derivations can be applied to (29) after some modifications.

TABLE II: Summary of Parameters

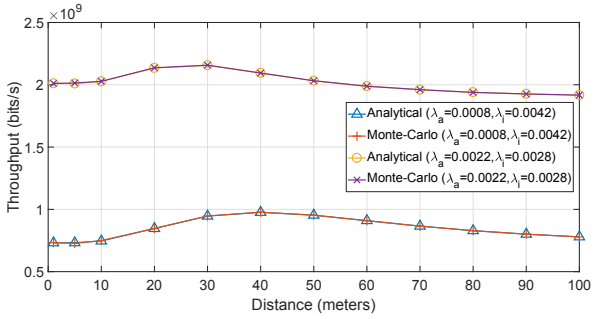
Parameters	Values/Descriptions
D; N; K; N_{BS}; $W_{\text{c,D2D}}$; W_{BS}	$D = 600$ m; $N = \frac{D}{d}$; $K = 16$; $N_{\text{BS}} = 100$; $W_{\text{c,D2D}} = 20$ MHz ; $W_{\text{BS}} = 20$ MHz
N_0 ; λ_{c}; α; d_0; ν	$N_0 = -174$ dBm/Hz; $\lambda_{\text{c}} = 2$ GHz; $\alpha = 3.68$; $d_0 = 5$ m; $\nu = 2^{\frac{\alpha}{2}} N_0 W$
λ_{a}; λ_{i}; λ_{u}	$\lambda_{\text{a}} = 0.0008, 0.0022, 0.0032\ m^{-2}$; $\lambda_{\text{i}} = \lambda_{\text{u}} - \lambda_{\text{a}}$; $\lambda_{\text{u}} = 0.0050\ m^{-2}$
T_{B}; T_{D}; T_{S}	$T_{\text{B}} = 200$ kbits/s; $T_{\text{D}} = 20$ Mbits/s; $T_{\text{S}} = T_{\text{D}}$
E_{B}; E_{D}	$E_{\text{B}} = 26$ dBm; $E_{\text{D}} \leq 23$ dBm is determined by the power control
S; M; γ ; q ; θ	$S = 10$; $M = 1000$; $\gamma = 0.6, 1.28$; $q = 0, 34$; $\theta = 1.8$

$D = 600$ m and $K = 16$. Also, we consider $d_0 = 5$ m, $\alpha = 3.68$, $\lambda_{\text{c}} = \frac{3\times10^8}{f_{\text{c}}}$, and $f_{\text{c}} = 2$ GHz in the path-loss model. The maximum allowable interference is set to be at the order of noise power, i.e., $\nu = 2^{\frac{\alpha}{2}} N_0 W_{\text{c,D2D}}$, where $N_0 = -174$ dBm/Hz is the noise power density and $W_{\text{c,D2D}} = 20$ MHz is the bandwidth for a D2D link. Thus the total bandwidth used for the D2D communications is 320 MHz. Although the theoretical framework in Secs. III, IV, and V does not consider the practical limit of the BS, in the simulations, unless otherwise indicated, we consider each BS link can use 200 kHz of bandwidth and 26 dBm of power consumption for transmission. Besides, we consider the BS to have 46 dBm total transmit power and $W_{\text{BS}} = 20$ MHz total bandwidth. Thus, the maximum number of users that can be served by the BS is $N_{\text{BS}} = 100$. By the aforementioned parameters, we consider $T_{\text{B}} = 200$ kbits/s and $T_{\text{D}} = T_{\text{S}} = 20$ Mbits/s; $E_{\text{B}} = 26$ dBm and $E_{\text{D}} \leq 23$ dBm is computed by (2). Thus, the cooperation distance d is within 100 meters. Note that here T_{B} and T_{D} are easily achievable in practice and $T_{\text{B}} = 200$ kbits/s can provide the video quality with 360p [42]. We consider $M = 1000$ and $S = 10$, and the request probabilities follow a MZipf distribution in [43], which has recently been extracted from a very large, real-world dataset, i.e., the BBC iPlayer dataset, with parameters γ and q:

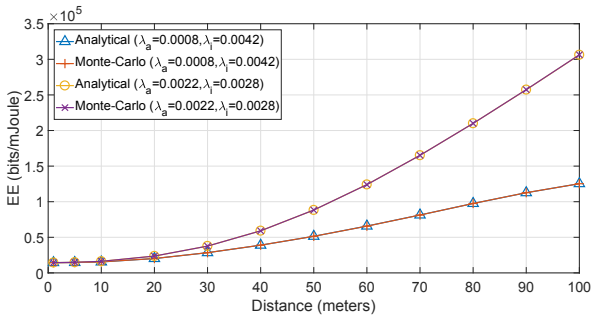
$$a_m = \frac{(m+q)^{-\gamma}}{\sum_{n=1}^{M}(n+q)^{-\gamma}}. \tag{30}$$

It can be seen that when $q = 0$, the MZipf distribution reduces to the commonly used Zipf distribution. To evaluate the proposed designs in practical situations, in the simulations, two parameter sets are used: $\gamma = 0.6$, $q = 0$ and $\gamma = 1.28$, $q = 34$. The first parameter set is from the UMass Amherst youtube experiment [16], which is widely used in the literature, and the second corresponds to the parameters reported in [43]. Furthermore, we adopt two density sets of users: $\lambda_{\text{a}} = 0.0008$ m^{-2} and $\lambda_{\text{i}} = 0.0042$ m^{-2}; $\lambda_{\text{a}} = 0.0022$ m^{-2} and $\lambda_{\text{i}} = 0.0028$ m^{-2}. Both of these have a considerable number of inactive users. These values were chosen because we are considering video streaming applications, in which each user could occupy a large amount of resources and even though the percentage of *data* that is used for video is very high, the percentage of *users* using video streaming at any time need not be; furthermore only a fraction of all cellphones in an area are active at all. Finally, for designs in prioritized-push networks, $\theta = 1.8$ is adopted according to some empirical experiences. The system parameters used in the simulations are summarized in Table II.

In Figs. 1, 2, and 3, to focus on evaluating the analysis results, we simulate the networks without considering the practical resource constraint, i.e., N_{BS} is temporarily assumed to be always sufficient in these figures. In Fig. 1a, we evaluate the proposed analyses of the random-push networks adopting $\gamma = 0.6$ and $q = 0$. When obtaining the results in Fig. 1, we adopt the caching policy designed by the proposed optimization in Sec. III.C for the random-push network. From the figures, we can observe that the analytical results are consistent with the Monte-Carlo results. Besides, it is interesting to observe that the EE increases with increasing cooperation distance. This is intuitive because when the cooperation distance, i.e., cluster size, increases, the probability that the user can find the desired file increases, leading to better EE. This is in contrast to the optimal throughput case where an increase of cooperation distance is not always good because it could decrease the number of clusters, leading to a lower total throughput. Note that although the increase of cooperation distance can also increase the power consumption of the D2D links and decrease the total throughput, the increase of probability of having the desired file in the D2D network, i.e., the hit-rate, is overwhelmingly important in the random-push network because the BS randomly picks one user to serve in this network and the BS power consumption is dominating. A

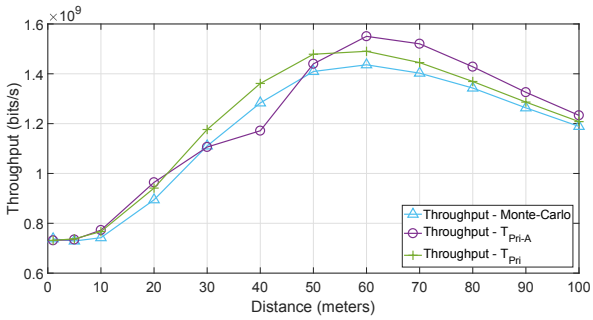


(a) Throughput.

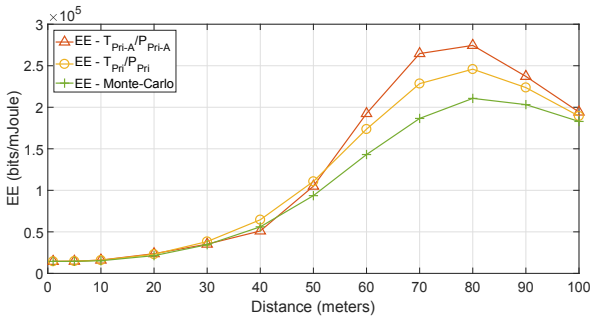


(b) EE.

Fig. 1: Evaluation of the proposed analyses in the random-push networks with $\gamma = 0.6$ and $q = 0$.



(a) Throughput.



(b) EE.

Fig. 2: Evaluation of the proposed analyses in the prioritized-push network with $\gamma = 0.6$, $q = 0$, $\lambda_a = 0.0008\ m^{-2}$, and $\lambda_i = 0.0042\ m^{-2}$.

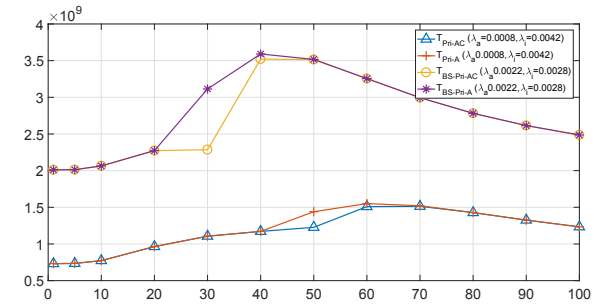


Fig. 3: Throughput comparisons between the approximations in prioritized-push networks with $\gamma = 0.6$ and $q = 0$.

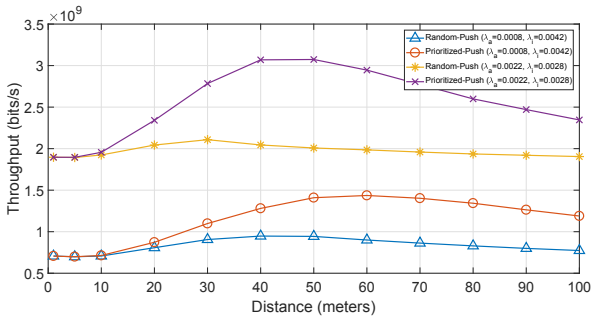


Fig. 4: Throughput comparisons between the random-push and prioritized-push networks with $\gamma = 0.6$ and $q = 0$.

different behavior is observed in the prioritized-push network.

In Figs. 2, we evaluate the proposed analytical results in the prioritized-push network adopting $\gamma = 0.6$, $q = 0$, $\lambda_a = 0.0008\ m^{-2}$, and $\lambda_i = 0.0042\ m^{-2}$. The adopted caching policy in the figure is designed by optimizing $N \cdot T_{\text{c,Pri-AC}}$ as discussed in Sec. III.C. In Fig. 2, curves labeled by $T_{\text{Pri}}(P_{\text{Pri}})$ are results of (12) and its power consumption counterpart; the curves labeled by $T_{\text{Pri-A}}(P_{\text{Pri-A}})$ are results jointly expressed by (15) and (17) and their power consumption counterparts. From the figure, we can observe that the proposed approximations can effectively characterize the trend of the Monte-Carlo result, though there is a gap between the analyses and the Monte-Carlo results. We note that for other combinations of densities and MZipf parameter set not shown here for space reasons, similar results are observed. In Fig. 3 we compare $T_{\text{c,Pri-A}}$ with $T_{\text{c,Pri-AC}}$ in the prioritized-push network adopting $\gamma = 0.6$ and $q = 0$ to validate the justification of using $T_{\text{c,Pri-AC}}$ for optimization. From the figure, we can observe that $T_{\text{c,Pri-AC2}}$ is obviously different only at $d = 30$ for $\lambda_a = 0.0022$ and $d = 50$ for $\lambda_a = 0.0008$, respectively, and these points are the closest points to the breaking-point and are not the optimal points. Note that although not shown here, we did compare the proposed designs with the designs obtained by directly optimizing $T_{\text{c,Pri-A}}$ using convex-concave procedure [40], which is a non-convex optimization, and saw no improvement.

In the remaining simulations, we consider the practical resource constraint and focus on the prioritized-push network since it is more spectrally efficient and practical. To validate that the prioritized-push network is more spectrally efficient than the random-push network. Fig. 4 compares their network throughput adopting $\gamma = 0.6$ and $q = 0$ and the same caching policy designed by optimizing $T_{\text{c,Pri-AC}}$. From the figure, we can readily see that the prioritized-push network offers better throughput. In Fig. 5, we consider $\gamma = 0.6$ and $q = 0$ and compare the prioritized-push networks with and without the practical resource constraint. The adopted caching policy in the figure is designed by the proposed throughput optimization in Sec. III.C. Note that since $N_{\text{BS}} = 100$, when considering the practical constraint, the BS can serve at most 100 users. The curves labeled by "Limited" indicate the results considering the practical resource constraint; the curves labeled by "Unlimited" indicate the results without considering the practical resource constraint. From the figure, we can observe that the difference between the curves are significant when d is less than 20 m, which are points that we are not interested in.[15]

In the following, we evaluate the proposed designs, i.e., the proposed throughout and EE designs, and compare them with the "Max-Hit-Rate" design proposed in [18] and the "selfish" design, in which each user caches the most popular files. By using the simulations, we also discuss the trade-off between throughput and EE. In Fig. 6, different caching designs are evaluated in terms of throughput and the adopted MZipf parameters are $\gamma = 0.6$ and $q = 0$. From the figures, we can observe that the proposed throughput and EE designs can provide very similar throughput performance.[16] The Max-

[15]Similar results can be observed for the EE case.

[16]The same results can be observed when considering the random-push networks.

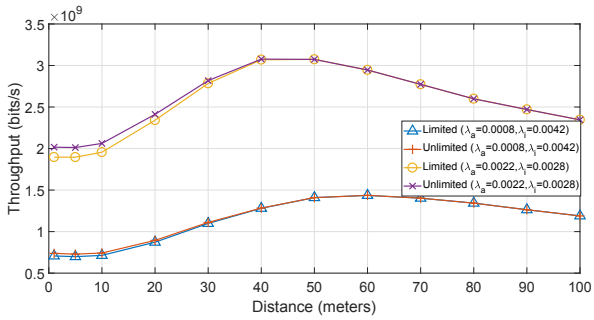
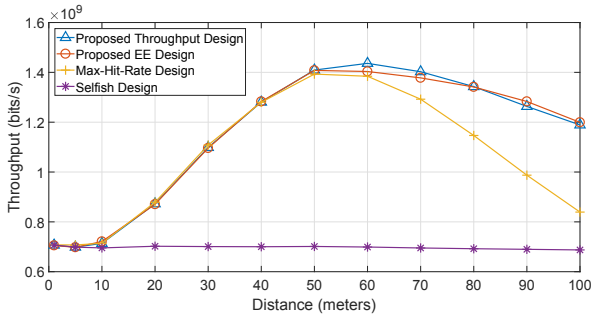
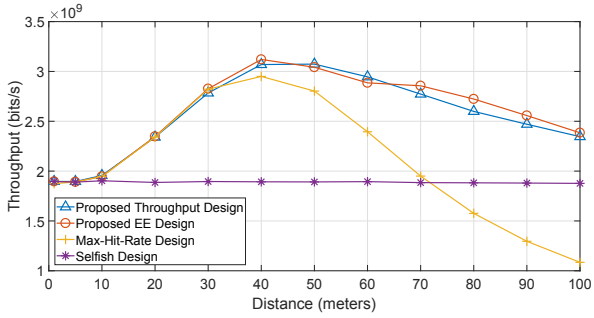


Fig. 5: Throughput comparisons between the networks with and without resource constraint in the prioritized-push network with $\gamma = 0.6$ and $q = 0$.



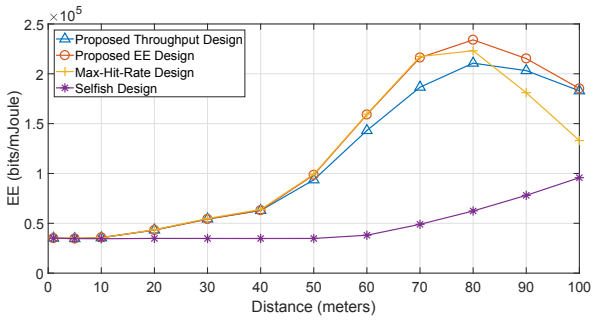
(a) $\lambda_a = 0.0008\ m^{-2}$ and $\lambda_i = 0.0042\ m^{-2}$.



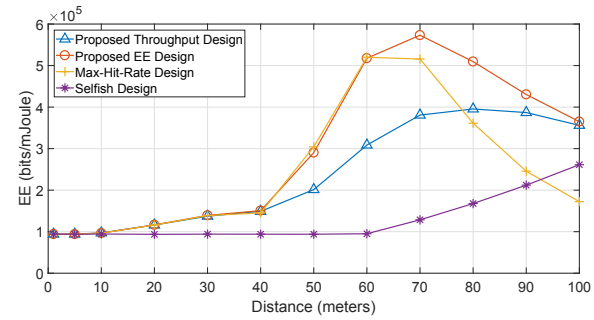
(b) $\lambda_a = 0.0022\ m^{-2}$ and $\lambda_i = 0.0028\ m^{-2}$.

Fig. 6: Throughput comparisons between different caching policies in the prioritized-push network with $\gamma = 0.6$ and $q = 0$.

Hit-Rate approach can provide an effective performance when the systems operate at a suitable cooperation distance, but its performance degrades significantly when the cooperation distance is large. This is because the Max-Hit-Rate approach cannot balance between using self-caching and D2D-caching, and thus the low frequency reuse gain due to the large cluster size could lead to a significant throughput degradation. This result indicates that the self-caching is influential and the effects of D2D-caching (with D2D communications) and self-caching should be jointly considered. This also indicates that when adopting the Max-Hit-Rate policy, it is safer to have a smaller cluster size rather than a larger cluster size to prevent the significant throughput degradation. The selfish approach performs poorly because it does not consider the benefits of D2D communications.



(a) $\lambda_a = 0.0008\ m^{-2}$ and $\lambda_i = 0.0042\ m^{-2}$.

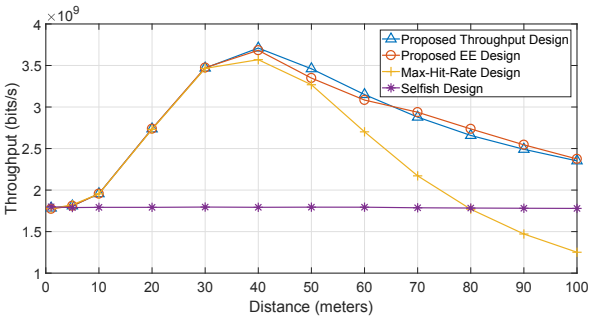


(b) $\lambda_a = 0.0022\ m^{-2}$ and $\lambda_i = 0.0028\ m^{-2}$.

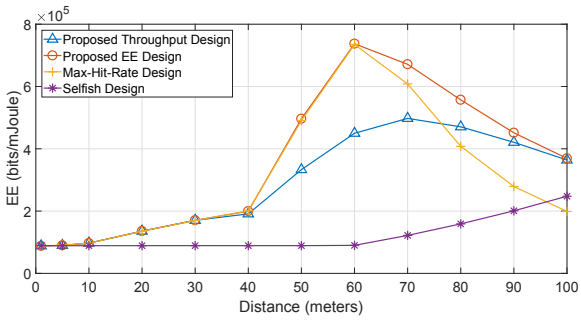
Fig. 7: EE comparisons between different caching policies in the prioritized-push network with $\gamma = 0.6$ and $q = 0$.

In Figs. 7, different caching designs are evaluated in terms of EE and the adopted MZipf parameters are $\gamma = 0.6$ and $q = 0$. From the figures, we can observe that the proposed EE design can offer the best EE performance. Again, the Max-Hit-Rate design is effective when the cooperation distance is appropriately selected, and the selfish design provides poor performance. By comparing between the throughput and EE evaluations, it can be observed that the optimal cooperation distances are different. This leads to the trade-off between throughput and EE when selecting different cooperation distances,[17] and the compromise can be taken by selecting the cooperation distances between these two optimal cooperation distances. We note that although the Max-Hit-Rate design could be effective in terms of throughput and EE when appropriately selecting the corresponding cooperation distances, respectively, it offers a less effective trade-off between throughput and EE. Besides, by comparing results in Figs. 6 and 7, we can see that the Max-Hit-Rate design starts to diverge from the best throughput and EE designs when the density of active users increases. This indicates that when the density of active users increases, the best policy starts to be different from pure cooperation. Finally, from all figures, we observe that when the density of active users increases, the optimal cooperation distance decreases, owing to the benefits of the frequency reuse and eliminating the necessity of having

[17]We also observe that, when considering a given cooperation distance, the proposed EE design can be near-optimal and optimal in terms of throughput and EE, respectively. This degrades the usefulness of the compromise design discussed in Sec. V. We thus omit the simulations using different weights of (27). That being said, we think that the provided mathematical framework in Sec. V could be useful in certain scenarios or other parameter sets.



(a) Throughput.



(b) EE.

Fig. 8: Performance comparisons between different caching policies in the prioritized-push network with $\gamma = 1.28$, $q = 34$, $\lambda_a = 0.0022\ m^{-2}$, and $\lambda_i = 0.0028\ m^{-2}$.

many active users in a cluster.

In Fig. 8, different caching designs are evaluated in the prioritized-push network adopting MZipf with $\gamma = 1.28$ and $q = 34$ and densities $\lambda_a = 0.0022$ and $\lambda_i = 0.0028$ in terms of throughput and EE. Similar results as in the previous figures can be observed, i.e., the effectiveness of the Max-Hit-Rate design and the trade-off between throughput and EE when adopting different cooperation distances. Overall, the simulation results show that, with practical popularity distributions, the trade-off between throughput and EE exists, and the trade-off can be adjusted by changing the cooperation distance. Besides, although the Max-Hit-Rate approach could provide good results when the cooperation distance is appropriately selected, it provides poor trade-off. Furthermore, the superior performance of the proposed designs as compared with the Max-Hit-Rate and selfish designs indicates that jointly considering the effects of D2D- and self-caching is important.

In Fig. 9, we compare between different policies using different densities of active and inactive users to see the influence in the same network as in Figs. 8. We can once again observe that the Max-Hit-Rate policy starts to diverge from the proposed throughput design when the density of active users increases, just as in Figs. 6 and 7. More interestingly, even when we increase the density of active users to a larger number, the selfish policy is still much worse than the others. This indicates that in the prioritized-push networks considering practical popularity distribution, a policy to be cooperative and to exploit the D2D communications should be more appropriate in terms of throughput. A similar result can also

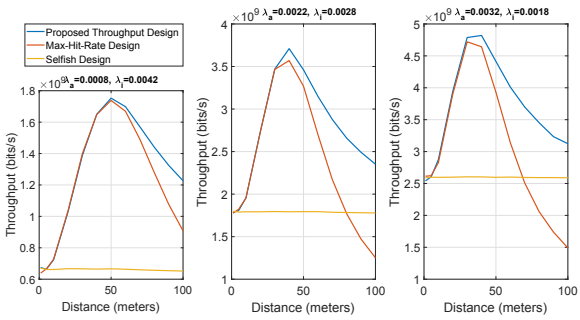


Fig. 9: Throughput comparisons between different caching policies and densities in the prioritized-push network with $\gamma = 1.28$ and $q = 34$.

be observed in terms of EE. We thus omit the corresponding figure for brevity.

VII. Conclusions

By considering the joint effects of BS-, D2D-, and self-caching and the impact of the cooperation distance, the design of caching policy and cooperation distance is investigated in the clustering BS-assisted wireless D2D caching network. Based on this setup, we analyze and optimize the network throughput and EE with two different network structures, i.e., random-push and prioritized-push networks. Note that the although the prioritized-push network is more spectrally efficient and practical, its analysis builds on the analysis of the random-push network. Since the throughput-based and EE-based designs could conflict with each other, to resolve this issue, we discuss the trade-off between them. From simulations, we conclude that the self-caching effect is influential and considering the joint effects of D2D- and self-caching is important. Besides, the proposed throughput and EE designs can outperform other designs and provide better trade-off because they can acquire the balance point between selfishness and cooperativeness. By comparing between the throughput and EE evaluations, it can be observed that their optimal cooperation distances are different. This leads to the trade-off between throughput and EE when selecting different cooperation distances.

This work focuses on the throughput- and EE-relevant designs and analyses. Consequently, the delay performance or any other delay-related constraints, such as outage performance, is not considered. The corresponding performance investigations and trade-offs are considered to be important future directions.

Appendix A
Proof of Lemma 1.2

Proof. Here we only prove the part regarding $\left[\sum_{m=1}^{M} a_m e^{-(\kappa_a+\kappa_i)b_m}\right]^{\kappa_a}$ because the part regarding $\left[\sum_{m=1}^{M} a_m e^{-\kappa_i b_m}\right]^{\kappa_a}$ can be similarly proved. Note that $\sum_{m=1}^{M} a_m e^{-(\kappa_a+\kappa_i)b_m}$ is convex and non-increasing with respect to $\mathcal{B}$, and x^{κ_a} is convex and non-decreasing when $\kappa_a \geq 1$ and $x \geq 0$. We denote

$\sum_{m=1}^{M} a_m e^{-(\kappa_a+\kappa_i)b_m}$ as $g(\mathbf{b})$, where $\mathbf{b} \in \mathcal{B}$; x^{κ_a} as $h(x)$. Thus, $\left[\sum_{m=1}^{M} a_m e^{-(\kappa_a+\kappa_i)b_m}\right]^{\kappa_a} = h(g(\mathbf{b}))$. Suppose $0 \leq \delta \leq 1$. We observe that

$$g\left(\delta \mathbf{b}_1 + (1-\delta)\mathbf{b}_2\right) \leq \delta g(\mathbf{b}_1) + (1-\delta) g(\mathbf{b}_2) \tag{31}$$

due to convexity, where $\mathbf{b}_1, \mathbf{b}_2, \delta\mathbf{b}_1 + (1-\delta)\mathbf{b}_2 \in \mathcal{B}$. Then noticing that $0 \leq g\left(\delta \mathbf{b}_1 + (1-\delta)\mathbf{b}_2\right) \leq 1$ and that $0 \leq \delta g(\mathbf{b}_1) + (1-\delta) g(\mathbf{b}_2) \leq 1$ due to the facts that $0 \leq g(\mathbf{b}_1) \leq 1$ and $0 \leq g(\mathbf{b}_2) \leq 1$, we know that

$$h\left(g\left(\delta \mathbf{b}_1 + (1-\delta)\mathbf{b}_2\right)\right) \leq h\left(\delta g(\mathbf{b}_1) + (1-\delta) g(\mathbf{b}_2)\right) \tag{32}$$

due to the non-decreasing property of $h(x), x \geq 0$. Finally, by combining the above results and that $h(x)$ is convex when $x \geq 0$, we have

$$\begin{aligned} h\left(g\left(\delta \mathbf{b}_1 + (1-\delta)\mathbf{b}_2\right)\right) &\leq h\left(\delta g(\mathbf{b}_1) + (1-\delta) g(\mathbf{b}_2)\right) \\ &\leq \delta h(g(\mathbf{b}_1)) + (1-\delta) h(g(\mathbf{b}_2)). \end{aligned} \tag{33}$$

This proves that $\left[\sum_{m=1}^{M} a_m e^{-(\kappa_a+\kappa_i)b_m}\right]^{\kappa_a}$ is convex. To prove that $\left[\sum_{m=1}^{M} a_m e^{-(\kappa_a+\kappa_i)b_m}\right]^{\kappa_a}$ is non-increasing, we simply notice that $\sum_{m=1}^{M} a_m e^{-(\kappa_a+\kappa_i)b_m}$ is non-increasing and x^{κ_a} is non-decreasing when considering the feasible set $\mathcal{B}$. ■

ACKNOWLEDGMENT

The authors would like to thank Professor Michael James Neely and Professor Chenyang Yang for helpful discussions.

REFERENCES

[1] "Cisco Virtual Networking Index: Global Mobile Data Traffic Forcast Updatae, 2016-2021," San Jose, CA, USA.

[2] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Commmun. Mag.*, vol. 51, no. 4, pp. 142-149, Apr. 2013.

[3] A. F. Molisch, G. Caire, D. Ott, J. R. Foerster, D. Bethanabhotla, and M. Ji, Caching eliminates the wireless bottleneck in video aware wireless networks, Adv. Elect. Eng., vol. 2014, Nov. 2014, Art. ID 261390.

[4] D. Liu, B. Chen, C. Yang, and A. F. Molisch, "Caching at the wireless edge: Design aspects, challenges, and future directions," *IEEE Commmun. Mag.*, vol. 54, no. 9, pp. 22-28, Sep. 2016.

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

[6] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," *IEEE ICC*, Jun. 2015.

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

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

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

[10] A. Liu and V. K. N. Lau, "Exploiting base station caching in MIMO cellular networks: Opportunistic cooperation for video streaming," *IEEE Trans. Sig. Process.*, vol. 63, no. 1, pp. 57-69, Jan. 2015.

[11] M. A. Maddah-Ali and U. Niessen, "Fundamental limits of caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2856-2867, May 2014.

[12] M. A. Maddah-Ali and U. Niesen, "Coding for caching: Fundamental limits and practical challenges," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 23-29, Aug. 2016.

[13] M. K. Kiskani and H. R. Sadjadpour, "Multihop caching-aided coded multicasting for the next generation of cellular networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 3, pp. 2576-2585, Mar. 2017

[14] K. Doppler, M. Rinne, C. Wijting, C. B. Ribeiro, and K. Hugl, "Device-to-device communication as an underlay to LTE-advanced networks,", *IEEE Commmun. Mag.,* vol. 47, no. 12, pp. 42-49, Dec. 2009.

[15] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Commmun. Mag.*, vol. 51, no. 4, pp. 142-149, Apr. 2013.

[16] N. Golrezaei, P. Mansourifard, A. F. Molisch, and A. G. Dimakis, "Base-Station assisted device-to-device communications for high-throughput wireless video networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, pp. 3665-3676, Jul. 2014.

[17] J. Rao, H. Feng, C. Yang, Z. Chen, and B. Xia, "Optimal caching placement for D2D assisted wireless caching networks," *in Proc. IEEE ICC,* May 2016.

[18] Z. Chen, N. Pappas, and M. Kountouris, "Probabilistic caching in wireless D2D networks: cache hit optimal vs. throughput optimal," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 584-587, Mar. 2017.

[19] B. Chen, C. Yang, G. Wang, "High-Throughput opportunistic cooperative device-to-device communications with caching," *IEEE Trans. Veh. Technol.,* vol. 66, no. 8, pp. 7527-7539, Aug. 2017.

[20] N. Golrezaei, A. D. Dimakis, A. F. Molisch, "Scaling behavior for device-to-device communications with distributed caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4286-4298, Ju. 2014.

[21] M. Ji, G. Caire, and A. F. Molisch "Wireless device-to-device caching networks: Basic principles and system performance," *IEEE J. Sel. Area Commun.*, vol. 34, no. 1, pp. 176-189, Jan. 2016.

[22] M. Ji, G. Caire, and A. F. Molisch "Fundamental limits of caching in wireless D2D networks," *IEEE Trans. Inf. Theory*, vol. 62, no. 2, pp. 849-864, Feb. 2016.

[23] M. Ji, G. Gaire, and A. F. Molisch, "The throughput-outage tradeoff of wireless one-hop caching networks," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6833-6859, Dec. 2015.

[24] M. Gregori, J. Gómez-Vilardebó, J. Matamoros, and D. Gündüz, "Wireless content caching for small cell and D2D networks," *IEEE J. Sel. Area Commun.,* vol. 34, no. 5, pp. 1222-1234, May 2016.

[25] D. Malak, M. Al-Shalash, and J. G. Andrews, "Optimizing content caching to maximize the density of successful receptions in device-to-device networking," *IEEE Trans. Commun.*, vol. 64, no. 10, pp. 4365-4380, Oct. 2016.

[26] Y. Wang, X. Tao, X. Zhang, and Y. Gu, "Cooperative Caching Placement in Cache-Enabled D2D Underlaid Cellular Network," *IEEE Commun. Lett.,* vol. 21, no. 5, pp. 1151-1154, May 2017.

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

[28] L. Zhang, M. Xiao, G. Wu, and S. Li, "Efficient scheduling and power allocation for D2D-assisted wireless caching networks," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2438-2452, June 2016

[29] K. Poularakis and L. Tassiulas, "Exploiting user mobility for wireless content delivery," *in Proc. IEEE ISIT*, Jul. 2013.

[30] R. Wang, J. Zhang, S. H. Song, and K. B. Letaief, "Mobility-Aware Caching in D2D Networks," *IEEE Trans. Wireless Commun.,* vol. 16, no. 8, pp. 5001-5015, Aug. 2017.

[31] D. Karamshuk, N. Sastry, M. Al-Bassam, A. Secker, and J. Chandaria, "Take-Away TV: Recharging wok commutes with predictive preloading of catch-up TV content," *IEEE J. Sel. Commun.*, vol. 34, no. 8, pp. 2091-2101, Aug. 2016.

[32] M.-C. Lee, A. F. Molisch, N. Sastry, and A. Raman, "Individual preference probability modeling for video content in wireless caching networks," *IEEE GLOBECOM*, Dec. 2017.

[33] M.-C. Lee and A. F. Molisch, "Individual preference aware caching policy design for energy-efficient wireless D2D communications," *IEEE GLOBECOM*, Dec. 2017.

[34] B. Chen and C. Yang, "Caching policy for cache-enabled D2D communications by learning user preference," arXiv preprint, arXiv:1707.08409v1, Jul. 2017.

[35] Y. Guo, L. Duan, and R. Zhang, "Cooperative local caching under heterogeneous file preferences," *IEEE Trans. Commun.*, vol. 65, no. 1, pp. 444-457, Jan. 2017.

[36] M.-C. Lee and A. F. Molisch, "On the caching policy and cooperation distance design in base station assisted wireless D2D networks," *in Proc. IEEE ICC*, May 2018.

[37] M. Ehrgott, *Multicriteria Optimization.* Springer-Verlag New York, 2005.

[38] R. W. Freund and F. Jarre, "Solving the sum-of-ratios problem by an interior-point method," *J. Global Optimization,* vol. 19, no. 1, pp. 83-102, 2001.
[39] A. F. Molisch, *Wireless Communications*, IEEE Press-Wiley, 2nd ed., 2011.
[40] B. K. Sriperumbudur and G. R. Lanckriet, "On the convergence of the concave-convex procedure," *Advances in Neural Information Processing Systems 22 (NIPS 2009)*, pp. 1759–1767, 2009.
[41] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.
[42] S. Lederer, C. Müller, and C. Timmerer, "Dynamic adaptive streaming over HTTP dataset," *in Proc. ACM the 3rd Multimedia Systems Conference*, pp. 89-94, 2012.
[43] M.-C. Lee, M. Ji, A. F. Molisch, and N. Sastry, "Performance of caching-based D2D video distribution with measured popularity distributions," *arXiv:1806.05380,* Jun. 2018.