Device-to-Device Load Balancing for Cellular Networks Technical Report

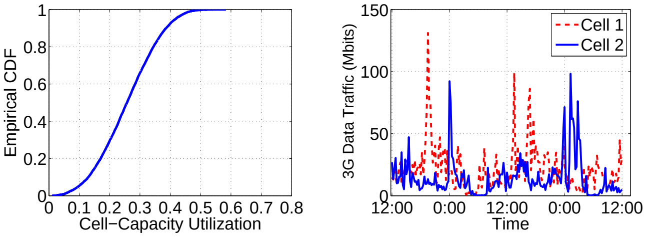
Lei Deng[*], Ying Zhang[*], Minghua Chen[*], Zongpeng Li[†], Jack Y. B. Lee[*], Ying Jun (Angela) Zhang[*], Lingyang Song[‡]
[*]Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong
[†]Department of Computer Science, University of Calgary, Canada
[‡]School of Electrical Engineering and Computer Science, Peking University, Beijing, China

Abstract*—Small-cell architecture is widely adopted by cellular network operators to increase network capacity. By reducing the size of cells, operators can pack more (low-power) base stations in an area to better serve the growing demands, without causing extra interference. However, this approach suffers from low spectrum temporal efficiency. When a cell becomes smaller and covers fewer users, its total traffic fluctuates significantly due to insufficient traffic aggregation and exhibiting a large "peak-to-mean" ratio. As operators customarily provision spectrum for peak traffic, large traffic temporal fluctuation inevitably leads to low spectrum temporal efficiency. In this work, we first carry out a case-study based on real-world 3G data traffic traces and confirm that 90% of the cells in a metropolitan district are less than 40% utilized. Our study also reveals that peak traffic of adjacent cells are highly asynchronous. Motivated by these observations, we advocate device-to-device (D2D) load-balancing as a useful mechanism to address the fundamental drawback of small-cell architecture. The idea is to shift traffic from a congested cell to its adjacent under-utilized cells by leveraging inter-cell D2D communication, so that the traffic can be served without using extra spectrum, effectively improving the spectrum temporal efficiency. We provide theoretical modeling and analysis to characterize the benefit of D2D load balancing, in terms of sum peak traffic reduction of individual cells. We also derive the corresponding cost, in terms of incurred D2D traffic overhead. We carry out empirical evaluations based on real-world 3G data traces to gauge the benefit and cost of D2D load balancing under practical settings. The results show that D2D load balancing can reduce the sum peak traffic of individual cells by 35% as compared to the standard scenario without D2D load balancing, at the expense of 45% D2D traffic overhead.

I. Introduction

The drastic growth in mobile devices and applications has triggered an explosion in cellular data traffic. According to Cisco [1], global cellular data traffic reached 2.5 exabytes per month in 2014 and will further witness a 10-fold increase in 2014-2019. Meanwhile, radio frequency remains a scarce resource for cellular communication. Supporting the fast-growing data traffic demands has become a central concern of cellular network operators.

There are mainly two lines of efforts to address this concern. The first is to serve cellular traffic by exploring additional spectrum, including offloading cellular traffic to WiFi [2] and the recent 60GHz millimeter-wave communication endeavor [3]. The second is to improve *spectrum spatial efficiency*. A common approach is to adopt a small-cell architecture, such as micro/pico-cell [4]. By reducing cell size, operators can pack more (low-power) base stations in an area and reuse radio frequencies more efficiently to increase network capacity.



(a) Empirical CDF for cell-capacity utilization of 194 cells for one month.

(b) 3G (aggregated) mobile data traffic of two adjacent cells in 48 hours.

Fig. 1. Real-world data traffic traces.

While the small-cell architecture improves the *spectrum spatial efficiency*, it comes at a price of degrading the *spectrum temporal efficiency*. When a cell becomes smaller and covers fewer users, there is less traffic aggregation. Consequently, the total traffic of a cell fluctuates significantly, exhibiting a large "peak-to-mean" ratio. As operators customarily provision spectrum to a cell based on peak traffic, high temporal fluctuation in traffic volumes inevitably leads to low spectrum temporal efficiency.

To see this concretely, we carry out a case-study based on cell-traffic traces from Smartone [14], a major cellular network operator in Hong Kong, a highly-populated metropolis. Smartone deploys 194 small-cell base stations in the case-study area of 22 square kilometers, with cell radii of 200-300 meters. [1] The traces include 3G data traffic for each cell, sampled at 15-minute intervals over a month in 2014. The data traffic is delay insensitive and can tolerate a couple of seconds delay. We have the following important observations.

- First, the empirical CDF of the cell-capacity utilization in Fig. 1(a) shows that the average cell-capacity utilization is 24.9%, and 90% of the cells are less than 40% utilized. This confirms that small-cell architecture indeed causes low spectrum temporal utilization, and it suggests ample

[1] The raw data covers 374 cell sectors. In this work, we do not consider cell sectorization. Sectors at the same site location are merged into one cell.

room to improve temporal utilization[2].

- Second, from the 48-hour traffic plot of two adjacent cells in Fig. 1(b), we observe that their peak traffic occurs at different time epochs. We remark that this observation is indeed common among the cells we studied. It implies that one may shift the peak traffic from a congested cell to its under-utilized neighbors, so as to serve the traffic without allocating extra spectrum, effectively improving the spectrum temporal utilization.

Motivated by the above observations, we advocate *device-to-device (D2D) load-balancing* as a useful mechanism to improve spectrum temporal efficiency. D2D communication [6] [7] is a promising paradigm for improving system performance in next generation cellular networks that enables direct communication between user devices (*e.g.*, smartphones) using cellular frequency. It is conceivable to relay (delay insensitive) traffic from congested cells to adjacent underutilized cells via inter-cell D2D communication, enabling load-balancing across cells at the expense of incurred inter-cell D2D traffic.

We remark that an idea of this kind was also studied by Liu *et al.* in their recent work [8]. They focus on important aspects of examining the technical feasibility of D2D load balancing and practical algorithm design in three-tier LTE-Advanced networks. This work is complement to their study and focuses on the following open questions:

- How much benefit can D2D load balancing bring to a cellular network, in terms of reduction in sum peak traffic of individual cells?
- What is the corresponding D2D traffic overhead for achieving the benefit?

Answers to these questions provide fundamental understanding of the viability of D2D load balancing in cellular networks. In this paper, we explore answers to the questions through both theoretical analysis and empirical evaluations based on real-world traces. We make the following contributions.

▷ In Sec. II, using perhaps the simplest possible example, we illustrate the concept of D2D load balancing and show that it can reduce peak traffic for two adjacent cells by 33%. We also compute the associated D2D traffic overhead.

▷ For general settings beyond the example, we provide reasonable and tractable models with some assumptions to analyze the performance of D2D load balancing in Sec. III. We also exploit the optimal solutions in both cases without and with D2D load balancing in Sec. IV and Sec. V, respectively.

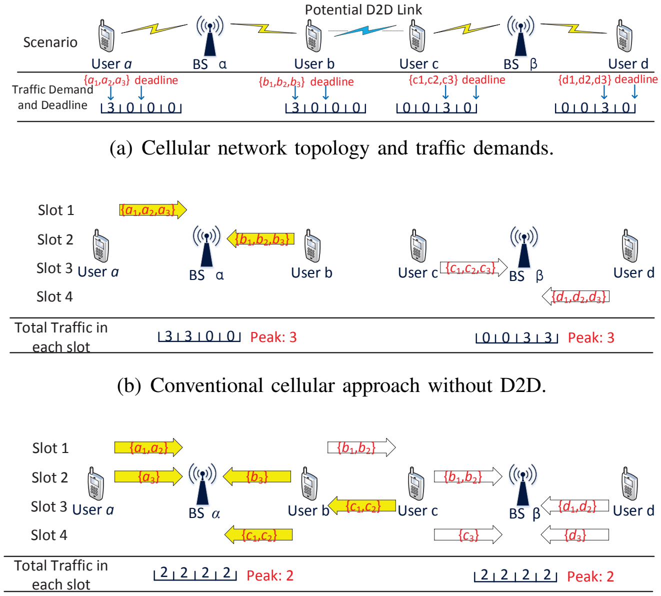
▷ Theoretically, for arbitrary settings, we derive an upper bound for the benefit of D2D load balancing, in terms of sum peak traffic reduction in Sec. VI-B. We show that the bound is asymptotically tight for a specified network scenario, where we further derive the corresponding overhead, in terms of incurred D2D traffic. Our bound and analysis reveal the insight behind the effectiveness of D2D load balancing: by aggregating traffic among adjacent cells via inter-cell D2D communication, the scheme can exploit statistical multiplexing gains to better serve the overall traffic without requiring extra network capacity.

▷ Empirically, in Sec. VIII, we use real-world 3G data traces to verify our theoretical analysis and reveal that D2D load balancing can reduce sum peak traffic of individual cells by 35%, at the cost of 45% D2D traffic overhead.

Throughout this paper, we assume that time is slotted into intervals of unit length, and each wireless hop incurs one-slot delay. We focus on uplink communication scenarios, while our analysis is also applicable to the downlink communication.

II. An Illustrating Example

We consider a simple scenario shown in Fig. 2(a), where 4 users are each aiming at transmitting 3 packets to two base stations (BS) subject to a delay constraint. We compare the peak traffic of both BSs for the case without D2D load balancing (Fig. 2(b)) and for the case with D2D load balancing (Fig. 2(c)). We illustrate the concept of D2D load balancing and show that it can reduce the peak traffic for two adjacent cells by 33%.



(a) Cellular network topology and traffic demands.

(b) Conventional cellular approach without D2D.

(c) Our approach with D2D load balancing.

Fig. 2. A simple example for demonstrating the concept of D2D load balancing, and that it can reduce the peak traffic for both cells by 33% (both from 3 to 2) at the expense of 4 inter-cell D2D transmissions.

Specifically, we consider a cellular network of two adjacent cells served by BS α and BS β, and four users a, b, c, d. BS α (resp. β) can directly communicate with only users a and b (resp. users c and d). BS α and BS β use orthogonal frequency bands. Due to proximity, users b and c can communicate with each other using frequency band of either BS α or β, creating inter-cell D2D links. Users a and b each generate 3 packets at the beginning of slot 1, and users c and d each generate 3 packets at the beginning of slot 3. All packets have the same size and a delay constraint of 2 slots, *i.e.*, a packet must reach BS α or β within 2 slots from its generation time. Note that

[2] The recent time-dependent pricing proposal is one endeavor to improve the spectrum temporal utilization by encouraging users to shift their traffic from peak hour to off-peak hour; see for example [5] and the references therein.

we assume that a packet is successfully delivered as long as it reaches any BS, since BSs today are connected by a high-speed optical backbone, supported by power clusters, and can coordinate to jointly process/forward packets for users.

In the conventional approach without D2D load balancing, a user only communicates with its own BS. It is straightforward to verify that the minimum peak traffic of both BS α and BS β is 3 (unit: packets), and can be achieved by the scheme in Fig. 2(b). For instance, the minimum peak traffic for BS α is achieved by user a (resp. user b) transmitting all its 3 packets to BS α in slot 1 (resp. slot 2).

With D2D load balancing, we can exploit the inter-cell D2D links between users b and c to perform load balancing and reduce the peak traffic for both BS α and BS β.

- In slot 1, user a transmits two packets a_1 and a_2 to BS α, and user b transmits two packets b_1 and b_2 to user c using the orthogonal frequency band of BS β. The traffic is 2 for both cells. In slot 2, users a and b transmit their remaining packets a_3 and b_3 to BS α, and user c relays the two packets it received in slot 1, *i.e.*, b_1 and b_2, to BS β. The traffic is again 2 for both cells. By the end of slot 2, we deliver 6 packets for users a and b to BSs.
- In slots 3 and 4, note that users c and d have the same traffic pattern as users a and b, but delayed by 2 slots. Thus we can also deliver 3 packets for both users c and d in two slots. The traffic of both BSs is 2 per slot.

Overall, with D2D load balancing, we can serve all traffic demands with peak traffic of 2 for both BSs, which is 33% reduced as compared to the case without D2D load balancing.

The intuition behind this example is that the peak traffic for the two cells occurs at different time instances. When users a and b transmit data to BS α in the first two slots, BS β is idle. Meanwhile, BS α is idle when users c and d transmit data to BS β in the last two slots. Therefore, D2D communication can help load balance traffic from the busy BS to the other idle BS, reducing the peak traffic for both BSs.

However, D2D load balancing also comes with cost, since it requires transmissions over the inter-cell D2D links. In the example, the total traffic is $8 \times 2 = 16$ packets and the D2D traffic is $2 \times 2 = 4$ packets, yielding an overhead traffic ratio of $\frac{4}{16} = 25\%$. Such D2D traffic is the overhead that we pay in return for peak traffic reduction.

III. System Model

In this section, we present the system model for a general network topology and a general traffic demand pattern beyond the simple example expounded in the previous section. Such models will be used to analyze the benefit of D2D load balancing in general settings, in terms of the sum peak traffic reduction, and the cost in terms of overhead D2D traffic ratio.

A. Cellular Network Topology

Consider an uplink wireless cellular network with multiple cells and multiple mobile users. We assume that each cell has one BS and each user belongs to one BS.[3] Define $\mathcal{B}$ as the set of all BSs, $\mathcal{U}_b$ as the set of users belonging to BS $b \in \mathcal{B}$, and $\mathcal{U} = \cup_{b \in \mathcal{B}} \mathcal{U}_b$ as the set of all users in the cellular network. Let $b_u \in \mathcal{B}$ denote the cell (or BS) where user $u \in \mathcal{U}$ belongs. We model the uplink cellular network topology as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertex set $\mathcal{V} = \mathcal{U} \cup \mathcal{B}$ and edge set $\mathcal{E}$ where $(u, v) \in \mathcal{E}$ if there is a wireless link from vertex (user) $u \in \mathcal{U}$ to vertex (BS or user) $v \in \mathcal{V}$.

In addition, we characterize the heterogeneous nature of wireless links in the cellular network with a link quality parameter R_{uv} for each link $(u, v) \in \mathcal{E}$. For instance, R_{uv} could be the link rate. Larger R_{uv} means better link condition, and larger transmitted volume within the same time slot and the same amount of allocated resources.

B. Traffic Model

We consider a time-slotted system with T slots in total. Each user can generate a delay-constrained traffic demand at the beginning of any slot. Specifically, we let the *volume-deadline tuple* $z^{s\tau} \triangleq (x^{s\tau}, d^{s\tau})$ be the traffic generated by user $s \in \mathcal{U}$ at the beginning of slot $\tau \in [1, T]$ where $x^{s\tau}$ is the traffic volume and $d^{s\tau}$ is the *deadline*. In this case, every packet[4] in the traffic $z^{s\tau}$ must reach a BS $b \in \mathcal{B}$ before the end of slot $d^{s\tau}$, along with a maximum allowable *delay* $d^{s\tau} - \tau + 1$. Let the interval $[\tau, d^{s\tau}]$ be the *lifetime* of the traffic $z^{s\tau}$. Thus, the input traffic demand pattern for all users is defined by the set $\{z^{s\tau} : s \in \mathcal{U}, \tau \in [1, T]\}$. Every user can transmit a packet either to the BS directly in a single hop or to another user via the D2D link between them such that the packet can reach another BS in multiple hops.

C. Performance Metrics

In this paper, we will use the following performance metrics to characterize both the benefit and the cost for D2D load balancing. As mentioned before, cellular network operators commonly provision spectrum to a cell according to its peak traffic. We thus consider the sum peak traffic of all the cells in the cellular network as the performance metric. Specifically, we define the *sum peak traffic reduction*[5] ρ as

$$\rho \triangleq \frac{P_{ND} - P_D}{P_{ND}} \in [0, 1), \tag{1}$$

where P_{ND} is the minimal sum peak traffic without D2D and P_D is the minimal sum peak traffic with D2D. We remark here that ρ is a simplified metric to facilitate analysis, but it captures the essential benefit of D2D load balancing and we will relate it to the practical spectrum reduction in Sec. VII.

D2D load balancing incurs cost in the sense that any traffic going through D2D links will consume spectrum resources but

[3] We say that user u belongs to BS b if user u is in the cellular cell covered by BS b. When a user can connect to multiple BSs, we force this user to belong to one BS. In the rest of this paper, we will also use the terminology, cell b, to represent the cell covered by BS b.

[4] In the rest of this paper, we assume that the packet size is infinitesimal.

[5] Precisely, we should say *the percentage of sum peak traffic reduction*. However, for simplicity we just call *sum peak traffic reduction* in this paper.

do not immediately reach any BS. This motivates us to define the *overhead ratio* η as

$$\eta \triangleq \frac{V_{D2D}}{V_{D2D}+V_{BS}} \in [0,1), \tag{2}$$

where V_{D2D} is the sum of all D2D traffic and V_{BS} is the sum of all traffic directly sent by cellular users to BSs.

Later on we will discuss how to obtain P_{ND} in Sec. IV and P_D in Sec. V. Then we will show the theoretical upper bounds for ρ in Sec. VI and empirical evaluations in Sec. VIII.

IV. Optimal Solution without D2D

In this section, we describe how to compute the sum peak traffic when D2D is not enabled. Since there are no D2D links, all BSs are independent from each other. We can calculate the peak traffic for each BS individually. Let us denote P_b^{ND} as the minimal peak traffic to satisfy all traffic demands within BS b. Then we can get the sum peak traffic without D2D as

$$P_{ND} = \sum_{b\in\mathcal{B}} P_b^{ND}. \tag{3}$$

A. Problem Formulation

For each BS $b \in \mathcal{B}$, we can formulate the problem to minimize the peak traffic, named as $\mathsf{PEAK\text{-}ND}^b$, as follows:

$$\begin{aligned}
\min \quad & P_b && \text{(4a)}\\
\text{s.t.} \quad & \sum_{t=\tau}^{d^{s\tau}} y_{sb}^{s\tau}(t) R_{sb} = x^{s\tau}, \forall s \in \mathcal{U}_b, \tau \in [1,T] && \text{(4b)}\\
& \sum_{s\in\mathcal{U}_b} \sum_{\tau:\tau\le t\le d^{s\tau}} y_{sb}^{s\tau}(t) = \alpha_b(t), \forall t \in [1,T] && \text{(4c)}\\
& \alpha_b(t) \le P_b, \forall t \in [1,T] && \text{(4d)}\\
& y_{sb}^{s\tau}(t) \ge 0, \forall s \in \mathcal{U}_b, \tau \in [1,T], t \in [\tau, d^{s\tau}] && \text{(4e)}\\
\text{var} \quad & y_{sb}^{s\tau}(t), \alpha_b(t), P_b
\end{aligned}$$

where the auxiliary variable $\alpha_b(t)$ is the total traffic from users to BS b at slot t, P_b is the peak traffic of BS b, and $y_{sb}^{s\tau}(t)$ is the traffic volume transmitted from user s to BS b at slot t for the source traffic demand $z^{s\tau}$.

Equation (4b) shows the volume requirement for any source traffic demand $z^{s\tau}$. Without D2D, users can only be served by its own BS. Note that we model the relationship between the effectively transmitted volume and the amount of allocated resources (which is peak traffic here) as a linear function with respect to the link quality parameter. Equation (4c) depicts the total allocated traffic for all users in the cell.

B. Optimal Solution

To solve $\mathsf{PEAK\text{-}ND}^b$, we can use standard linear programming (LP) solvers. However, LP solvers cannot exploit the structure of this problem. We next propose a combinatorial algorithm that exploits the problem structure and achieves lower complexity than general LP algorithms.

We note that $\mathsf{PEAK\text{-}ND}^b$ resembles a uniprocessor scheduling problem for preemptive tasks with hard deadlines [11]. Indeed, we can attach each task $z^{s\tau}$ with an arrival time τ and a hard deadline $d^{s\tau}$ and the requested service time $\frac{x^{s\tau}}{R_{sb}}$. Then for a given peak traffic P_b (which resembles the maximal speed of the processor), we can use earliest deadline first (EDF) scheduling algorithm [15] to check its feasibility. Since we can easily get an upper bound for the optimal peak traffic, we can use binary search to find the minimal peak traffic, supported by the EDF feasibility-check subroutine.

More interestingly, we can even get a semi-closed form for P_b^{ND}, inspired by [16, Theorem 1]. Specifically, let us define the *intensity* [16] of an interval $I = [z, z']$[6] to be

$$g_b(I) = \frac{\sum_{(s,\tau)\in\mathcal{A}_b(I)} \frac{x^{s\tau}}{R_{sb}}}{z'-z+1} \tag{5}$$

where $\mathcal{A}_b(I) = \{(s,\tau) : s \in \mathcal{U}_b, \tau \in [1,T], x^{s\tau} > 0, [\tau, d^{s\tau}] \subset [z,z']\}$ is the set of all active traffic demands whose lifetime is within the interval $I = [z,z']$. Then we have the following theorem.

Theorem 1: $P_b^{ND} = \max_{I\subset[1,T]} g_b(I)$.

Proof: Since the proof of Theorem 1 was omitted in [16] and this theorem is not directly mapped to the peak traffic minimization problem, we give a full proof in Appendix B for completeness. ■

Based on Theorem 1, we adapt the YDS algorithm originally developed for solving the job scheduling problem in [16] to our peak-traffic minimization problem. The time complexity of the YDS algorithm is related to the total number of possible intervals. Clearly the optimal interval can only begin from the generation time of a demand and end at the deadline of a demand. So the total number of intervals needed to be checked is $O(n^2)$ where n is the total number of traffic demands within cell b. Thus the time complexity of a straightforward implementation of our adaptive YDS algorithm is $O(n^2)$ [16], which is lower than general LP algorithms [12].

V. Optimal Solution With D2D

In this section, we formulate the optimization problem to compute the minimal sum peak traffic P_D when D2D communication is enabled. In this case, since the traffic can be directed to other BSs via inter-cell D2D links, all BSs are coupled with each other and need to be considered as a whole. We will first define the traffic scheduling policy with D2D and then formulate the problem as a LP.

A. Traffic Scheduling Policy

Given a traffic demand pattern, we need to find a routing policy to forward each packet to BSs before the deadline, which is the *traffic scheduling problem.* Since we should consider the traffic flow in each slot, we will use the *time-expanded graph* to model the traffic flow over time [10]. Specifically, denote $y_{uv}^{s\tau}(t)$ as the traffic volume delivered from node u to node v at slot t for the source traffic $z^{s\tau}$. Note that the notation also includes self-link traffic, *i.e.*, $y_{uu}^{s\tau}(t)$ is the traffic volume stored in node u at slot t for the source traffic

[6] This interval is from the beginning of slot z to the end of slot z'.

$z^{s\tau}$ and we let the self-link quality to be $R_{uu} = 1$ for ease of formulation. All traffic flows over time are precisely captured by the time-expanded graph and $y_{uv}^{s\tau}(t)$. Then we can define the *traffic scheduling policy* as follows.

Definition 1: A traffic scheduling policy is the set $\{y_{uv}^{s\tau}(t) : (u,v) \in \mathcal{E}, s \in \mathcal{U}, \tau \in [1,T], t \in [\tau, d^{s\tau}]\} \cup \{y_{uu}^{s\tau}(t) : u \in \mathcal{V}, s \in \mathcal{U}, \tau \in [1,T], t \in [\tau, d^{s\tau}]\}$ such that

$$\sum_{v \in \text{out}(s)} y_{sv}^{s\tau}(\tau) R_{sv} = x^{s\tau}, \forall s \in \mathcal{U}, \tau \in [1,T] \tag{6a}$$

$$\sum_{b \in \mathcal{B}} \sum_{v \in \text{in}(b)} y_{vb}^{s\tau}(d^{s\tau}) R_{vb} = x^{s\tau}, \forall s \in \mathcal{U}, \tau \in [1,T] \tag{6b}$$

$$\sum_{v \in \text{in}(u)} y_{vu}^{s\tau}(t) R_{vu} = \sum_{v \in \text{out}(u)} y_{uv}^{s\tau}(t+1) R_{uv}, \\ \forall s \in \mathcal{U}, \tau \in [1,T], u \in \mathcal{V}, t \in [\tau, d^{s\tau} - 1] \tag{6c}$$

$$y_{uv}^{s\tau}(t) \geq 0, \forall (u,v) \in \mathcal{E}, (u,v) = (u,u) \text{ where } u \in \mathcal{V}, \\ s \in \mathcal{U}, \tau \in [1,T], t \in [\tau, d^{s\tau}] \tag{6d}$$

where $\text{in}(u) = \{v : (v,u) \in \mathcal{E}\} \cup \{u\}$ and $\text{out}(u) = \{v : (u,v) \in \mathcal{E}\} \cup \{u\}$ are the incoming neighbors and outgoing neighbors of node $u \in \mathcal{V}$ in the time-expanded graph.

Constraint (6a) shows the flow balance in the source node while (6b) shows the flow balance in the destination nodes such that all traffic can reach BSs before their deadlines. Equality (6c) is the conservation constraint for each intermediate node in the time-expanded graph. Also note that we assume that all BSs and all users have sufficient number of radios such that they can transmit data to and receive data from multiple BSs (or users). This is a strong assumption for mobile users because currently mobile devices are not equipped with sufficient number of radios. However, multi-radio mobile devices could be a trend and there are actually substantial research work in multi-radio wireless systems (see a survey in [9] and the references therein). We made this assumption here because *wireless scheduling problem* for single-radio users is generally intractable and we want to avoid detracting our attention and focus on how to characterize the benefit of D2D load balancing and get a first-order understanding.

B. Problem Formulation

Then we can formulate the problem of computing the minimal sum peak traffic with D2D, named as PEAK-D2D, as follows:

$$\min \quad \sum_{b \in \mathcal{B}} P_b \tag{7a}$$

$$\text{s.t.} \quad \text{(6a), (6b), (6c), (6d)},$$

$$\sum_{v \in \mathcal{U}_b} \sum_{s \in \mathcal{U}} \sum_{\tau : \tau \leq t \leq d^{s\tau}} y_{vb}^{s\tau}(t) = \alpha_b(t), \\ \forall b \in \mathcal{B}, t \in [1,T] \tag{7b}$$

$$\sum_{u \in \mathcal{U}_b} \sum_{v \in \text{in}(u) \backslash \{u\}} \sum_{s \in \mathcal{U}} \sum_{\tau : \tau \leq t \leq d^{s\tau}} y_{vu}^{s\tau}(t) = \beta_b(t), \\ \forall b \in \mathcal{B}, t \in [1,T] \tag{7c}$$

$$\alpha_b(t) + \beta_b(t) \leq P_b, \forall b \in \mathcal{B}, t \in [1,T] \tag{7d}$$

$$\text{var} \quad y_{uv}^{s\tau}(t), \alpha_b(t), \beta_b(t), P_b$$

where the auxiliary variable $\alpha_b(t)$ is the total traffic from users to BS b at slot t, the auxiliary variable $\beta_b(t)$ is the total traffic dedicated to all users in BS b at slot t, and P_b is the peak traffic of BS b.

Note that we assume a *receiver-takeover* scheme in the sense that any traffic will consume resources (peak traffic) of the receiver's BS. Equalities (7b) and (7c) show that BS b is responsible for all traffic dedicated to itself and to its users.

Among the optimal solution to PEAK-D2D, we denote P_b^D as the optimal peak traffic for each BS b, and thus the sum peak traffic is

$$P_D = \sum_{b \in \mathcal{B}} P_b^D. \tag{8}$$

And the total D2D traffic and total user-to-BS traffic are

$$V_{D2D} = \sum_{t=1}^{T} \sum_{s \in \mathcal{U}} \sum_{\tau : \tau \leq t \leq d^{s\tau}} \sum_{v \in \mathcal{U}} \sum_{u : (u,v) \in \mathcal{E}} y_{uv}^{s\tau}(t), \tag{9}$$

$$V_{BS} = \sum_{t=1}^{T} \sum_{s \in \mathcal{U}} \sum_{\tau : \tau \leq t \leq d^{s\tau}} \sum_{b \in \mathcal{B}} \sum_{u \in \mathcal{U}_b} y_{ub}^{s\tau}(t), \tag{10}$$

which are used to calculate the overhead ratio η in (2).

Although PEAK-D2D is a LP problem, it is challenging to solve it due to its large scale. We solve it with the state-of-the-art LP solver, Gurobi [13]. For more discussions and techniques to reduce the complexity, please see Appendix A.

VI. Theoretical Results

From the two preceding sections, we can compute P_{ND} with the (adaptive) YDS algorithm (Theorem 1) and P_D by solving the large-scale LP problem PEAK-D2D (Sec. V-B). Hence, numerically we can get the sum peak traffic reduction and the overhead ratio. In this section, however, we seek to derive theoretical upper bounds on the sum peak traffic reduction. The theoretical upper bound on the benefit helps determine whether it is worthwhile to implement D2D load balancing scheme in real-world cellular systems.

A. A Trivial Upper Bound

We can get a trivial upper bound for P_D by assuming no cost for D2D communication in the sense that any D2D communication will not consume bandwidth and will not incur delays. Then we can construct a virtual grand BS where all users $\mathcal{U}$ are in this BS. Then the system becomes similar to the case without D2D. We can apply the YDS algorithm to compute the minimal peak traffic, which is a lower bound for P_D, *i.e.*,

$$P_D^{lb} = \max_{I \subset [1,T]} g(I), \tag{11}$$

where

$$g(I) = \frac{\sum_{(s,\tau)\in\mathcal{A}(I)} \frac{x^{s\tau}}{R_{\max}}}{z' - z + 1} \tag{12}$$

and $\mathcal{A}(I) = \{(s,\tau) : s \in \mathcal{U}, \tau \in [1,T], x^{s\tau} > 0, [\tau, d^{s\tau}] \subset [z,z']\}$ is the set of all active traffic demands whose lifetime is within the interval $I = [z,z']$ and $R_{\max} = \max_{s\in\mathcal{U}} R_{sb_s}$ is the best user-BS link. Then we have the following theorem.

***Theorem** 2:* $\rho \le \frac{P_{ND} - P_D^{lb}}{P_{ND}}$.

Proof: Omitted due to space limitations. ■

Note that both P_D^{lb} and P_{ND} can be computed by the YDS algorithm, much easier than solving the large-scale LP PEAK-D2D. Therefore, numerically we can get a quick understanding of the maximum benefit that can be achieved with D2D.

B. A General Upper Bound

We next describe another general upper bound for any arbitrary topology and any arbitrary traffic demand pattern. We will begin with some preliminary notations.

1) Preliminary Notations: Let $N = |\mathcal{B}|$ be the number of BSs and we define a directed *D2D communication graph* $\mathcal{G}_{D2D} = (\mathcal{B}, \mathcal{E}_{D2D})$ where the vertex set is the BS set $\mathcal{B}$ and $(i,j) \in \mathcal{E}_{D2D}$ if there exists at least one inter-cell D2D link from user $u \in \mathcal{U}_i$ in BS $i \in \mathcal{B}$ to user $v \in \mathcal{U}_j$ in BS $j \in \mathcal{B}$. Denote d_i^- as the in-degree of BS i in the graph $\mathcal{G}_{D2D}$ and define the maximal in-degree of the graph $\mathcal{G}_{D2D}$ as $\Delta^- = \max_{i\in\mathcal{B}} d_i^-$. In addition, we define some notations in Tab. I to capture the discrepancy of D2D links and non-D2D links for users and BSs. Note that these definitions will be used thoroughly in Appendix C to prove Theorem 3.

TABLE I
DISCREPANCY NOTATIONS

$r_s = \max_{v:(s,v)\in\mathcal{E}, v\in\mathcal{U}_{b_s}} \frac{R_{sv}}{R_{sb_s}}, \quad \forall s \in \mathcal{U}$
$\tilde{r}_s^j = \max_{v:(s,v)\in\mathcal{E}, v\in\mathcal{U}_j} \frac{R_{sv}}{R_{sb_s}}, \quad \forall s \in \mathcal{U}, j \in \mathcal{B}$
$r_i = \max_{s\in\mathcal{U}_i} r_s, \quad \forall i \in \mathcal{B}$
$\tilde{r}_{ij} = \max_{s\in\mathcal{U}_i} \tilde{r}_s^j, \quad \forall i \in \mathcal{B}, j \in \mathcal{B}$
$r = \max_{i\in\mathcal{B}} r_i$
$\tilde{r} = \max_{(i,j)\in\mathcal{E}_{D2D}} \tilde{r}_{ij}$

2) Main Result:

***Theorem** 3:* For an arbitrary network topology $\mathcal{G}$ associated with a D2D communication graph $\mathcal{G}_{D2D} = (\mathcal{B}, \mathcal{E}_{D2D})$ and an arbitrary traffic demand pattern, the sum peak traffic reduction is upper bounded by

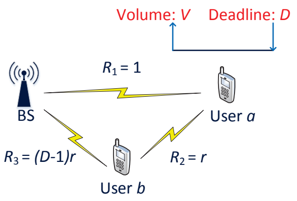


Fig. 3. The benefit of intra-cell D2D communications.

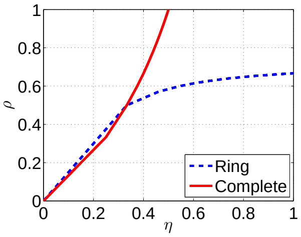


Fig. 4. Tradeoff between ρ and η.

$$\rho \le \frac{\max\{r,1\} + \tilde{r}\Delta^- - 1}{\max\{r,1\} + \tilde{r}\Delta^-}. \tag{13}$$

Proof: See Appendix C. ■

Based on this upper bound, we observe that the benefit of D2D load balancing comes from two parts: intra-cell D2D and inter-cell D2D. More interestingly, we can obtain the individual benefit of intra-cell D2D and inter-cell D2D separately, as shown in the following two subsections along with Corollaries 1 and 2. One can go through the proof for Theorem 3 by disabling inter-cell or intra-cell D2D communication and get the proof of these two corollaries.

3) Benefit of Intra-cell D2D:

***Corollary** 1:* If only intra-cell D2D communication is enabled, the sum peak traffic reduction is upper bounded by

$$\rho \le \frac{\max\{r,1\} - 1}{\max\{r,1\}}. \tag{14}$$

This upper bound is quite intuitive. When $r \le 1$, then for any user s, there does not exist any intra-cell D2D link with better link quality than its direct link to BS b_s. Therefore, using the user-BS link is always the optimal choice. Thus the peak traffic reduction is 0. When $r > 1$, larger r means more advantages for intra-cell D2D links over the user-BS links. Therefore, D2D can exploit more benefit.

Moreover, this upper bound can be achieved by the simple example in Fig. 3. Suppose that user a generates one traffic demand with volume V and delay $D \ge 2$ at slot 1. Suppose $R_1 = 1, R_2 = r, R_3 = (D-1)r$. Then without intra-cell D2D, the peak traffic is $P_1 = \frac{V}{D}$. With intra-cell D2D, user a transmits $\frac{V}{D-1}$ traffic to user b from slot 1 to slot $D-1$ and then user b transmits all V traffic to BS at slot D. The peak traffic is $P_2 = \max\{\frac{V}{(D-1)R_2}, \frac{V}{R_3}\} = \frac{V}{(D-1)r}$. Then the peak traffic reduction is

$$\frac{P_1 - P_2}{P_1} = 1 - \frac{\frac{V}{(D-1)r}}{\frac{V}{D}} \to \frac{r-1}{r}, \text{as } D \to \infty. \tag{15}$$

The benefit of intra-cell D2D communication is widely studied (see [6] [7]). However, in this paper, we mainly focus on the benefit of inter-cell D2D load balancing. Indeed, in our simulation settings in Sec. VIII, the intra-cell D2D brings negligible benefit.

4) *Benefit of Inter-cell D2D:*

***Corollary** 2:* If only inter-cell D2D communication is enabled, the sum peak traffic reduction is upper bounded by

$$\rho \leq \frac{\tilde{r}\Delta^-}{1+\tilde{r}\Delta^-}. \tag{16}$$

The intuition behind the parameter $\tilde{r}$ is similar to the effect of parameter r in the intra-cell D2D case. In what follows, we will only discuss the effect of parameter Δ^-, which actually reveals the insight of our advocated D2D load balancing scheme. Now suppose that all the links have the same quality and *w.l.o.g.* let $R_{uv} = 1, \forall (u,v) \in \mathcal{E}$. Then $r = \tilde{r} = 1$, meaning that no intra-cell D2D benefit exists. And the benefit of inter-cell D2D is reduced to the following upper bound

$$\rho \leq \frac{\Delta^-}{1+\Delta^-}. \tag{17}$$

The rationale to understand this upper bound is as follows. On a high level of understanding, the main idea for load balancing is traffic aggregation. If each BS can aggregate more traffic from other BSs, it can exploit more statistical multiplexing gains to server more traffic with the same peak traffic. Since the in-degree for each BS indeed measures its capacity of traffic aggregation, it is not surprising that the upper bound for ρ is related to maximal in-degree Δ^-.

To evaluate how good the upper bound in (17) is, two natural questions can be asked. The first is: *Is this upper bound tight*? Another observation is that if we want to achieve unbounded benefit, *i.e.*, $\rho \to 1$, it is necessary to let $\frac{\Delta^-}{\Delta^-+1} \to 1$, which means that $\Delta^- \to \infty$. Then the second question is: *Can ρ indeed approach 100% as $\Delta^- \to \infty$*?

In the rest of this subsection, we will answer these two questions by artificially constructing a specified network and traffic demand pattern. Specifically, we consider $N = |\mathcal{B}|$ BSs each serving one user only. To facilitate analysis, let b_i be the i-th BS and u_i be the user in BS i, for all $i \in [1, N]$. We consider a *singleton-decoupled* traffic demand pattern as follows. Each user has one and only one traffic demand with the same volume V and the same delay $D \geq 2$. Let $T = ND$ and the traffic generation time of user i is slot $D(i-1)+1$. Therefore, the lifetime of user u_i's traffic is $[D(i-1)+1, Di]$, during which there is only one such traffic. This is why we call it singleton-decoupled traffic.

Under such settings, we will vary the user-connection pattern such that the D2D communication graph is different. Specifically, we will prove that this upper bound is asymptotically tight in the ring topology for $\Delta^- = 2$ in Fact 1, and $\rho \to$ 100% in the complete topology as the number of BSs $N \to \infty$ in Fact 2. Moreover, we will also mention the overhead ratio for these two special topologies.

***Fact** 1:* If $N = 2D-1$ and the D2D communication graph forms a bidirectional ring graph, then there exists a traffic scheduling policy to make the sum peak traffic reduction

$$\rho = \frac{2(D-1)}{3D-2} \to \frac{2}{3} = \frac{\Delta^-}{\Delta^-+1}, \text{as } D \to \infty. \tag{18}$$

Besides, the overhead ratio in this case is

$$\eta = \frac{D(D-1)}{D^2+2D-2}. \tag{19}$$

Proof: See Appendix D. ■

***Fact** 2:* If the D2D communication graph forms a bidirectional complete graph, then there exists a traffic scheduling policy to make the sum peak traffic reduction

$$\rho = \frac{N-1}{N+1} \to 100\%, \text{as } N \to \infty. \tag{20}$$

Besides, the overhead ratio in this case is

$$\eta = \frac{N-1}{2N}. \tag{21}$$

Proof: See Appendix E. ■

Remark: (i) Fact 1 shows the tightness of the upper bound in (17) for the ring topology when $\Delta^- = 2$. (ii) **Fact 2 shows that ρ can indeed approach 100%, implying that in the best case, ρ goes to 100%. Therefore, we show that D2D load balancing is *ultra powerful*, giving us strong motivation to investigate D2D load balancing scheme both theoretically and practically.** (iii) For the complete topology, the upper bound $\frac{\Delta^-}{\Delta^-+1}$ is not tight. Indeed, since $\Delta^- = N-1$ in the complete topology, we have

$$\frac{\Delta^-}{\Delta^-+1} = \frac{N-1}{N} > \frac{N-1}{N+1}. \tag{22}$$

(iv) Let us revisit the toy example in Fig. 2 which forms a complete topology with $N = 2$. It verifies the sum peak traffic reduction and overhead ratio in Fact 2, *i.e.*, $\rho = \frac{1}{3} = \frac{N-1}{N+1}$ and $\eta = \frac{1}{4} = \frac{N-1}{2N}$. (v) We also highlight the tradeoff between the benefit ρ and the cost η, as illustrated in Fig. 4. Furthermore, Fig. 4 shows that the complete topology outperforms the ring topology asymptotically because $\rho \to \frac{2}{3}$ and $\eta \to 1$ for the ring topology but $\rho \to 1 > \frac{2}{3}$ (larger benefit) and $\eta \to \frac{1}{2} < 1$ (smaller cost) for the complete graph.

VII. Towards Spectrum Reduction

In this paper, we use sum peak traffic to capture how many resources are needed to serve all users' traffic demands in cellular networks. This may not directly reflect the total required spectrum for cellular operators, because the same spectrum can be spatially reused by multiple BSs sufficiently far away from each other. The benefit of spectrum spatial reuse is characterized by the frequency reuse factor K, which represents the proportion of the total spectrum that one cell can utilize. For instance, $K = 1$ means that any cell can use all spectrum, and $K = 1/7$ means that one cell can only utilize $1/7$ of the total spectrum, to avoid excessive interference among adjacent cells. A *back-of-the-envelope* calculation suggests that, if the total number of required channels for all N BSs is C, then $\frac{C/N}{K}$ distinct radio channels are needed to serve the entire cellular network.

In the case without D2D, the sum peak traffic of all BSs is P_{ND}, which corresponds to the total number of channels for all cells. Thus, with frequency reuse factor K, $\frac{P_{ND}}{NK}$ distinct channels are needed without D2D.

In the case with D2D, D2D communication can degrade the original frequency reuse pattern if they are sharing the same spectrum with cellular users (which is called underlay D2D [6]). Given the new frequency reuse factor $K_D(\leq K)$. A back-of-the-envelope analysis suggests that $\frac{P_D}{NK_D}$ distinct radio channels are needed with D2D load balancing. Consequently, the spectrum reduction can be estimated as

$$\frac{\frac{P_{ND}}{NK}-\frac{P_D}{NK_D}}{\frac{P_{ND}}{NK}} = 1-\frac{K}{K^D}\times\frac{P_D}{P_{ND}} = 1-\frac{K}{K^D}(1-\rho), \quad (23)$$

which gives us a first-order understanding of how much spectrum reduction can be achieved by D2D load balancing.

VIII. Empirical Evaluations

We use real-world 3G uplink traffic traces from Smartone, a major cellular network operator in Hong Kong, to evaluate the performance of D2D load balancing.

A. Methodology

Dataset: Our dataset contains 374 cell sectors covering a highly-populated area of 22 km^2 in Hong Kong. We merge them based on their unique site locations and get 194 cells. The data traffic traces are sampled every 15 minutes, spanning an one-month period from 2014/06/16 to 2014/07/15.

Network Topology: We use the site location to be the BS's location (194 BSs in total). We let each BS cover a circle area with radius 300m. In each BS, 5 users are uniformly distributed in the coverage circle. Assume that the communication range for both user-to-BS links and D2D links is 300m. For each link (u, v) with distance $d(u, v)$, we use Shannon capacity to be the link quality, *i.e.* $R_{uv} = \log_2(1 + \frac{P_t d(u,v)^{-3.5}}{N})$, where $P_t = 21$dBm is the transmit power and $N = -102$dBm is the noise power.

Traffic Model: We let each slot last for 2 seconds and thus we have $T = 24\times 3600/2 = 43200$ slots in each day. For each raw traffic trace, which is the aggregate traffic volume sampled every 15 minutes, we average it into 15 traffic demands and randomly assign each traffic to different users (in total 5 users) with random start time (in total $15 \times 60/2 = 450$ slots). The delay for all traffic demands is set to 3 slots.

Tools: We use the state-of-the-art LP solver Gurobi [13] and implement all evaluations with C++ language (4K+ lines of code). All evaluations are running in a cluster of 15 computers, each of which has a 8-core Intel Core-i7 3770 3.4Ghz CPU with 8GB memory, running CentOS 6.4.

B. Sum Peak Traffic Reduction of D2D Load Balancing

We evaluate the sum peak traffic reduction for all 194 BSs in the area of 22km^2. However, due to the large-scale LP and our computational resource limit, we divide the entire 22km^2 area into 27 smaller regions, and the number of BSs in each region ranges from 3 to 11. We evaluate the peak traffic reduction of D2D load balancing for each region individually. We then sum up the individual reductions to obtain the overall reduction for the entire area. Since we essentially limit the D2D load balancing opportunities by this area dividing approach, the obtained result gives a conservative estimate on the maximum possible sum peak traffic reduction achievable by D2D load balancing in the whole area. In the case with D2D, we get the optimal peak traffic for all BSs in each small region, and then sum up all 194 BSs to get the sum peak traffic reduction. Fig. 5(a) shows the sum peak traffic reduction for 30-day traffic traces after employing D2D load balancing. It reveals that D2D load balancing can reduce sum peak traffic by 35.27% on average, while the average D2D traffic overhead ratio is 45.05%. We also remark here that most of the benefit comes from inter-cell D2D communication, based on our separated simulation by disabling all inter-cell D2D links but only enabling all intra-cell D2D links, which shows negligible benefit. Furthermore, Fig. 5(a) verifies the upper bound, represented in Theorem 2 and Theorem 3.

TABLE II
Three different-level instances.

Instance	$\lvert\mathcal{B}\rvert$	$\lvert\mathcal{U}\rvert$	$\lvert\mathcal{E}\rvert$	# of demands	T
S1 (Light)	3	15	139	4035	43200
S2 (Medium)	6	30	344	6945	43200
S3 (Heavy)	9	45	1083	10095	43200

We also evaluate the effects of traffic delay sensitivity and the communication range. From Fig. 5(b), we can observe that D2D load balancing brings more benefit with larger delay traffic demands and/or larger communication range. The reason is as follows. Larger delay means that more traffic can be load-balanced with more freedom, and larger communication range means better network connectivity, both of which enable D2D load balancing to exploit more benefit.

C. Running Time and Memory Usage

We also use the following three different-level instances in Tab. II to show the computational cost of LP problem PEAK-D2D. Fig. 5(c) and Fig. 5(d) show the respective running time and the memory usage for solving the three instances. Clearly, the light instance S1 with only 3 BSs can be solved quickly. The medium instance S2 with 6 BSs takes around 20 minutes and consumes around half of the memory (8GB in total). However, for the heavy instance S3 with 9 BSs, it takes about 2.7 hours by occupying almost all memory. This confirms that solving PEAK-D2D is quite challenging. Thus how to design low-complexity algorithms to solve it (either optimally or approximately) deserves further research efforts.

IX. conclusion and Future Work

To the best of our knowledge, this is the first work to characterize the system-level benefit and cost of D2D load balancing, through both theoretical analysis and empirical evaluations. We show that D2D load balancing can substantially reduce sum peak traffic (and thus spectrum reduction), which provides strong support to standardize D2D in the coming cellular systems. This work servers as a position paper and aims to provide performance metrics/benchmarks and call for investigation participation on the D2D load balancing scheme.

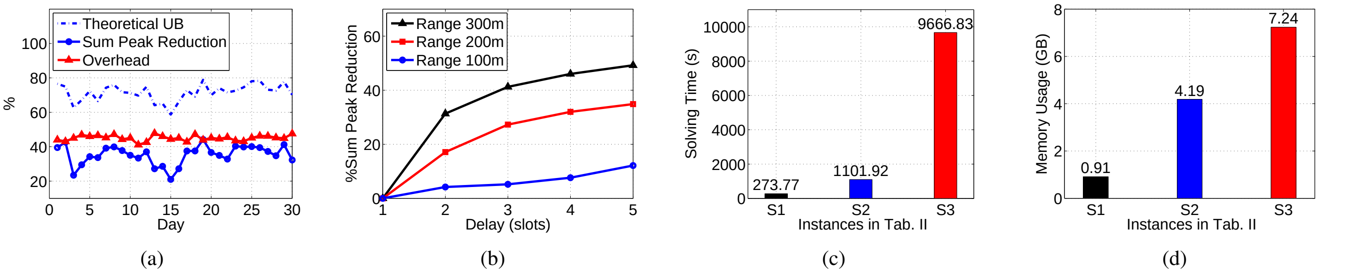


Fig. 5. Simulation Results: (a) Sum peak traffic reduction and overhead ratio in 30 days (194 BSs) (b) Effects of traffic delay sensitivity and the communication range (c) Running time for solving PEAK-D2D for three instances (d) Memory usage for solving PEAK-D2D for three instances.

Furthermore, we list some important problems here. (i) Since D2D communication will consume energy of mobile devices, it is always a challenge to design incentive mechanisms to encourage mobile users to participate such D2D load balancing. (ii) In this paper, we provide fundamental understandings based on the offline problem. However, in real-world systems, future traffic arrivals cannot be known (or predicted precisely) in advance. This spurs us to study the online counterpart. (iii) In this paper, we model the heterogenous wireless links only in terms of channel qualities. It requires more efforts to investigate D2D load balancing under more practical PHY layer wireless models.

Acknowledgment

We would like to thank CUHK Multimedia Lab for providing computing resources. The first author would like to thank Shaoquan Zhang and Boya Di for discussing and proofreading the paper. The work described in this paper was partially supported by National Basic Research Program of China (Project No. 2013CB336700) and the University Grants Committee of the Hong Kong Special Administrative Region, China (Area of Excellence Grant Project No. AoE/E-02/08 and General Research Fund No. 14201014).

References

[1] Cisco, “Cisco visual networking index: global mobile data traffic forecast update, 2014-2019,” *White Paper*, Feb. 2015.

[2] K. Lee, I. Rhee, J. Lee, S. Chong, and Y. Yi, “Mobile data offloading: how much can wifi deliver?” in *Proc. ACM CoNEXT*, 2010.

[3] R. C. Daniels, J. N. Murdock, T. S. Rappaport, and R. W. Heath, “60 GHz wireless: up close and personal,” *IEEE Microw. Mag.*, vol. 11, no. 7, pp. 44-50, Dec. 2010.

[4] A. Ghosh, N. Mangalvedhe, R. Ratasuk, B. Mondal, M. Cudak, E. Visotsky, T. A. Thomas, J. G. Andrews, P. Xia, H. S. Jo, H. S. Dhillon, and T. D. Novlan, “Heterogeneous cellular networks: from theory to practice,” *IEEE Commun. Mag.*, vol. 50, no. 6, pp. 54-64, Jun. 2012.

[5] S. Ha, S. Sen, J. Carlee, Y. Im, and M. Chiang, “TUBE: time dependent pricing for mobile data,” in *Proc. ACM SIGCOMM*, 2012.

[6] K. Doppler, M. Rinne, C. Wijting, C. B. Ribeiro, and K. Hugl, “Device-to-device communication as an underlay to LTE-advanced networks,” *IEEE Commun. Mag.*, vol. 7, no. 12, pp. 42-49, Dec. 2009.

[7] G. Fodor, E. Dahlman, G. Mildh, S. Parkvall, N. Reider, G. Miklós, and Z. Turányi, “Design aspects of network assisted device-to-device communications,” *IEEE Commun. Mag.*, vol. 50, no. 3, pp. 170-177, Mar. 2012.

[8] J. Liu, Y. Kawamoto, H. Nishiyama, N. Kato, and N. Kadowaki, “Device-to-device communications achieve efficient load balancing in LTE-advanced networks,” *IEEE Wireless Commun.*, vol. 52, no. 4, pp. 56-65, Apr. 2014.

[9] W. Si, S. Selvakennedy, and A. Y. Zomaya, “An overview of channel assignment methods for multi-radio multi-channel wireless mesh networks,” *J. Parallel Distrib. Comput.*, vol. 70, no. 5, pp. 505-524, May 2010.

[10] M. Skutella, “An introduction to network flows over time,” *Research Trends in Combinatorial Optimization*, pp. 451-482, 2009.

[11] G. C. Buttazzo, *Hard Real-Time Computing Systems: Predictable Scheduling Algorithms and Applications*. Kluwer Academic Publishers, Norwell, MA, 1997.

[12] M. S Bazaraa, J. J. Jarvis, and H. D. Sherali, *Linear Programming and Network Flows*. John Wiley & Sons, 2011, chapter 8.

[13] Gurobi, http://www.gurobi.com.

[14] Smartone, http:www.smartone.com.

[15] C. L. Liu, and J. W. Layland, “Scheduling algorithms for multiprogramming in a hard-real-time environment,” *J. ACM*, vol. 20, no. 1, pp. 46-61, Jan. 1973.

[16] F. Yao, A. Demers and S. Shenker, “A scheduling model for reduced CPU energy,” in *Proc. IEEE FOCS*, 1995.

Appendix

A. *Reduce Complexity of* PEAK-D2D

To solve PEAK-D2D faster, we will use the following two implementation techniques in space domain and time domain, respectively. In space domain, consider the memory usage for the optimization variable $y^{s\tau}_{uv}(t)$. Consider a network with hundreds of links, *e.g.*, 200, and tens of thousands of traffic demands, *e.g.*, 10,000, and several-slot delay, *e.g.*, 5. We should allocate memory to create at least $200 \times 10,000 \times 5 = 10$ million variables. In Gurobi C++ interface, a variable should be an object of class `GRBVar`. Since we cannot see the source implementation code, we use test codes to evaluate the memory usage for one `GRBVar` object (double type), which is around 100$\sim$150bytes. Thus 10 million variables will consume more than 1 GB memory. One way to reduce the memory usage is to maintain an available link list for each traffic demand $z^{s\tau}$. Since $z^{s\tau}$ has a deadline requirement $d^{s\tau}$ and thus a maximal delay $d^{s\tau} - \tau + 1$, such traffic cannot reach too far away links. Specifically, link (u, v) is available for traffic demand $z^{s\tau}$ only if the shortest path of node s and node u is not larger than $d^{s\tau} - \tau$. Therefore, we only need to create the variable $y^{s\tau}_{uv}(t)$ for those available links (u, v).

In the time domain, we can use multi-thread to speed up model-building time when running in multi-processor operating system. For the traffic scheduling policy constraints in (6a), (6b), (6c), (6d), different traffic demands $z^{s\tau}$ can run concurrently. For the peak traffic constraints in (7b) and (7c), different BSs can run concurrently. Therefore, we can parallelize the constraint-building process. Note that the

Gurobi does not support multi-thread programming for a single environment. One way to use multi-thread is to store a set of `GRBLinExpr` objects and return to the main thread and pass them to the `GRBModel.addConstr()` function.

B. Proof of Theorem 1

Denote

$$I^* = \arg\max_{I \subset [1,T]} g_b(I) = [z_1, z_1']. \tag{24}$$

First, we show that $P_b^{ND} \geq g_b(I^*)$. This is true because the feasible peak traffic P_b^{ND} can finish all traffic demands in the interval I^*, *i.e.*, we must have

$$(z_1' - z_1 + 1)P_b^{ND} \geq \sum_{(s,\tau) \in \mathcal{A}_b(I^*)} \frac{x^{s\tau}}{R_{sb}}. \tag{25}$$

Second, we show that $g_b(I^*)$ can finish all traffic in the interval $[1,T]$ with EDF, *i.e.*, $P_b^{ND} \leq g_b(I^*)$. This can be proved by contradiction. Suppose $g_b(I^*)$ cannot finish all traffic in the interval $[1,T]$. Then we record the time when EDF returns false as z_f, which must be the deadline of a valid yet uncompleted traffic. For any $t \in [1, z_f]$, we define a binary variable h_t to indicate whether or not the peak traffic is fully utilized as follows,

$$h_t = \begin{cases} 1, & \text{if } \alpha_b(t) = g_b(I^*); \\ 0, & \text{otherwise.} \end{cases} \tag{26}$$

Clearly we must have $h_{z_f} = 1$. Now let us define z_0 as the latest time such that $h_t = 0$, *i.e.*, $z_0 = \max_{t \in [1,z_f]: h_t = 0} t$. If $h_t = 1$ for any $t \in [1, z_f]$, then we let $z_0 = 0$. Since $h_{z_0} = 0$, we conclude that all traffic demands whose deadlines are not larger than z_0 have been completed at the end of slot z_0 with EDF algorithm. Then we consider the interval $I' = [z_0+1, z_f]$. Since $h_t = 1$ for any $t \in [z_0+1, z_f]$, we obtain that the total traffic volume delivered in the interval I' is $(z_f - z_0)g(I^*)$. Since EDF returns false at the end of slot z_f, we must have

$$(z_f - z_0) g_b(I^*) < \sum_{(s,\tau) \in \mathcal{A}_b(I')} \frac{x^{s\tau}}{R_{sb}}, \tag{27}$$

which yields to

$$g_b(I') = \frac{\sum_{(s,\tau) \in \mathcal{A}_b(I')} \frac{x^{s\tau}}{R_{sb}}}{z_f - z_0} > g_b(I^*). \tag{28}$$

This is a contradiction to the fact that I^* maximize $g_b(I)$.

Therefore, $P_b^{ND} = g_b(I^*)$.

C. Proof of Theorem 3

The proof logic is to construct a feasible solution to PEAK-ND^b based on the optimal solution with D2D.

Let us denote the optimal traffic scheduling policy for PEAK-D2D as $y_{uv}^{s\tau}(t)$ and the optimal peak traffic for each BS i as P_i^D. Then consider BS $i \in \mathcal{B}$. For each traffic demand $z^{s\tau}$, user $s \in \mathcal{U}_i$ must transmit all volume $x^{s\tau}$ either to BS i directly or any other neighbour users. Thus $\forall s \in \mathcal{U}_i, \tau \in [1,T]$, the following equality holds,

$$x^{s\tau} = \sum_{t=\tau}^{d^{s\tau}} [y_{si}^{s\tau}(t) R_{si} + \sum_{v: v \in \mathcal{U}_i, (s,v) \in \mathcal{E}} y_{sv}^{s\tau}(t) R_{sv} \tag{29a}$$

$$+ \sum_{j:(i,j) \in \mathcal{E}_{D2D}} \sum_{v: v \in \mathcal{U}_j, (s,v) \in \mathcal{E}} y_{sv}^{s\tau}(t) R_{sv}], \tag{29b}$$

In addition, the peak traffic requirement should be satisfied,

$$\sum_{s \in \mathcal{U}_i} \sum_{\tau: \tau \leq t \leq d^{s\tau}} y_{si}^{s\tau}(t) + \sum_{u \in \mathcal{U}_i} \sum_{v \in \text{in}(u) \setminus \{u\}} \sum_{s \in \mathcal{U}} \sum_{\tau: \tau \leq t \leq d^{s\tau}} y_{vu}^{s\tau}(t) \leq P_i^D,$$

Now we construct a feasible solution to PEAK-ND^i, *i.e.*,

$$\bar{y}_{si}^{s\tau}(t) = [y_{si}^{s\tau}(t) + \sum_{v: v \in \mathcal{U}_i, (s,v) \in \mathcal{E}} y_{sv}^{s\tau}(t) \frac{R_{sv}}{R_{si}} \tag{30a}$$

$$+ \sum_{j:(i,j) \in \mathcal{E}_{D2D}} \sum_{v: v \in \mathcal{U}_j, (s,v) \in \mathcal{E}} y_{sv}^{s\tau}(t) \frac{R_{sv}}{R_{si}}], \tag{30b}$$

Thus we have

$$\begin{aligned}
\alpha_i(t) &= \sum_{s \in \mathcal{U}_i} \sum_{\tau: \tau \leq t \leq d^{s\tau}} \bar{y}_{si}^{s\tau}(t) \\
&= \sum_{s \in \mathcal{U}_i} \sum_{\tau: \tau \leq t \leq d^{s\tau}} [y_{si}^{s\tau}(t) + \sum_{v: v \in \mathcal{U}_i, (s,v) \in \mathcal{E}} y_{sv}^{s\tau}(t) \frac{R_{sv}}{R_{si}} \\
&+ \sum_{j:(i,j) \in \mathcal{E}_{D2D}} \sum_{v: v \in \mathcal{U}_j, (s,v) \in \mathcal{E}} y_{sv}^{s\tau}(t) \frac{R_{sv}}{R_{si}}] \\
&\leq \sum_{s \in \mathcal{U}_i} \sum_{\tau: \tau \leq t \leq d^{s\tau}} [y_{si}^{s\tau}(t) + r_s \sum_{v: v \in \mathcal{U}_i, (s,v) \in \mathcal{E}} y_{sv}^{s\tau}(t) \\
&+ \sum_{j:(i,j) \in \mathcal{E}_{D2D}} \tilde{r}_s^j \sum_{v: v \in \mathcal{U}_j, (s,v) \in \mathcal{E}} y_{sv}^{s\tau}(t)] \\
&\overset{(a)}{\leq} \max\{r, 1\} \sum_{s \in \mathcal{U}_i} \sum_{\tau: \tau \leq t \leq d^{s\tau}} [y_{si}^{s\tau}(t) + \sum_{v: v \in \mathcal{U}_i, (s,v) \in \mathcal{E}} y_{sv}^{s\tau}(t)] \\
&+ \tilde{r} \sum_{s \in \mathcal{U}_i} \sum_{\tau: \tau \leq t \leq d^{s\tau}} [\sum_{j:(i,j) \in \mathcal{E}_{D2D}} \sum_{v: v \in \mathcal{U}_j, (s,v) \in \mathcal{E}} y_{sv}^{s\tau}(t)] \\
&\leq \max\{r, 1\} P_i^D + \tilde{r} \sum_{j:(i,j) \in \mathcal{E}_{D2D}} P_j^D,
\end{aligned} \tag{31}$$

where (a) trivially holds for $r > 1$ and also holds for $r \leq 1$ by noting that there is no intra-cell D2D traffic when $r \leq 1$.

Thus $\max\{r,1\}P_i^D + \tilde{r} \sum_{j:(i,j) \in \mathcal{E}_{D2D}} P_j^D$ is a feasible peak traffic for BS i, thus we must have

$$P_i^{ND} \leq \max\{r, 1\} P_i^D + \tilde{r} \sum_{j:(i,j) \in \mathcal{E}_{D2D}} P_j^D. \tag{32}$$

Then we do summation over all BSs and get

$$\begin{aligned}
P_{ND} &= \sum_{i\in\mathcal{B}} P_i^{ND} \le \sum_{i\in\mathcal{B}} \max\{r,1\} P_i^D + \tilde{r} \sum_{i\in\mathcal{B}} \sum_{j:(i,j)\in\mathcal{E}_{D2D}} P_j^D \\
&\overset{(b)}{=} \max\{r,1\} \sum_{i\in\mathcal{B}} P_i^D + \tilde{r} \sum_{j\in\mathcal{B}} \sum_{i:(i,j)\in\mathcal{E}_{D2D}} P_j^D \\
&= \max\{r,1\} \sum_{i\in\mathcal{B}} P_i^D + \tilde{r} \sum_{j\in\mathcal{B}} d_j^- P_j^D \\
&\le \max\{r,1\} \sum_{i\in\mathcal{B}} P_i^D + \tilde{r} \sum_{j\in\mathcal{B}} \Delta^- P_j^D \\
&= [\max\{r,1\} + \tilde{r}\Delta^-] P_D, \qquad (33)
\end{aligned}$$

where (b) holds because any $(i,j) \in \mathcal{E}_{D2D}$ contributes one $\tilde{r}P_j^D$ on both sides. Thus, we conclude that

$$\rho = \frac{P_{ND} - P_D}{P_{ND}} \le \frac{\max\{r,1\} + \tilde{r}\Delta^- - 1}{\max\{r,1\} + \tilde{r}\Delta^-}. \qquad (34)$$

D. Proof of Fact 1

In the ring topology, we assume the BS is indexed from 1 to $N = 2D-1$ counterclockwise. In the case without D2D load balancing, the minimal peak traffic for any BS $i \in [1,N]$ is

$$P_i^{ND} = \frac{V}{D} \triangleq P^{nd}. \qquad (35)$$

In the case with D2D load balancing, we will construct a traffic scheduling policy to achieve the peak traffic for any BS $i \in [1,N]$,

$$P_i^D = \frac{V}{3D-2} \triangleq P^d. \qquad (36)$$

Let us consider BS 1 firstly. For the traffic in BS 1, we first consider the counterclockwise side, *i.e.*, $1 \to 2 \to 3 \to \cdots \to D$. We construct the following traffic scheduling policy from slot 1 to slot D where b_i means BS i and the t-th entry in the braces is the traffic volume at slot t on that link:

- $u_1 \to u_2 : \{\underbrace{P^d,\cdots,P^d}_{D-1},0\}$, $u_2 \to b_2 : \{\underbrace{0,\cdots,0}_{D-1},P^d\}$,
- $u_2 \to u_3 : \{0,\underbrace{P^d,\cdots,P^d}_{D-2},0\}$, $u_3 \to b_3 : \{\underbrace{0,\cdots,0}_{D-1},P^d\}$,
- $\cdots\cdots$
- $u_{D-1} \to u_D : \{\underbrace{0,\cdots,0}_{D-2},P^d,0\}$, $u_D \to b_D : \{\underbrace{0,\cdots,0}_{D-1},P^d\}$.

Clearly, the counterclockwise side BSs can help transfer $(D-1)P^d$ traffic for user u_1. We can construct the same traffic scheduling for the clockwise side, *i.e.*, $1 \to (2D-1) \to (2D-2) \to \cdots \to (D+1)$ such that they also help transfer $(D-1)P^d$ traffic for user u_1. In addition, user u_1 can directly transmit DP^d traffic to BS 1 as

- $u_1 \to b_1 : \{\underbrace{P^d,P^d,\cdots,P^d}_{D}\}$.

Hence, all the traffic for user u_1 has been finished before its deadline (slot D) because

$$DP^d + (D-1)P^d + (D-1)P^d = (3D-2)P^d = V.$$

Furthermore, we can check that the peak traffic for all N BSs is $P^d = \frac{V}{3D-2}$.

In addition, since the ring topology is symmetric and all traffic is decoupled, we immediately get that all other traffic can be satisfied when the peak traffic for all BSs is P^d.

Therefore, we get the sum peak traffic reduction

$$\rho = \frac{P_{ND} - P_D}{P_{ND}} = \frac{NP^{nd} - NP^d}{NP^{nd}} = \frac{2(D-1)}{3D-2} \to \frac{2}{3}\ (D \to \infty).$$

In addition, the sum D2D traffic for all users is,

$$\begin{aligned}
V_{D2D} &= N \cdot 2(P^d + 2P^d + \cdots + (D-1)P^d) = 2NP^d \sum_{i=1}^{D-1} i \\
&= (D-1)D \cdot \frac{NV}{3D-2},
\end{aligned}$$

and the sum traffic directly sent by users to BSs is the total traffic volume for all users in the given traffic demand pattern, *i.e.*, $V_{BS} = NV$. Thus, the overhead ratio is

$$\eta = \frac{V_{D2D}}{V_{D2D} + V_{BS}} = \frac{D(D-1)}{D^2 + 2D - 2}.$$

The proof is completed.

E. Proof of Fact 2

In the case without D2D load balancing, the minimal peak traffic for any BS $i \in [1,N]$ is

$$P_i^{ND} = \frac{V}{D} \triangleq P^{nd}. \qquad (37)$$

In the case with D2D load balancing, we will construct a traffic scheduling policy to achieve the peak traffic for any BS $i \in [1,N]$,

$$P_i^D = \frac{2V}{(N+1)D} \triangleq P^d. \qquad (38)$$

We first consider the traffic for user u_1 and construct the following traffic scheduling policy:

- Case 1 when D is even: $\forall i \in [2,N]$,
  $u_1 \to u_i : \quad \{\underbrace{P^d,\cdots,P^d}_{D/2},\underbrace{0,\cdots,0}_{D/2}\}$,
  $u_i \to b_i : \quad \{\underbrace{0,\cdots,0}_{D/2},\underbrace{P^d,\cdots,P^d}_{D/2}\}$.
- Case 2 when D is odd: $\forall i \in [2,N]$,
  $u_1 \to u_i : \quad \{\underbrace{P^d,\cdots,P^d}_{(D-1)/2},\frac{P^d}{2},\underbrace{0,\cdots,0}_{(D-1)/2}\}$,
  $u_i \to b_i : \quad \{\underbrace{0,\cdots,0}_{(D-1)/2},\frac{P^d}{2},\underbrace{P^d,\cdots,P^d}_{(D-1)/2}\}$.

In both cases, any other BS $i \in [2,N]$ can help transfer $\frac{D}{2}P^d$ traffic for user u_1. Besides, user u_1 can transmit DP^d traffic to BS 1 as:

- $u_1 \to b_1 : \quad \{\underbrace{P^d,\cdots,P^d}_{D}\}$.

Then we can check all traffic for user u_1 has been finished before the deadline (slot D) because

$$DP^d + (N-1)\frac{D}{2}P^d = \frac{N+1}{2}DP^d = V.$$

In addition, we can see that the peak traffic for all BSs is P^d.

Since the complete topology is symmetric and all traffic is decoupled, the traffic for all other users can be satisfied when the peak traffic for all BSs is P^d.

Therefore, the sum peak traffic reduction is

$$\rho = \frac{P_{ND} - P_D}{P_{ND}} = \frac{NP^{nd} - NP^d}{NP^{nd}} = \frac{N-1}{N+1} \to 1 \ (N \to \infty).$$

In addition, the sum D2D traffic for all users is,

$$V_{D2D} = N \cdot (N-1)\frac{D}{2}P^d = \frac{N(N-1)V}{N+1},$$

and the sum traffic directly sent by users to BSs is the total traffic volume for all users in the given traffic demand pattern, *i.e.*, $V_{BS} = NV$. Thus, the overhead ratio is,

$$\eta = \frac{V_{D2D}}{V_{D2D} + V_{BS}} = \frac{\frac{N(N-1)V}{N+1}}{\frac{N(N-1)V}{N+1} + NV} = \frac{N-1}{2N}. \qquad (39)$$

The proof is completed.