Buffer-Aided Device-to-Device Communication: Opportunities and Challenges

Although buffer-aided protocols may provide significant throughput gains in wireless networks, the opportunities and challenges of buffer-aided D2D communications are not yet fully understood. Differing from most existing works that focus on investigating buffering policy, the authors analyze the fundamental impact of the constrained buffers on the D2D communication underlaying cellular system by an optimization framework.

Haoming Zhang, Yong Li, Depeng Jin, Mohammad Mehedi Hassan, Abdulhameed Alelaiwi, and Sheng Chen

Haoming Zhang, Yong Li, and Depeng Jin are with Tsinghua University. Yong Li is the corresponding author.

Mohammad M. Hassan and Abdulhameed Alelaiwi are with King Saud University.

Sheng Chen is with the University of Southampton, and also with King Abdulaziz University..

The authors would like to extend their sincere appreciation to the Deanship of Scientific Research at King Saud University for its funding of this research through the International Research Group project No. IRG-17.

COMMUNICATIONS STANDARDS

Abstract

To meet the increasing demands for popular content downloading services in next-generation cellular networks, device-to-device (D2D) communication was proposed to enable user equipments (UEs) to communicate directly over the D2D links in addition to traditional cellular operation by base stations (BSs), which is capable of utilizing the available cellular network's resource more efficiently to enhance content downloading performance. Although buffer-aided protocols may provide significant throughput gains in wireless networks, the opportunities and challenges of buffer-aided D2D communications are not yet fully understood. Differing from most existing works that focus on investigating buffering policy, we analyze the fundamental impact of the constrained buffers on the D2D communication underlaying cellular system by an optimization framework. Our study quantitatively reveals the positive correlation between the buffer sizes of BSs and UEs and the overall system performance, as well as further revealing the opportunities created by buffer-aided D2D communications for bandwidth conservation. In addition, we discuss practical challenges inherent in buffer-limited D2D communication underlaying next generation cellular networks, including increased transmission delay and optimal bandwidth allocation.

Introduction

As an underlay to LTE-A and fifth generation (5G) cellular networks, device-to-device (D2D) communication was introduced in Proximity Services (ProSe) in LTE Release-12 issued by 3GPP. D2D communications enhance many proximity-related services and applications, including content sharing and social networks. For local area services of popular content downloading, a few contents may be requested by a large number of users. Meeting this type of content downloading demand by cellular direct transmissions is extremely costly [1]. D2D communication enables user equipments (UEs) to communicate with each other directly on cellular resources [2], and may offer a high bit-rate and low power-consumption alternative. Specifically, D2D communications, in which UEs remain under the control of base stations (BSs) [2], take advantage of the physical proximity of communicating devices [1] and good channel conditions between them to better utilize the available resources.

Although D2D communication may enhance the performance of content-downloading systems, it can only take place when a UE is within the communication range of another UE or a BS that has the desired content, which indicates that the helper UE or BS must have stored the contents in its buffer in order to participate in D2D content downloading [3]. Therefore, the buffer sizes of both the BSs and UEs that serve as "relays" in the content-downloading paths play significant roles in the system performance and user experience, simply because the popular content must be stored in their limited storages so that the content can be transmitted to other UEs on appropriate occasions.

Nonetheless, existing studies [4–6] have not focused on the impact of limited buffer, a natural and indispensable attribute of UEs such as mobile phones, on the overall system performance. For example, in [4] D2D discovery processes are classified as either evolved packet core (EPC) network assisted discovery or direct discovery, and an energy-efficient D2D direct discovery is proposed, which facilitates D2D communications. Furthermore, current works fail to consider large-scale systems with hundreds of UEs [7, 8], and quantitative observations and conclusions are often reached under the unrealistic assumption that BSs and UEs have infinite storage. Thus, aiming to reveal the fundamental impact of the buffer on D2D communication underlaying cellular networks, we propose an optimization framework, a dynamic graph model that facilitates the analysis of system performance under optimal storage resource allocation and transmission control [9]. Based on this framework, we carry out the investigation under a practical network scenario with hundreds of UEs and multiple BSs. In addition to variable but limited buffer sizes of BSs and helpers, we also modulate the ratio of helpers to subscribers, and the allocation of the system bandwidth for cellular and D2D communications, which influence system performance as well. From the results and analysis, we draw conclusions regarding the opportunities and challenges created by the buffer, including boosting system performance, conserving bandwidth resources, and increasing transmission delay.

This article is structured as follows. We first provide an overview of D2D communication underlaying cellular networks. Then we propose a dynamic graph model and analyze the system constraints to form a weighed directional graph optimization model. With this optimization framework, we present our simulation results and analyze the positive impacts of the enlarged buffer, focusing on its theoretical performance bound. Next, we quantitatively analyze the boost-

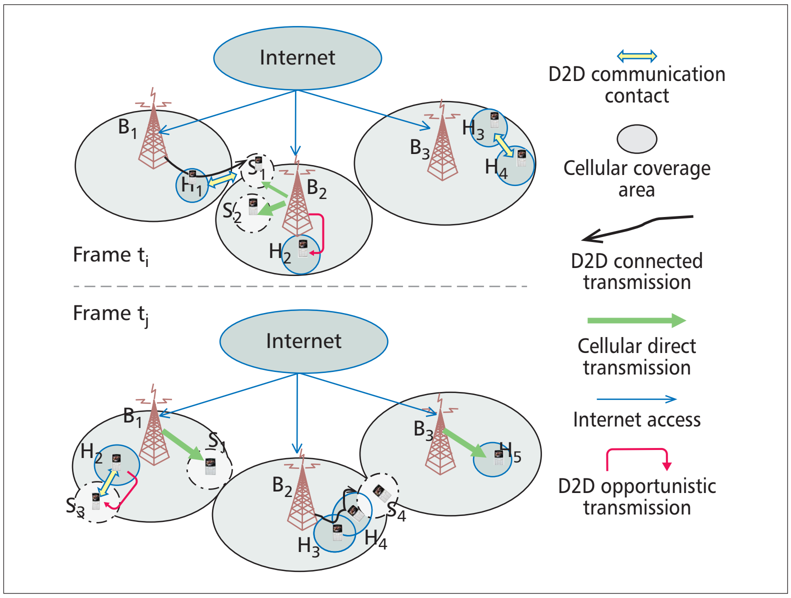
Internet
B1
B2
B3
H1
H2
H3
H4
S1
S2
Frame t_i
Frame t_j
Internet
B1
B2
B3
H2
H3
H4
H5
S1
S3
S4
D2D communication contact
Cellular coverage area
D2D connected transmission
Cellular direct transmission
Internet access
D2D opportunistic transmission

Figure 1. Illustration of D2D communication underlaying a cellular system, where UEs gain access to the cellular BSs or establish D2D communication.

ed system performance and the conserved bandwidth as well as the increased delay brought by enlarging the buffer. We also analyze the influence of the helper to subscriber ratio and the allocation of the system bandwidth to cellular and D2D communications on the achievable system performance. Finally, the article is concluded and further works are pointed out.

System Overview

A typical scenario of D2D underlaying content downloading cellular networks is illustrated in Fig. 1, where the BSs, whose coverage areas and buffers are circumscribed, are connected to the Internet to provide service to UEs. The buffer-constrained UEs are mobile nodes whose positions and access states change over time. Therefore, at different time frames, their physical locations and access relations are different. Here, a time frame is loosely used to mark a system time period during which access and physical relationships remain unchanged. For example, two different time frames, t_i and t_j $(i < j)$, are indicated in Fig. 1. In content sharing systems, UEs are naturally divided into two different groups in the time frames considered: the UEs that are requesting and downloading content are called subscribers, while other UEs that currently are not retrieving content for themselves are referred to as helpers. Helpers may participate in data transmission by receiving some content, storing it in their buffer, and then transmitting it to the relevant subscribers via D2D communication. For the example depicted in Fig. 1, there are three BSs denoted by B_1 to B_3, five helpers denoted by H_1 to H_5, and four subscribers denoted by S_1 to S_4, whose requested content is delivered to them from BSs and helpers. The dotted thin circles denote the communication ranges of subscribers, while those of helpers are denoted by solid thin circles. Apart from the original way of cellular direct transmission trough BSs, UEs can also receive data from helpers in the two D2D transmission modes defined below.

D2D Connected Transmission*:* Utilizing the physical proximity of user devices, connected transmission paths from BSs via some helpers to subscribers can be established. In Fig. 1, S_1 and H_1 have established D2D communication contact, and a connected path from B_1 via H_1 to S_1 is established so that B_1 is able to transmit content to S_1 with the aid of H_1, during time frame t_i. Similarly, B_2 is transmitting content to S_4 via the D2D connected path $B_2 \rightarrow H_3 \rightarrow H_4 \rightarrow S_4$, during time frame t_j. D2D connected communication is also known as relay assisted communication.

D2D Opportunistic Transmission*:* As UEs are naturally mobile, a D2D connected path is prone to be broken and the channel conditions always fluctuate. Nevertheless, a helper is able to store some content in its finite buffer and wait for the opportunity to transmit the data to a subscriber when it establishes a communication contact with the subscriber under good channel conditions. For example, in Fig. 1, H_2 has received data from B_2 during time frame t_i and has stored the data in its buffer. During time frame t_j, when H_2 establishes a contact with S_3, it transmits the data to S_3. D2D opportunistic communication is based on the store-carry-forward mechanism that exploits opportunistic connectivity and UE mobility.

In the system, the content is available at the initial period to the BSs, and the BSs, whose buffers are also far from unconstrained, are able to store the data temporally and deliver the content to the subscribers under appropriate circumstances, by means of cellular direct transmission, D2D connected transmission, and/or D2D opportunistic transmission. In order to model this sophisticated scenario and to analyze the impact of buffering on the D2D underlaying cellular network, we develop an optimization framework for evaluating the theoretical performance bound of the D2D underlaying content downloading cellular network.

Model and Analysis Framework

Graph Model and Objective

There are five types of network events (start of cellular accessing, start of D2D contact, end of cellular accessing, end of D2D contact, and change in link quality) that may affect the access relationship and D2D contact in the network. Consequently, continuous time can be divided into n time periods, and within each time period the access states of all the network participating nodes remain unchanged. In other words, during a time period between two successive events, called a time frame, neither any contact event nor any change in link quality occurs.

Clearly, we can acquire a static graph similar to Fig. 1 for every time frame. In order to include all potential transmission modes, the graph model should include all BSs and UEs in the network. Assume that there are b BSs labeled by the set of $\mathcal{B} = \{B_1, B_2, \cdots B_b\}$, h helpers labeled as $\mathcal{H} = \{H_1, H_2, \cdots, H_h\}$, and s subscribers labeled as $\mathcal{S} = \{S_1, S_2, \cdots, S_s\}$. We can use a node to represent a BS or a UE in a given time frame. Then a static graph model of each time frame includes $b + h + s$ nodes. The data flows between nodes (BSs, helpers, and subscribers) within the time frame can be represented by directed edges, among which the edges of D2D opportunistic transmissions are from helpers to subscribers (or other helpers) and the edges of D2D connected transmissions are from BSs via some helpers to subscribers, while those of cellular direct transmissions are directly from BSs to subscribers.

Because the buffer-aided D2D mechanism enables helpers and BSs to store the content in their local buffers at certain time frames and then transmit it in the coming frames, this mechanism based on finite data buffering enables data flows across time frames and accordingly makes it possible for us to model the time evolution of this time-varying system by static graphs. When we take n time frames into consideration, we can first put the n graphs of a single time frame together and then use directed edges across time frames to represent data flows in buffers. In other words, because data flows can transmit across time frames (but only from a time frame to its successive time frame), the static graph becomes a connected digraph. For example, in Fig. 2 all the possible transmission modes, cellular direct transmissions, D2D connected transmissions, and D2D opportunistic transmissions, are included. In Fig. 2, BSs and UEs are represented by vertices, and directed edges are added to UE vertices to represent the data flows by the cellular direct transmission and/or the D2D communication.

Next we can model data flows by attributing weights to the directed edges and make the connected digraph weighted. For instance, each directed edge in the same row, green arrows for direct cellular transmission and blue arrows for D2D transmissions in Fig. 2, is associated with a positive value representing the data flow transmitted within this time frame, whose upper bound is the product of the temporal link transmission rate and the time-frame duration. It should be emphasized that the directed edges from BSs and helpers to themselves between two successive time frames (red arrows) represent the data buffering of these nodes across the two successive time frames, and the positive weights associated with these directed edges correspond to their finite buffer capacities, i.e. the finite amounts of the data stored.

Furthermore, to model the Internet access of BSs, all the content is distributed to the BSs by the Internet source, denoted as ***S*** in Fig. 2, at the initial period before time frame t_0, which represents the content downloaded from the Internet during the time period considered. Similarly, the total amount of the data received by the subscribers, which is represented by $s \times (n + 1)$ directed edges with the infinite-large transmission rate from the subscribers to the imaginary destination, denoted as ***D*** in Fig. 2, can be used to evaluate system performance [9]. When participating in D2D opportunistic communication, helpers can use cellular resources to selectively download content from BSs and store it in their buffer, and then share it. As a result of a limited buffer, helpers cannot store all content desired by subscribers. Furthermore, BSs and helpers can keep the stored content in the selected time periods, which depends on the obtained system optimization results.

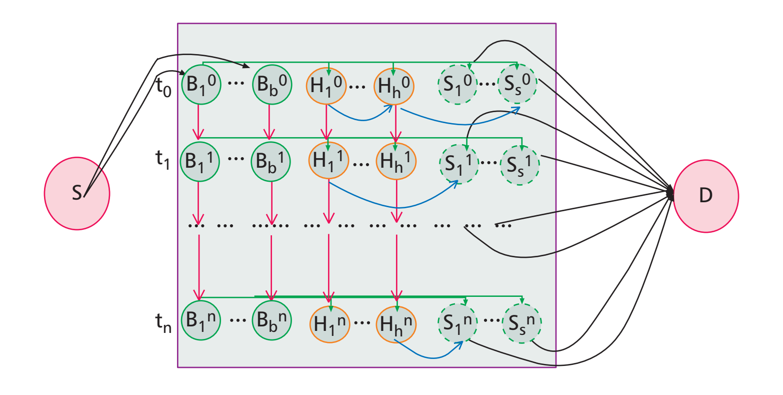
S
t_0
B_1^0 ··· B_b^0 H_1^0 ··· H_h^0 S_1^0 ··· S_s^0
t_1
B_1^1 ··· B_b^1 H_1^1 ··· H_h^1 S_1^1 ··· S_s^1
t_n
B_1^n ··· B_b^n H_1^n ··· H_h^n S_1^n ··· S_s^n
D

Figure 2. Static weighed directional graph model of the buffer-limited D2D communication underlaying cellular system.

To recap, all the BSs and UEs are involved in this weighted directed graph that models the temporal and spatial distributions of the network topology. Although the accessing relationships between UEs and BSs are dynamic and the communication contacts are time-varying, each row in the graph has the static topology for the duration of one time frame since the access states of all the participating network nodes remain unchanged for the duration of each frame. The objective of our optimization framework [9] is to maximize the total amount of data received by all subscribers, which is equal to the total amount of the flows to the destination ***D*** of Fig. 2.

System Constraints and Solutions

There exist three key system constraints in this buffer-limited D2D cellular network:

Flow Conservation*:* For any vertex in the graph, the amount of incoming flows must equal the amount of outgoing flows plus the amount of data stored if the vertex is a BS or helper.

Transmission Rate and Channel Access*:* Given the limited spectral resources for the D2D and cellular direct communications, the weight of each edge is directly associated with the allocated resource. Specifically, the total transmitted flow of each edge must meet the transmission bandwidth constraint. Moreover, the transmitted content flows must be strictly circumscribed within the connected UEs at each time frame, and they must also meet the interference requirements for channel access.

Finite Buffer*:* The buffer of a BS is constrained and the buffer of a helper is limited. Also, a BS typically has larger buffering capacity than a helper. To investigate the impact of a finite buffer on the theoretical performance bound of D2D communication underlaying cellular networks, we set the upper bound of the BS buffering data flows and that of the helper buffering data flows separately under the realistic assumption that every BS or helper has a limited buffer size.

After combining the above-introduced objective and constraints, we form a maximization

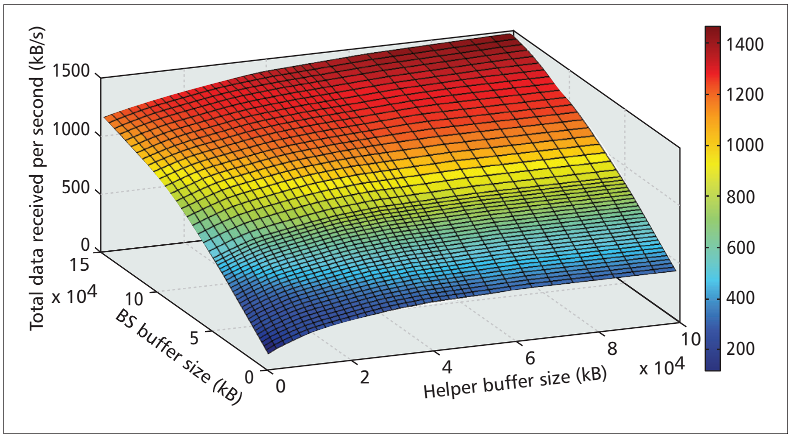

Figure 3. General trend of the total data received per second when both BS buffer size and helper buffer size are constrained and variable.

problem with the decision variables denoted by the set $\mathcal{C}$, which consists of all the data flows, i.e. the weights of all the directional edges in Fig. 2. A challenge is that not all the associated constraints are linear constraints, and thus the problem does not belong to the category of linear programming problems. Nevertheless, these nonlinear constraints can be transformed into linear expressions using the reformulation linearization technique (RLT) [10]. Consequently, this maximization problem can be solved using the existing optimization tool kits, such as CPLEX [11] and YALMIP [12].

Utilization of the Buffer

Intuitively, enlarging the buffer contributes to system performance and is accordingly a potential way of conserving bandwidth, but it will also result in an unavoidable increase in content delivering delay. Specifically, larger BS buffer sizes enable the BSs to receive more content from the source initially and to wait for appropriate opportunities to transmit them, while enlarging the buffers of helpers enables them to store sufficient amounts of data and to wait for appropriate D2D transmission opportunities to transmit more data to subscribers. By contrast, more limited buffer capacities will restrict the achievable system performance more severely.

To quantitatively exemplify the positive impacts of enlarging the buffer on the total amount of data received by subscribers, we implement simulations under a network scenario with 15 BSs and 100 UEs, among which 25 UEs are subscribers and the others are helpers. In order to yield general results, the network begins in zero-state, meaning that helpers have not retrieved content in the past and begin with an empty buffer. The number of UEs is sufficient for establishing D2D communications, and the human mobility model self-similar least action walk (SLAW) [13] is used to implement the traces of the simulated UEs. We use the typical settings in SLAW [13], where the Hurst parameter for self-similarity of waypoints is set to 0.75, the clustering range is set to 50 m, the Levy exponent for pause time is set to 1, the minimum pause time is set to 30 s, and the maximum pause time is set to 3600 s. The cell radius is 400 m and the D2D communication distance is 50 m. Since each LTE physical resource block (PRB) consists of 12 subcarriers with typically 15 kHz spacing, we allocate each UE with 800 kHz of bandwidth resources (approximately equal to four to five PRBs) to participate in cellular direct and D2D communications. Seventy percent of bandwidth resources is used for cellular direct transmissions, while the other 30 percent is allocated for D2D transmissions. In the simulation, we concentrate on investigating the influence of buffering, and we only consider the intra-cell interference, i.e. calculating the link transmission rate by only considering the interference caused by the nodes sharing the same spectrum resources [14]. We point out that there exist physical-layer techniques that can effectively manage inter-cell interference [6].

Figure 3 shows the general trend in the impacts of BS buffering and helper buffering on the capability of the system. It can be seen from Fig. 3 that there exists a significant positive correlation between the BS buffer size and the total data received per second (TDRPS) by all the subscribers, which indicates that enlarging the BS buffer contributes strongly to the enhanced performance of the entire system in the 1000-second simulation period. On the other hand, although enlarging the helper storage also has a positive impact on the system's achievable performance, it is much less effective compared to increasing the BS storage, especially when the helper buffer size is more than 50 MB. More specifically, given 100 MB of BS buffer, the TDRPS ascends only approximately 2.3 percent when the helper buffer increases from 51 MB to 100 MB, as can be clearly seen in Fig. 3. This is in contrast to more than 39 percent performance improvement due to increasing the BS buffer from 51 MB to 100 MB, with a fixed 100 MB helper buffer.

Since the significant performance improvement results from enlarging the BS buffer, a D2D content-downloading system can achieve the same required TDRPS performance with less bandwidth resources by increasing the BS buffer size. In Fig. 4a each line fitted to the selected simulation points has approximately a constant TDRPS. The results of Fig. 4a clearly demonstrate that the demand for bandwidth drops sharply with the increase in BS buffer size, given the same TDRPS requirement. This indicates that we can trade off the BS buffer size with the bandwidth. For example, with a 15 MB BS buffer, the total cellular bandwidth required is more than 700 kHz to achieve the 2 MB TDRPS, while with the 81 MB BS buffer, the system only needs 440 kHz bandwidth to achieve the same TDRPS performance. Thus, enlarging the BS buffer size can be utilized to enable the system to maintain the same level of TDRPS performance with less bandwidth. Of course, the BS buffer costs much less than cellular bandwidth.

Although enlarging the buffer offers an effective means of enhancing system performance, it will also increase data delivery delay. The average data delay in this 1000-second simulation period is calculated by computing the weighted arithmetic mean of the mid-times of every time frame, with the normalized weights set according to the total amount of data received

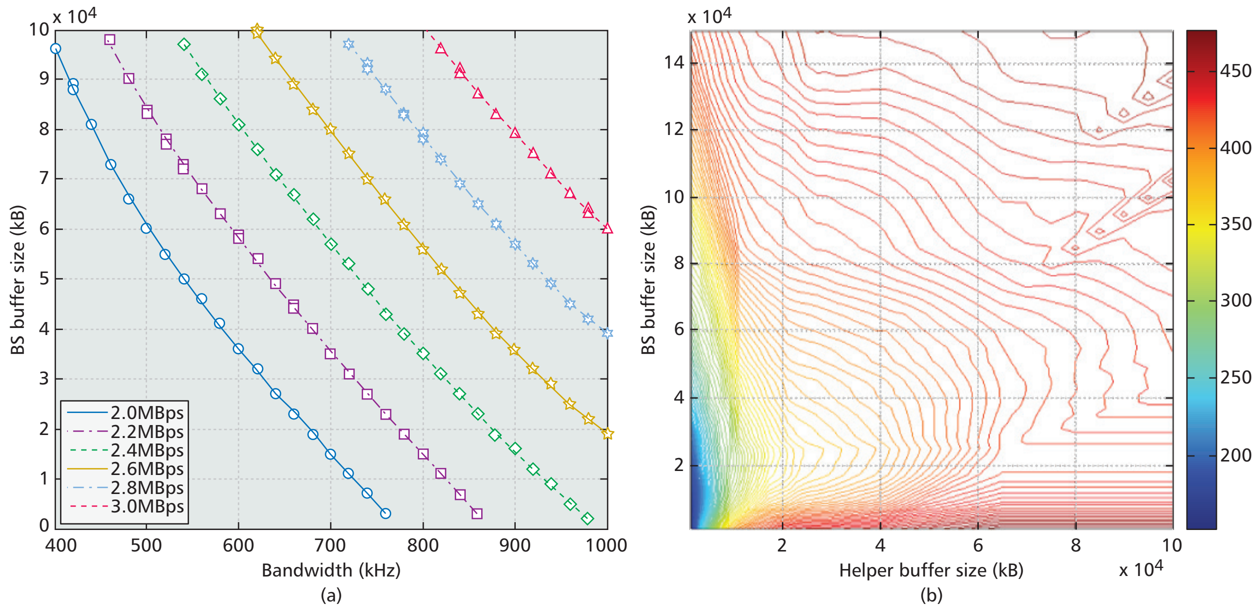

Figure 4. a) Relationship between the BS buffer size and the demand for bandwidth to meet the given TDRPS requirement; b) average data delivery delay as a function of the BS and helper buffer sizes, in the buffer-aided D2D content downloading.

in each frame, respectively. As shown in Fig. 4b, an increase in either the BS buffer or the helper buffer will result in a longer delay. The main reason for this unavoidable delay is that D2D opportunistic communication, which relies on the mobility of mobile devices, requires the helpers to store content temporally and to wait for opportunistic communication contacts.

Clearly, a long delay is always undesired as delay also impacts the user experience. For cellular services that are sensitive to both transmission delay and throughput, special protocols should be designed to circumscribe buffer size as well as the proportion of the data delivered by D2D transmission. In particular, for real-time applications, users should rely on cellular direct transmission instead of the D2D option in order to meet quality of service (QoS) requirements. However, certain delay is permissible in content downloading because this content is not real-time sensitive. Specifically, most users care more about the downloading rate but pay less attention to how long the data has stayed in the buffer of another device. In other words, it is the TDRPS instead of delay that mainly determines system performance and user experience in content-downloading services. Furthermore, with more users involved in D2D opportunistic communications, communication contacts occur more frequently, which can significantly accelerate the downloading speed of popular content. With a large proportion of content downloading services shifted to relying on D2D transmission, the network can in turn free more cellular direct transmission resources for real-time applications.

Further Discussions

In a buffer-limited D2D content-downloading underlaying cellular system, how the total system bandwidth is divided between cellular direct communication and D2D communication as well as the ratio of helpers to subscribers given the total number of UEs also influence the achievable performance. By carrying out further simulations to study the influence of these two parameters, our empirical results suggest that to achieve a reasonable optimal value of TDRPS, the proportion of the cellular direct-transmission bandwidth over the total system bandwidth should be in the range of 0.6 to 0.8, while the proportion of subscribers given the total number of UEs should be in the range of 0.5 to 0.65, respectively. Furthermore, other important issues, such as UE requirements, UE compensation, security, and energy consumption, are also discussed here.

Bandwidth Allocation

In contrast to the traditional D2D technologies that usually work on the crowded 2.4 GHz unlicensed band, in the D2D underlaying cellular network, D2D communication shares the bandwidth with cellular direct transmission. An appropriate allocation of the system bandwidth between these two communication modes is important for meeting the required system performance. After performing the simulation study under the same practical network setting (15 BSs, 25 subscribers, and 75 helpers), we acquire the results depicted in Fig. 5a and Fig. 5b for the variable BS buffer size and variable helper buffer size, respectively. The bandwidth resources allocated to each UE is also 800 kHz. Additionally, in Fig. 5a the helper buffer size is fixed (60 MB), while in Fig. 5b the BS buffer size is fixed (60 MB). Although the absolute measures may slightly fluctuate due to different mobility patterns, it is clear that a cellular direct-transmission bandwidth proportion in the range of 0.6 to 0.8 achieves the highest TDRPS. In this range, the impact of the helper buffer size is important when it is smaller than 30 MB. But a comparison between Fig. 5a

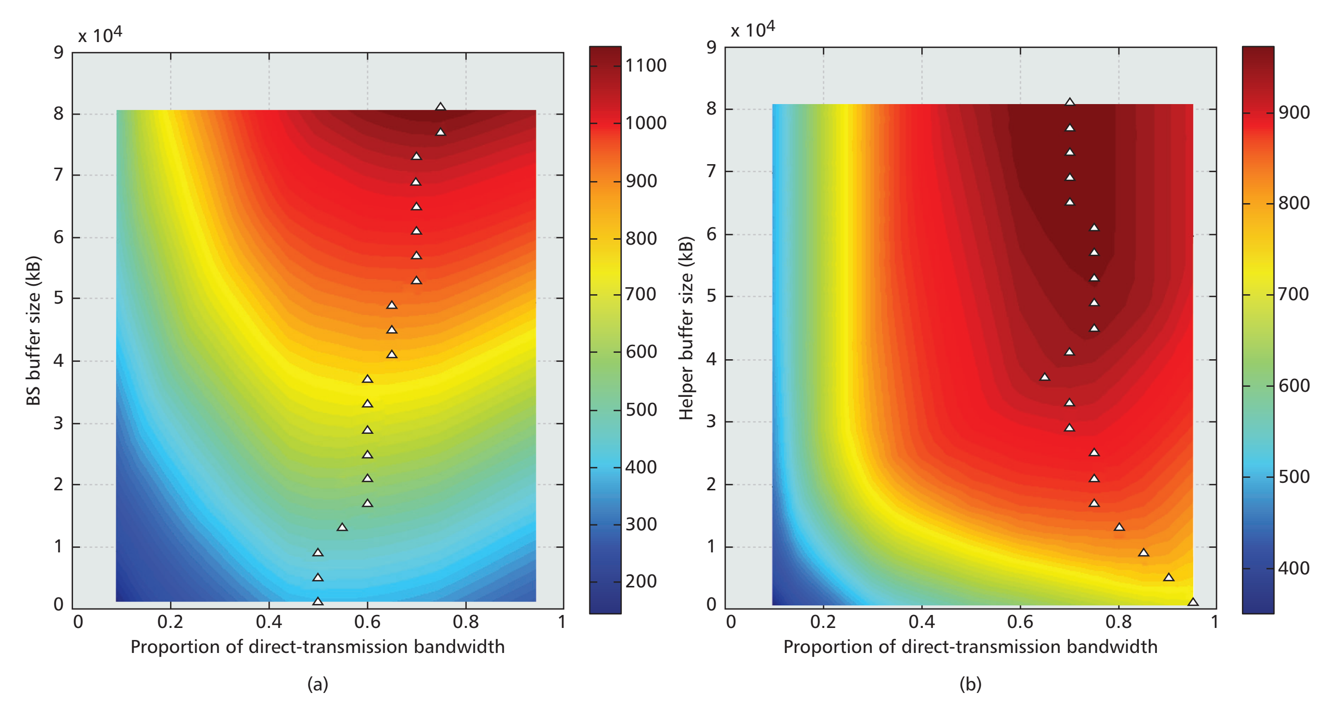

Figure 5. Total data received per second achieved for different allocated cellular direct-transmission bandwidth fractions, given fixed helper buffer size (60MB).

and Fig. 5b indicates that when the buffer size is large (larger than 30 MB in this case), or when the cellular direct transmission bandwidth proportion is small (smaller than 0.3 in this case), the impact of the helper buffer size on system performance is much less than that of the BS buffer size. For instance, given that the cellular direct-transmission bandwidth fraction is 0.2, the TDRPS remains a constant 686.3 kB when the helper buffer size increases from 33 MB to 81 MB, while the same growth in the BS buffer size leads to 47.6 percent TDRPS improvement. This observation is reasonable in that an ample D2D-transmission bandwidth fraction, which is equal to 1 minus the cellular direct-transmission bandwidth fraction, enables helpers to transmit their stored data at a rapid rate and to clear their buffers in a timely manner, and consequently the helper buffer size is less influential.

In Fig. 5 the optimal proportions of cellular direct-transmission bandwidth are marked by small black triangles. Obviously, the optimal proportion of cellular direct-transmission bandwidth has a positive correlation with the BS buffer size, which indicates that enlarging the BS buffer size contributes more to cellular direct transmission than to D2D communication. By contrast, the optimal proportion of cellular direct-transmission bandwidth tends to decrease with the increase in the helper buffer size when the helper buffer size is small, but the trend fluctuates when the helper buffer size becomes large. Considering that the allocation of resources does not vary among time frames in our graph model, we can only draw the conclusion that the optimal allocation of resources is influenced by both BS buffer size and helper buffer size. Although our scalable graph model is able to optimize the time-varying allocation of resources for different time frames, the linear programming problem will turn into a complex nonlinear programming problem and consequently reduce the efficiency of this model. Therefore, more flexible models are required to better investigate the optimal resource allocation and practical resource scheduling for D2D communication underlaying cellular networks, which calls for considerable future work, with buffer size taken into consideration.

Proportion of Subscribers

In a D2D content-downloading underlaying cellular system, the ratio of helpers to subscribers will naturally impact system performance in terms of achievable TDRPS. Figure 6 depicts the TDRPS as the function of the proportion of subscribers and the BS buffer size, given a fixed helper buffer size of 60 MB. It can be seen from Fig. 6 that when the subscriber fraction is less than 0.3, the TDRPS increases quickly as the subscriber fraction increases, and the TDRPS attains the highest values when the subscriber fraction is approximately in the range of 0.5 to 0.65. Further increasing the proportion of subscribers leads to a reduction in the TDRPS. Based on these results, we may conclude that in an optimal D2D content-downloading underlaying cellular system under the previously-mentioned practical assumptions, approximately 50 percent to 65 percent of all UEs should be subscribers, i.e. the ratio of helpers to subscribers should be approximately in the range of 0.54 to 1.

Other Issues

As revealed in our analysis framework and simulation, helpers are required to devote sufficient storage to D2D transmission in order to ensure

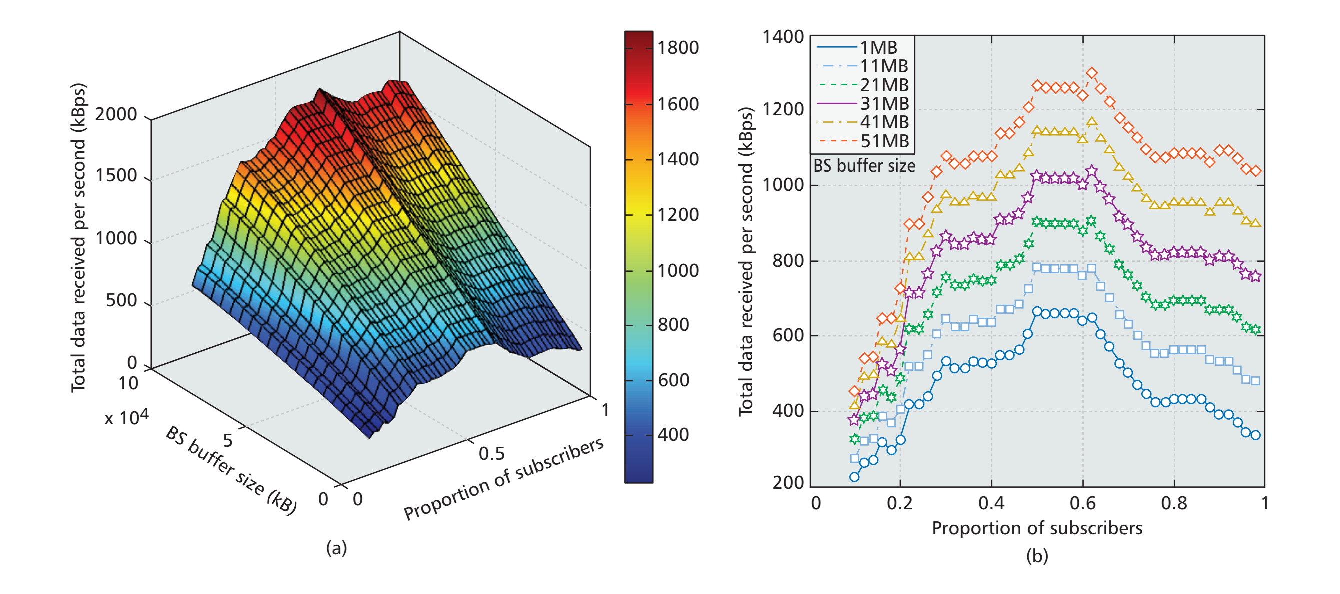

Figure 6. Total data received per second as: a) the function of the proportion of subscribers and the BS buffer size; b) the functions of the proportion of subscribers for various BS buffer sizes. The helper buffer size is fixed to 60 MB.

good system performance. In practice, subscribers may be asked to pay compensation to helpers and service providers for the related D2D data-transmission data flow and direct-transmission data flow, respectively. Furthermore, as in other forms of collaborative communication, buffer-aided D2D communication may raise security problems as well. The security issue has been discussed and potential solutions have been provided in [15]. In addition, energy consumption is also a challenging issue for D2D content-downloading underlaying cellular systems, but an energy-efficient device discovery radio with cellular network assistance has been proposed in [8]. However, it is still open to debate whether buffer size will influence energy consumption. If this influence is not negligible, related research on the trade off between buffer size and energy consumption will also be promising.

Conclusions

We have proposed an optimization framework for analyzing the performance of a buffer-limited D2D content-downloading underlaying cellular system. In particular, we have quantitatively evaluated the positive impact of enlarging the BS and helper buffer sizes on enhancing achievable content downloading performance. Moreover, we have demonstrated that enlarging the BS buffer size leads to a significant performance enhancement and, consequently, it can be utilized as an effective means of saving the required system bandwidth, while maintaining the same level of performance. We have also investigated the negative impact of enlarging the buffer size, which may increase content-downloading delay. Furthermore, based on the proposed optimization framework, we have investigated the optimal bandwidth allocation between the cellular direct communication and the D2D communication, as well as the optimal ratio of helpers to subscribers for the simulated buffer-limited D2D content-downloading underlaying cellular system under realistic assumptions. Similar to other forms of collaborative communications, mobile users are required to operate cooperatively and unselfishly to transmit the data for other users in this framework. However, if we consider the social-domain features, most users behave in a more or less selfish way. Thus, social altruism is another key factor that needs to be considered in future work. Thus, our study also opens a new research direction for bandwidth conserving, delay control, and altruistic preserving in cellular networks.

References

[1] L. Lei *et al.*, "Operator Controlled Device-to-Device Communications in LTE-Advanced Networks," *IEEE Wireless Commun.*, vol. 19, no. 3, June 2012, pp. 96–104.

[2] K. Doppler *et al.*, "Device-to-Device Communication as an Underlay to LTE-Advanced Networks," *IEEE Commun. Mag.*, vol. 47, no. 12, Dec. 2009, pp. 42–49.

[3] N. Golrezaei, A. G. Dimakis, and A. F. Molisch, "Device-to-Device Collaboration Through Distributed Storage," *Proc. Globecom 2012*, Anaheim, CA, Dec. 3–7, 2012, pp. 2397–402.

[4] A. Prasad *et al.*, "Energy-Efficient D2D Discovery for Proximity Services in 3GPP LTE Advanced Networks: ProSe Discovery Mechanisms," *IEEE Vehic. Tech. Mag.*, vol. 9, no. 4, Dec. 2014, pp. 40–50.

[5] M. Chen, *et al.*, "On the Computation Offloading at Ad Hoc Cloudlet: Architecture and Service Modes," *IEEE Commun. Mag.*, vol. 53, no. 6, June 2015, pp. 18-24.

[6] P. Pahlevan *et al.*, "Novel Concepts for Device-to-Device Communication Using Network Coding," *IEEE Commun. Mag.*, vol. 52, no. 4, Apr. 2014 , pp. 32–39.

[7] D. H. Lee *et al.*, "Two-Stage Semi-Distributed Resource Management for Device-to-Device Communication in Cellular Networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, Apr. 2014, pp. 1908–20.

[8] H. Nishiyama, M. Ito, and N. Kato, "Relay-by-Smartphone: Realizing Multihop Device-to-Device Communications," *IEEE Commun. Mag.*, vol. 52, no. 4, Apr. 2014 , pp. 56–65.

[9] Y. Li *et al.*, "Optimal Mobile Content Downloading in Device-to-Device Communication Underlaying Cellular Networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, July 2014, pp. 3596–608.

[10] H. D. Sherali and C. H. Tuncbilek, "A Global Optimization Algorithm for Polynomial Programming Problems Using a Reformulation-Linearization Technique,"*J. Global Optimization*, vol. 2, no. 1, Mar. 1992, pp. 101–12.

[11] *CPLEX: Linear Programming Solver*, available: http://www.ilog.com/.

[12] J. Löfberg, "YALMIP: A Toolbox for Modeling and Optimization in MATLAB," *Proc. 2004 IEEE Int'l. Symp. Computer Aided Control Systems Design*, Taipei, China, Sept. 2–4, 2004, pp. 284–89.

[13] K. Lee *et al.*, "SLAW: A New Mobility Model for Human Walks," *Proc. INFOCOM 2009*, Rio de Janeiro, Brazil, Apr. 19–25, 2009, pp. 855–63.

[14] C. Xu *et al.*, "Efficiency Resource Allocation for Device-to-Device Underlay Communication Systems: A Reverse Iterative Combinatorial Auction Based Approach," *IEEE JSAC*, vol. 31, no. 9, Sept. 2013, pp. 348–58.

[15] M. Alam *et al.*, "Secure Device-to-Device Communication in LTE-A," *IEEE Commun. Mag.*, vol. 52, no. 4, Apr. 2014, pp. 66–73.

Biographies

Haoming Zhang received the B.S. degree from Tsinghua University, Beijing, China, in 2015. His research interests are in the areas of mobile computing and wireless communications. He is currently pursuing his master's degree at Carnegie Mellon University.

Yong Li [M'09] (liyong07@tsinghua.edu.cn) received the B.S. and Ph.D. degrees from Huazhong University of Science and Technology and Tsinghua University in 2007 and 2012, respectively. From 2012 to 2013 he was a visiting research associate with Telekom Innovation Laboratories and Hong Kong University of Science and Technology, respectively. From 2013 to 2014 he was a visiting scientist with the University of Miami. He is currently a faculty member in the Department of Electronic Engineering, Tsinghua University. His research interests are in the areas of mobile computing and social networks, urban computing and vehicular networks, and network science and future Internet. He has served as general chair, technical program committee (TPC) chair, and TPC member for several international workshops and conferences. He is currently an associate editor of the *Journal of Communications and Networking* and the *EURASIP Journal of Wireless Communications and Networking*.

Depeng Jin received the B.S. and Ph.D. degrees from Tsinghua University, Beijing, China, in 1995 and 1999, respectively, both in electronics engineering. He is an associate professor at Tsinghua University and vice chair of the Department of Electronic Engineering. He was awarded the National Scientific and Technological Innovation Prize (Second Class) in 2002. His research fields include telecommunications, high-speed networks, ASIC design, and future Internet architecture.

Mohammad Mehedi Hassan [M'12] is an assistant professor in the Information Systems Department at the College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia. He received his Ph.D. degree in computer engineering from Kyung Hee University, South Korea in February 2011. He has authored and co-authored more than 70 publications in refereed IEEE/ACM/Springer journals, conference papers, books, and book chapters. His research interests include cloud collaboration, media cloud, sensor-cloud, mobile cloud, IPTV, and wirless sensor networks.

Abdulhameed Alelaiwi [M'12] is an assistant professor in the Software Engineering Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia. He received his Ph.D. degree in software engineering from the College of Engineering, Florida Institute of Technology-Melbourne, USA in 2002. His research interests include software testing analysis and design, cloud computing, and multimedia.

Sheng Chen [M'90, SM'97, F'08] obtained his B.Eng. degree from the East China Petroleum Institute, Dongying, China, in January 1982, and his Ph.D. degree from City University, London, in September 1986, both in control engineering. In 2005 he was awarded the higher doctorate degree, doctor of sciences (DSc), from the University of Southampton, Southampton, UK. From 1986 to 1999 he held research and academic appointments at the Universities of Sheffield, Edinburgh, and Portsmouth, all in the UK. Since 1999 he has been with the Department of Electronics and Computer Science, University of Southampton, UK, where he currently holds the post of professor in intelligent systems and signal processing. He is a distinguished adjunct professor at King Abdulaziz University, Jeddah, Saudi Arabia. He is a chartered engineer (CEng) and a Fellow of IET (FIET). His recent research interests include adaptive signal processing, wireless communications, modelling and identification of nonlinear systems, neural network and machine learning, intelligent control system design, evolutionary computation methods and optimization. He has published more than 470 research papers. Dr. Chen is an ISI highly cited researcher in the engineering category (March 2004).