D2D Communications Assisted Traffic Offloading in Integrated Cellular-WiFi Networks

Bing Feng, Chi Zhang, *Member, IEEE*, Jianqing Liu, and Yuguang Fang, *Fellow, IEEE*

***Abstract*—Offloading cellular traffic to WiFi networks plays an important role in alleviating the increasing burden on cellular networks. However, excessive traffic offloading brings severe packet collisions into a WiFi network due to its contention-based medium access scheme, which significantly reduces the WiFi network's throughput. In this paper, we propose DAO, a device-to-device (D2D) communications assisted traffic offloading scheme to improve the amount of traffic offloaded from cellular to WiFi in integrated cellular and WiFi networks. Specifically, in an integrated cellular-WiFi network, the cellular network exploits D2D communications in licensed cellular bands to aggregate traffic from cellular users before offloading it to the WiFi network to reduce the number of contending users in WiFi access. The traffic offloading process in DAO is formulated as an optimization problem that jointly takes into account the activations of aggregation nodes (ANs) and the connections between ANs and offloading users to maximize the offloaded traffic while guaranteeing the long-term data rates required by the offloading users. Extensive simulation results reveal the significant performance gain achieved by DAO over the existing schemes.**

***Index Terms*—Traffic offloading, integrated cellular-WiFi networks, D2D communications.**

I. Introduction

With the popularity of smart devices and mobile applications, wireless traffic is explosively growing, which raises significant challenges to cellular networks. The emergence of Internet-of-Things (IoT) [1] where hundreds of billions of IoT devices are connected will further add traffic burden on cellular networks. To support the explosion of wireless traffic, cellular operators are continuously upgrading existing cellular networks to next generation communications systems (5G and beyond) [2]. However, the evolvement of a cellular system may not keep up with the dramatic growth of wireless traffic. In addition, upgrading cellular systems requires deploying and maintaining expensive infrastructure and adding more spectrum.

Offloading cellular traffic to existing radio access networks, such as WiFi network, has been proposed as a cost-effective and practical solution to reducing the stress on cellular networks [3]–[5]. Currently, as WiFi access points (APs) are cheap and easy to deploy, cellular operators (e.g., Verizon, AT&T, and Vodafone) have already deployed WiFi networks in densely populated areas [6]. There are have been some works [6] [7] focusing on the framework design for the integration of cellular and WiFi networks. In addition, a few papers [8]–[10] investigate traffic offloading in integrated cellular-WiFi networks. However, in the existing works, cellular users who intend to offload traffic to a WiFi network (i.e., offloading users) directly connect to a WiFi AP using their WiFi interfaces. As a result, the WiFi network has poor throughput due to severe access collisions when there exist a large number of offloading users. Therefore, how to enhance the throughput of the WiFi network in an integrated cellular-WiFi network in traffic offloading is still an open problem.

The throughput of a WiFi network depends on many factors. As we know, the contention-based medium access control (MAC) scheme, distributed coordinated function (DCF), is employed in current WiFi networks. Previous performance analysis [11]–[14] has shown that the DCF throughput drops sharply as the access contention (i.e., the number of contending nodes) increases. In addition, since different WiFi nodes have different data rates in a practical WiFi network, Joshi et al. [15] has shown that, in a multi-rate WiFi network, the low-rate nodes lower the overall throughput of the WiFi network. To mitigate throughput degradation induced by low-rate nodes, the relay-enabled DCF [16] has been proposed to exploit relay nodes to assist low-rate nodes' transmissions.

Motivated by the above observations, we propose DAO (D2D communications Assisted traffic Offloading), which exploits D2D communications in licensed cellular bands to assist traffic offloading from cellular to WiFi in integrated cellular-WiFi networks. D2D communications is a promising technique in 5G to support direct communications between two cellular users without traversing the core network by reusing cellular spectrum (i.e., underlay mode). In DAO, with D2D communications, an offloading user (i.e., D2D transmitter) directly communicates with an aggregation node (i.e., D2D receiver). Each aggregation node (AN) has two radio interfaces and establishes associations with both cellular and WiFi (i.e., dual connectivity) [8], so it can simultaneously receive traffic offloaded from cellular users via licensed D2D communications and transmit aggregated traffic to WiFi APs

This work was partially supported by the Natural Science Foundation of China under grant 61871362 and by the National Key Research and Development Program of China under grant 2017YFB0802202. The work of J. Liu and Y. Fang was partially supported by the US National Science Foundation under grants CNS-1717736 and CNS-1409797.

B. Feng and C. Zhang are with the School of Information Science and Technology, University of Science and Technology of China, Hefei 230027, P. R. China (e-mail: fengice@mail.ustc.edu.cn, chizhang@ustc.edu.cn).

J. Liu is with the Department of Electrical and Computer Engineering, University of Alabama in Huntsville, Huntsville, Alabama 35899, USA (e-mail: jianqing.liu@uah.edu).

Y. Fang is with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, Florida 32611, USA (e-mail: fang@ece.ufl.edu).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2019.2922550, IEEE Internet of Things Journal

via unlicensed WiFi links. The proposed DAO achieves traffic aggregation through D2D communications in licensed cellular bands, which reduces the number of offloading users involved in access contention in the WiFi network. Notice that the traffic aggregation in DAO is also beneficial to transmit large MAC frames in WiFi networks by frame aggregation [17] to improve WiFi networks' MAC efficiency. In addition, the WiFi network in DAO avoids the multi-rate scenario because the ANs can be properly selected to ensure that all of them have high data rates. The throughput of the WiFi network in DAO is enhanced so that more cellular traffic can be migrated to the WiFi network. To the best of our knowledge, DAO is the first scheme in the literature that exploits D2D communications in licensed cellular bands to assist traffic offloading in integrated cellular-WiFi networks.

In the design of DAO, we investigate the practical problems of AN activation, offloading user connection, power control, and resource allocation. The set of activated ANs determines the number of contending nodes in the WiFi network. When performing offloading user connection, an offloading user can be connected to an activated AN only if the amount of resource allocated by the activated AN can satisfy its quality of service (QoS) constraint in terms of the long-term data rate. In addition, in the process of offloading user connection, power control [18] [19] is crucial to avoid interferences caused by D2D communications to cellular users because D2D communications reuse the licensed bands of the cellular users. Therefore, the process of offloading user connection in DAO implicitly takes into account resource allocation and power control. To avoid WiFi network congestion, the process of offloading user connection should ensure that the average arrival rate (i.e., the average rate of aggregated cellular traffic) at each activated AN should be smaller than its average service rate (i.e., the average per-AN throughput in the WiFi network).

We formulate the traffic offloading process in DAO as a mixed integer nonlinear programming (MINLP) problem. As the time-scale for AN activation is much larger than that for offloading user connection, the problem is decomposed into two subproblems, namely, AN activation problem and offloading user connection problem. An exhaustive search algorithm is proposed to solve the AN activation problem because the number of ANs in DAO is small. Given the result of AN activation, the offloading user connection problem is a mixed integer linear programming (MILP) problem. We first relax the integer variables to find an upper bound for the offloading user connection problem and use this upper bound as a performance measure. Then, we propose a sequential fixing with checking (SFC) algorithm with polynomial time to derive a near-optimal solution to the offloading user connection problem. Simulation results show that the lower bound obtained by the SFC algorithm is very close to the upper bound.

The rest of this paper is organized as follows. Section II gives an overview of related work. Section III introduces the system of the proposed traffic offloading scheme and formulates the optimization problem. Section IV presents the algorithm design. Section V presents extensive simulation results and analysis. Finally, Section VI concludes this paper.

II. Related Work

There is a large body of work attempting to offload cellular traffic to WiFi networks. Poularakis et al. [20] proposed to offload mobile data by exploiting the bandwidth and cache resources in residential WiFi APs. Gao et al. [9] proposed to offload delay-sensitive traffic to WiFi networks and designed a dynamic traffic offloading algorithm. Sou et al. [21] presented an analytical model for multipath offloading where TCP packets are transmitted seamlessly across multiple wireless interfaces. Poularakis et al. [22] investigated the optimal deployment of WiFi offloading infrastructure to maximize carrier profits. The congestion-aware network selection schemes [4] [10] had been proposed to avoid excessive traffic offloading by limiting the amount of offloaded traffic to the WiFi network. Chen et al. [5] developed a hybrid method that offloads cellular traffic to WiFi networks and simultaneously enables cellular network to exploit licensed-assisted access (LAA) to transmit traffic in unlicensed bands. Fan et al. [23] investigated the tradeoff between throughput and power consumption when designing an association scheme in WLAN/cellular integrated networks. Different from the existing traffic offloading schemes where offloading users directly connect to a WiFi AP, DAO aggregates offloading users' cellular traffic via D2D communications in licensed cellular bands before offloading it to the WiFi network.

D2D communications as a promising offloading solution in cellular networks has been widely studied. Liu et al. [24] exploited inter-cell D2D communications to offload traffic from a congested cell to its adjacent cells that are very lightly loaded. For a group of cellular users that share similar interests, Wu et al. [25] utilized D2D communications to form a local *ad hoc* network to distribute common content, which reduces traffic burden on the base stations (BSs). Asadi et al. [26] proposed an architecture to leverage D2D communications in unlicensed spectrum to relay cellular traffic. To offload traffic from macro BSs to small BSs, Cao et al. [27] exploited D2D communication to relay data transmissions of users that are inside the coverage range of macro BSs but outside the coverage range of small BSs. Different from these works, DAO exploits D2D communications in licensed bands to assist traffic offloading in integrated cellular-WiFi networks.

To mitigate the impact of low-rate nodes on the capacity of multi-rate WiFi networks, relay-aided media access schemes have been proposed. Zhu et al. [16] proposed a relay-enabled DCF where after low-rate nodes obtain the chance through DCF to access the channel, they exploit relay nodes to convey data packets instead of directly transmitting. Lim et al. [28] proposed to deploy Proxy Relay Points (PRPs) to perform relay communications in practical WiFi networks. In relay-aided MAC schemes, all nodes still need to perform DCF to contend for the channel, which cannot reduce transmission collisions. On the contrary, DAO exploits licensed D2D communications to reduce the number of contending users in WiFi access.

III. System Model and Problem Formulation

In DAO, we assume that both cellular network and WiFi network are owned and managed by a single cellular operator.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2019.2922550, IEEE Internet of Things Journal

TABLE I
PARAMETER NOTATIONS

Parameter	Description
$\mathcal{N}$	The set of offloading users
$\mathcal{M}$	The set of aggregation nodes (ANs)
y_m	The indicator for the activation of AN m
x_{nm}	The indicator for the connection between offloading user n and AN m
l^D_{nm}	The D2D link from offloading user n to AN m
α_{nm}	The fraction of allocated time resource to offloading user n by AN m
$\mathcal{K}$	The set of cellular users (CUs)
p^D_{nm}	The transmit power of offloading user n on l^D_{nm}
R^D_{nm}	The achievable data rate of offloading user n on l^D_{nm}
P^{max}_n	The maximum transmit power of offloading user n
R^{min}_m	The minimum data rate required by cellular user m
R^C_m	The achievable data rate of cellular user m
γ_n	The long-term data rate required by offloading user n
$\mathcal{M}_a$	The set of activated ANs
M_a	The number of activated ANs
λ_m	The average rate of aggregated traffic at activated AN m
$\mu^{M_a}_m$	The average service rate of activated AN m in an unsaturated WiFi network with M_a activated ANs
$\hat{\mu}^{M_a}_m$	The average service rate of activated AN m in a saturated WiFi network with M_a activated ANs

The improved capacity of offloaded traffic achieved by DAO motivates cellular operator to exploit licensed D2D communications to assist traffic offloading. The traffic offloading process in DAO is under the control of BS. Different from the existing traffic offloading schemes where offloading users directly connect to a WiFi AP, offloading users in DAO still use their cellular interfaces to establish D2D connections, so they do not need to perform association with a WiFi network. In addition, with the assistance of D2D communications, the offloading users in DAO do not have to be in the coverage of a WiFi network. Table 1 presents the notations used in this paper.

In Fig. 1, we use $\mathcal{N} = \{1, ..., N\}$ to denote the set of offloading users that shift their uplink cellular traffic to a WiFi network, and $\mathcal{M} = \{1, ..., M\}$ to denote the set of aggregation nodes (ANs) that can simultaneously associate with a WiFi AP and a cellular BS via dual connectivity. In DAO, the number of activated ANs determines the number of contending nodes in the WiFi network. For each AN $m \in \mathcal{M}$, we define binary variable $y_m \in \{0, 1\}$ to indicate whether AN m is activated or not (i.e., AN activation),

$$y_m = \begin{cases} 1, & \text{AN } m \text{ is activated} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Instead of directly connecting to the WiFi AP, each offloading user n is connected to an AN m via D2D communications, which forms a D2D link from offloading user n to AN m, denoted by l^D_{nm}. For each D2D link, a binary variable $x_{nm} \in \{0, 1\}$ is introduced to indicate whether offloading user $n \in \mathcal{N}$ is connected to AN $m \in \mathcal{M}$ or not (i.e., offloading user connection),

$$x_{nm} = \begin{cases} 1, & \text{offloading user } n \text{ is connected to AN } m \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

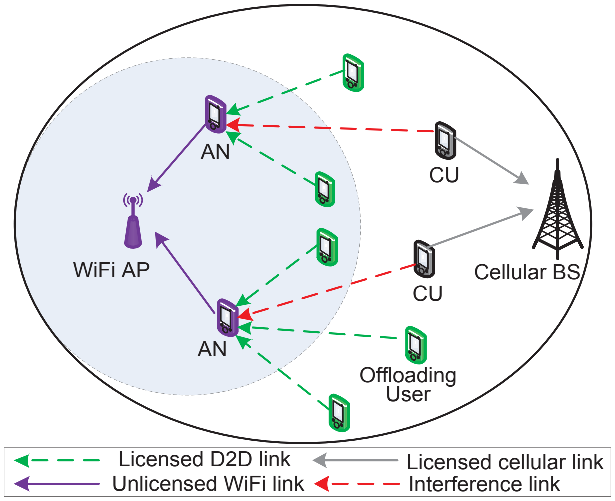


Fig. 1. The system model of DAO.

Obviously, $x_{nm} \le y_m$. With x_{nm}, BS is responsible for directing offloading users to establish D2D connections with corresponding ANs. Moreover, each offloading user can be connected to at most one AN. Thus, we have

$$\sum_{m \in \mathcal{M}} x_{nm} \le 1, \quad \forall n \in \mathcal{N}. \tag{3}$$

For convenience, we denote offloading user connection matrix and AN activation vector as $\boldsymbol{x} = [x_{nm}]_{\forall n,m}$ and $\boldsymbol{y} = [y_1, ..., y_M]$, respectively.

In DAO, the offloading users connected to an AN adopt time division mode to achieve multiple access. Let $\alpha_{nm} \in [0, 1]$ denote the fraction of time resource allocated to offloading user n connected to AN m. $\alpha_{nm} > 0$ if and only if offloading user n is connected to AN m that is activated, which implies that $\alpha_{nm} \le y_m x_{nm}$. We denote the resource allocation matrix as $\boldsymbol{\alpha} = [\alpha_{nm}]_{\forall n,m}$. The total time resource allocated to offloading users connected to AN m is limited by

$$\sum_{n \in \mathcal{N}} \alpha_{nm} \le 1, \quad \forall m \in \mathcal{M}. \tag{4}$$

A. D2D Communications

In DAO, D2D communications reuse cellular users' uplink cellular channels. In Fig. 1, we use $\mathcal{K} = \{1, ..., M\}$ to denote the set of cellular users (CUs) whose uplink cellular channels are reused by D2D communications. Without loss of generality, we assume that AN $m \in \mathcal{M}$ is allocated with the uplink channel of cellular user $m \in \mathcal{K}$. That is, the D2D links of offloading users connected to AN $m \in \mathcal{M}$ all reuse the uplink channel of cellular user $m \in \mathcal{K}$. Thus, the D2D link from offloading user n to AN m, l^D_{nm}, only has interference from the cellular user m whose uplink channel is reused by this D2D link. Then, the signal to interference plus noise ratio (SINR) at the D2D receiver of the D2D link l^D_{nm} from offloading user $n \in \mathcal{N}$ to AN $m \in \mathcal{M}$, Γ^D_{nm}, is given by

$$\Gamma^D_{nm} = \frac{p^D_{nm} h^D_{nm}}{p^C_m h^I_{mm} + N_0}, \tag{5}$$

where p^D_{nm} is the transmit power of offloading user n on D2D link l^D_{nm}, h^D_{nm} is the channel gain of D2D link l^D_{nm}, p^C_m is the transmit power of cellular user $m \in \mathcal{K}$, h^I_{mm} is the channel gain of the interference link from cellular user m to AN m, and N_0 is the noise power. We denote the transmit power matrix as $\boldsymbol{p^D} = [p^D_{nm}]_{\forall n,m}$. Similarly, the SINR at the receiver of the uplink channel from cellular user $m \in \mathcal{K}$ to its associated BS is

$$\Gamma^C_m = \frac{p^C_m h^C_m}{p^D_{nm} h^I_{nm} + N_0}, \tag{6}$$

where h^C_m is the channel gain of the uplink channel from cellular user m to its associated BS and h^I_{nm} is the channel gain of the interference link from offloading user n to the BS associated by cellular user m.

The achievable data rate of cellular user m on its uplink channel can be calculated as

$$R^C_m = B_0 \text{log}_2(1 + \Gamma^C_m), \tag{7}$$

where B_0 is the bandwidth of the uplink channel. Similarly, the achievable data rate of offloading user n on D2D link l^D_{nm} can be calculated as

$$R^D_{nm} = \alpha_{nm} r^D_{nm}, \tag{8}$$

where $r^D_{nm} = B_0 \text{log}_2(1 + \Gamma^D_{nm})$.

In DAO, we assume each cellular user $m \in \mathcal{K}$ has a fixed value of transmit power p^C_m. To guarantee the minimum data rate R^{min}_m required by the cellular user m, it is important to control the offloading user's transmit power p^D_{nm}. The rate constraint is $R^C_m \geq R^{min}_m, \forall m \in \mathcal{K}$. The corresponding constraint of transmit power p^D_{nm} on the D2D link l^D_{nm} is given as $p^D_{nm} \leq (1/h^I_{mn})((p^C_m h^C_m/(2^{(R^{min}_m/B_0)} - 1)) - N_0), \forall n \in \mathcal{N}$. In addition, the maximum transmit power of offloading user n, P^{max}_n, also limits the transmit power p^D_{nm} on D2D link l^D_{nm}, i.e., $p^D_{nm} \leq P^{max}_n, \forall n \in \mathcal{N}$. Thus,

$$p^D_{nm} \leq p^{max}_{nm}, \quad \forall n \in \mathcal{N}, \tag{9}$$

where

$$p^{max}_{nm} = \min \left\{ P^{max}_n, \frac{1}{h^I_{mm}} \left(\frac{p^C_m h^C_m}{2^{(R^{min}_m/B_0)}} - N_0 \right) \right\}. \tag{10}$$

In DAO, the process of offloading user connection implicitly takes into account power control. The transmit power p^D_{nm} is fixed to p^{max}_{nm} if $x_{nm} = 1$, and $p^D_{nm} = 0$ otherwise. We rewrite (8) as

$$\hat{R}^D_{nm} = \alpha_{nm} \hat{r}^D_{nm}, \tag{11}$$

where $\hat{r}^D_{nm} = B_0 \text{log}_2(1 + \frac{p^{max}_{nm} h^D_{nm}}{p^C_m h^I_{mm} + N_0})$.

B. QoS Constraint

In the process of offloading user connection, an offloading user will not be connected to an AN if the AN cannot satisfy the offloading user's QoS requirement in terms of the long-term data rate. Let γ_n denote the long-term data rate required by offloading user n. Then, we have

$$\hat{R}^D_{nm} \geq \gamma_n, \quad \forall n \in \mathcal{N}. \tag{12}$$

In DAO, the process of offloading user connection implicitly takes into account resource allocation. If $x_{nm} = 1$, the fraction of time resource allocated to offloading user n is $\alpha_{nm} = \alpha^R_{nm}$, where the minimum fraction of time resource required by offloading user n to satisfy its QoS requirement, α^R_{nm}, is

$$\alpha^R_{nm} = \frac{\gamma_n}{\hat{r}^D_{nm}}. \tag{13}$$

C. WiFi Throughput Constraint

The ANs in DAO play a key role in performing admission control that limits the amount of traffic admitted into the WiFi network. Both AN activations and offloading user connections are used to regulate the traffic to the WiFi network such that the WiFi network can achieve the maximum throughput while avoiding network congestion. Let $\mathcal{M}_a = \{1, ..., M_a\}$ denote the set of activated ANs. Obviously, $\mathcal{M}_a \subseteq \mathcal{M}$. The number of activated ANs is $M_a = \sum_{m \in \mathcal{M}} y_m$. To characterize the traffic in a WiFi network, each activated AN $m \in \mathcal{M}_a$ in the WiFi network is modeled as a single-server queueing system, whose average arrival rate and average service rate are denoted by λ_m and $\mu^{M_a}_m$, respectively. We assume each activated AN m itself does not generate traffic and only aggregates traffic from offloading users connected to it. Thus, the average arrival rate (i.e., the average rate of aggregated traffic) at activated AN m, λ_m, is given by

$$\lambda_m = \sum_{n \in \mathcal{N}} x_{nm} \hat{R}^D_{nm}, \quad \forall m \in \mathcal{M}_a. \tag{14}$$

If too many offloading users are connected to activated AN m, i.e., $\lambda_m > \mu^{M_a}_m$, then $\mu^{M_a}_m$ becomes the bottleneck in offloading cellular traffic to the WiFi AP, which results in unacceptable delay at activated AN m. Therefore, the average arrival rate of activated AN m, λ_m, should be kept below its average service rate, $\mu^{M_a}_m$, i.e.,

$$\lambda_m < \mu^{M_a}_m, \quad \forall m \in \mathcal{M}_a. \tag{15}$$

Thus, the WiFi network is not congested (i.e., unsaturated). However, previous works [29] [30] have shown that it is complex to obtain the value of the average service rate $\mu^{M_a}_m$ in an unsaturated WiFi network where the contention-based DCF scheme is employed. In addition to the MAC parameters of DCF and the WiFi network size M_a, the average service rate $\mu^{M_a}_m$ of activated AN m depends on the average arrival rates of the other $M_a - 1$ activated ANs. Since the result of offloading user connection affects the average arrival rates of the other $M_a - 1$ activated ANs, the average service rate $\mu^{M_a}_m$ depends on the result of offloading user connection. On the other hand, the value of $\mu^{M_a}_m$ affects the result of offloading user connection. Therefore, $\mu^{M_a}_m$ is coupled with offloading user connection.

To decouple $\mu^{M_a}_m$ from the process of offloading user connection, $\hat{\mu}^{M_a}_m$, a lower bound on the average service rate $\mu^{M_a}_m$ is used to limit the amount of aggregated traffic at each activated AN. $\hat{\mu}^{M_a}_m$ is the average service rate of activated AN m in a saturated WiFi network with M_a activated ANs. Notice that the value of $\hat{\mu}^{M_a}_m$ is equal to the throughput of activated AN m in a saturated WiFi network. Given the MAC parameters of DCF, we will show that $\hat{\mu}_m^{M_a}$ only depends on M_a. Let P_{tr} and P_s be the probabilities that at least one activated AN transmits and that exactly one activated AN transmits, respectively. Then,

$$P_{tr} = 1 - (1-\tau)^{M_a} \tag{16}$$

$$P_s = M_a \tau (1-\tau)^{M_a - 1} / P_{tr} \tag{17}$$

where τ is the transmission probability of each activated AN. Let $\hat{\mu}^{M_a}$ denote the system throughput of a saturated WiFi network where each activated AN always has packets to transmit. Then, $\hat{\mu}^{M_a}$ can be expressed as [12] [31]

$$\hat{\mu}^{M_a} = \frac{P_s P_{tr} E[P]}{(1-P_{tr})T_i + P_{tr} P_s T_s + P_{tr}(1-P_s)T_c} \tag{18}$$

where $E[P]$ is the average duration of data packet payload, T_i is the duration of an idle time slot, T_s is the average duration of a successful transmission, and T_c is the average duration of a frame collision. From (18), we can observe that, given the MAC parameters of DCF, $\hat{\mu}^{M_a}$ is only a function of the number of activated ANs, M_a.

DCF provides equal opportunities to all contending nodes in a WiFi network for channel access in a long run [31]. If all contending nodes in the WiFi network have the same data packet length, the throughput of activated AN m in a saturated WiFi network (i.e., the per-AN throughput) is $\hat{\mu}_m^{M_a} = \frac{1}{M_a}\hat{\mu}^{M_a}$ [23]. Given the MAC parameters of DCF and the WiFi network size M_a, $\hat{\mu}_m^{M_a}$ is a fixed value that does not depend on the result of offloading user connection.

With different network parameters (data rate and data packet length) configured in a WiFi network, the maximum values of $\hat{\mu}^{M_a}$ are achieved at different values of M_a. Notice that, when the WiFi network only includes one node, i.e., $M_a = 1$, the corresponding system throughput $\hat{\mu}^{M_a}$ is not the maximum value due to the contention access scheme DCF. With data rate of 48 Mbps and data packet length of 2046 Bytes, the maximum $\hat{\mu}^{M_a}$ is achieved when $M_a = 5$ [31]. The optimal set of activated ANs should be determined based on practical network scenarios.

If λ_m is limited to be less than $\hat{\mu}_m^{M_a}$ at each activated AN, the WiFi network is guaranteed not in saturation, which can avoid WiFi network congestion. Therefore, the constraint in (15) is relaxed to

$$\lambda_m < \hat{\mu}_m^{M_a}, \quad \forall m \in \mathcal{M}_a. \tag{19}$$

Notice that (19) means each activated AN has the same average service rate, which implies that, compared with the constraint (15), the constraint (19) ensures a certain level of load balance among activated ANs.

In an unsaturated WiFi network satisfying the constraint (15) or (19), the throughput of activated AN m, denoted by $S_m^{M_a}$, is the average rate of traffic transmitted from activated AN m to the WiFi AP (i.e., the average departure rate), which is equal to the average arrival rate, λ_m, i.e., $S_m^{M_a} = \lambda_m$. Then, the system throughput of an unsaturated WiFi network with M_a activated ANs, denoted by S^{M_a}, is

$$S^{M_a} = \sum_{m \in \mathcal{M}_a} S_m^{M_a} = \sum_{m \in \mathcal{M}_a} \lambda_m. \tag{20}$$

Obviously, $S^{M_a} \le \hat{\mu}^{M_a}$ according to (19). Notice that $\sum_{m \in \mathcal{M}_a} \lambda_m$ is the total amount of traffic aggregated from all offloading users. Thus, under unsaturation conditions, the total amount of offloaded traffic in DAO is equal to the system throughput of the unsaturated WiFi network in DAO.

D. Problem Formulation

Based on the system model, the traffic offloading process in DAO is formulated as an optimization problem. In the process of offloading user connection, if offloading user $n \in \mathcal{N}$ is admitted to connect to AN $m \in \mathcal{M}$, i.e., $x_{nm} = 1$, the minimum time resource is allocated to offloading user n, i.e., $\alpha_{nm} = \alpha_{nm}^R$, to satisfy its QoS constraint. In addition, the power control $p_{nm}^D = p_{nm}^{max}$ implies $r_{nm}^D = \hat{r}_{nm}^D$. Thus, $\alpha_{nm} r_{nm}^D = \alpha_{nm}^R \hat{r}_{nm}^D = \gamma_n$. Our objective is to maximize the amount of offloaded traffic, which is equivalent to maximizing the amount of traffic aggregated from all offloading users. Notice that the objective function implicitly takes into account power control $p_{nm}^D = p_{nm}^{max}$ and resource allocation $\alpha_{nm} = \alpha_{nm}^R$. The detailed problem formulation is given by

$$\mathbf{P1}: \quad \max_{\boldsymbol{y},\boldsymbol{x}} \quad \sum_{m \in \mathcal{M}} \sum_{n \in \mathcal{N}} y_m x_{nm} \alpha_{nm}^R \hat{r}_{nm}^D \tag{21}$$

$$s.t. \quad x_{nm} \le y_m \quad \forall n \in \mathcal{N} \quad \forall m \in \mathcal{M} \tag{21a}$$

$$\sum_{m \in \mathcal{M}} x_{nm} \le 1 \quad \forall n \in \mathcal{N} \tag{21b}$$

$$\sum_{n \in \mathcal{N}} x_{nm} \alpha_{nm}^R \le y_m \quad \forall m \in \mathcal{M} \tag{21c}$$

$$\sum_{n \in \mathcal{N}} x_{nm} \alpha_{nm}^R \hat{r}_{nm}^D \le y_m \hat{\mu}_m^{M_a} \quad \forall m \in \mathcal{M} \tag{21d}$$

$$y_m, x_{nm} \in \{0,1\} \quad \forall n \in \mathcal{N} \quad \forall m \in \mathcal{M} \tag{21e}$$

where (21a) indicates that offloading user n can connect to AN m only if AN m is activated; (21b) implies that each offloading user can be connected to at most one AN; (21c) accounts for the constraint of time resource available at AN m, and (21c) also ensures $\sum_{n \in \mathcal{N}} x_{nm} \alpha_{nm}^R = 0$ if AN m is not activated ($y_m = 0$); (21d) accounts for the constraint of average service rate of AN m in the WiFi network, and (21d) also ensures $\sum_{n \in \mathcal{N}} x_{nm} \alpha_{nm}^R \hat{r}_{nm}^D = 0$ if AN m is not activated ($y_m = 0$); (21e) indicates that y_m and x_{nm} are binary variables. Notice that the value of $\hat{\mu}_m^{M_a}$ depends on the number of activated ANs M_a ($M_a = \sum_{m \in \mathcal{M}} y_m$) and $\hat{\mu}_m^{M_a}$ is a non-linear function in y_m.

IV. Algorithm Design

The problem **P1** is an MINLP problem, which is NP-hard in general. The complexity of problem **P1** arises from both the binary variables (x_{nm} and y_m) and the complex coupling between offloading user connection and AN activation. Simply relaxing binary variables x_{nm} and y_m is still a non-linear programming (NLP) problem. Based on the observation that the period on which the set of activated ANs are determined is larger than the time-scale for performing offloading user connection, we decompose the original problem **P1** into a higher level AN activation problem and a lower level offloading user connection problem. The higher level AN activation problem is solved at a slower time-scale than the lower level offloading user connection problem. Under a given set $\boldsymbol{y}$, a near-optimal solution to the lower level offloading user connection problem is obtained by the proposed sequential fixing with checking (SFC) algorithm with polynomial time. Then, we can utilize the exhaust algorithm to obtain the optimal set $\boldsymbol{y}$ among all possible sets of $\boldsymbol{y}$ because the number of ANs, M, is a small value.

A. Offloading User Connection

With fixed AN activation $\boldsymbol{y}$, i.e., given the set of activated ANs, $\mathcal{M}_a$, the problem **P1** is transformed into

$$\textbf{P2}: \quad \max_{\boldsymbol{x}} \quad \sum_{m\in\mathcal{M}_a}\sum_{n\in\mathcal{N}} x_{nm}\alpha_{nm}^{R}\hat{r}_{nm}^{D} \tag{22}$$

$$s.t. \quad x_{nm}\in\{0,1\} \quad \forall n\in\mathcal{N} \quad \forall m\in\mathcal{M}_a \tag{22a}$$

$$\sum_{m\in\mathcal{M}_a} x_{nm}\le 1 \quad \forall n\in\mathcal{N} \tag{22b}$$

$$\sum_{n\in\mathcal{N}} x_{nm}\alpha_{nm}^{R}\le 1 \quad \forall m\in\mathcal{M}_a \tag{22c}$$

$$\sum_{n\in\mathcal{N}} x_{nm}\alpha_{nm}^{R}\hat{r}_{nm}^{D}\le \hat{\mu}_{m}^{M_a} \quad \forall m\in\mathcal{M}_a \tag{22d}$$

Observe that, the problem **P2** is a mixed integer linear programming (MILP) problem with respect to x_{nm}. It can be approximately solved by the branch-and-bound algorithm that has exponential complexity in the worst-case. Therefore, we propose the SFC algorithm, a heuristic algorithm with polynomial time, to derive a near-optimal solution to the offloading user connection problem.

Specifically, we relax the binary variable x_{nm} to continuous value in $[0,1]$. Thus, each offloading user is allowed to connect to multiple ANs instead of just one AN. The relaxed problem is given by

$$\textbf{P2R}: \quad \max_{\boldsymbol{x}} \quad \sum_{m\in\mathcal{M}_a}\sum_{n\in\mathcal{N}} x_{nm}\alpha_{nm}^{R}\hat{r}_{nm}^{D} \tag{23}$$

$$s.t. \quad x_{nm}\in[0,1] \quad \forall n\in\mathcal{N} \quad \forall m\in\mathcal{M}_a \tag{23a}$$

$$(22b)-(22d)$$

The problem **P2R** is a linear programming (LP) problem that can be solved in polynomial time. Let $\hat{\boldsymbol{x}}=[\hat{x}_{nm}]_{\forall n,m}$ denote the optimal solution of the problem **P2R**. The fraction solution $\hat{x}_{nm}$ represents the fraction of offloading user n's traffic that is expected to transmit to AN m. Notice that the solution to the problem **P2R** yields an upper bound on the original offloading user connection problem **P2** because the feasibility region of the problem **P2** is a subset of that of the problem **P2R**. The upper bound offers a benchmark to measure the performance of the proposed SFC algorithm.

Then, the SFC algorithm is to sequentially fix the binary variables of x_{nm} through iteratively solving a series of problems **P2R** [32] [33]. M_a binary variables x_{nm} are fixed in each iteration. The proposed SFC algorithm is described in Algorithm 1. Let $X=\{x_{nm}|n\in\mathcal{N},m\in\mathcal{M}_a\}$ denote the set of unfixed binary variables. Specifically, in the first iteration, all unfixed binary variables are relaxed to continuous values in $[0,1]$ to obtain the problem **P2R**. Then, we solve the problem **P2R**. Let B denote the set where all fractions $\hat{x}_{nm}$ of the solution to the problem **P2R** are sorted in non-increasing order. We select $\hat{x}_{nm}$ with the largest value from the set B. Let $A=\{x_{nm}=0|n\in\mathcal{N},m\in\mathcal{M}_a\}$ denote the initial feasible solution to the problem **P2**. We set the integer variable x_{nm} corresponding to the selected fraction $\hat{x}_{nm}$ to 1. Then, we replace the $x_{nm}=0$ in the A with $x_{nm}=1$, and check the constraints (22c) and (22d). If the newly obtained A is not a feasible solution, we reset $x_{nm}=0$ in the A. Then, we select $\hat{x}_{nm}$ with the second largest value from the set B to repeat the above procedure until the A changed by the corresponding $x_{nm}=1$ can produce a feasible solution. Thus, we fix an integer variable x_{nm} in the first iteration. Notice that, once the integer variable x_{nm} is fixed to 1, we should fix x_{ij} to 0 for $i=n$ and $j\neq m$ based on the constraint (22b). Therefore, M_a integer variables x_{nm} are fixed in the first iteration. In the second iteration, we remove all the fixed integer variables from the set X and update the problem **P2R** to a new one. Then, we solve the new problem **P2R** and selects a new $\hat{x}_{nm}$ among all the remaining unfixed variables to fix its corresponding integer variable x_{nm} based on the same process. The iteration in the proposed SFC algorithm continues until all offloading users are connected to ANs or no new feasible offloading user connection can be found. In the latter case, we set all the remaining unfixed variables to 0.

Algorithm 1 SFC Algorithm for Offloading User Connection

```
1:  Initialize a feasible solution $A=\{x_{nm}=0|n\in\mathcal{N},m\in\mathcal{M}_a\}$.
2:  Initialize a set of unfixed variables, denoted by $X=\{x_{nm}|n\in\mathcal{N},m\in\mathcal{M}_a\}$.
3:  while the set $X$ is not empty do
4:      Relax the binary variables in $X$ and formulate the LP problem P2R.
5:      Sort the fractions $\hat{x}_{nm}$ of the solution of the problem P2R in non-increasing order, denoted by $B$.
6:      for $\hat{x}_{nm}\in B$ do
7:          Set $x_{nm}=1$ in $A$.
8:          if $A$ satisfies the constraints (22c) and (22d) then
9:              Fix the found $x_{nm}=1$ and also fix $x_{ij}=0$ for ($i=n$, $j\in M_a$, $j\neq m$).
10:             Remove the fixed $x_{ij}$ ($i=n$, $j\in M_a$) from $X$.
11:             Go to Step 3. /*end the for loop*/
12:         else
13:             Reset $x_{nm}=0$ in $A$.
14:         end if
15:     end for
16: end while
```

Since M_a integer variables x_{nm} are fixed in each iteration, the number of iterations in the SFC algorithm is at most N. In each iteration, the time complexity is bounded by the complexity of the LP algorithm. Because the complexity of an LP algorithm is polynomial, the proposed SFC algorithm has a polynomial complexity.

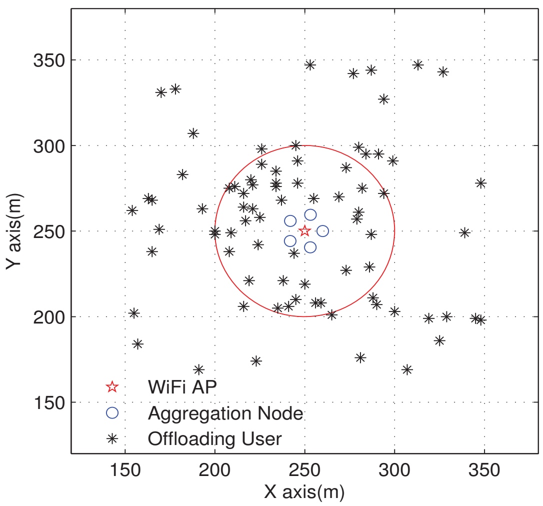


Fig. 2. Simulation topology.

B. AN Activation

In this subsection, we determine the optimal set of activated ANs, $\mathcal{M}_a$. Notice that the average service rate of an activated AN in the WiFi network, $\hat{\mu}_m^{M_a}$, depends on the number of activated ANs, M_a. The algorithm for AN activation should take into account all values of M_a from 1 to M. In addition, we notice that the number of ANs in DAO, M, is a small value. Therefore, we adopt an exhaust algorithm that searches over 2^M possible cases to obtain the optimal $\boldsymbol{y}$. For each possible case, we update the value of $\hat{\mu}_m^{M_a}$ according to $M_a = \sum_{m \in \mathcal{M}} y_m$. Let $G^*(\boldsymbol{y})$ denote the optimal value of problem **P2** for given $\boldsymbol{y}$. The optimal $\boldsymbol{y}$ can be obtained by

$$\mathbf{P3}: \boldsymbol{y}^* = \arg\max_{\boldsymbol{y}} G^*(\boldsymbol{y}). \tag{24}$$

V. Performance Evaluation

In this section, we present numerical simulations to evaluate the performance of traffic offloading achieved by DAO.

A. Network Parameters

We consider a single cellular network. The BS is located at the center of the cell with coordinate (0,0). We place 5 cellular users (CUs) whose uplink channels are reused by D2D communications, and consider two scenarios for the location distribution of the 5 CUs. Unless otherwise specified, the coordinates are in meters. In the scenario 1, the coordinates of the 5 CUs, CU$_1$, CU$_2$, CU$_3$, CU$_4$, and CU$_5$, are (0,150), (150,-50), (30,100), (-50,-250), and (260,250), respectively. In the scenario 2, the coordinates of the 5 CUs are (300,100), (-50,200), (-100,-500), (200,-100), and (-150,-250), respectively. Unless otherwise specified, the default scenario in our simulations is scenario 2. For the WiFi network, as shown in Fig. 2, the coordinate of the WiFi AP is (250,250). We set the coverage of the WiFi AP to 50m. The number of ANs, M, is set to 5, and the 5 ANs, AN$_1$, AN$_2$, AN$_3$, AN$_4$, and AN$_5$, with fixed coordinates, are distributed around the WiFi AP. We assume that the uplink channel of CU$_i$ is reused by AN$_i$ ($i \in \{1,2,3,4,5\}$). As shown in Fig. 2, 80 offloading users are randomly located around the WiFi AP. We can see that some offloading users are not in the coverage of the WiFi AP. The minimum rate requirements of all offloading users are set to the same value in our simulations.

TABLE II
SIMULATION PARAMETERS

Network	Parameter	Value
Cellular	Cell radius	500 m
	Number of offloading users	80
	Number of ANs	5
	Uplink cellular bandwidth	4 MHz
	SINR threshold of cellular user	15 dB
	Noise spectral density	-174 dBm/Hz
	Shadowing standard deviation	8 dB
	Transmit power of cellular user	30 dBm
	Transmit power of offloading user	13 dBm
	Path loss model for cellular links	$128.1 + 36.7\log_{10}(d)$
	Path loss model for D2D links	$148 + 40\log_{10}(d)$
WiFi	Coverage radius	50 m
	Maximum backoff stage	5
	Minimal backoff window size	32
	Idle slot time	20 μs
	DIFS	50 μs
	SIFS	10 μs
	Payload size	3000 Bytes
	Transmit rate	130 Mbps
	PHY header	128 bits
	ACK	240 bits

We set the transmit power of cellular users to 30 dBm. The path loss between the BS and the cellular users is $128.1 + 37.6\log_{10}(d[\text{km}])$. We set the transmit power of offloading users to 13 dBm. The path loss between a D2D transmitter and a D2D receiver is $148 + 40\log_{10}(d[\text{km}])$ [34]. For all communication links, shadowing is a log-normal distribution with a standard deviation of 8 dB and noise power is -174 dBm. In Table II, the most cellular network parameters are set based on [34]. The DCF scheme is employed by all aggregation nodes to contend for the channel in the WiFi network. The RTS/CTS scheme is not adopted in DCF. There are no channel errors in the WiFi network. Table II presents parameters for DCF adopted in the WiFi network.

B. Results

In this subsection, we present simulation results for the optimization problem **P1** formulated for the joint AN activation and offloading user connection problem.

1) Performance of Algorithm SFC: Given a case $\boldsymbol{y}$, we present the result of offloading user connection problem **P2** obtained by the SFC algorithm and compare it with the upper bound obtained by solving the relaxed problem **P2R**. Notice that the proposed SFC algorithm yields a lower bound to the offloading user connection problem **P2**. Since the optimal solution $\boldsymbol{y} = [y_1, y_2, ..., y_5]$ of the AN activation problem **P3** is obtained by searching over $2^5 - 1$ possible cases (the number of ANs is 5 in our simulations), in Table III, we list 31 cases of the results for two scenarios.

In Table III, we can observe that the optimal solution of the AN activation problem **P3** is $\boldsymbol{y} = [1,1,1,1,0]$ in scenario 1 while that is $\boldsymbol{y} = [1,1,1,1,1]$ in scenario 2. In scenario 1, we deliberately set that the position of CU$_5$ is near to AN$_5$. Thus, offloading users prefer to connect to other ANs

TABLE III
SIMULATION RESULTS OF THE OFFLOADING USER CONNECTION PROBLEM FOR 31 $\boldsymbol{y}$ CASES IN TWO SCENARIOS

$\boldsymbol{y}$	scenario 1		scenario 2	
	Upper Bound	Result by SFC	Upper Bound	Result by SFC
[0,0,0,0,1]	0.036	0	27.47	27.20
[0,0,0,1,0]	26.78	26.40	19.99	19.20
[0,0,0,1,1]	20.00	19.20	39.93	38.40
[0,0,1,0,0]	16.28	16.00	24.71	24.00
[0,0,1,0,1]	16.31	16.00	39.93	38.40
[0,0,1,1,0]	36.24	35.20	38.65	37.60
[0,0,1,1,1]	28.53	27.20	42.74	40.80
[0,1,0,0,0]	21.08	20.80	19.58	19.20
[0,1,0,0,1]	20.00	19.20	39.54	38.40
[0,1,0,1,0]	39.93	38.40	35.25	34.40
[0,1,0,1,1]	28.53	27.20	42.74	40.80
[0,1,1,0,0]	32.93	32.0	38.39	37.60
[0,1,1,0,1]	28.53	27.20	42.74	40.80
[0,1,1,1,0]	42.74	40.80	42.74	40.80
[0,1,1,1,1]	33.05	31.20	44.02	41.60
[1,0,0,0,0]	17.44	16.80	11.74	11.20
[1,0,0,0,1]	17.48	16.80	31.70	30.40
[1,0,0,1,0]	37.41	36.00	29.13	28.80
[1,0,0,1,1]	28.53	27.20	40.24	38.40
[1,0,1,0,0]	30.76	30.40	31.70	30.40
[1,0,1,0,1]	28.53	27.20	40.24	38.40
[1,0,1,1,0]	42.74	40.80	40.24	38.40
[1,0,1,1,1]	33.05	31.20	44.02	41.60
[1,1,0,0,0]	33.57	32.80	26.89	25.60
[1,1,0,0,1]	28.53	27.20	40.24	38.40
[1,1,0,1,0]	42.74	40.80	40.04	38.40
[1,1,0,1,1]	33.05	31.20	44.02	41.60
[1,1,1,0,0]	41.72	40.80	40.24	38.40
[1,1,1,0,1]	33.05	31.20	44.02	41.60
[1,1,1,1,0]	44.02	41.60	44.02	41.60
[1,1,1,1,1]	35.73	35.20	44.62	44.00

rather than AN$_5$ because the strong interference from CU$_5$ to AN$_5$. Therefore, the optimal AN activation is achieved when AN$_5$ is inactivated, i.e., $y_5 = 0$. We can also observe that the result of the AN activation problem **P3** is close to 0 when $\boldsymbol{y} = [0, 0, 0, 0, 1]$ in scenario 1, which also verifies our analysis. In scenario 2, the positions of CUs are far away from ANs, so ANs do not suffer from strong interference from CUs, and the maximum result of the AN activation problem **P3** is found when all ANs are activated, i.e., $\boldsymbol{y} = [1, 1, 1, 1, 1]$. Notice that with the WiFi network parameters configured in our simulations, the maximum system throughput of the saturated WiFi network is achieved when the number of contending nodes in the WiFi network is 5.

In Table III, we can also observe that, given a case $\boldsymbol{y}$, the result of the offloading user connection problem **P2** obtained by the SFC algorithm (a lower bound to the offloading user connection problem **P2**) is close to the upper bound obtained by solving the relaxed problem **P2R**, i.e., the lower bound obtained by the SFC algorithm is close to the upper bound. To show the closeness between these bounds, in Table IV, we further present the ratio of the upper bound to lower bound for 31 $\boldsymbol{y}$ cases in two scenarios, based on Table III. In scenario 1, the average ratio for all the 31 cases is 1.03 and the standard derivation is 0.016. In scenario 2, the average ratio for all the 31 cases is 1.04 and the standard derivation is 0.014. All the statistical results show that the ratio of the upper bound to lower bound is close to 1. Since the optimal solution lies between the lower bound obtained by the SFC algorithm and the upper bound, the solution found by the SFC algorithm to the offloading user connection problem **P2** is a near-optimal solution.

TABLE IV
RATIO OF THE UPPER BOUND TO THE LOWER BOUND FOR 31 $\boldsymbol{y}$ CASES IN TWO SCENARIOS

$\boldsymbol{y}$	ratio		$\boldsymbol{y}$	ratio	
	scenario 1	scenario 2		scenario 1	scenario 2
[0,0,0,0,1]	-	1.01	[1,0,0,0,1]	1.04	1.04
[0,0,0,1,0]	1.01	1.04	[1,0,0,1,0]	1.03	1.01
[0,0,0,1,1]	1.04	1.03	[1,0,0,1,1]	1.04	1.04
[0,0,1,0,0]	1.01	1.02	[1,0,1,0,0]	1.01	1.04
[0,0,1,0,1]	1.01	1.03	[1,0,1,0,1]	1.04	1.04
[0,0,1,1,0]	1.02	1.02	[1,0,1,1,0]	1.04	1.04
[0,0,1,1,1]	1.04	1.04	[1,0,1,1,1]	1.05	1.05
[0,1,0,0,0]	1.01	1.01	[1,1,0,0,0]	1.02	1.05
[0,1,0,0,1]	1.04	1.02	[1,1,0,0,1]	1.04	1.04
[0,1,0,1,0]	1.03	1.02	[1,1,0,1,0]	1.04	1.04
[0,1,0,1,1]	1.04	1.04	[1,1,0,1,1]	1.05	1.05
[0,1,1,0,0]	1.02	1.02	[1,1,1,0,0]	1.02	1.04
[0,1,1,0,1]	1.04	1.04	[1,1,1,0,1]	1.05	1.05
[0,1,1,1,0]	1.04	1.04	[1,1,1,1,0]	1.05	1.05
[0,1,1,1,1]	1.05	1.05	[1,1,1,1,1]	1.01	1.01
[1,0,0,0,0]	1.03	1.04	-	-	-

2) Performance of Admission Control: In DAO, the admission control is incorporated in the process of offloading user connection to accommodate a certain number of offloading users while satisfying their QoS constraints in terms of the minimum rate requirement. In the default scenario 2, given the optimal $\boldsymbol{y} = [1, 1, 1, 1, 1]$, Fig. 3 plots the number of accommodated offloading users as a function of γ (the minimum rate requirement of offloading users) for different values of P_n^{max} (the maximum transmit power of offloading users). In Fig. 3, for both $P_n^{max} = 13$ dBm and $P_n^{max} = 3$ dBm, we can observe that the number of accommodated offloading users decreases with γ because a larger value of γ increases the amount of time resource required to meet an offloading user's minimum rate requirement. In Fig. 3, we can also observe that given γ, the number of accommodated offloading users increases with larger P_n^{max}. The increase of P_n^{max} reduces the amount of time resource required to meet an offloading user's minimum rate requirement, and consequently increases the number of accommodated offloading users.

Fig. 4 plots the number of accommodated offloading users as a function of D (the data rate of ANs in the WiFi network) for different values of P_n^{max} and γ. We can observe that, given $\gamma = 0.8$ Mbps or $\gamma = 1.4$ Mbps, the number of accommodated offloading users does not increase with larger P_n^{max} when D is small. For small values of D, accommodating offloading users is constrained by the service rate of activated ANs in the WiFi network, so the increase of P_n^{max} cannot increase the number of accommodated offloading users. With $P_n^{max} = 13$ dBm, the number of accommodated offloading users increases with D because the increase of D increases the system throughput of the WiFi network. However, in some cases, the increase of D does not increase the number of accommodated offloading users. The reason is that the increase of D leads to a marginal increase in the system throughput of the WiFi network. The increased system throughput of the WiFi network does not satisfy the minimum rate requirement of an offloading user. For instance, with D increasing from 170 Mbps to 210 Mbps, the system throughput of the WiFi network increases from

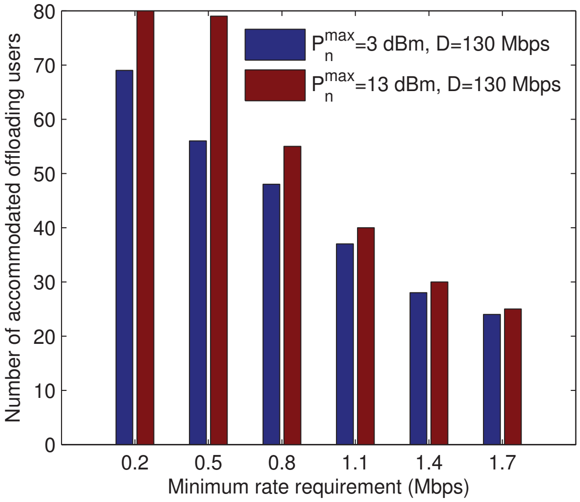


Fig. 3. Number of accommodated offloading users versus minimum rate requirement of offloading users.

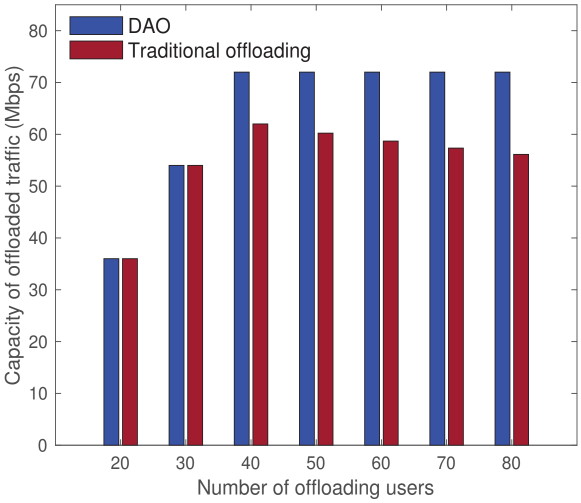


Fig. 5. Capacity of offloaded traffic versus number of offloading users. The minimum rate requirement γ is 1.8 Mbps.

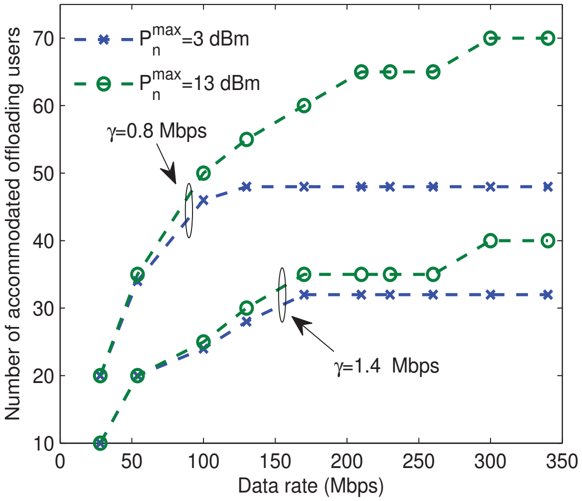


Fig. 4. Number of accommodated offloading users versus data rate of ANs.

46.90 Mbps to 48.24 Mbps. The increased system throughput is only 1.34 Mbps, which cannot accommodate a new offloading user when $\gamma = 1.4$ Mbps. The increased system throughput can accommodate a new offloading user when $\gamma = 0.8$ Mbps. In addition, in Fig. 4, we can also observe that, with $P_n^{max} = 3$ dBm, when D is large enough, the number of accommodated offloading users reaches a saturation point and does not increase with the further increase of D. With $P_n^{max} = 3$ dBm, the dominating constraint for accommodating offloading users is the time resource when D is a large value, and the further increase of D only increases the system throughput of the WiFi network. Therefore, the number of accommodated offloading users does not increase with the further increase of D.

3) Performance Comparisons: We compare the capacity of offloaded traffic achieved by DAO with that achieved by the traditional offloading where offloading users directly connect to the WiFi AP. In the simulations, the default packet payload size is 8,000 Bytes. The data rate of offloading users in the traditional offloading is set to the same value with the data rate of ANs in DAO.

Fig. 5 plots the capacity of offloaded traffic achieved by DAO compared with that achieved by the traditional offloading for different numbers of offloading users. In the simulations, the minimum rate requirement γ is 1.8 Mbps. When the number of offloading users is small, the WiFi network in the traditional offloading is unsaturated. The DAO and traditional offloading achieve the same capacity of offloaded traffic. When the number of offloading users is large, the WiFi network in the traditional offloading is saturated (i.e., congested). The capacity of offloaded traffic achieved by the traditional offloading decreases with the number of offloading users because the WiFi network' throughput decreases with the number of offloading users involved in access contention. However, the increase of the number of offloading users does not affect the capacity of offloaded traffic achieved by DAO because the number of users involved in access contention in the WiFi network is fixed to a small value (i.e., the number of activated ANs). In Fig. 5, we can clearly observe the significant performance gain achieved by DAO over the traditional offloading when the network load is heavy. For example, when the number of offloading users is 70, the capacity of offloaded traffic achieved by DAO is 72 Mbps while that achieved by the traditional offloading is 57.34 Mbps.

Fig. 6 plots the capacities of offloaded traffic achieved by DAO and traditional offloading for different minimum rate requirements of offloading users. In the simulations, the number of offloading users is 70. In the traditional offloading, as the minimum rate requirement of offloading users increases, the WiFi network's throughput increases until it saturates. Since the saturated throughput of the WiFi network depends on the number of users involved in access contention, the saturated throughput of the WiFi network in DAO is larger than that in the traditional offloading. When the WiFi network in the traditional offloading is saturated (i.e., overloaded), the WiFi network in DAO may be unsaturated. In other words, the throughput of the WiFi network in the traditional offloading

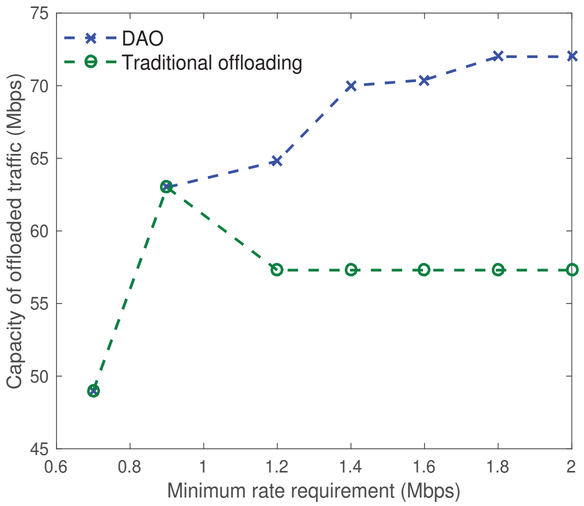


Fig. 6. Capacity of offloaded traffic versus minimum rate requirement of offloading users. The number of offloading users is 70.

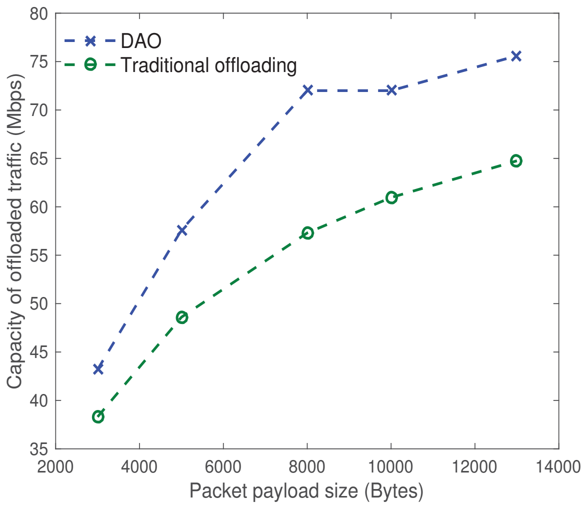


Fig. 7. Capacity of offloaded traffic versus packet payload size. The number of offloading users is 70 and the minimum rate requirement γ is 1.8 Mbps.

saturates early. For example, when the minimum rate requirements of offloading users are 1.6 Mbps, the WiFi network in the traditional offloading is saturated and the achieved capacity of offloaded traffic is 57.3 Mbps while the WiFi network in DAO is still unsaturated and the achieved capacity of offloaded traffic is 70.4 Mbps. Notice that the saturated throughput of the WiFi network in DAO is 72.88 Mbps.

When offloading traffic from cellular to WiFi, the WiFi network's throughput determines the capacity of offloaded traffic. The frame aggregation [17] is widely adopted in the current WiFi standards, which affects the throughput of WiFi networks. However, frame aggregation may cause delays as nodes need to wait until enough packets arrive to form a large MAC frame. The traffic aggregation in DAO can mitigate delays induced by frame aggregation. In DAO, each AN can aggregate cellular traffic to transmit with a large packet payload in the WiFi network. Fig. 7 plots the capacity of offloaded traffic as a function of packet payload size. In the simulations, the number of offloading users is 70 and the minimum rate requirement of offloading users is 1.8 Mbps. As the increase of packet payload size increases the WiFi network's throughput, the capacities of offloaded traffic achieved by both DAO and traditional offloading increase. Since the number of users involved in the WiFi network's access contention in DAO is small, DAO achieves a larger capacity of offloaded traffic than the traditional offloading.

VI. Conclusion

In this paper, we have developed DAO, a novel traffic offloading scheme in integrated cellular-WiFi networks. Different from the traditional offloading where offloading users directly connect to a WiFi network, DAO exploits D2D communications in licensed cellular bands to aggregate cellular traffic to reduce the number of offloading users contending for access in the WiFi network, and hence the capacity of traffic offloaded to the WiFi network in DAO can be significantly improved. To come up with the optimal scheme, the offloading process in DAO is formulated as an optimization problem that jointly takes into account the activations of aggregation nodes and the connections between aggregation nodes and offloading users. Unfortunately, the optimization tends to be NP-hard and we have solved the resulting joint optimization problem by exploiting its layered-structure and decomposing it into subproblems based on time-scale separation. We have conducted extensive study and show that DAO does support significantly higher offloaded traffic compared with the traditional offloading, especially in heavy traffic load scenarios.

References

[1] J. Liu, C. Zhang, and Y. Fang, "Epic: A differential privacy framework to defend smart homes against internet traffic analysis," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 1206–1217, 2018.

[2] D. Liu, L. Wang, Y. Chen, M. Elkashlan, K.-K. Wong, R. Schober, and L. Hanzo, "User association in 5g networks: A survey and an outlook," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1018–1044, 2016.

[3] K. Lee, J. Lee, Y. Yi, I. Rhee, and S. Chong, "Mobile data offloading: How much can wifi deliver?" *IEEE/ACM Trans. Netw.*, vol. 21, no. 2, pp. 536–550, 2013.

[4] B. H. Jung, N.-O. Song, and D. K. Sung, "A network-assisted user-centric wifi-offloading model for maximizing per-user throughput in a heterogeneous network," *IEEE Trans. Veh. Technol.*, vol. 63, no. 4, pp. 1940–1945, 2014.

[5] Q. Chen, G. Yu, H. Shan, A. Maaref, G. Y. Li, and A. Huang, "Cellular meets wifi: Traffic offloading or resource sharing?" *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3354–3367, 2016.

[6] R. Mahindra, H. Viswanathan, K. Sundaresan, M. Y. Arslan, and S. Rangarajan, "A practical traffic management system for integrated lte-wifi networks," in *Proc. ACM MobiCom*, 2014, pp. 189–200.

[7] A. Balasubramanian, R. Mahajan, and A. Venkataramani, "Augmenting mobile 3g using wifi," in *Proc. ACM MobiSys*, 2010, pp. 209–222.

[8] Y. Wu, Y. He, L. P. Qian, J. Huang, and X. S. Shen, "Optimal resource allocations for mobile data offloading via dual-connectivity," *IEEE Trans. Mobile Comput.*, 2018.

[9] G. Gao, M. Xiao, J. Wu, K. Han, L. Huang, and Z. Zhao, "Opportunistic mobile data offloading with deadline constraints," *IEEE Trans. Parallel Distrib. Syst.*, vol. 28, no. 12, pp. 3584–3599, 2017.

[10] M. H. Cheung, F. Hou, J. Huang, and R. Southwell, "Congestion-aware dns for integrated cellular and wi-fi networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 6, pp. 1269–1281, 2017.

[11] Y. Kwon, Y. Fang, and H. Latchman, "Fast collision resolution (fcr) mac algorithm for wireless local area networks," in *Proc. IEEE GLOBECOM*, vol. 3. IEEE, 2002, pp. 2250–2254.

[12] H. Zhai, Y. Kwon, and Y. Fang, "Performance analysis of ieee 802.11 mac protocols in wireless lans," *Wireless communications and mobile computing, Wiley Online Library*, vol. 4, no. 8, pp. 917–931, 2004.

[13] Y. Kwon, Y. Fang, and H. Latchman, "Design of mac protocols with fast collision resolution for wireless local area networks," *IEEE Trans. Wireless Commun.*, vol. 3, no. 3, pp. 793–807, 2004.

[14] H. Zhai, J. Wang, X. Chen, and Y. Fang, "Medium access control in mobile ad hoc networks: challenges and solutions," *Wireless Communications and Mobile Computing, Wiley Online Library*, vol. 6, no. 2, pp. 151–170, 2006.

[15] T. Joshi, A. Mukherjee, Y. Yoo, and D. P. Agrawal, "Airtime fairness for ieee 802.11 multirate networks," *IEEE Trans. Mobile Comput.*, vol. 7, no. 4, pp. 513–527, 2008.

[16] H. Zhu and G. Cao, "rdcf: A relay-enabled medium access control protocol for wireless ad hoc networks," *IEEE Trans. Mobile Comput.*, vol. 5, no. 9, pp. 1201–1214, 2006.

[17] D. Skordoulis, Q. Ni, H.-H. Chen, A. P. Stephens, C. Liu, and A. Jamalipour, "Ieee 802.11 n mac frame aggregation mechanisms for next-generation high-throughput wlans," *IEEE Wireless Communications*, vol. 15, no. 1, pp. 40–47, 2008.

[18] Z. Zhou, M. Dong, K. Ota, G. Wang, and L. T. Yang, "Energy-efficient resource allocation for d2d communications underlaying cloud-ran-based lte-a networks," *IEEE Internet of Things Journal*, vol. 3, no. 3, pp. 428–438, 2016.

[19] Z. Zhou, K. Ota, M. Dong, and C. Xu, "Energy-efficient matching for resource allocation in d2d enabled cellular networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 6, pp. 5256–5268, 2017.

[20] K. Poularakis, G. Iosifidis, I. Pefkianakis, L. Tassiulas, and M. May, "Mobile data offloading through caching in residential 802.11 wireless networks," *IEEE Trans. Netw. Service Manage.*, vol. 13, no. 1, pp. 71–84, 2016.

[21] S.-I. Sou and Y.-T. Peng, "Performance modeling for multipath mobile data offloading in cellular/wi-fi networks," *IEEE Trans. Commun.*, vol. 65, no. 9, pp. 3863–3875, 2017.

[22] K. Poularakis, G. Iosifidis, and L. Tassiulas, "Deploying carrier-grade wifi: offload traffic, not money," in *Proc. ACM MobiHoc*, 2016, pp. 131–140.

[23] Q. Fan, H. Lu, P. Hong, and Z. Zhu, "Throughput–power tradeoff association for user equipment in wlan/cellular integrated networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 4, pp. 3462–3474, 2017.

[24] 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. 21, no. 2, pp. 57–65, 2014.

[25] Y. Wu, J. Chen, L. P. Qian, J. Huang, and X. S. Shen, "Energy-aware cooperative traffic offloading via device-to-device cooperations: An analytical approach," *IEEE Trans. Mobile Comput.*, vol. 16, no. 1, pp. 97–114, 2017.

[26] A. Asadi and V. Mancuso, "Network-assisted outband d2d-clustering in 5g cellular networks: theory and practice," *IEEE Trans. Mobile Comput.*, vol. 16, no. 8, pp. 2246–2259, 2017.

[27] W. Cao, G. Feng, S. Qin, and M. Yan, "Cellular offloading in heterogeneous mobile networks with d2d communication assistance," *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 4245–4255, 2017.

[28] W.-S. Lim, D.-W. Kim, and Y.-J. Suh, "Pr-mac: a practical approach for exploiting relay transmissions in multi-rate wlans," *IEEE Trans. Wireless Commun.*, vol. 9, no. 1, pp. 66–71, 2010.

[29] D. Malone, K. Duffy, and D. Leith, "Modeling the 802.11 distributed coordination function in nonsaturated heterogeneous conditions," *IEEE/ACM Trans. Netw.*, vol. 15, no. 1, p. 159172, 2007.

[30] L. X. Cai, X. Shen, J. W. Mark, L. Cai, and Y. Xiao, "Voice capacity analysis of wlan with unbalanced traffic," *IEEE Trans. Veh. Technol.*, vol. 55, no. 3, pp. 752–761, 2006.

[31] G. Bianchi, "Performance analysis of the ieee 802.11 distributed coordination function," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 3, pp. 535–547, 2000.

[32] M. Pan, C. Zhang, P. Li, and Y. Fang, "Joint routing and link scheduling for cognitive radio networks under uncertain spectrum supply," in *Proc. IEEE INFOCOM*, 2011, pp. 2237–2245.

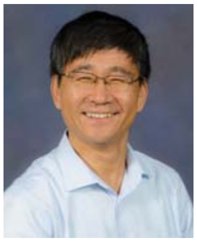
[33] ——, "Spectrum harvesting and sharing in multi-hop crns under uncertain spectrum supply," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 2, pp. 369–378, 2012.

[34] G. Yu, L. Xu, D. Feng, R. Yin, G. Y. Li, and Y. Jiang, "Joint mode selection and resource allocation for device-to-device communications," *IEEE Trans. Wireless Commun.*, vol. 62, no. 11, pp. 3814–3824, 2014.

Bing Feng is currently a Ph.D. student at the University of Science and Technology of China. He received the B.E. degrees from the Anhui University, Hefei, China, in 2013. His research interests include wireless communication and wireless local area networks.

Chi Zhang received the B.E. and M.E. degrees in electrical and information engineering from the Huazhong University of Science and Technology, China, in 1999 and 2002, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Florida in 2011. He joined the School of Information Science and Technology, University of Science and Technology of China, as an Associate Professor in 2011. His research interests include the areas of network protocol design and performance analysis and network security particularly for wireless networks and social networks.

Jianqing Liu received the Ph.D. degree from University of Florida in 2018 and the B.Eng. degree from University of Electronic Science and Technology of China in 2013. He is currently a tenure-track assistant professor in the Department of Electrical and Computer Engineering at University of Alabama in Huntsville. His research interest is to apply cryptography, differential privacy and convex optimization to design secure and efficient protocols for various IoT systems. He is the recipient of the 2018 Best Journal Paper Award from IEEE Technical Committee on Green Communications & Computing (TCGCC) and the Best Paper Award from 2012 IEEE Workshop on Microwave and Millimeter-Wave Circuits and Systems (MMWCST).



Yuguang "Michael" Fang (F'08) received an MS degree from Qufu Normal University, Shandong, China in 1987, a PhD degree from Case Western Reserve University in 1994 and a PhD degree from Boston University in 1997. He joined Department of Electrical and Computer Engineering at University of Florida in 2000 and has been a full professor since 2005. He held a University of Florida Research Foundation (UFRF) Professorship (2006-2009), a Changjiang Scholar Chair Professorship (Xidian University, Xi'an, China, 2008-2011; Dalian Maritime University, Dalian, China, 2015-present), and a Guest Chair Professorship with Tsinghua University, China (2009-2012). Dr. Fang received the US National Science Foundation Career Award in 2001, the Office of Naval Research Young Investigator Award in 2002, the 2015 IEEE Communications Society CISTC Technical Recognition Award, the 2014 IEEE Communications Society WTC Recognition Award, the Best Paper Award from IEEE ICNP (2006), and the 2010-2011 UF Doctoral Dissertation Advisor/Mentoring Award. He is the Editor-in-Chief of IEEE Transactions on Vehicular Technology, was the Editor-in-Chief of IEEE Wireless Communications (2009-2012), serves/served on several editorial boards of technical journals. He is a fellow of the IEEE and the AAAS.