Hindawi
Mobile Information Systems
Volume 2017, Article ID 8012517, 12 pages
https://doi.org/10.1155/2017/8012517

Research Article

A Device-to-Device Multicast Scheme for Delay-Constraint Content Delivery

Yanli Xu

Department of Information Engineering, Shanghai Maritime University, Shanghai 201306, China

Correspondence should be addressed to Yanli Xu; xylzoe1@163.com

Received 9 September 2016; Revised 6 January 2017; Accepted 15 January 2017; Published 8 February 2017

Academic Editor: Michele Garetto

Motivated by the explosive increase of mobile traffic, study on the device-to-device (D2D) communication is kicked off for content delivery through proximal transmission among users. D2D multicast has advantage on serving multiple users simultaneously with less resource cost. However, when D2D multicast is appropriate for content delivery and how to make it serve delay-constraint traffic are still unclear. In this paper, parameters impacting on D2D multicast content delivery is investigated to find good chances for utilizing D2D multicast. Furthermore, some rules to be obeyed are proposed for the content caching and delivery of D2D multicast to satisfy delay constraints. Based on these analyses, a delay-aware multicast scheme is proposed to maximize the network performance utility while satisfying delay constraints of contents. Simulations results verify our analyses and show that the proposed scheme can significantly improve multicast efficiency with guaranteed delay.

1. Introduction

Considering video streaming currently accounts for almost 50 percent of mobile data traffic with a 500-fold increase over the next 10 years [1], the idea of caching contents proactively at mobile users to offload network traffic has received great attentions recently [2–4]. As device-to-device (D2D) communication is standardized by 3GPP, proximate users are permitted to be connected with each other under the assistance of cellular network, which motivates the study of D2D communication in many ways as reviewed in [5]. For example, information delivery mechanism is studied to offload data from base station (BS) in [6]. Interference mitigation and utilization are studied to integrate D2D to cellular networks in [7]. The performance of D2D-aided cellular networks is tried to be optimized in [8]. One of these hot topics on D2D is the content dissemination where user equipment (UE) can obtain required contents from proximal UE which cache them in advance.

To introduce D2D content delivery, some beginning works have been done such as the design of architecture, caching, and delivery strategy to realize proactive caching and intelligent delivery at UE. For example, by assuming that D2D UE with common interests constructs a cluster for D2D content sharing, the average number of served requests per cluster and optimal cluster size are analyzed in [9, 10]. A tractable closed-form equation to find when redundant caching should be used in order to minimize the expected energy consumption was derived in [11]. With the consideration of D2D interference and fading, some works analyze the D2D content delivery performance in [12, 13]. Being aware that the social relationships among mobile users may be benefit for D2D content delivery, some researches investigate content delivery protocol by combining social and communication layers [14].

Most of existing works on D2D content delivery are based on the architecture of D2D unicast with request-and-response pattern, whereas multicast also has its advantages for proactive caching in 5G networks due to its better usage of wireless broadcast character as advocated in [15]. As an assistant method for cellular multicast, some works investigate D2D multicast content delivery for content recovery with the help of proximal UE in the same multicast group. For example, some agents are selected for multicasting contents and the selection scheme is proposed to offload data from BS in [16]. A jointly using D2D multicast and unicast scheme

is proposed to satisfy different requirements of UE in [17]. In [18], the grouping of D2D UE for content delivery is investigated to improve the energy efficiency. In [19], D2D communications are used for aiding cellular multicast to improve the performance of cell-edge devices.

However, D2D multicast pushes a content to multiple UE simultaneously and the content delivery cannot be adjusted by each content request like unicast and the quality of service (QoS) of delivery could not be guaranteed. Although a delay bound is necessary to ensure user experience for real-time video traffic with stringent delay requirements [20–22], most of existing works on D2D-aided content delivery analyze or design D2D delivery mechanisms without considering a delay constraint. Thereby, whether D2D multicast is appropriate for content delivery and how to schedule the multicast intelligently to realize the D2D multicast gain are still unclear to the best of our knowledge. This work aims at designing a delay-guaranteed multicast scheme by intelligently scheduling the content delivery and caching to fulfill the D2D multicast content delivery in cellular networks. To achieve this goal, a D2D multicast scheme is proposed in which contents of UE are delivered through D2D multicast periodically. In this scheme, the interval between two multicast periods of a D2D transmitter and the caching strategy of a D2D receiver should satisfy some conditions. With study of these conditions, the multicast performance can be improved under guaranteed delay constraints.

The remaining parts of the paper are organized as follows. In Section 2, the system model is set up and the D2D multicast delivery problem is formulated. In Section 3, the management of content delivery and caching are studied to satisfy improve D2D multicast performance while satisfying delay constraints. Based on these analyses, an optimal D2D multicast content delivery scheme is proposed which can not only maximize the network utility but also satisfy delay constraints of contents. Simulation results and related discussion are presented in Section 4. Finally, Section 5 concludes this paper. List of Symbols Section lists main symbols used in the paper.

2. System Model

2.1. Network Architecture. We consider an infinite planer cellular network where UE transmits data with two alternative transmission modes, that is, cellular mode in which transmission traverses BS and D2D mode in which proximal UE communicated with each other directly without the relay of BS. In the D2D mode, neighboring UE can be detected based on Received Signal Strength Indicator (RSSI) by listening signals from other UE in the discovery period. For UE, another UE can be regarded as a neighbor if the RSSI from this UE is beyond a threshold. Based on this method, some works discuss the discovery procedure in LTE networks [23, 24]. For each communication link, a better transmission mode can be selected to achieve higher system performance according to a certain mode selection scheme [25, 26]. D2D communication links employ spatial reuse for interference mitigation and higher resource efficiency thanks to proximal transmission. Communication links under different transmission modes use orthogonal resources for interference mitigation.

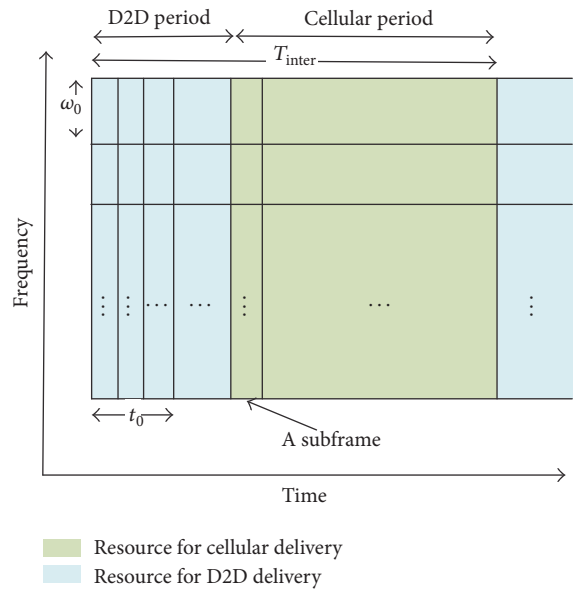


Figure 1: Resource allocation for D2D multicast.

Without loss of generality, we study a cell of the cellular network with radius R_0 as an interested system and study its content delivery performance for easy elaboration. Impacts from other cells such as interference are also considered for analyses. UE can obtain its required contents by listening to multicast contents from proximal transmitters. Transmitters are assumed to be distributed according to a Poisson point process (PPP) which is a popular model for characterizing locations of nodes in wireless environment and widely used for the analysis of D2D communication [27–29].

2.2. Channel Model. For communication link $i \to j$, the transmission over it is regarded to be successful when the SIR at the receiver is larger than threshold γ_{th}. We assume that the thermal noise is negligible and this assumption may be easily relaxed (e.g., see [30, 31]) but at the cost of complicating the derived expressions without providing additional insight. For the power control of transmitter, the transmitter chooses its transmission power such that the signal power at the intended receiver will be some designated constant η_0. For example, for a transmitter expected to cover a distance R, the transmission power is $\eta_0 R^\alpha$, where α is the path loss exponent. Some periodic resources with interval T_{inter} are allocated to D2D multicast for the content delivery (as shown in Figure 1). A multicast period includes several resource blocks (RBs) with bandwidth ω_0 and lasts several subframes with time t_0 in the frequency and time domains, respectively. For a multicast period, the volume of transmitted data over a link $i \to j$ is

$$R_{ij} = \omega_0 \log\left(1 + \gamma_j\right) t_0, \tag{1}$$

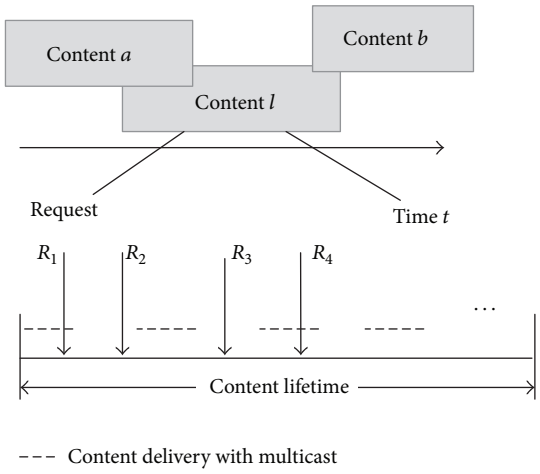


Figure 2: An example for D2D multicast content delivery operation: during lifetime of content l, user randomly generates request R_k and the request is served via D2D periodic multicast.

where γ_j is the received signal-interference-ratio (SIR) at the receiver j, which can be written as

$$\gamma_j = \frac{E_i d_{ij}^{-\alpha} H_{ij}}{I_j} = \frac{\eta_0 R^{\alpha} d_{ij}^{-\alpha} H_{ij}}{\sum_{k \in \Omega_j} E_k d_{kj}^{-\alpha} H_{kj}}, \quad (2)$$

where E_i denotes the transmission power of UE i, d_{ij} is the distance from the transmitter i to the receiver j, I_j is the interference at the receiver j from the simultaneous transmitter set Ω_j and H_{ij} characterizes the fast fading power from i to j.

2.3. Content Request and UE Experienced Delay. Each UE has a cache which can be populated with some video contents. Assume that the size of content l is F_l, l is successfully delivered during a multicast period if $R_{ij} \geq F_l$; that is, $\gamma_j \geq \gamma_{\text{th}}$, where $\gamma_{\text{th}} = 2^{F_l/\omega_0 t_0 - 1}$. As shown in Figure 2, multiple contents coexist in the network during an observation period. During the lifetime of l, request for the content is modeled by Poisson arrival process with arriving rate λ, which can be easily generalized for more practical cases based on queuing theory. UE can obtain the required content by listening to multicast contents from proximal UE. Different from request-and-response in unicast content delivery, the multicast content may be not required by UE receiving it right now. UE cached the multicast content in case that it needs this content in the future. Considering the limited capacity of a cache, UE discards cached content l with θ_l. Clearly, θ_l affects the multicast performance since it determines the amount of UE caching l and transmitting l. Thereby, the performance of D2D multicast can be improved by well-designed caching management policies in which UE chose caching time of video contents intelligently. UE may depart the system before it acquires its requested packet due to mobility. We assume that each UE leaves the system with rate μ.

UE may obtain a required content from precaching or from a multicast latter. Hence, whether UE can be served by D2D multicast is uncertain. As a key metric of real-time video traffic, the time delay experienced by UE requiring a content determines whether D2D multicast is appropriate for content delivery or not. Here, the time delay per content request t_d is used for characterizing the access delay affected by scheduling and the transmission time affected by channel quality and link congestion.

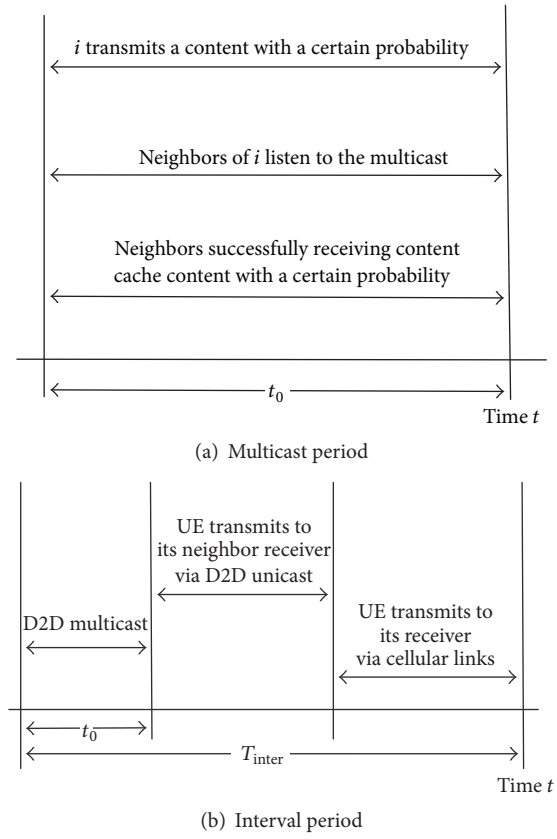


Figure 3: System behaviors of the D2D multicast scheme.

3. D2D Multicast for Delay-Constraint Content Delivery

In this section, we investigate how to apply D2D multicast to the content delivery with delay constraints. To solve this problem, the first focus is the problem formulation and system architecture setup for the D2D multicast. Then impacts induced by delay constraints on D2D multicast are studied in terms of transmission and interval periods, respectively. Lastly, a delay-constraint multicast scheme is proposed.

3.1. Preliminary of the Multicast Content Delivery Scheme. Unlike unicast, content delivery through multicast is a push strategy. Multicast contents of UE are transmitted to proximal UE periodically in allocated D2D resources. Taking the multicast of D2D UE i as an example, the system behaviors during the multicast and interval periods are illustrated in Figures 3(a) and 3(b), respectively. In the multicast scheme, D2D UE i which caches some contents transmits multicast

content at the beginning of a D2D period lasting t_0 (shown in Figure 3(a)). During t_0, BS schedules idle neighbor UE of i to listen to the multicast content. These neighbors of UE are searched during a neighbor discovery period in advance [32]. When these neighbors receive the multicast content, they determine whether to cache the content based on a caching strategy. During the interval period T_{inter} (shown in Figure 3(b)), the UE i is scheduled to communicate with other UE after the multicast period t_0 if there are communication requirements. Which mode is used for the communication, that is, D2D unicast or cellular, and the corresponding communication time under each mode are determined by a predefined mode selection scheme.

To maximize the delivery efficiency served by D2D multicast while guaranteeing the delay constraints of contents, we denote a utility as $U = \sum_{l=1}^{L} P_{D_l} g_l$ and formulate the optimal problem as follows:

$$\max \quad U = \sum_{l=1}^{L} P_{D_l} g_l, \tag{3a}$$

$$\text{s.t.} \quad \sum_{l=1}^{L} \theta_l \geq \theta_0, \tag{3b}$$

$$\theta_l \leq \theta_1, \tag{3c}$$

where L is the number of coexisting contents, P_{D_l} is the successful probability of UE receiving a content l under constraint D_l, and g_l is the probability that content l is required by the UE. Hence, the utility denotes the expectation value of a request served by D2D multicast. The caching time constrained by the cache capacity is shown in (3b). That is, there is a lower bound, θ_0, of the discarding rate of contents in a cache to avoid overflow. (3c) shows the constraint to caching time which guarantees enough cache contents at UE to satisfy content requirements through D2D multicast; that is, the discarding rate of contents is less than threshold θ_1.

3.2. The Transmission Management for D2D Multicast. From the optimal problem under given constraints in (3a), (3b), and (3c), we observe that a key factor impacting on the delay-constraint multicast is P_{D_l}, which is mainly determined by the transmission strategy in the multicast scheme. Thereby, we study how to manage the transmission of multicast UE in this subsection. Different from unicast, UE in the D2D multicast scheme received contents passively. UE is served by multicast if the received multicast content is required by this UE. For example, when UE j requires a content at time t, this content request is regarded to be satisfied by D2D multicast if the required content has been proactively cached by this UE and still not discarded at t. Thereby, the service quality of D2D multicast for content requests depends on the probability of UE successfully receiving its required content. This probability is determined by two parts: (1) the matching degree between multicast and required contents and (2) the probability that UE can successfully receive multicast content from a transmitter. Firstly, we present the successful probability that UE can obtain content l as follows.

Lemma 1. *The successful probability P_l of UE receiving a content l during a multicast period is*

$$P_l = q_l \frac{1 - \exp\left(-\kappa \rho_s R^2\right)}{\kappa \rho_s R^2}, \tag{4}$$

where q_l is the probability that l is transmitted by UE during an arbitrary multicast period, ρ_s is the transmitter density sharing the resource, $\kappa = \pi m \Gamma(m)(1-m)\gamma_{\text{th}}^m$, $m = 2/\alpha$, and $\Gamma(\cdot)$ is the Gamma function.

Proof. UE j can obtain l if and only if its transmitter sends l and this UE successfully receives it; P_l can be written as

$$P_l = \sum_{i \in \Omega_j} \Pr\{S_i = l\} \Pr\left\{tx_j = i\right\} \Pr\left\{F_{ij} \mid tx_j = i,\ S_i = l\right\}, \tag{5}$$

where $tx_j = i$ denotes the event that i is the transmitter UE j listening to, F_{ij} denotes an event that j successfully receives a content from UE i, and $S_i = l$ denotes an event that the multicast content from i is l. Since F_{ij} only depends on channel quality between i and j, we have [29]

$$\begin{aligned} \Pr\left\{F_{ij} \mid tx_j = i, S_i = l\right\} &= \Pr\left\{F_{ij} \mid tx_j = i\right\} \\ &= \exp\left(-\kappa \rho_s d_{ij}^2\right). \end{aligned} \tag{6}$$

Assume that the transmission probability of a content is independent identically distributed at each transmitter and denote $q_l = \Pr\{S_i = l\}$. Since potential transmitters of j are distributed uniformly with density ρ_s in the circle with radius R and origin j, we have

$$\Pr\left\{tx_j = i\right\} = \frac{1}{\int_0^{2\pi} \int_0^R \rho_s dr}, \tag{7}$$

P_l can be calculated by including (6) and (7) into (5) as follows:

$$\begin{aligned} P_l &= \sum_{i \in \Omega_j} q_l \frac{1}{\int_0^{2\pi} \int_0^R \rho_s dr} \exp\left(-\kappa \rho_s d_{ij}^2\right) \\ &= \frac{q_l}{\pi \rho_s R^2} \int_0^{2\pi} \int_0^R \rho_s \exp\left(-\kappa \rho_s r^2\right) r\, dr \\ &= q_l \frac{1 - \exp\left(-\kappa \rho_s R^2\right)}{\kappa \rho_s R^2}. \end{aligned} \tag{8}$$

□

For a multicast lasting t_0, the expected amount of UE that successfully received the multicast content is NP_l with the average rate $R_c = NP_l/t_0$, where $N = \pi \rho_r R_0^2$ is the expected number of receivers and ρ_r is the density of receivers. The probability P_D that UE successfully caches a multicast content before it is required by this UE depends on the successful acquisition probability P_l of the content in each multicast period. Since P_l in each multicast period is independent of each other and once UE obtains the content, it will cache it. Then it will not receive this content in the following time. Thus, P_D follows a geometric distribution.

Lemma 2. *Without considering the discarded contents, the probability that a request for l is covered by the D2D multicast before its delay constraint D can be expressed by*

$$P_D = 1 - \sum_{k=D}^{\infty} \left(1 - P_l\right)^k P_l = 1 - \left(1 - P_l\right)^D = 1 - \left[1 - q_l \frac{1 - \exp\left(-\kappa \rho_s R^2\right)}{\kappa \rho_s R^2}\right]^D. \tag{9}$$

P_D reflects the probability that a content requirement is covered by a multicast; thereby, it is also called a multicast coverage probability (MCP) here. As indicated, one significant advantage of multicast is that multiple UE may benefit from once multicast. With the perspective of efficiency, a multicast is expected to satisfy more current and future content requirements. Here we denote a multicast coverage threshold P_0 and guarantee the probability that UE located in the coverage of a transmitter benefited from a multicast via $P_D \geq P_0$. P_0 can be seen as the coverage requirement for a D2D multicast from UE. To achieve this goal, we present Theorem 3 as follows.

Theorem 3. *The transmission scheduling should satisfy (10) to guarantee that the MCP is larger than a threshold P_0:*

$$\kappa \rho_s R^2 \leq x_0, \tag{10}$$

and x_0 satisfies

$$\frac{1 - e^{-x_0}}{x_0} - \frac{1 - \left(1 - P_0\right)^{1/D}}{q_l} = 0. \tag{11}$$

Proof. Including P_D in (9) into inequality $P_D \geq P_0$, we have

$$1 - \left[1 - q_l \frac{1 - \exp\left(-\kappa \rho_s R^2\right)}{\kappa \rho_s R^2}\right]^D \geq P_0; \tag{12}$$

that is,

$$\frac{1 - \exp\left(-\kappa \rho_s R^2\right)}{\kappa \rho_s R^2} \geq \frac{1 - \left(1 - P_0\right)^{1/D}}{q_l}. \tag{13}$$

Let $x = \kappa \rho_s R^2$. Since

$$\frac{d\left(\left(1 - e^{-x}\right)/x\right)}{dx} = \frac{x e^{-x} - 1 + e^{-x}}{x^2} < 0, \tag{14}$$

the left part of above formulary decreases with x, and there is an upper bound x_0 for x to satisfy (13), that is, the coverage constraint P_0. Then x_0 is the solution of equation in (13); that is,

$$\frac{1 - e^{-x_0}}{x_0} - \frac{1 - \left(1 - P_0\right)^{1/D}}{q_l} = 0. \tag{15}$$

□

x_0 can be seen as a threshold for the multiplication of ρ_s, κ, and R^2. Similarly, there are bounds for the density ρ_s (determined by simultaneous transmitters which share the same resource), targeted distance R (determined by transmission power), and κ (determined by channel quality and decoding threshold at a receiver) due to the fact that x increases with the above parameters. In addition, from (15), we see that these upper bounds depend on the delay constraint D, the transmission strategy (q_l) at a transmitter, and targeted MCP threshold P_0.

3.3. The Interval Management for D2D Multicast. From the analyses in the last subsection, we see that the multicast performance of contents is affected by the management of transmission in terms of simultaneous transmitter density and transmission range as shown in Theorem 3. To further study the multicast performance in a time flow view and reveal the impact of delay constraints on the multicast scheme, the interval between two multicast periods in the multicast scheme is investigated in this subsection. Without loss of generality, we take the delivery of an arbitrary content l in the system as an example. The subscript l of some symbols is omitted for simplification of mathematical expressions. For D2D-supported content delivery, the delivery performance depends on the number of UE caching contents, $X(t)$, to a large extent, and the delay performance is dominatingly determined by queuing delay which depends on the amount of UE requesting this content, $Y(t)$. $X(t)$ decreases when UE with content leaves the network or discards content and $X(t)$ increases when UE successfully receives multicast content and caches it. Thereby, the variation of $X(t)$ can be expressed by (16) during an infinite small time slot dt:

$$\frac{dX(t)}{dt} = R_c - \mu X(t) - \theta X(t), \tag{16}$$

which is a first-order linear nonhomogeneous differential equation with general solution as follows:

$$X(t) = k \exp\left[-\left(\mu + \theta\right) t\right] + \frac{R_c}{\mu + \theta}, \tag{17}$$

where k can be derived by substituting $C(0) = c_0$ into (17) with the assumption that there is c_0 UE caching content l at the start time $t = 0$. Then we have $k = c_0 - (R_c/(\mu + \theta))$ and

$$X(t) = \left(c_0 - \frac{R_c}{\mu + \theta}\right) \exp\left[-\left(\mu + \theta\right) t\right] + \frac{R_c}{\mu + \theta}. \tag{18}$$

In the multicast strategy, a request for content l from UE at t is served if this UE caches l before t. Denoting the probability that l is requested at t as q, the number of served requests by multicast is $qX(t)$. Since the request arrival rate is λ and decreases with leaving UE with rate μ, the variation of $Y(t)$ can be expressed by

$$\frac{dY(t)}{dt} = \lambda - qX(t) - \mu Y(t) = \lambda - q\left\{\left(c_0 - \frac{R_c}{\mu + \theta}\right) \exp\left[-\left(\mu + \theta\right) t\right] + \frac{R_c}{\mu + \theta}\right\} - \mu Y(t). \tag{19}$$

Solving (19), we obtain

$$Y(t) = ke^{-\mu t} + \frac{\lambda - q\left(R_c/(\mu+\theta)\right)}{\mu} - \frac{q\left(c_0 - \left(R_c/(\mu+\theta)\right)\right)}{\theta} e^{-(\mu+\theta)t}. \tag{20}$$

Since UE which requests l at any time with probability q, the amount of UE requesting l follows binomial distribution at an instantaneous time. Due to the fact that the request arrival rate is λ, we have the relationship $\lambda \approx Nq$ based on the probability theory [33]. Then $Y(t)$ can be written by (21) by including $R_c = NP_l/t_0$ and $q = \lambda/N$ into (20):

$$Y(t) = ke^{-\mu t} + \frac{\lambda}{\mu}\left[1 - \frac{P_l}{(\mu+\theta)t_0}\right] - \frac{\lambda}{\theta}\left[\frac{c_0}{N} - \frac{P_l}{(\mu+\theta)t_0}\right] e^{-(\mu+\theta)t}. \tag{21}$$

Using $Y(0) = 0$, we have

$$k = \frac{\lambda}{\theta}\left[\frac{c_0}{N} - \frac{P_l}{(\mu+\theta)t_0}\right] - \frac{\lambda}{\mu}\left[1 - \frac{P_l}{(\mu+\theta)t_0}\right]. \tag{22}$$

Thus, $Y(t)$ can be written as

$$\begin{aligned} Y(t) &= \left\{\frac{\lambda}{\theta}\left[\frac{c_0}{N} - \frac{P_l}{(\mu+\theta)t_0}\right] - \frac{\lambda}{\mu}\left[1 - \frac{P_l}{(\mu+\theta)t_0}\right]\right\} \\ &\cdot e^{-\mu t} - \frac{\lambda}{\theta}\left[\frac{c_0}{N} - \frac{P_l}{(\mu+\theta)t_0}\right] e^{-(\mu+\theta)t} \\ &+ \frac{\lambda}{\mu}\left[1 - \frac{P_l}{(\mu+\theta)t_0}\right]. \end{aligned} \tag{23}$$

The steady state of D2D multicast content delivery can be obtained by

$$\frac{dX(t)}{dt} = \frac{dY(t)}{dt} = 0. \tag{24}$$

Denoting equilibrium values of $X(t)$ and $Y(t)$ as $\overline{X}$ and $\overline{Y}$, these equilibrium values can be obtained by including (16) and (19) into (24):

$$\overline{X} = \frac{R_c}{\mu+\theta}, \tag{25a}$$

$$\overline{Y} = \frac{\lambda}{\mu}\left[1 - \frac{P_l}{(\mu+\theta)t_0}\right]. \tag{25b}$$

According to Little's law which indicates that the average number of requests is equal to the multiplier of valid arrival rate and served time per request [34], the served time t_s per content request can be derived from

$$\overline{Y} = \lambda t_s. \tag{26}$$

Submitting (25b) into (26), t_s can be written as

$$t_s = \frac{\overline{Y}}{\lambda} = \frac{1}{\mu}\left[1 - \frac{P_l}{(\mu+\theta)t_0}\right]. \tag{27}$$

Considering that D2D UE is only scheduled in a multicast period for content delivery, the total delay t_d from UE requiring a content to it obtaining this content is affected by the scheduled period T_{inter} for D2D multicast. Given a service time t_s for a content request and a multicast period lasting t_0, the number of used multicast periods is $\mathcal{N}_{\text{mp}} = \lfloor t_s/t_0 \rfloor$ and $\lfloor \cdot \rfloor$ is the round down function. Then t_d can be written as

$$t_d = \mathcal{N}_{\text{mp}} T_{\text{inter}} + \left(t_s - \mathcal{N}_{\text{mp}} t_0\right). \tag{28}$$

Equation (28) shows that T_{inter} and t_0 affect the delay of a content served by D2D multicast. Hence, time resources should be intelligently allocated to D2D multicast to satisfy different delay constraints of contents. For an intelligent scheduling, we present Theorem 4 here.

Theorem 4. *For the D2D multicast content delivery, the resource allocation of D2D multicast should make the interval T_{inter} satisfy (29) to guarantee the delay QoS of a content request:*

$$T_{\text{inter}} \le \begin{cases} \dfrac{D - \left(t_s - \mathcal{N}_{\text{mp}} t_0\right)}{\mathcal{N}_{\text{mp}}} & t_s \ge t_0 \\ \infty & t_s < t_0, \end{cases} \tag{29}$$

where $\mathcal{N}_{\text{mp}} = \lfloor t_s/t_0 \rfloor$.

Proof. To guarantee that the delay for obtaining a required content via the service of D2D multicast satisfies delay constraint D, we need $t_d \le D$. Including (28) into this inequality can easily get the conclusion. □

Discussion. From Theorem 4, we can find that the interval between two multicast periods has an upper bound which supplies a reference for the practical resource allocation for D2D multicast according to a given content QoS requirement. In addition, it can be used for evaluating whether D2D multicast is appropriate for the delivery of content. For example, multicast is inappropriate to worse transmission environment due to larger time delay and more energy consumption caused by frequent transmissions. Furthermore, multicast is also inappropriate to the content with less popularity due to fewer number of benefit UE from the multicast while severer interference is induced by multicast intending to cover the farthest receiver.

3.4. A Delay-Aware Multicast Scheme for D2D Content Delivery. In above two subsections, the impact of delay constraint on the transmission and interval management are analyzed for the multicast scheme. Based on these analyses, we solve the optimal problem presented at the beginning of this section and propose a delay-aware multicast scheme for the D2D content delivery in this subsection.

From the optimal target in (3a), it is shown that the utility depends on the content requirement except the MCP characterizing the service performance of D2D multicast. To describe the probability of a content to be required by UE, the required statistics of contents is modeled by a Zipf distribution, which has been shown to fit well with video requirement [35, 36] and be widely used for the analysis of D2D content delivery [37, 38]. In the Zipf distribution, the probability g_l that content l is required by UE can be expressed by

$$g_l = \frac{1/l^\beta}{\sum_{k=1}^{L}\left(1/k^\beta\right)}, \tag{30}$$

where β is the Zipf exponent.

Institutively, content with more popularity should have longer cached time. However, the cache capacity of a device is limited. On one hand, it is expected that UE caches more contents for a longer time to make D2D multicast serve more content requests under different delay constraints. On the other hand, the cache capacity of a device is limited and it is impossible to cache abundant content without discarding them. These cache capacity and delay constraints are shown in (3b) and (3c), respectively. Given a service time bound $t_{s,l}^{\dagger}$ to constrain the service time for a request of content l, the constraint on θ_l in (3c) can be written as (31) by instituting t_s in (27) into inequality $t_s \le t_{s,l}^{\dagger}$ and solving the inequality

$$\theta_l \le \frac{P_l}{\left(1-\mu t_{s,l}^{\dagger}\right)t_0} - \mu. \tag{31}$$

The utility of (3a) can be written as follows by including (9) and (30) into (3a):

$$\begin{aligned} U &= \sum_{l=1}^{L}\left[1-\left(1-P_l\right)^{D_l}\right]\frac{1/l^\beta}{\sum_{k=1}^{L}\left(1/k^\beta\right)} \\ &= 1-\sum_{l=1}^{L}\left(1-P_l\right)^{D_l}\frac{1/l^\beta}{\sum_{l=1}^{L}\left(1/l^\beta\right)}. \end{aligned} \tag{32}$$

Letting $U' = \sum_{l=1}^{L}(1-P_l)^{D_l}((1/l^\beta)/\sum_{l=1}^{L}(1/l^\beta))$, we can maximize U through minimizing U'. The corresponding Lagrangian function $\mathscr{L}$ of the optimization problem is as follows:

$$\begin{aligned} \mathscr{L} = &\sum_{l=1}^{L}\left(1-P_l\right)^{D_l}\frac{1/l^\beta}{\sum_{l=1}^{L}\left(1/l^\beta\right)} - \phi\left(\sum_{l=1}^{L}\theta_l - \theta_0\right) \\ &+ \sum_{l=1}^{L}\varphi_l\left(\theta_l - \frac{P_l}{\left(1-\mu t_{s,l}^{\dagger}\right)t_0} + \mu\right). \end{aligned} \tag{33}$$

The KKT conditions are

$$\left.\frac{\partial\mathscr{L}}{\partial\theta_l}\right|_{\theta_l=\theta_l^*} = 0 \quad \forall l, \tag{34a}$$

$$\sum_{l=1}^{L}\theta_l^* = \theta_0, \tag{34b}$$

$$\frac{P_l}{\left(1-\mu t_{s,l}^{\dagger}\right)t_0} - \mu = \theta_l^* \quad \forall l. \tag{34c}$$

Taking derivation of $\mathscr{L}$ with respect to θ_l, we have

$$\begin{aligned} \frac{\partial\mathscr{L}}{\partial\theta_l} = &D_l\left(1-P_l\right)^{D_l-1}\frac{\partial P_l}{\partial\theta_l}\frac{1/l^\beta}{\sum_{l=1}^{L}\left(1/l^\beta\right)} - \phi \\ &+ \varphi_l\left(1 - \frac{1}{\left(1-\mu t_{s,l}^{\dagger}\right)t_0}\frac{\partial P_l}{\partial\theta}\right). \end{aligned} \tag{35}$$

Substituting (35) into (34a), we obtain

$$\begin{aligned} &\left.\frac{\partial P_l}{\partial\theta_l}\right|_{\theta_l=\theta_l^*} \\ &= \frac{\phi-\varphi_l}{D_l\left(1-P_l\right)^{D_l-1}\left(\left(1/l^\beta\right)/\sum_{l=1}^{L}\left(1/l^\beta\right)\right) - \varphi_l/\left(1-\mu t_{s,l}^{\dagger}\right)t_0}. \end{aligned} \tag{36}$$

Since

$$\begin{aligned} \left.\frac{\partial P_l}{\partial\theta_l}\right|_{\theta_l=\theta_l^*} &= q_l\frac{P_l e^{-P_l} - 1 + e^{-P_l}}{P_l^2}\left.\frac{\partial\rho_s}{\partial\theta_l}\right|_{\theta_l=\theta_l^*} \\ &= -q_l\frac{P_l e^{-P_l} - 1 + e^{-P_l}}{P_l^2}\frac{\kappa R^2 R_c}{\pi R_0^2\left(\mu+\theta_l^*\right)^2}, \end{aligned} \tag{37}$$

θ_l^* needs to satisfy (38) by including (37) into (36).

$$\begin{aligned} &\frac{\phi-\varphi_l}{D_l\left(1-P_l\right)^{D_l-1}\left(\left(1/l^\beta\right)/\sum_{l=1}^{L}\left(1/l^\beta\right)\right) - \left(\varphi_l/\left(1-\mu t_{s,l}^{\dagger}\right)t_0\right)} \\ &+ q_l\frac{P_l e^{-P_l} - 1 + e^{-P_l}}{P_l^2}\frac{\kappa R^2 R_c}{\pi R_0^2\left(\mu+\theta_l^*\right)^2} = 0. \end{aligned} \tag{38}$$

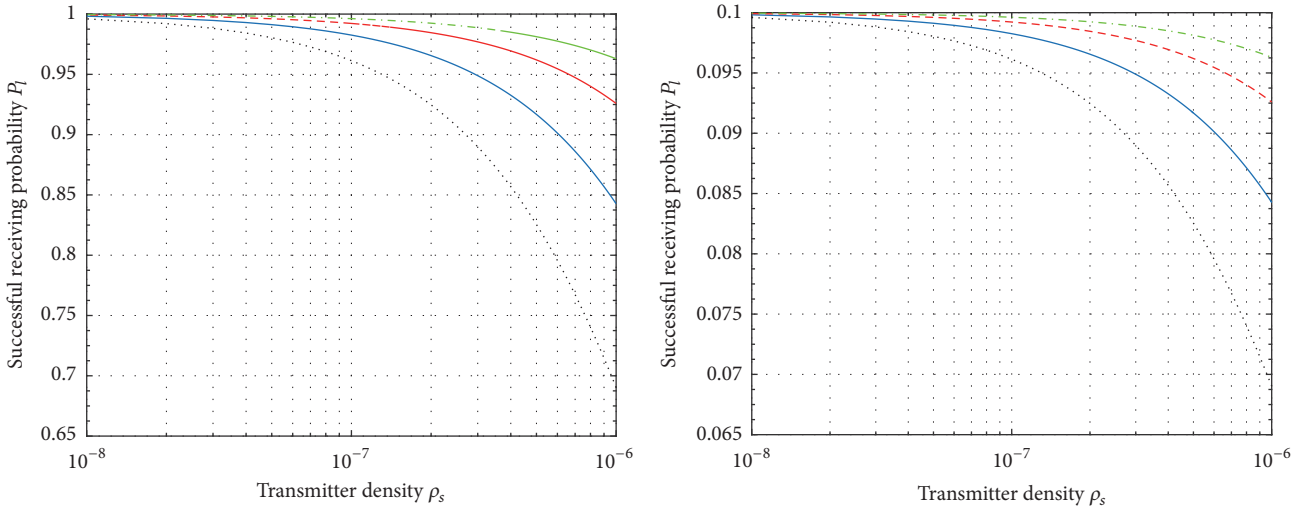
From (38), we see that θ_l^* is a function of ϕ and φ_l, which is denoted by $\theta_l^* = g(\phi^*, \varphi_l^*)$. Including this relationship into (34b) and (34c), we can obtain $L+1$-dimensional equation set as

$$\sum_{l=1}^{L} g\left(\phi^*, \varphi_l^*\right) = \theta_0, \tag{39a}$$

$$\frac{P_l}{\left(1-\mu t_{s,l}^{\dagger}\right)t_0} - \mu = g\left(\phi^*, \varphi_l^*\right) \quad \forall l. \tag{39b}$$

Solving the equation set yields to ϕ^* and φ_l^*; thereby, the optimal θ_l^* can be obtained.

From (38), we observe that the decreasing rate of $\overline{Y}$ with respect to θ_l needs to satisfy a given rate to guarantee the delivery performance while avoiding overflow as listed in Theorem 5.



(a) $q_l = 1$ (b) $q_l = 0.1$

FIGURE 4: Successful probability P_l of UE receiving a multicast content versus transmitter density ρ_s.

Theorem 5. *To maximize the expectation probability of a request served by D2D multicast with guaranteed delay QoS and avoid cache overflow, the decreasing rate of $\overline{Y}$ with respect to θ_l needs to satisfy*

$$\frac{\phi - \varphi_l}{D_l\left(1-P_l\right)^{D_l-1}\left(\left(1/l^{\beta}\right)/\sum_{l=1}^{L}\left(1/l^{\beta}\right)\right) - \left(\varphi_l/\left(1-\mu t_{s,l}^{\dagger}\right)t_0\right)} + q_l\frac{P_l e^{-P_l} - 1 + e^{-P_l}}{P_l^2}\kappa R^2\frac{\partial\overline{Y}}{\theta_l} = 0, \tag{40}$$

where ϕ and φ_l are constants.

4. Performance Evaluation

In this section, some simulation results will be presented based on analyses in Section 3 to evaluate the performance of D2D multicast and validate proposed schemes. The simulation experiments are performed on MATLAB platform. In the simulation, a cellular cell is targeted to be evaluated with radius R_0. In the cell, BS is located in the center and UE is uniformly distributed in the cell with a certain density and this UE density is the sum of transmitter density ρ_s and receiver density ρ_r. Content request arrivals at the simulation system with rate λ and UEs leaving the system with rate μ. D2D transmitters send multicast content every T_{inter} which lasts t_0. Receivers randomly select a transmitter within distance R and cache the successfully received contents. A communication link is regarded to be successful if the receiving SIR is larger than a threshold. Cached contents will be discarded with a certain rate, for example, θ_l for content l. The default used simulation parameters are shown in Table 1. These sets of parameters can be referred to simulation settings recommended by 3GPP for D2D communication in [32] or prior related works [39, 40].

TABLE 1: Simulation parameters.

Simulation parameters	Value
Density of D2D transmitters ρ_s	10^{-8}~10^{-5}
Density of receivers ρ_r	10^{-5}
Path loss factor α	3
SIR threshold at a receiver γ_{th}	10 dB
Allocated bandwidth ω_0	10 RB
Request rate λ	10^{-3}/s
UE leaving rate μ	10^{-3}/s
Multicast time in a multicast period t_0	10 ms
Multicast interval T_{inter}	50 ms
Radius of a cell R_0	1732 m
D2D cooperation distance R	100 m
Zipf exponent β	1
Delay constraint for a content D	5
Transmitting probability q_l	$1/L$
Number of contents L	10

P_l is used for evaluating the coverage probability of once multicast for requests of content l and results are shown in Figure 4. As shown in this figure, P_l decreases with the increase of R, ρ_s, and κ (led by increase of γ_{th} or decrease of α), which is in accordance with that of Theorem 3. Thus, we can determine whether the performance of D2D multicast is appropriate to content delivery or not based on Theorem 3 under different transmission environments. In addition, the comparison between two subfigures shows that the transmitting probability q_l of content l affects the performance significantly. Thereby, a popular content required by more UE is expected to transmitted with larger probability.

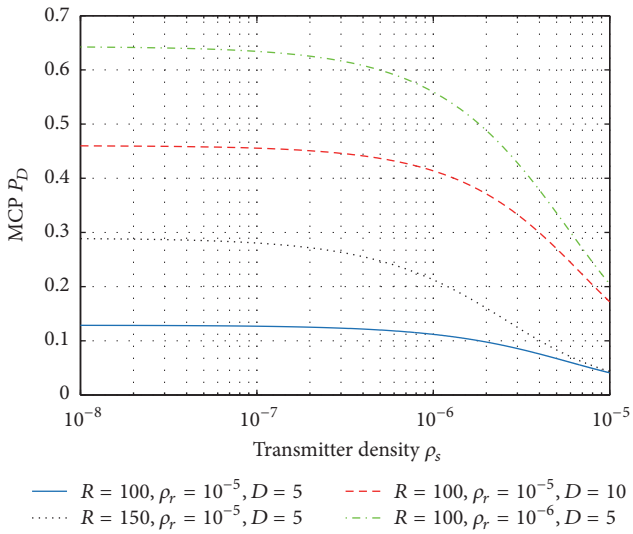


FIGURE 5: MCP P_D versus transmitter density ρ_s.

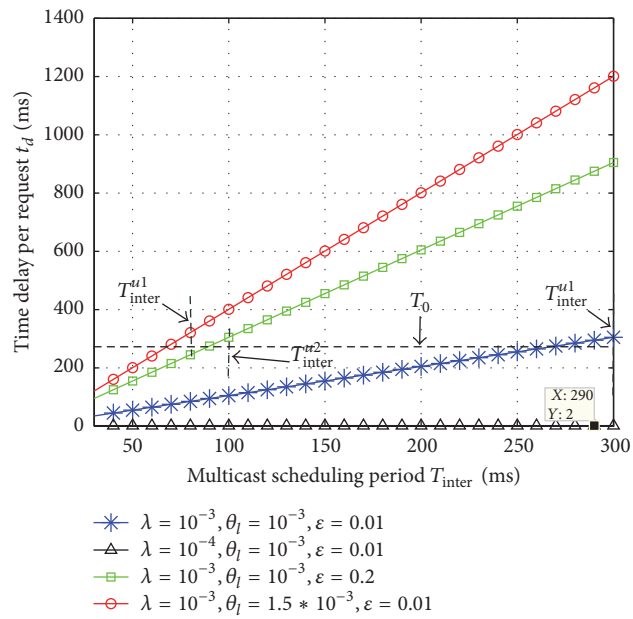


FIGURE 6: Comparison of time delay per request t_d for D2D multicast content delivery against the scheduled multicast period T_{inter}.

Figure 5 shows MCP P_D under different parameters to evaluate the service quality of D2D multicast when delay constraint is considered. D is the constraint to the number of multicast periods used for receiving a content. P_D also decreases with ρ_s like P_l due to the fact that P_D is determined by P_l. From this perspective, the improvement of P_l is valid for improving the performance of multicast content delivery. In addition, D2D multicast is not appropriate to higher density of simultaneous transmitters case due to severe interference. Furthermore, it is also indicated that P_D increases with ρ_r which shows a possible scenario for D2D multicast content delivery, that is, higher density of UE. This criteria of UE density can be easily derived from our analysis on P_D. Results of this figure also show that P_D performs better at larger delay constraint; that is to say, multicast is more appropriate for delay tolerant traffic. It does not mean that multicast is not suitable for delay-constraint traffic. However, some rules should be obeyed to guarantee a delay QoS as analyzed, which are evaluated in Figure 6.

The time delay per request t_d is used for evaluating multicast performance in terms of delay. t_d is the time experienced by UE from the start of its requirement of a content to its acquisition of this content. With different scheduled multicast periods T_{inter}, the variation of t_d is shown in Figure 6. Here we define $\varepsilon = 1 - P_l$ as the probability that the UE requesting a content fails to receive the content in a multicast period. It is shown that t_d is affected by many parameters such as θ_l, μ, λ, and P_l as analyzed. As illustrated in the figure, increasing P_l can reduce the time delay given certain values of μ, λ, and θ_l. Hence, we can adjust P_l by scheduling simultaneous transmitters ρ_s and transmission time t_0 of each multicast resource to obtain targeted delivery performance based on Theorem 3. Furthermore, it is shown that t_d increases with T_{inter}. Taking a time delay QoS constraint $T_0 = 300$ ms as an example (labeled by a dotted line in this figure), we can see that T_{inter} has upper bounds which can be calculated by Theorem 4 (labeled by T^{u1}_{inter}, T^{u2}_{inter}, and T^{u3}_{inter}, respectively) under different cases. When T_{inter} is smaller than these bounds, the time delay QoS requirement can be guaranteed under different cases. In addition, the served performance of content request is reduced by the increase of θ_l which indicates the importance of proactive caching for D2D content delivery. Thus, the caching time should be assigned to contents according to their popularity in practical applications. For example, considering the limitation of the cache capability of UE, the less popular content should be scheduled with shorter caching time, that is, larger θ_l.

Based on the proposed multicast scheme, Table 2 shows optimal θ_l^* under different parameters. Here $L = 2$ is used to evaluate the scheme. From this table, we observe that optimal θ_l^* depends on delay constraints for different contents, transmitted probability, and requested probability. θ_l^* increases with the increase q_l. That is because more frequent transmission leads to more caching UE which may excess the requirement to support requests for this content. Thus, θ_l^* increases to abort redundant contents for more useful contents. For similar reasons, θ_l^* is larger for contents with larger delay constraint. In addition, larger β means less popular of contents, which also increase the optimal aborting rate of contents. Hence, the proposed scheme provides a feasible way for determining caching time of contents to guarantee different delay constraints of contents.

The utility U is compared for different schemes as shown in Table 3. Schemes 1 and 3 are equally caching schemes; that is, θ_l is equal for different contents. In Scheme 1, $\theta_l = 10^{-4}$ for $l = 1, \ldots, 10$. In Scheme 3, $\theta_l = 0.5 * 10^{-4}$ for $l = 1, \ldots, 10$. Scheme 2 is the proposed multicast scheme

Table 2: Optimal θ_l^* based on proposed multicast scheme.

θ_1^*	θ_2^*	$ts_1^\dagger$	$ts_2^\dagger$	β	q_1	q_2
$0.3906 * 10^{-3}$	$0.3907 * 10^{-3}$	0.5	1	1	0.1	0.1
$0.3906 * 10^{-3}$	$0.3906 * 10^{-3}$	0.5	0.5	1	0.1	0.1
$0.3916 * 10^{-3}$	$0.3916 * 10^{-3}$	0.5	0.5	2	0.1	0.1
$0.4276 * 10^{-3}$	$0.0001 * 10^{-3}$	0.5	0.5	2	0.1	0.01

Table 3: Utility and constraint values for different schemes.

Scheme	U	Constraint values									
1	0.5509	−31.5087	−11.8975	−3.6336	1.7957	6.0873	9.8069	13.1995	16.3925	19.4475	22.4125	−142.0341
2	0.5405	−33.7779	−16.8240	−11.2045	−5.6443	−6.4263	−5.5780	−4.6975	−4.0801	−3.4157	−2.4422	−2.5717
3	0.4092	−7.7678	−7.7600	−7.7600	−7.7608	−7.7664	−7.7635	−7.7682	−7.7593	−7.7617	−7.7684	−22.3014

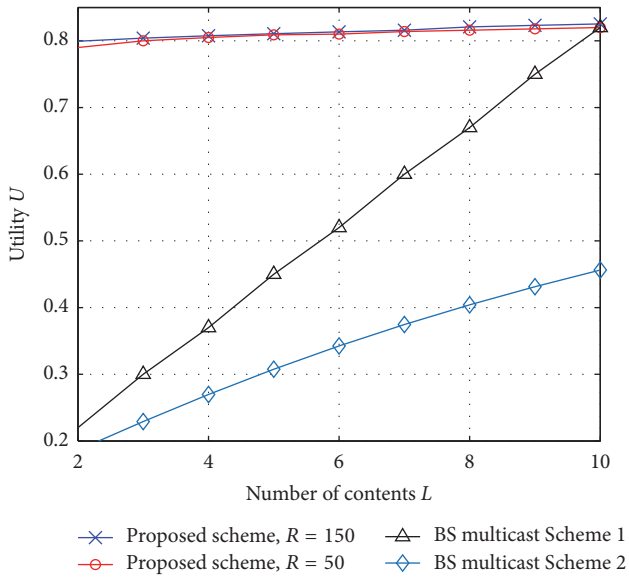


Figure 7: Comparison of utility among proposed D2D multicast content delivery scheme and BS multicast schemes.

where caching strategy is constrained by delay and optimized to maximize U. Thereby, we present the constraint values to show whether constraints are satisfied (calculated based on (3b) and (3c)). Negative values of these constraint values mean that the corresponding constraint is satisfied. From this table, we see that optimal scheme provides maximal utility value when constraints are satisfied. Therefore, the proposed multicast scheme can improve multicast efficiency for D2D content delivery with guaranteed delay constraints.

In Figure 7, the proposed scheme is compared to two BS-based multicast schemes when delay-constraints are satisfied by these schemes. In BS multicast Scheme 1, advantages of BS are considered; that is, BS caches all contents and knows some a priori knowledge such as content requirements. Hence, the multicast content from BS is assumed to be required by receiving UE. In BS multicast Scheme 2, BS only caches all contents without preknown content requirements like D2D UE. From this figure, we see that the proposed scheme significantly improves the BS multicast performance under different parameters and the performance can be improved up to three times. As the number of contents increases, the constraint of the UE cache and the decrease of q_l for each content lead to the performance reduction, while it is also better than BS multicast schemes.

5. Conclusion

In this paper, device-to-device (D2D) multicast is investigated to efficiently integrate D2D multicast for content delivery in cellular networks. To achieve this goal, a delay-constraint multicast scheme is proposed. The D2D multicast performance is analyzed in terms of once multicast transmission and time flow, respectively. Successful delivery probability of once multicast, multicast coverage probability (MCP) of serving content requirements with delay constraints, and the serving time per request are provided which give an insight into the relationship between different parameters and the service quality of D2D multicast. Based on these analyses, rules for multicast scheduling to obey are proposed, which provides qualitative rules of the delivery scheduling for contents with different delay constraints. The distributed caching is also studied, which provides aborting rate for contents, that is, staying time of each contents cached by UE. Based on the proposed D2D multicast scheme, better delivery performance is achieved with guaranteed delay. Simulation results show that the delivery performance (MCP) is dependent on many factors such as transmission performance P_l, cooperation range R, transmitter density ρ_s, and transmission probability q_l. The quality of D2D multicast can be improved by adjusting these parameters. Furthermore, proposed schemes can support a delay-guaranteed content delivery with three times performance improvement compared to cellular multicast. In conclusion, the work presented in this paper will be useful for the efficient implementation of content delivery via D2D multicast in cellular networks.

Symbols

η_0: Intended receiving power strength
α: Path loss factor
T_{inter}: Interval period between two multicasts
R_{ij}: The volume of transmitted data over link i to j
ω_0: Allocated bandwidth
d_{ij}: Distance between node i and j
γ_j: Received SIR at D2D UE j
H_{ij}: Fast fading power of the channel between i and j
R: Coverage distance of a multicast UE
F_l: Size of content l
γ_{th}: Decoding SIR threshold
E_i: Transmission power of UE i
θ_l: Discarding rate of content l at UE
λ: Arriving rate of content requests
μ: UE leaving rate
t_d: Time delay per content request
t_0: Multicast time per D2D period
P_{D_l}: The successful probability of UE receiving a content l under constraint D_l
P_l: Successful probability of UE receiving l
R_c: Average rate of UE receiving l
ρ_s: Transmitter density sharing the resource
q_l: The probability that l is transmitted by a UE
ρ_r: Density of receivers
P_D: Multicast coverage probability
$X(t)$: Number of pieces of UE caching l at t
$Y(t)$: Number of pieces of UE requesting l at t
t_s: Served time per content request.

Competing Interests

The author declares that there are no competing interests regarding the publication of this paper.

Acknowledgments

This work was supported in part by the Project of National Natural Science Foundation of China (no. 61601283, no. 61301110, and no. 61571108) and by the China Postdoctoral Science Foundation (no. 2016M590349).

References

[1] "Cisco visual networking index: Global mobile data traffic forecast update, 2015–2020 white paper," Cisco, white paper, 2016 http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html.

[2] B. Han, P. Hui, V. A. Kumar, M. V. Marathe, G. Pei, and A. Srinivasan, "Cellular traffic offloading through opportunistic communications: a case study," in *Proceedings of the 5th ACM Workshop on Challenged Networks (CHANTS '10)*, pp. 31–38, ACM, Chicago, Ill, USA, September 2010.

[3] H.-H. Cheng and K. C.-J. Lin, "Source selection and content dissemination for preference-aware traffic offloading," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 11, pp. 3160–3174, 2015.

[4] V. Sciancalepore, D. Giustiniano, A. Banchs, and A. Hossmann-Picu, "Offloading cellular traffic through opportunistic communications: analysis and optimization," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 1, pp. 122–137, 2016.

[5] A. Asadi, Q. Wang, and V. Mancuso, "A survey on device-to-device communication in cellular networks," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 4, pp. 1801–1819, 2014.

[6] M. Dai, B. Mao, D. Shen, X. Lin, H. Wang, and B. Chen, "Incorporating D2D to current cellular communication system," *Mobile Information Systems*, vol. 2016, Article ID 2732917, 7 pages, 2016.

[7] Y. Guo, J. Gao, and J. Hao, "Exploiting the user-level interference based on network coding in D2D underlaid cellular networks," *Mobile Information Systems*, vol. 2015, Article ID 142967, 8 pages, 2015.

[8] B. Fang, Z. Qian, W. Zhong, W. Shao, and H. Xue, "Coordinated precoding for D2D communications underlay uplink mimo cellular networks," *Mobile Information Systems*, vol. 2016, Article ID 1901952, 11 pages, 2016.

[9] A. Altieri, P. Piantanida, L. R. Vega, and C. G. Galarza, "A stochastic geometry approach to distributed caching in large wireless networks," in *Proceedings of the 11th International Symposium on Wireless Communications Systems (ISWCS '14)*, Barcelona, Spain, August 2014.

[10] 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 Transactions on Wireless Communications*, vol. 13, no. 7, pp. 3665–3676, 2014.

[11] J. Paakkonen, C. Hollanti, and O. Tirkkonen, "Device-to-device data storage for mobile cellular systems," in *Proceedings of the IEEE Globecom Workshops (GC Wkshps '13)*, pp. 671–676, Atlanta, Ga, USA, December 2013.

[12] D. Malak and M. Al-Shalash, "Optimal caching for device-to-device content distribution in 5G networks," in *Proceedings of the IEEE Globecom Workshops (GC Wkshps '14)*, pp. 863–868, IEEE, Austin, Tex, USA, December 2014.

[13] H. J. Kang, K. Y. Park, K. Cho, and C. G. Kang, "Mobile caching policies for device-to-device (D2D) content delivery networking," in *Proceedings of the IEEE Conference on Computer Communications Workshops*, pp. 299–304, IEEE, Toronto, Canada, May 2014.

[14] Z. Zheng, T. Wang, L. Song, Z. Han, and J. Wu, "Social-aware multi-file dissemination in device-to-device overlay networks," in *Proceedings of the IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS '14)*, pp. 219–220, IEEE, Ontario, Canada, May 2014.

[15] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: the role of proactive caching in 5G wireless networks," *IEEE Communications Magazine*, vol. 52, no. 8, pp. 82–89, 2014.

[16] M. Zulhasnine, C. Huang, and A. Srinivasan, "Exploiting cluster multicast for P2P streaming application in cellular system," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC '13)*, pp. 4493–4498, IEEE, Shanghai, China, April 2013.

[17] Y. Cao and A. Maaref, "Device cooperation-assisted scalable video multicast with heterogeneous QoE guarantees," in *Proceedings of the 11th International Symposium on Wireless Communications Systems (ISWCS '14)*, pp. 733–738, Barcelona, Spain, August 2014.

[18] Y. Zhang, F. Li, X. Ma, K. Wang, and X. Liu, "Cooperative energy-efficient content dissemination using coalition formation game over device-to-device communications," *Canadian Journal of Electrical and Computer Engineering*, vol. 39, no. 1, pp. 2–10, 2016.

[19] L. Militano, M. Condoluci, G. Araniti, A. Molinaro, A. Iera, and G.-M. Muntean, "Single frequency-based device-to-device-enhanced video delivery for evolved multimedia broadcast and multicast services," *IEEE Transactions on Broadcasting*, vol. 61, no. 2, pp. 263–278, 2015.

[20] G. Piro, L. A. Grieco, G. Boggia, R. Fortuna, and P. Camarda, "Two-level downlink scheduling for real-time multimedia services in LTE networks," *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 1052–1065, 2011.

[21] W. K. Lai and C.-L. Tang, "QoS-aware downlink packet scheduling for LTE networks," *Computer Networks*, vol. 57, no. 7, pp. 1689–1698, 2013.

[22] S. E. Ghoreishi, A. Aijaz, and A. H. Aghvami, "Delay-constrained video transmission: a power-efficient resource allocation approach for guaranteed perceptual quality," in *Proceedings of the 58th IEEE Global Communications Conference (GLOBECOM '15)*, pp. 1–7, IEEE, San Diego, Calif, USA, December 2015.

[23] H. Tang, Z. Ding, and B. C. Levy, "Enabling D2D communications through neighbor discovery in LTE cellular networks," *IEEE Transactions on Signal Processing*, vol. 62, no. 19, pp. 5157–5170, 2014.

[24] K. W. Choi and Z. Han, "Device-to-device discovery for proximity-based service in LTE-advanced system," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 1, pp. 55–66, 2015.

[25] D. D. Penda, L. Fu, and M. Johansson, "Mode selection for energy efficient D2D communications in dynamic TDD systems," in *Proceedings of the IEEE International Conference on Communications Workshops (ICC '15)*, pp. 5404–5409, London, UK, June 2015.

[26] Y. Xu, "A mode selection scheme for D2D communication in heterogeneous cellular networks," in *Proceedings of the IEEE Global Communications Conference (GLOBECOM '15)*, pp. 1–6, IEEE, San Diego, Calif, USA, December 2015.

[27] A. Al-Hourani, S. Kandeepan, and A. Jamalipour, "Stochastic geometry study on device-to-device communication as a disaster relief solution," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 5, pp. 3005–3017, 2016.

[28] A. Altieri, P. Piantanida, L. R. Vega, and C. G. Galarza, "On fundamental trade-offs of device-to-device communications in large wireless networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 9, pp. 4958–4971, 2015.

[29] Y. Xu and S. Wang, "Mode selection for energy efficient content delivery in cellular networks," *IEEE Communications Letters*, vol. 20, no. 4, pp. 728–731, 2016.

[30] S. Weber, J. G. Andrews, and N. Jindal, "The effect of fading, channel inversion, and threshold scheduling on ad hoc networks," *IEEE Transactions on Information Theory*, vol. 53, no. 11, pp. 4127–4149, 2007.

[31] N. Jindal, S. Weber, and J. G. Andrews, "Fractional power control for decentralized wireless networks," *IEEE Transactions on Wireless Communications*, vol. 7, no. 12, pp. 5482–5492, 2008.

[32] "3gpp technical report 36.843," Tech. Rep. 3GPP WG RAN1, 2014, http://www.3gpp.org/ftp/Specs/archive/36_series/36.843/.

[33] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*, Tata McGraw-Hill Education, 2002.

[34] D. Chhajed and T. J. Lowe, Eds., *Building Intuition: Insights from Basic Operations Management Models and Principles*, Springer US, 1st edition, 2008.

[35] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: evidence and implications," in *Proceedings of the 18th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '99)*, vol. 1, pp. 126–134, New York, NY, USA, March 1999.

[36] M. Busari and C. Williamson, "On the sensitivity of web proxy cache performance to workload characteristics," in *Proceedings of the 12th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '01)*, vol. 3, pp. 1225–1234, Anchorage, Alaska, USA, April 2001.

[37] E. Bastug, M. Bennis, and M. Debbah, "Social and spatial proactive caching for mobile data offloading," in *Proceedings of the IEEE International Conference on Communications Workshops (ICC '14)*, pp. 581–586, IEEE, Sydney, Australia, June 2014.

[38] K. Poularakis and L. Tassiulas, "Optimal selfishness-aware device-assisted content delivery in cellular networks," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC '14)*, pp. 2288–2293, Istanbul, Turkey, April 2014.

[39] D. Qiu and R. Srikant, "Modeling and performance analysis of bittorrent-like peer-to-peer networks," *SIGCOMM Computer Communication Review*, vol. 34, no. 4, pp. 367–378, 2004.

[40] A. Pyattaev, O. Galinina, S. Andreev, M. Katz, and Y. Koucheryavy, "Understanding practical limitations of network coding for assisted proximate communication," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 2, pp. 156–170, 2015.