Article:
Chen, B., Pappas, N., Chen, Z. et al. (2019) Throughput and delay analysis of LWA with bursty traffic and randomized flow splitting. IEEE Access, 7. pp. 24667-24678. ISSN: 2169-3536

https://doi.org/10.1109/ACCESS.2019.2897017

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 5, 2018, accepted December 31, 2018, date of current version March 7, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2897017

Throughput and Delay Analysis of LWA With Bursty Traffic and Randomized Flow Splitting

BOLIN CHEN[1], NIKOLAOS PAPPAS[2], (Member, IEEE), ZHENG CHEN[3], (Member, IEEE), DI YUAN[2], (Senior Member, IEEE), AND JIE ZHANG[1]

[1]Department of Electronic and Electrical Engineering, The University of Sheffield, Sheffield S10 2TN, U.K.
[2]Department of Science and Technology, Linköping University, 602 21 Norrköping, Sweden
[3]Department of Electrical Engineering, Linköping University, 581 83 Linköping, Sweden

Corresponding author: Bolin Chen (bolinchenuos@gmail.com)

This work was supported in part by the Swedish Foundation for Strategic Research (SSF), in part by the Swedish Research Council (VR), in part by the Excellence Center at Linköping–Lund in Information Technology, and in part by the Joint Research Project Deploying High Capacity Dense Small Cell Heterogeneous Networks (DECADE), within the Research and Innovation Staff Exchange (RISE) Scheme of the European Horizon 2020 Framework Program, under Grant 645705.

ABSTRACT We investigate the effect of bursty traffic in a long term evolution (LTE) and Wi-Fi aggregation (LWA)-enabled network. The LTE base station routes packets of the same IP flow through the LTE and Wi-Fi links independently. We motivate the use of superposition coding at the LWA-mode Wi-Fi access point (AP) so that it can serve LWA users and Wi-Fi users simultaneously. A random access protocol is applied in such system, which allows the native-mode AP to access the channel with probabilities that depend on the queue size of the LWA-mode AP to avoid impeding the performance of the LWA-enabled network. We analyze the throughput of the native Wi-Fi network and the delay experienced by the LWA users, accounting for the native-mode AP access probability, the traffic flow splitting between LTE and Wi-Fi, and the operating mode of the LWA user with both LTE and Wi-Fi interfaces. Our results show some fundamental tradeoffs in the throughput and delay behavior of LWA-enabled networks, which provide meaningful insight into the operation of such aggregated systems.

INDEX TERMS LTE and Wi-Fi aggregation, shared access, throughput, delay, queueing analysis.

I. INTRODUCTION

A. BACKGROUND

One possible solution to address the increasing wireless data demand is traffic offloading from licensed Long Term Evolution (LTE) networks to the unlicensed spectrum [1]. One common approach for LTE to use the unlicensed band is to interwork with Wi-Fi. The third generation partnership project (3GPP) has defined a tight interworking solution called LTE and Wi-Fi aggregation (LWA) since Release 13 to support the access to both LTE and Wi-Fi networks simultaneously. LWA splits packet data convergence protocol (PDCP) packets of the same IP flow through both the LTE and Wi-Fi links, and is also able to aggregate received packets from both LTE and Wi-Fi at the user PDCP layer.

B. RELATED WORKS AND MOTIVATION

Early studies of LWA mainly focus on the prototype and architecture design [2], [3]. The feasibility of licensed and unlicensed carriers aggregation has been verified experimentally in [4]. Lopez-Perez *et al.* [5] present a traffic aggregation-based LWA flow control algorithm. Reference [6] implements the radio resource management layer for LWA. The layer 2 structure for LWA to achieve the compatibility with Wi-Fi is proposed in [7]. Singh *et al.* [8] investigate the load balancing and user assignment solutions for LWA. Techniques for traffic splitting and aggregation at the radio layer have also been considered in the literature. References [9], [10] investigate aggregation and path selection mechanisms that maximize the network utility. The LWA and Wi-Fi offloading scheme are jointly considered in [11], which also strikes the balance between user payment and quality of service (QoS).

The aforementioned LWA studies are based on one common assumption: a Wi-Fi access point (AP) with LWA capability is only able to offload bearers from LTE, and does not have its own user equipments (UEs) to serve. In reality, a Wi-Fi AP can operate in both the native mode and the LWA mode simultaneously [2]. Specifically, the LWA-mode

The associate editor coordinating the review of this manuscript and approving it for publication was Nan Zhao.

Wi-Fi AP cooperates with the LTE base station (BS) to transmit bearers to the LWA UE, which aggregates packets from both LTE and Wi-Fi. The native-mode Wi-Fi AP transmits Wi-Fi packets to those native Wi-Fi UEs that are not with LWA capability. Hence the problem arises of how to transmit different packets for one AP to different UEs. The conventional approach is to set up orthogonal channels in terms of time/frequency etc [12]. However, this approach is inefficient and not optimal in terms of achievable rates [13]. More importantly, collisions may be inevitable because of imperfect knowledge of the channel occupancy state. As an alternative, a method called superposition coding (SC) proposed in [14] and [15] can be used to remove the orthogonality constraint in a transmission by one AP to both the LWA UE and the native Wi-Fi UE. The SC method is considered as a promising technique for enhancing resource efficiency, and it achieves the capacity on a scalar Gaussian broadcast channel [16].

Nevertheless, spectrum sharing between the native-mode AP and the LWA-mode AP inevitably creates interference among concurrent transmissions. Accounting for the interference caused by the native Wi-Fi network and that affects the LWA UE, an appropriate access protocol needs to be carefully designed such that the QoS of the LWA UE will not be adversely degraded. Zhao *et al.* [17], Urgaonkar and Neely [18] develop the scheduling polices for the low-priority node under partial channel state information. In [19], a random access protocol is proposed, where low-priority nodes make transmission attempts with a given probability. The study [19] is based on the general multi-packet reception (MPR) channel model proposed in [20] and [21], which captures the interference at the physical layer more efficiently compared to the traditional channel model, as in the former a transmission may still succeed even in the presence of interference. References [22], [23] study the interference created by the spectrum sharing between high-priority and low-priority nodes in the MPR channel among concurrent transmissions. Reference [24] analyzes the throughput of the low-priority network where MPR capability is adopted in a cognitive network with the high-priority node under certain conditions. Ewaisha and Tepedelenlioglu [25] optimize the throughput with deadline constraints on a single low-priority node accessing a multi-channel system.

In this paper, we consider a shared access Wi-Fi network inspired by the cognitive radio network paradigm. More specifically, the high-priority node (i.e., the LWA-mode Wi-Fi AP) is allowed to access the channel whenever it is needed. However, the low-priority node (i.e., the native-mode Wi-Fi AP) will randomly access the channel when the queue size of the LWA-mode Wi-Fi AP is below a congestion limit, so as not to create harmful interference to the LWA UE. How to investigate the performance of such systems remains open. Recently, there is growing interest in the delay analysis or the combination of throughput and delay analysis in LWA-enabled networks [26]–[29]. However, most of the existing studies focus on the case under the saturated traffic assumption. In fact, based on the queueing theory, the analysis of the delay of networks with bursty sources cannot be easily seen with the saturated traffic assumption [30]. In general, how to design the random access protocol accounting for both the throughput of the native Wi-Fi network and the delay of the LWA-enabled network with bursty LWA traffic has not been addressed yet.

C. MAIN RESULTS AND PAPER ORGANIZATION

The main contributions of this paper can be summarized as follows. We investigate the effect of bursty LWA traffic on the throughput and the delay performance in an LWA-enabled network, where the LWA-mode Wi-Fi AP can simultaneously operate as the native-mode AP with the help of SC. With congestion control on the LWA-mode AP, the native Wi-Fi AP not only utilizes the idle slots, but also transmits along with the LWA-mode AP by randomly accessing the channel. In this paper, we first analyze the characteristics of the queues at the LTE BS transmitter and the LWA-mode Wi-Fi AP transmitter. We model those queues as discrete time Markov Chains and obtain their stationary distributions. We then characterize the performance of the considered network in terms of the native Wi-Fi throughput and the LWA UE delay. More specifically, we derive the native Wi-Fi throughput and the delay of the LWA UE as functions of the native Wi-Fi AP access probability, the probability that the LWA UE chooses the LTE or Wi-Fi interface at one time slot, and the probability that an LTE packet to be routed through the LTE or the Wi-Fi link. To the best of our knowledge, similar results to this work have not been reported yet. Although our study builds on a simple network with four nodes (i.e., one LTE BS, one Wi-Fi AP, one LWA UE, and one native Wi-Fi UE), the analysis can be used for further investigations in larger topologies.

The rest of this paper is organized as follows. In Section II, we present the considered system model including the network model and the priority based Wi-Fi transmission scheme. In Section III and Section IV, we include the analysis for the queues, and show how to derive the LWA UE delay and the native Wi-Fi network throughput. Then we provide numerical evaluation of the presented results in Section V. Finally we conclude the paper in Section VI.

II. SYSTEM MODEL

A. NETWORK MODEL

As shown in Fig. 1, we consider an LWA-enabled network with one LTE BS and one Wi-Fi AP, which operate in different frequency bands. In the following, we refer to the BS and the Wi-Fi AP as L and W respectively. The time domain is divided into equal-length time slots. The data traffic arrives at L and W in the form of fixed-length packet and the transmission of a packet requires one time slot. The acknowledgements (ACKs) are received instantaneously and error-free. We assume that the packet arrives at L according to a Bernoulli process with arrival rate λ_L. Note that L and W are connected via a wireless non-ideal backhaul, which

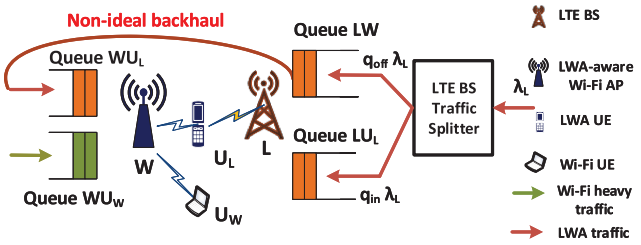


FIGURE 1. An example LWA network. LWA splits data units of the same bearer over the LTE and Wi-Fi link simultaneously. The AP W serves LWA purpose, and also operates as the native-mode Wi-Fi AP. The BS transmitter has queue LW and queue LU_L with packets intended to W and U_L, respectively. The AP transmitter has queue WU_L and queue WU_W containing the messages that are destined to U_L and U_W, respectively.

TABLE 1. Notation for probabilities.

Probability	Explanation
q_{in}	probability that a packet generated at L to be routed through the LTE link
q_{off}	probability that a packet generated at L to be routed through the Wi-Fi link
$q_{\text{U}_L,\text{L}}$	probability that U_L chooses the LTE interface
$q_{\text{U}_L,\text{W}}$	probability that U_L chooses the Wi-Fi interface
$q_{\text{W,W}}$	probability that W serves the native Wi-Fi user

is used to offload packets from LTE to Wi-Fi. When a packet arrives at L, it has probability q_{in} to be routed through the LTE link, and q_{off} to be offloaded to W through the backhaul. To avoid the in-band interference, we further assume that the BS operates on 2.4 GHz to transmit packets through the LTE link, and offloads packets to W using the 3.5 GHz band. The BS maintains two different queues for packets intended for different receivers. Specifically, queue LW and queue LU_L contain packets that arrive at L, which will be transmitted through the LWA-mode Wi-Fi AP and the LTE BS, respectively. Note that the arrival rate at each queue denotes the probability of a new packet arrival in a time slot without accounting for the packets that are already in the queue. Obviously, the packets that enter LW and LU_L form two Bernoulli processes with arrival rates $\lambda_{LW} = q_{\text{off}}\lambda_L$ and $\lambda_{LU_L} = q_{\text{in}}\lambda_L$, respectively.

As a result of the LWA functionality, the WiFi AP can operate in two modes. On the one hand, it can assist L's transmissions by keeping the offloaded packets in its queue WU_L, and trying to transmit them to the LWA UE in a later time slot. It is obvious that the packets enter WU_L form a Bernoulli process with arrival rates $\lambda_{WU_L} = \lambda_{LW}$. Note U_L is equipped with both LTE and Wi-Fi receivers, such like the current smartphone, and has the capability to aggregate traffic over L and W with LWA capability. On each time slot, U_L may access either LTE or Wi-Fi or both, and thus is assumed to have two options for receptions:

1) Both LTE and Wi-Fi receivers are activated, i.e., U_L can receive packets through both interfaces simultaneously.
2) UE U_L chooses randomly the LTE or the Wi-Fi receiver on each time slot.

Denote by $q_{\text{U}_L,\text{L}}$ and $q_{\text{U}_L,\text{W}}$ the probabilities that U_L chooses the LTE and Wi-Fi interfaces on each time slot, respectively. For the first case, $q_{\text{U}_L,\text{L}} = q_{\text{U}_L,\text{W}} = 1$. For the second case, $q_{\text{U}_L,\text{L}} + q_{\text{U}_L,\text{W}} = 1$.

In addition, W also has its own messages to transmit to the native Wi-Fi UE. Denote by WU_W the queue that contains the packets destined to the native Wi-Fi UE, which can be served by W only. In this paper, we assume heavy traffic between them, i.e., the queue WU_W never empties.

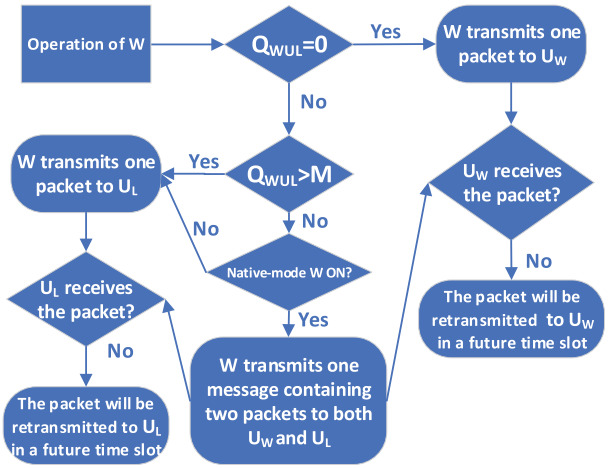


FIGURE 2. The operation of W in the described protocol.

B. PRIORITY BASED Wi-Fi TRANSMISSION SCHEME

As illustrated in Fig. 2, a priority-based Wi-Fi transmission scheme is considered in this paper. Specifically, whether the native-mode Wi-Fi AP will access the channel depends on the size of queue WU_L, such that the native-mode Wi-Fi AP will not deteriorate the performance of the LWA-mode AP [24]. Denote by Q_i the size of queue $i \in \{LW, LU_L, WU_L, WU_W\}$, measured in number of packets. We introduce a threshold M, which plays the role of a congestion limit for WU_L, and the activities of the native-mode and LWA-mode AP in a time slot are programmed in the following cases:

1) When $Q_{WU_L} = 0$, the Wi-Fi AP has no packet to transmit to the LWA user U_L. In such case, it transmits a packet to the native Wi-Fi user U_W with probability 1.
2) When $1 \leqslant Q_{WU_L} \leqslant M$, the Wi-Fi AP transmits one packet to U_L, and it transmits one packet to U_W with probability $q_{\text{W, W}}$.
3) When $Q_{WU_L} > M$, the Wi-Fi AP transmits one packet to U_L only.

For the second case, the Wi-Fi AP will adopt the SC scheme to transmit one message containing two packets, intended for the LWA UE and the native-mode Wi-Fi UE, respectively. We consider a decoding strategy where the UE with the better channel applies successive decoding and the other one treats interference as noise. More specifically, we assume that the channel from W to U_L is better than that to U_W. When U_L receives the signal transmitted from W,

it decodes first the message intended for the native Wi-Fi UE, subtracting it from the received signal, then decodes its own packet. The native Wi-Fi UE decodes its packets treating the superimposed additional layer just as noise. For more information about how to deploy the SC method in Wi-Fi networks, see [14], [15]. Please refer to Table 1 for the notation for probabilities.

Given a set of non-empty queues denoted by $\mathcal{T}$, let $\mathcal{S}_{j/\mathcal{T}}$ denote the event that UE j successfully decodes the packet transmitted from the queue that contains packets intended for j. For example, $\mathcal{S}_{U_W/WU_L,WU_W}$ refers to the event that UE U_W can decode the message from W when both queues WU_L and WU_W are not empty, i.e., $\mathcal{T} = \{WU_L, WU_W\}$. Let $\mathcal{P}(\mathcal{B})$ represent the probability of occurrence of the event $\mathcal{B}$. Obviously, we always have $\mathcal{P}(\mathcal{S}_{U_W/WU_L,WU_W}) \leqslant \mathcal{P}(\mathcal{S}_{U_W/WU_W})$ and $\mathcal{P}(\mathcal{S}_{U_L/WU_L,WU_W}) \leqslant \mathcal{P}(\mathcal{S}_{U_L/WU_L})$.

In general, our scheme can be regarded as an extension of the Aloha random access scheme by adding a coordinator which exchanges the information between the native-mode and LWA-mode APs. The coordinator is also in charge of broadcasting the respective activities of the native-mode and LWA-mode AP depending on the queue size of WU_L. Although the exchange of information will introduce extra overhead to the LWA system operation, the proposed scheme is rather flexible.

III. NETWORK PERFORMANCE METRICS

In this section, we define several relevant metrics for the performance evaluation of the considered LWA-enabled network with the priority based Wi-Fi transmission scheme.

A. ANALYSIS OF THE QUEUES AT THE LTE BS

The service probability can be defined as the probability of a successful packet transmission per time slot. Recall that the packet transmission from queues LU_L and LW are interference-free because of the orthogonal frequency bands. The service probability for queue LU_L at the LTE BS is

$$\mu_{LU_L} = q_{\mathrm{U_L,L}} \cdot \mathcal{P}(\mathcal{S}_{U_L/LU_L}). \tag{1}$$

The event $\mathcal{S}_{U_L/LU_L}$ denotes the successful transmission of a packet in the queue LU_L, which means that the received signal-to-noise ratio (SNR) of link $L \rightarrow U_L$ is above a certain threshold γ_{U_L}, i.e.,

$$\mathcal{P}(\mathcal{S}_{U_L/LU_L}) = \mathcal{P}(\mathrm{SNR}_{LU_L} \geqslant \gamma_{U_L}). \tag{2}$$

Particularly, the SNR of link $L \rightarrow U_L$ can be represented as

$$\mathrm{SNR}_{LU_L} = \frac{P_{LU_L}|h_{LU_L}|^2 d_{LU_L}^{-\alpha}}{\sigma^2}, \tag{3}$$

where P_{LU_L} denotes the transmission power of node L while serving U_L; h_{LU_L} refers to the small-scale channel fading from the transmitter L to the receiver U_L, which follows Rayleigh fading with unit mean; σ^2 is the noise power. Here we assume a standard distance-dependent power law pass loss attenuation $d_{LU_L}^{-\alpha}$, where d_{LU_L} denotes the distance from the transmitter L to the receiver U_L, and α with $\alpha > 2$ refers to the pathloss exponent. Combined with (1), we have

$$\mu_{LU_L} = q_{\mathrm{U_L,L}} \cdot \exp\left(-\frac{\gamma_{U_L} d_{LU_L}^{\alpha}}{P_{LU_L}}\right). \tag{4}$$

Similarly, for queue LW, the service rate is represented as

$$\mu_{LW} = \mathcal{P}(\mathcal{S}_{W/LW}) = \exp\left(-\frac{\gamma_W d_{LW}^{\alpha}}{P_{LW}}\right), \tag{5}$$

where γ_W refers to the SNR threshold for successful packet transmission to W; d_{LW} denotes the distance from L to W; P_{LW} denotes the transmission power of node L while offloading packets to W.

B. ANALYSIS OF THE QUEUES AT THE Wi-Fi AP

In the following, we will show how to compute the service rates for queue WU_L and queue WU_W, respectively, depending on the value of Q_{WU_L}.

1) When $Q_{WU_L} = 0$, AP W has no data to transmit to U_L. In such case, the service rate seen at queue WU_W is

$$\mu_{WU_W,1} = \mathcal{P}(\mathcal{S}_{U_W/WU_W}) = \exp\left(-\frac{\gamma_{U_W} d_{WU_W}^{\alpha}}{P_{WU_W}}\right), \tag{6}$$

where γ_{U_W} refers to the SNR threshold; d_{WU_W} denotes the distance from W to U_W; P_{WU_W} denotes the power of node W while transmitting packets to U_W.

2) When $1 \leqslant Q_{WU_L} \leqslant M$, the service rate seen at queue WU_W and queue WU_L are given by

$$\mu_{WU_W,2} = q_{\mathrm{W,W}} \cdot \mathcal{P}(\mathcal{S}_{U_W/WU_L,WU_W}), \tag{7}$$

$$\mu_{WU_L,1} = (1 - q_{\mathrm{W,W}}) \cdot q_{\mathrm{U_L,W}} \cdot \mathcal{P}(\mathcal{S}_{U_L/WU_L}) + q_{\mathrm{W,W}} \cdot q_{\mathrm{U_L,W}} \cdot \mathcal{P}(\mathcal{S}_{U_L/WU_L,WU_W}). \tag{8}$$

In order to compute (7) and (8), we need to derive $\mathcal{P}(\mathcal{S}_{U_L/WU_L,WU_W})$ and $\mathcal{P}(\mathcal{S}_{U_W/WU_L,WU_W})$ first. Take the event $\mathcal{S}_{U_L/WU_L,WU_W}$ for example. Recall that since WU_W is saturated, $Q_{WU_W} > 0$ always holds. The event $\mathcal{S}_{U_L/WU_L,WU_W}$ is feasible when the received signal-to-interference-plus-noise ratio (SINR) is above a threshold γ_{U_L} and can be expressed by

$$\mathcal{S}_{U_L/WU_L,WU_W} = \left\{ \frac{P_{WU_L}|h_{WU_L}|^2 d_{WU_L}^{-\alpha}}{1 + P_{WU_W}|h_{WU_L}|^2 d_{WU_L}^{-\alpha}} \geqslant \gamma_{U_L} \right\}, \tag{9}$$

where P_{WU_L} and P_{WU_W} denote the allocated transmission power of node W for the packets intended to reach U_L and U_W, respectively; d_{WU_W} denotes the distance from W to U_W; h_{WU_L} refers to the small-scale channel fading from the transmitter W to the receiver U_L, whose distribution also follows exp(1).

Since we consider a decoding strategy where the UE with the better channel (i.e., U_L) applies successive decoding, i.e., it decodes first the message of U_W, subtracting it from the received signal, and then decodes

its own message. The other user (i.e., U_W) treats the message of U_L as noise.

From (18) in [31], if $\frac{P_{WU_W}}{P_{WU_L}} > \frac{\gamma_{U_W}(1+\gamma_{U_L})}{\gamma_{U_L}}$, the successful decoding probability of U_L is given by

$$\mathcal{P}(\mathcal{S}_{U_L/WU_L,WU_W}) = \exp\left(-\frac{\gamma_{U_L} d_{WU_L}^{\alpha}}{P_{WU_L}}\right). \quad (10)$$

Otherwise, if $\gamma_{U_W} < \frac{P_{WU_W}}{P_{WU_L}} \leqslant \frac{\gamma_{U_W}(1+\gamma_{U_L})}{\gamma_{U_L}}$, we have

$$\mathcal{P}(\mathcal{S}_{U_L/WU_L,WU_W}) = \exp\left(-\frac{\gamma_{U_W} d_{WU_L}^{\alpha}}{P_{WU_W} - \gamma_{U_W} P_{WU_L}}\right). \quad (11)$$

From (15) in [31], the successful decoding probability of U_W can be derived as

$$\begin{aligned}\mathcal{P}(\mathcal{S}_{U_W/WU_L,WU_W}) &= \mathbb{1}\{P_{WU_W} > \gamma_{U_W} P_{WU_L}\} \\ &\quad \times \exp\left(-\frac{\gamma_{U_W} d_{WU_W}^{\alpha}}{P_{WU_W} - \gamma_{U_W} P_{WU_L}}\right). \quad (12)\end{aligned}$$

For the sake of simplicity, in the reminder of this paper, we assume that $\gamma_{U_W} < \frac{P_{WU_W}}{P_{WU_L}} \leqslant \frac{\gamma_{U_W}(1+\gamma_{U_L})}{\gamma_{U_L}}$ always holds.

3) When $Q_{WU_L} > M$, the service rate seen at queue WU_L can be represented by

$$\begin{aligned}\mu_{WU_L,2} &= q_{\mathrm{U_L,W}} \cdot \mathcal{P}(\mathcal{S}_{U_L/WU_L}) \\ &= \exp\left(-\frac{\gamma_{U_L} d_{WU_L}^{\alpha}}{P_{WU_L}}\right). \quad (13)\end{aligned}$$

Recall that by definition, the service probability for WU_W only accounts for the case with $Q_{WU_L} \leqslant M$.

In summary, the average service rate seen at queue WU_L is given by

$$\bar{\mu}_{WU_L} = \frac{T_{WU_L}}{\mathcal{P}(1 \leqslant Q_{WU_L} \leqslant M) + \mathcal{P}(Q_{WU_L} > M)}, \quad (14)$$

where

$$\begin{aligned}T_{WU_L} &= \mathcal{P}(1 \leqslant Q_{WU_L} \leqslant M) \cdot \mu_{WU_L,1} \\ &\quad + \mathcal{P}(Q_{WU_L} > M) \cdot \mu_{WU_L,2}. \quad (15)\end{aligned}$$

C. NATIVE Wi-Fi THROUGHPUT

The throughput of the native Wi-Fi link, denoted by T_{WU_W}, can be represented as

$$\begin{aligned}T_{WU_W} &= \mathcal{P}(Q_{WU_L} = 0) \cdot \mu_{WU_W,1} \\ &\quad + \mathcal{P}(1 \leqslant Q_{WU_L} \leqslant M) \cdot \mu_{WU_W,2}. \quad (16)\end{aligned}$$

Since we assume heavy data traffic requested by the native Wi-Fi user, the throughout of this link is limited by the service rate of queue WU_W.

D. LWA UE DELAY

The delay experienced by the LWA UE is a critical metric for the performance of the LWA system with delay-sensitive applications. Denote by $\bar{D}$ the average delay per packet at LWA UE U_L, which is the averaged over the possibilities that the packet will be transmitted through the LTE link or the LWA Wi-Fi link. The formal definition of D is

$$\bar{D} = q_{\mathrm{in}}\bar{D}_L + q_{\mathrm{off}}\bar{D}_W, \quad (17)$$

where $\bar{D}_L$ and $\bar{D}_W$ denote the delay per packet through the LTE link and the LWA Wi-Fi link, respectively. $\bar{D}_L$ can be represented as

$$\bar{D}_L = D_{LU_L}. \quad (18)$$

D_W equals to the sum of delay at queue LW and WU_L. Denote by D_i the average delay at queue i ($i \in \{LW, LU_L, WU_L, WU_W\}$) per packet, thus we have

$$\bar{D}_W = D_{LW} + D_{WU_L}. \quad (19)$$

Note that D_i ($i \in \{LU_L, WU_L, WU_W, LW\}$) consists of both queueing delay and transmission delay. From Little's law [32], we obtain the queueing delay as the average queue size per packet arrival. The transmission delay is inversely proportional to the service rate. In general, the following equation holds.

$$D_i = \frac{\bar{Q}_i}{\lambda_i} + \frac{1}{\bar{\mu}_i}, \quad i \in \{LU_L, WU_L, WU_W, LW\}, \quad (20)$$

where $\bar{Q}_i$ and $\bar{\mu}_i$ are the average queue size and the average service probability of the i-th queue, respectively.

IV. ANALYSIS OF NATIVE Wi-Fi THROUGHPUT AND LWA UE DELAY

From the performance metrics defined in Section III, the delay seen at the LWA UE U_L depends on $\bar{Q}_{WU_L}$, $\bar{Q}_{LW}$ and $\bar{Q}_{LU_L}$. In addition, the native Wi-Fi throughput depends on the state of the queue size of queue WU_L. In this section, we first derive $\mathcal{P}(Q_i = 0)$ and $\mathcal{P}(1 \leqslant Q_i \leqslant M)$, $i \in \{LU_L, LW, WU_L\}$, based on which we obtain the average queue size of LU_L, LW and WU_L. At last, we derive the native Wi-Fi throughput T_{WU_W} and the LWA UE delay $\bar{D}$.

A. ANALYSIS OF THE QUEUES

We first provide the definition of queue stability.

Definition 1: Denote by Q_i^t the length of queue i at the beginning of time slot t. The queue is said to be stable if $\lim_{t\to 0} \mathcal{P}(Q_i^t < x) = F(x)$ *and* $\lim_{x\to\infty} F(x) = 1$.

Although we will not make explicit use of this definition, here we take advantage of its corollary, namely Loynes' theorem [33], which states that if the average arrival rate is less than the average service rate, the queue will be stable. Otherwise, the queue is unstable and the value of Q_i^t approaches infinity.

From the system model, all the queues $i \in \{LU_L, WU_L, LW\}$ can be modeled as a Discrete Time Markov Chain

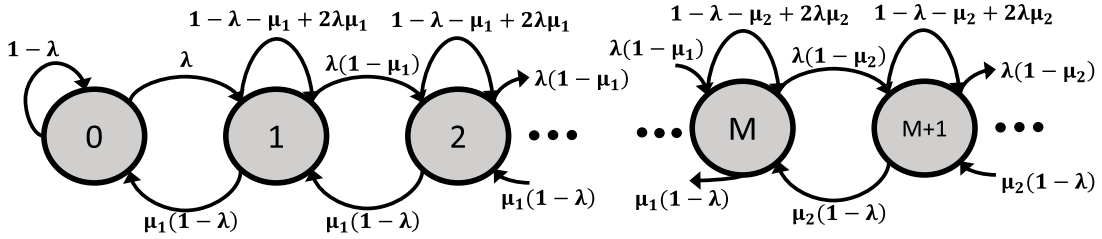


FIGURE 3. The Discrete Time Markov Chain which models the i-th queue evolution ($i \in \{LU_L, LW, WU_L\}$). When $i \in \{LU_L, LW\}$, $M \to \infty$ holds.

(DTMC), as presented in Fig. 3. Here, λ represents the arrival rate, μ_1 and μ_2 represent the service rates when $1 \leq Q \leq M$ and when $Q > M$, respectively. Note that for $i \in \{LU_L, LW\}$, $M \to \infty$ always holds since the congestion control is only applied to the native-mode Wi-Fi node. Thus, $\mu_1 = \mu_2$ for the queues $i \in \{LU_L, LW\}$.

Each state is denoted by an integer and represents the queue size. The metrics related to the rate are measured by the average number of packets per time slot.

Denote by π the stationary distribution of the DTMC, where $\pi(m) = \mathcal{P}(Q = m)$ is the probability that the queue Q has m packets in its steady state.

Lemma 1: The stationary distribution of the DTMC described in Fig. 3 is

1) *For $1 \leq Q \leq M$, we have*

$$\pi(m) = \frac{\lambda^m (1-\mu_1)^{m-1}}{(1-\lambda)^m \mu_1^m} \pi(0). \tag{21}$$

2) *For $Q > M$, we have*

$$\pi(m) = \frac{\lambda^m (1-\mu_1)^M (1-\mu_2)^{m-M-1}}{(1-\lambda)^m \mu_1^m \mu_2^{m-M}} \pi(0), \tag{22}$$

where $\pi(0)$ is the probability that the queue is empty, given by

1) *If $\lambda \neq \mu_1$, we have*

$$\pi(0) = \frac{(\mu_1 - \lambda)(\mu_2 - \lambda)}{\mu_1\mu_2 - \lambda\mu_1 - \lambda\left[\frac{\lambda(1-\mu_1)}{(1-\lambda)\mu_1}\right]^M (\mu_2 - \mu_1)}. \tag{23}$$

2) *If $\lambda = \mu_1$, we have*

$$\pi(0) = \frac{\mu_2 - \mu_1}{\mu_1 + (\mu_2 - \mu_1)\frac{M+1-\mu_1}{1-\mu_1}}. \tag{24}$$

Proof: Refer to Appendix A. ∎

Lemma 2: The queue in Fig. 3 is stable if and only if $\lambda < \mu_2$ holds.

Proof: Refer to Appendix B. ∎

With Lemma 1, the probability for $1 \leqslant Q \leqslant M$ and $Q > M$ when the queue is stable can be derived as in the following theorem. For the sake of simplicity, in the rest of this work, we only consider the case where $\lambda \neq \mu_1$. The result for $\lambda = \mu_1$ can be derived in a similar way. For convenience, in the reminder of this paper, let $\phi \triangleq \frac{\lambda(1-\mu_1)}{(1-\lambda)\mu_1}$.

Theorem 1: When the queue in Fig. 3 is stable, i.e., $\lambda < \mu_2$, and $\lambda \neq \mu_1$, the following two equations hold:

$$\mathcal{P}(1 \leqslant Q \leqslant M) = \frac{\lambda(1-\phi^M)(\mu_2 - \lambda)}{\mu_1\mu_2 - \lambda\mu_1 - \lambda\phi^M(\mu_2 - \mu_1)}. \tag{25}$$

$$\mathcal{P}(Q > M) = \frac{\lambda\phi^M(\mu_1 - \lambda)}{\mu_1\mu_2 - \lambda\mu_1 - \lambda\phi^M(\mu_2 - \mu_1)}. \tag{26}$$

Proof: The proof can be found in our previous paper [24]. ∎

Corollary 1: When $M \to \infty$, $\mathcal{P}(Q \geqslant 1)$ is given by

$$\mathcal{P}(Q \geqslant 1) = \frac{\lambda}{\mu_1}. \tag{27}$$

Theorem 2: The average queue size of the queue in Fig. 3 is given by

$$\bar{Q} = \frac{K_1 + K_2}{\mu_1\mu_2 - \lambda\mu_1 - \lambda\phi^M(\mu_2 - \mu_1)}, \tag{28}$$

where

$$K_1 = \phi^M \lambda(\mu_1 - \lambda)\left[M + \frac{(1-\lambda)\mu_2}{\mu_2 - \lambda}\right], \tag{29}$$

and

$$K_2 = \lambda(1-\lambda)\mu_1 \frac{\mu_2 - \lambda}{\mu_1 - \lambda}\left[M\phi^{M+1} - \phi^M(M+1) + 1\right]. \tag{30}$$

Proof: The proof is similar to that of Theorem 1 in our previous work [34]. ∎

Corollary 2: The average queue size of the queue in Fig. 3 when $M \to \infty$ is given by

$$\bar{Q} = \frac{\lambda(1-\lambda)}{\mu_1 - \lambda}. \tag{31}$$

Proof: Refer to Appendix C. ∎

B. ANALYSIS OF THE NATIVE Wi-Fi THROUGHPUT

From Lemma 1, Theorem 2 and (16), we have

$$\begin{aligned} T_{WU_W} &= \mathcal{P}(Q_{WU_L} = 0) \cdot \mu_{WU_W,1} \\ &\quad + \mathcal{P}(1 \leqslant Q_{WU_L} \leqslant M) \cdot \mu_{WU_W,2} \\ &= \frac{N_1 \cdot (N_2 + N_3)}{N_4 - N_5}, \end{aligned} \tag{32}$$

where

$$N_1 = \mu_{WU_L,2} - \lambda_{WU_L}, \tag{33}$$

$$N_2 = \mu_{WU_W,1}(\mu_{WU_L,1} - \lambda_{WU_L}), \tag{34}$$

$$N_3 = \mu_{WU_W,2}\lambda_{WU_L}(1-\xi^M), \tag{35}$$
$$N_4 = \mu_{WU_L,1}(\mu_{WU_L,2}-\lambda_{WU_L}), \tag{36}$$

and

$$N_5 = \lambda_{WU_L}\xi^M(\mu_{WU_L,2}-\mu_{WU_L,1}). \tag{37}$$

The entity ξ is represented by

$$\xi \triangleq \frac{\lambda_{WU_L}(1-\mu_{WU_L,1})}{(1-\lambda_{WU_L})\mu_{WU_L,1}}. \tag{38}$$

A special case is when $M \to \infty$. In such case, by Corollary 1 and (16), equation (32) can be transformed to

$$T_{WU_W} = (1-\frac{\lambda_{WU_L}}{\mu_{WU_L,1}})\cdot\mu_{WU_W,1} + \frac{\lambda_{WU_L}}{\mu_{WU_L,1}}\cdot\mu_{WU_W,2} \tag{39}$$

C. ANALYSIS OF THE LWA UE DELAY

From Theorem 2 and Corollary 2, we have

$$\begin{aligned} D_{LU_L} &= \frac{\bar{Q}_{LU_L}}{\lambda_{LU_L}} + \frac{1}{\mu_{LU_L}} \\ &= \frac{1-\lambda_{LU_L}}{\mu_{LU_L}-\lambda_{LU_L}} + \frac{1}{\mu_{LU_L}}, \end{aligned} \tag{40}$$

$$\begin{aligned} D_{LW} &= \frac{\bar{Q}_{LW}}{\lambda_{LW}} + \frac{1}{\mu_{LU_L}} \\ &= \frac{1-\lambda_{LW}}{\mu_{LW}-\lambda_{LW}} + \frac{1}{\mu_{LW}}, \end{aligned} \tag{41}$$

$$\begin{aligned} D_{WU_L} &= \frac{\bar{Q}_{WU_L}}{\lambda_{WU_L}} + \frac{1}{\bar{\mu}_{WU_L}} \\ &= \frac{L_1+L_2}{L_3} + \frac{1}{\bar{\mu}_{WU_L}}, \end{aligned} \tag{42}$$

where

$$L_1 = \xi^M(\mu_{WU_L,1}-\lambda_{WU_L})\left[M + \frac{(1-\lambda_{WU_L})\mu_{WU_L,2}}{\mu_{WU_L,2}-\lambda_{WU_L}}\right], \tag{43}$$

$$L_2 = (1-\lambda_{WU_L})\mu_{WU_L,1}\frac{\mu_{WU_L,2}-\lambda_{WU_L}}{\mu_{WU_L,1}-\lambda_{WU_L}} \cdot\left[M\xi^{M+1} - \xi^M(M+1)+1\right], \tag{44}$$

$$L_3 = \mu_{WU_L,1}\mu_{WU_L,2} - \lambda_{WU_L}\mu_{WU_L,1} - \lambda_{WU_L}\xi^M(\mu_{WU_L,2}-\mu_{WU_L,1}). \tag{45}$$

Also, remark that $\lambda_{WU_L} \neq \mu_{WU_L,1}$, from (14) and the result of Theorem 1, we have

$$\bar{\mu}_{WU_L} = \frac{\mu_{WU_L,1}H_1 + \mu_{WU_L,2}H_2}{H_1+H_2}, \tag{46}$$

where

$$H_1 = \lambda_{WU_L}(1-\xi^M)(\mu_{WU_L,2}-\lambda_{WU_L}) \tag{47}$$
$$H_2 = \lambda_{WU_L}\xi^M(\mu_{WU_L,1}-\lambda_{WU_L}) \tag{48}$$

Please note that when $M \to \infty$, similar to (40), equation (42) can be transformed to

$$\begin{aligned} D_{WU_L} &= \frac{\bar{Q}_{WU_L}}{\lambda_{WU_L}} + \frac{1}{\mu_{WU_L,1}} \\ &= \frac{1-\lambda_{WU_L}}{\mu_{WU_L,1}-\lambda_{WU_L}} + \frac{1}{\mu_{WU_L,1}}, \end{aligned} \tag{49}$$

Remark that the LWA UE delay $\bar{D}$ can be computed using (17), (18), and (19).

V. NUMERICAL RESULTS

In this section, we provide numerical evaluation of the analytical results presented in the previous sections. To be more specific, we plot the native Wi-Fi throughput and the delay of the LWA UE as functions of the native Wi-Fi AP access probability, the probability that the LWA UE chooses the LTE or Wi-Fi interface at one time slot, and the probability that an LTE packet to be routed through the LTE or the Wi-Fi link. The values of the simulation parameters are given in Table 2. Note that all the results below are obtained where the queue stability condition is satisfied.

TABLE 2. System parameters.

Parameters	Values
LTE link distance (d_{LU_L})	300 m
Native Wi-Fi link distance (d_{WU_W})	20 m
LWA-mode Wi-Fi link distance (d_{WU_L})	20 m
Backhaul distance (d_{LW})	200 m
Pathloss exponent (α)	4
LTE BS transmission power (P_{LU_L}, P_{LW})	200 mw
Wi-Fi AP transmission power (P_{WU_L}, P_{WU_W})	20 mw
Noise power (σ^2)	-113.97 dbm
SNR target (γ_j, $\forall j$)	0 dB

A. NATIVE Wi-Fi THROUGHPUT

From (16), the native Wi-Fi throughput depends heavily on the value of $\mathcal{P}(Q_{WU_L}=0)$ and $\mathcal{P}(1 \leqslant Q_{WU_L} \leqslant M)$. In Fig. 4, we plot the probability to have $Q_{WU_L}=0$ with respect to $q_{W,W}$ for the cases with $M=\{1,3\}$. The results are generated from (23). As expected, with smaller q_{off}, the probability that queue WU_L is empty is higher. We also observe that larger M leads to higher $\mathcal{P}(Q_{WU_L}=0)$. This is due to the fact that with weaker congestion control, the LWA-mode Wi-Fi AP will remain silent with higher probability. In addition, when $q_{W,W}$ is larger, $\mathcal{P}(Q_{WU_L}=0)$ has larger variation with respect to the variations of M. This means that when the probability that W serves the native Wi-Fi UE is low enough, M does not really affect the value of $\mathcal{P}(Q_{WU_L}=0)$.

We also plot the probability $\mathcal{P}(1 \leqslant Q_{WU_L} \leqslant M)$ with respect to $q_{W,W}$, as illustrated in Fig. 5. The results are generated from (25). The first important observation is that $\mathcal{P}(1 \leqslant Q_{WU_L} \leqslant M)$ is not always a monotonic function of $q_{W,W}$. In addition, it is not always the highest q_{off} that gives the largest $\mathcal{P}(1 \leqslant Q_{WU_L} \leqslant M)$. Moreover, as expected, larger M leads to smaller $\mathcal{P}(1 \leqslant Q_{WU_L} \leqslant M)$.

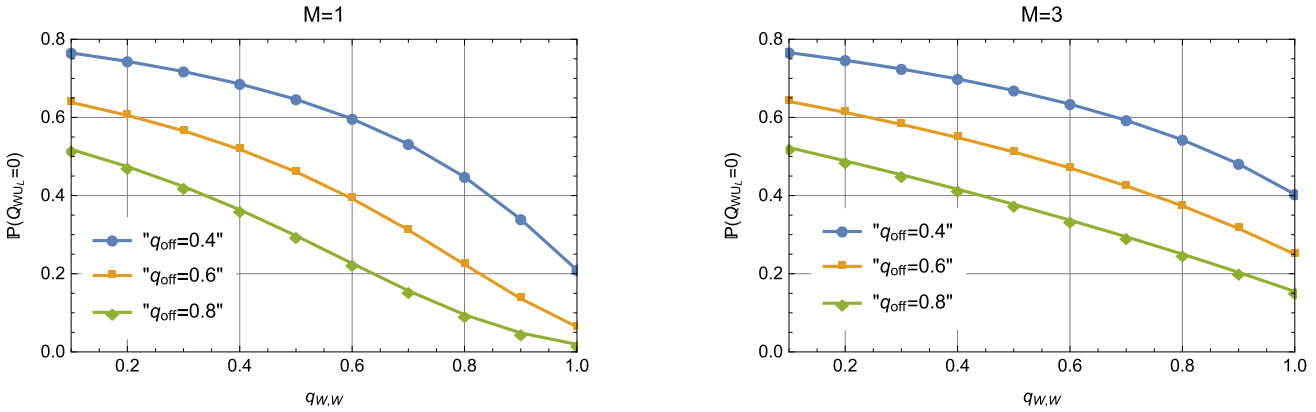


FIGURE 4. $\mathcal{P}(Q_{WU_L} = 0)$ vs. $q_{W,W}$.

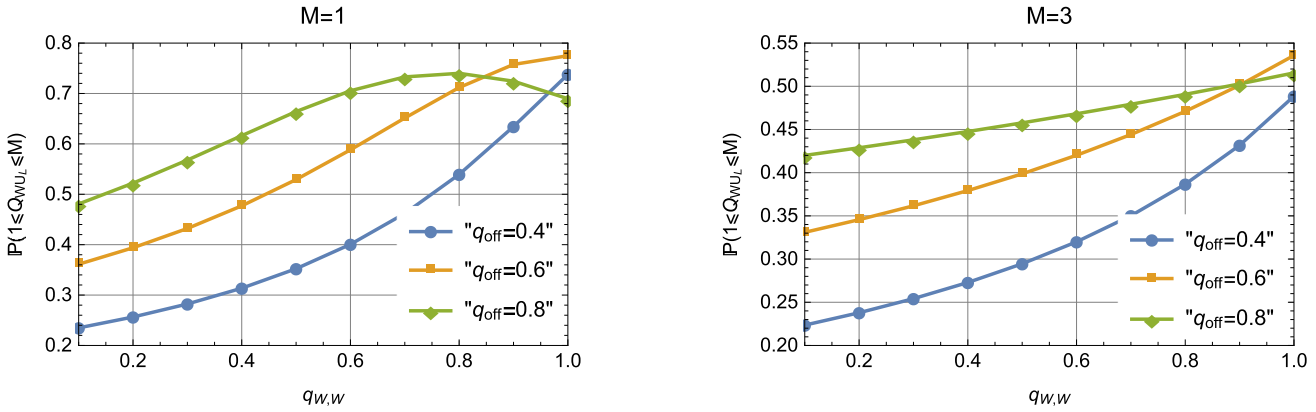


FIGURE 5. $\mathcal{P}(1 \leqslant Q_{WU_L} \leqslant M)$ vs. $q_{W,W}$.

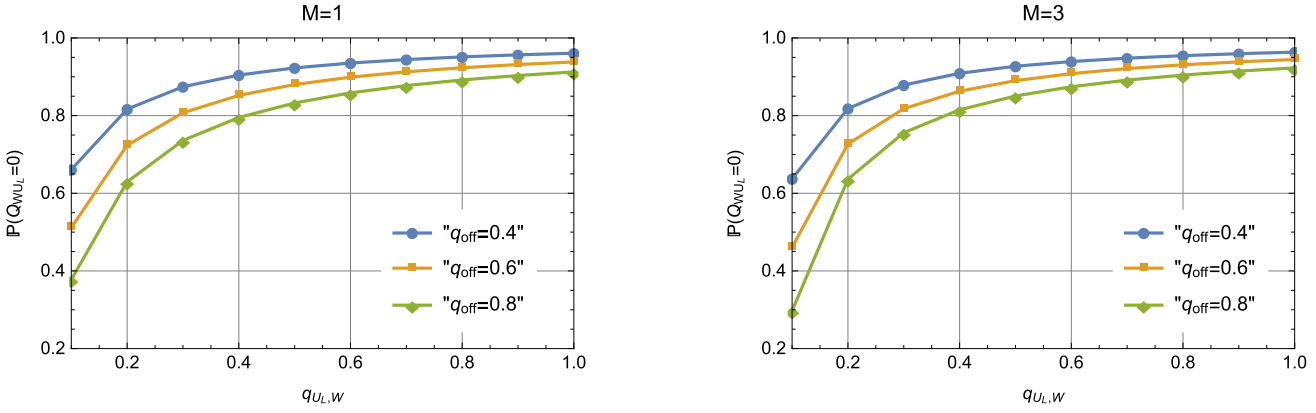


FIGURE 6. $\mathcal{P}(Q_{WU_L} = 0)$ vs. $q_{U_L,W}$.

Similarly, in Fig. 6, we present the probability to have $Q_{WU_L} = 0$ with respect to $q_{U_L,W}$ for the cases with $M = \{1, 3\}$. From this figure, with $q_{U_L,W}$ increasing, $\mathcal{P}(Q_{WU_L} = 0)$ increases at first, then saturates. Another important observation is that larger M does not increase the maximum value of $\mathcal{P}(Q_{WU_L} = 0)$. Fig. 7 shows the relationship between $\mathcal{P}(1 \leqslant Q_{WU_L} \leqslant M)$ and $q_{U_L,W}$. As expected, larger $q_{U_L,W}$ leads to lower $\mathcal{P}(1 \leqslant Q_{WU_L} \leqslant M)$. However, when the value of $q_{U_L,W}$ becomes higher, $\mathcal{P}(1 \leqslant Q_{WU_L} \leqslant M)$ has smaller variation with respect to the variations of M. This means that when the probability that W serves the native Wi-Fi UE is low enough, M does not really affect the value of $\mathcal{P}(1 \leqslant Q_{WU_L} \leqslant M)$.

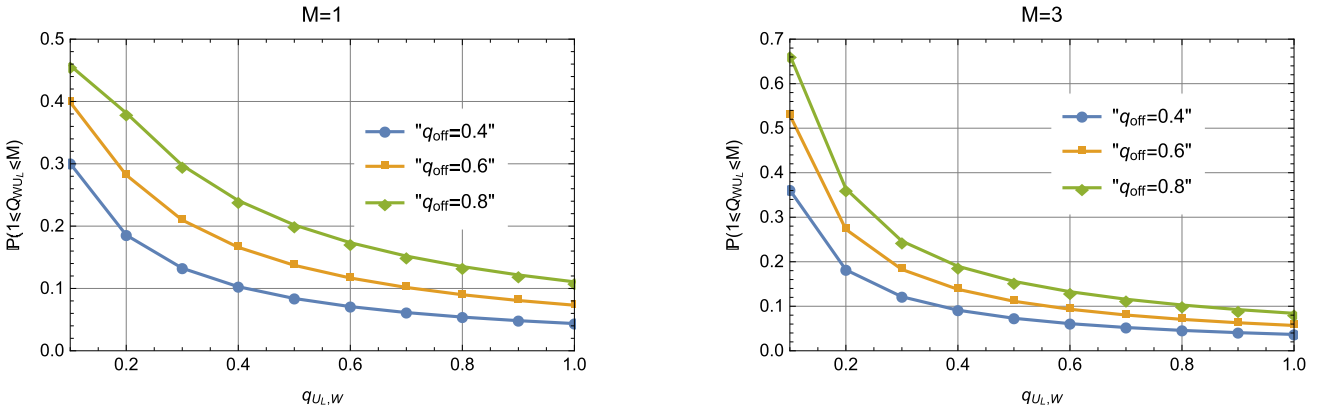


FIGURE 7. $\mathcal{P}(1 \leqslant Q_{WU_L} \leqslant M)$ vs. $q_{U_L,W}$.

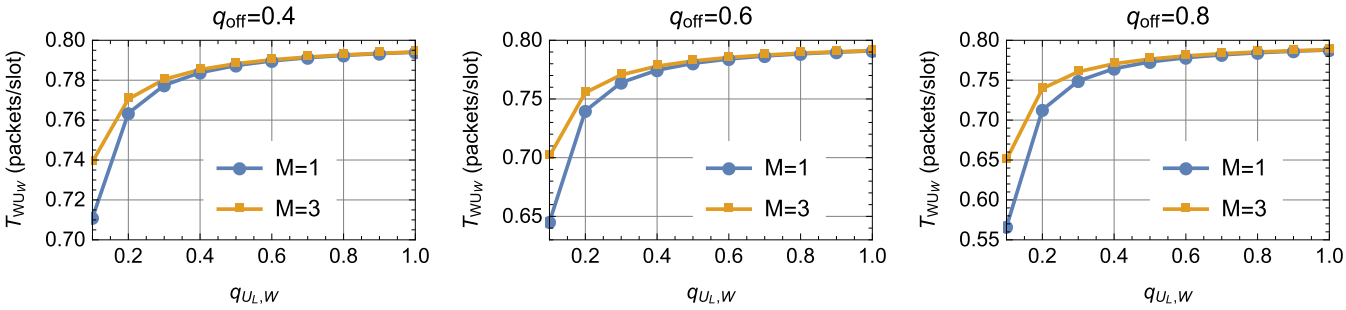


FIGURE 8. Native Wi-Fi throughput vs. $q_{U_L,W}$.

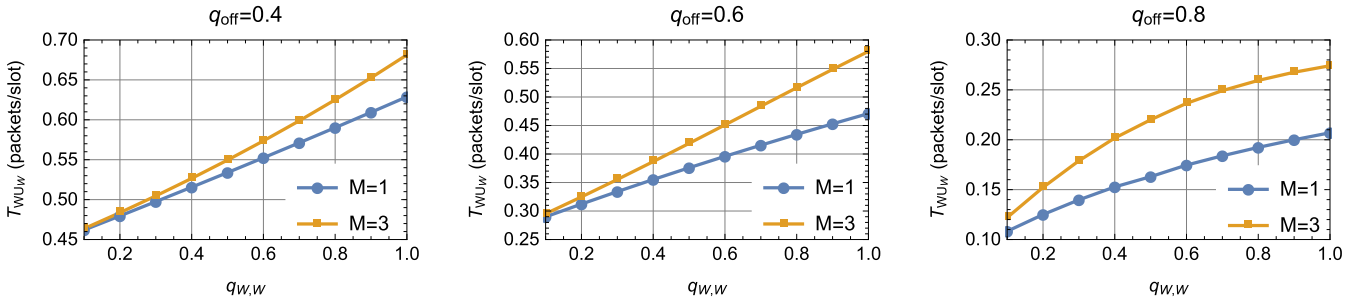


FIGURE 9. Native Wi-Fi throughput vs. $q_{W,W}$.

In Fig. 8, we plot the native Wi-Fi throughput with respect to the probability that the LWA UE activates the LTE interface only to receive packet in each time slot. The results are presented with congestion threshold $M = \{1, 3\}$ and $q_{\text{off}} = \{0.4, 0.6, 0.8\}$. Our first remark is that, with $q_{U_L,W}$ increasing, the native Wi-Fi throughput increases rapidly at first, then saturates. We also observe that larger M provides higher potential improvement for the native Wi-Fi throughput, as the native Wi-Fi link is more likely to be active. However, when $q_{U_L,W}$ becomes larger, T_{WU_W} has smaller variation with respect to variations of M, since the probability that queue WU_L is empty increases. In addition, comparing the sub-figures in Fig. 8, the maximum throughput of the native Wi-Fi network remains the same with different q_{off}.

In Fig. 9 we draw the native Wi-Fi throughput with respect to $q_{W,W}$. We observe that the native Wi-Fi throughput increases with $q_{W,W}$. Another interesting observation is that for the same q_{off}, the difference between the native Wi-Fi throughput with $M = 1$ and that with $M = 3$ increases with $q_{W,W}$, since with M increasing, the probability that queue WU_L is empty decreases. It can be observed from Fig. 9 that, for the same value of $q_{W,W}$, increasing q_{off} will also increase the difference between the native Wi-Fi throughput with $M = 1$ and that with $M = 3$. This is because when q_{off} is relatively low, the arrival rates λ_{WU_W} is also low, in which case M does not really affect the system, since the probability that WU_L is not empty decreases.

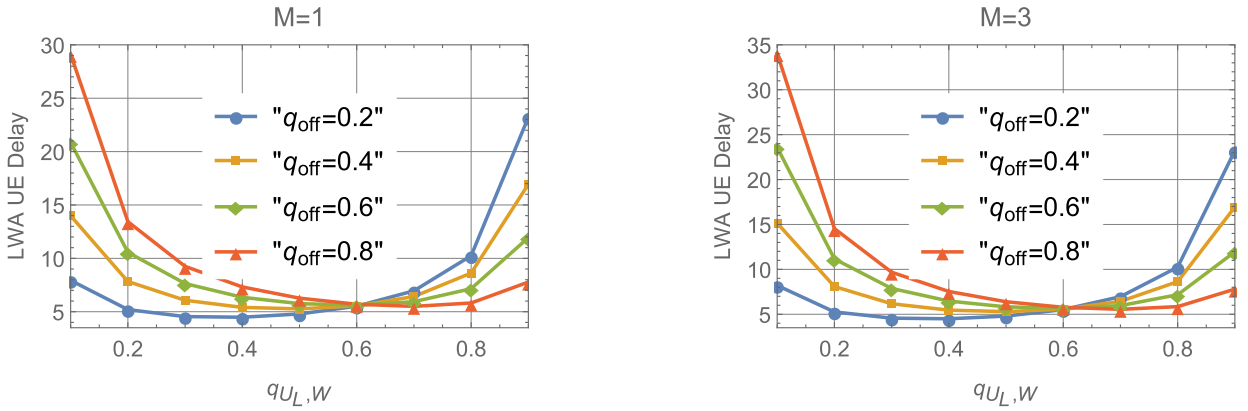


FIGURE 10. LWA UE delay vs. $q_{U_L,W}$.

B. LWA UE DELAY

To illustrate the impact of different parameters on the LWA UE delay. We first plot Fig. 10 to present the LWA UE delay as a function of $q_{U_{L,W}}$ for different values of M and q_{off}. The first observation is that the LWA UE delay is not a monotonic function of $q_{U_{L,W}}$. There exists an optimal point that gives the minimum LWA UE delay among the feasible choice of $q_{U_{L,W}}$. Second, comparing the sub-figures in Fig. 10, we observe that larger M results in higher LWA UE delay. However, once $q_{U_{L,W}}$ reaches a certain level, e.g. $q_{U_{L,W}} = 0.2$ in Fig. 10, $\bar{D}$ has very little variation with respect to the variation of M. The reason is the probability that queue WU_L is empty increases with $q_{U_{L,W}}$. In such case, due to the low utilization, choosing $M = 1$ in our protocol is beneficial. Third, it is not always the highest q_{off} that gives the largest $\bar{D}$. The first reason is that larger q_{off} leads to higher D_{WU_L}, but smaller D_{LU_L} and D_{LW}. Another reason is that the success probability for the link $W \rightarrow U_L$ is not constant, but depends on the specific value of M and $q_{U_{L,W}}$, as described in Section III. To be more specific, the value of M affects the probability of the queue size of WU_L to fall in the three different cases, and the value of $q_{U_{L,W}}$ affects the queue size of WU_L in cases $1 \leqslant Q_{WU_L} \leqslant M$ and $Q_{WU_L} > M$.

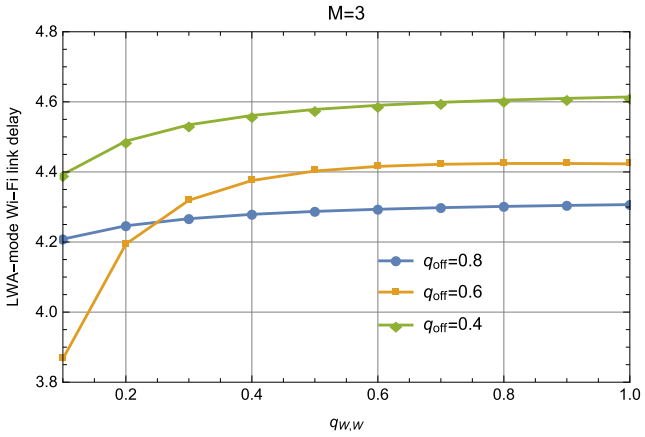


FIGURE 11. LWA UE delay vs. $q_{U_L,W}$.

Fig. 11 presents the delay $\bar{D}_W$ as a function of $q_{W,W}$ for different values of q_{off}. An interesting observation is that the LWA UE delay increases rapidly at first, then saturates. Higher q_{off} leads to lower saturated delay. The reason is when q_{off} is very high, the native Wi-Fi AP will not be allowed to transmit with high probability, as the queue size of WU_L falls in the case $Q_{WU_L} > M$ with high probability. Then, with high success probability, the LWA-mode Wi-Fi transmission is almost interference-free.

VI. CONCLUSION

This paper investigated an LWA-enabled network consisting of an LTE BS and a Wi-Fi AP. The LTE BS has bursty arrivals, and transmits packets to the Wi-Fi AP through a non-ideal backhaul. The AP can operate in LWA mode and native Wi-Fi mode simultaneously with the help of SC. We proposed a priority-based Wi-Fi transmission scheme with congestion control and studied the throughput of the native Wi-Fi network, as well as the LWA UE delay when the native Wi-Fi UE is under heavy traffic conditions. We further studied the impact of the scheme design parameters on the throughput and delay performance. Our results provide fundamental insights in the throughput and delay behavior of the considered network, which are essential for further investigation of this topic in larger topologies.

A. PROOF OF LEMMA 1

The derivation follows the similar techniques as in [24]. From the DTMC described in Fig. 3, we obtain the following balance equations:

$$\begin{aligned}
&\pi(0) = \pi(0)(1-\lambda) + \pi(1)\mu_1(1-\lambda) \\
\Leftrightarrow\ &\pi(1) = \frac{\lambda}{\mu_1(1-\lambda)}\pi(0). \\
&\pi(1) = \pi(0)\lambda + \pi(1)(1-\lambda-\mu_1+2\lambda\mu_1) \\
&\qquad + \pi(2)\mu_1(1-\lambda) \\
\Leftrightarrow\ &\pi(2) = \left(\frac{\lambda}{\mu_1(1-\lambda)}\right)^2 (1-\mu_1)\pi(0). \qquad (50)
\end{aligned}$$

In summary, for $1 \leq m \leq M$ we can derive (21), and for $m > M$, the (22) follows.

In addition, we have that $\sum_{m=0}^{\infty} \pi(m) = 1$ holds. This, together with (21) and (22), shows that the (23) holds when $\lambda \neq \mu_1$. When $\lambda = \mu_1$, denote by $x(\lambda)$ and $y(\lambda)$ the nomination and denominator of $\pi(0)$. We can derive $\pi(0)$ as

$$\pi(0) = \lim_{\lambda \to \mu_1} \frac{x'(\lambda)}{y'(\lambda)}. \tag{51}$$

Then equation (24) follows by applying 1'Hôpital's rule.

B. PROOF OF LEMMA 2

The derivation follows the similar techniques as in [34]. From $\sum_{m=0}^{\infty} \pi_i(m) = 1$, the condition that the series is converging when $\lambda < \mu_2$, which is also the condition that the DTMC is an aperiodic irreducible Markov Chain, showing that the queue is stable.

In addition, the condition $0 \leqslant \pi(0) \leqslant 1$ should also be satisfied. In the following, we consider the three specific cases:

1) If $\lambda < \mu_1$, consider equation (23), obviously $\pi(0) > 0$. In addition, we have the denominator $y(\lambda) > \mu_2(\mu_1 - \lambda)$, thus $\pi(0) < \frac{\mu_2 - \lambda}{\mu_2} < 1$ holds.
2) If $\lambda = \mu_1$, consider equation (24), obviously $0 < \pi(0) < 1$ holds.
3) If $\mu_1 < \lambda < \mu_2$, consider equation (23), obviously the nominator $x(\lambda) < 0$. we also have $\frac{\lambda(1-\mu_1)}{(1-\lambda)\mu_1} > 1$. Since both $\left(\frac{\lambda}{\mu_1}\right)^{M+1} > 1$ and $\left(\frac{1-\lambda}{1-\mu_1}\right)^{M} \frac{\mu_2 - \lambda}{\mu_2 - \mu_1} < 1$ hold, we have $y(\lambda) < 0$. Therefore $\pi(0) > 0$. Similar to the case $\lambda < \mu_1$, we still have $\pi(0) < 1$.

Therefore the conclusion that $0 < \pi(0) < 1$ always holds.

C. PROOF OF COROLLARY 2

The average queue size of the queue is

$$\bar{Q} = \sum_{m=1}^{\infty} m\pi(m). \tag{52}$$

Combined with (21), we have

$$\bar{Q} = \frac{\mu_1 - \lambda}{\mu_1} \cdot \frac{\lambda}{\mu_1(1-\lambda)} \cdot \sum_{m=1}^{\infty} m \left(\frac{\lambda(1-\mu_1)}{\mu_1(1-\lambda)}\right)^{m-1}. \tag{53}$$

Note that $\sum_{m=1}^{\infty} m\alpha^{m-1} = \frac{1}{(1-\alpha)^2}$ holds for $\alpha < 1$. Therefore (31) follows.

REFERENCES

[1] B. Chen, J. Chen, Y. Gao, and J. Zhang, "Coexistence of LTE-LAA and Wi-Fi on 5 GHz with corresponding deployment scenarios: A survey," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 1, pp. 7–32, 1st Quart., 2017.

[2] P. Nuggehalli, "LTE-WLAN aggregation [industry perspectives]," *IEEE Wireless Commun.*, vol. 23, no. 4, pp. 4–6, Aug. 2016.

[3] D. Laselva, D. López-Pérez, M. Rinne, and T. Henttonen, "3GPP LTE-WLAN aggregation technologies: Functionalities and performance comparison," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 195–203, Mar. 2018.

[4] Q. Zhu *et al.*, "A digital polar transmitter with DC–DC converter supporting 256-QAM WLAN and 40-MHz LTE-A carrier aggregation," *IEEE J. Solid-State Circuits*, vol. 52, no. 5, pp. 1196–1209, Mar. 2017.

[5] D. López-Pérez *et al.*, "Long term evolution-wireless local area network aggregation flow control," *IEEE Access*, vol. 4, pp. 9860–9869, 2016.

[6] Y.-B. Lin, Y.-J. Shih, and P.-W. Chao, "Design and implementation of LTE RRM with switched LWA policies," *IEEE Trans. Veh. Technol.*, vol. 67, no. 2, pp. 1053–1062, Feb. 2018.

[7] Y. Ohta, N. Michiharu, S. Aikawa, and T. Ode, "Link layer structure for LTE-WLAN aggregation in LTE-advanced and 5G network," in *Proc. IEEE Conf. Standards Commun. Netw. (CSCN)*, Oct. 2015, pp. 83–88.

[8] S. Singh, S.-P. Yeh, N. Himayat, and S. Talwar, "Optimal traffic aggregation in multi-RAT heterogeneous wireless networks," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC)*, May 2016, pp. 626–631.

[9] H. Wang, C. Rosa, and K. I. Pedersen, "Inter-eNB flow control for heterogeneous networks with dual connectivity," in *Proc. IEEE 81st Veh. Technol. Conf. (VTC)*, May 2015, pp. 1–5.

[10] S. Borst, A. Ö. Kaya, D. Calin, and H. Viswanathan, "Optimal path selection in multi-RAT wireless networks," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr. 2016, pp. 592–597.

[11] B. Liu, Q. Zhu, and H. Zhu, "Delay-aware LTE WLAN aggregation in heterogeneous wireless network," *IEEE Access*, vol. 6, pp. 14544–14559, 2018.

[12] *IEEE Standard for Information Technology–Telecommunications and Information Exchange Between Systems Local and Metropolitan Area Networks–Specific Requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, IEEE Standard 802.11, Apr. 2012. doi: 10.1109/IEEESTD.2012.6178212.

[13] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[14] S. Vanka, S. Srinivasa, Z. Gong, P. Vizi, K. Stamatiou, and M. Haenggi, "Superposition coding strategies: Design and experimental evaluation," *IEEE Trans. Wireless Commun.*, vol. 11, no. 7, pp. 2628–2639, Jul. 2012.

[15] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 721–742, 2nd Quart., 2017.

[16] T. Cover, "An achievable rate region for the broadcast channel," *IEEE Trans. Inf. Theory*, vol. IT-21, no. 4, pp. 399–404, Jul. 1975.

[17] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: A POMDP framework," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 3, pp. 589–600, Apr. 2007.

[18] R. Urgaonkar and M. J. Neely, "Opportunistic scheduling with reliability guarantees in cognitive radio networks," *IEEE Trans. Mobile Comput.*, vol. 8, no. 6, pp. 766–777, Jun. 2009.

[19] A. Fanous and A. Ephremides, "Stable throughput in a cognitive wireless network," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 3, pp. 523–533, Mar. 2013.

[20] S. Ghez, S. Verdu, and S. C. Schwartz, "Stability properties of slotted Aloha with multipacket reception capability," *IEEE Trans. Autom. Control*, vol. 33, no. 7, pp. 640–649, Jul. 1988.

[21] V. Naware, G. Mergen, and L. Tong, "Stability and delay of finite-user slotted ALOHA with multipacket reception," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2636–2656, Jul. 2005.

[22] A. Rabbachin, T. Q. S. Quek, H. Shin, and M. Z. Win, "Cognitive network interference," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 2, pp. 480–493, Feb. 2011.

[23] S. Kompella, G. D. Nguyen, C. Kam, J. E. Wieselthier, and A. Ephremides, "Cooperation in cognitive underlay networks: Stable throughput tradeoffs," *IEEE/ACM Trans. Netw.*, vol. 22, no. 6, pp. 1756–1768, Dec. 2014.

[24] N. Pappas and Kountouris, "Throughput of a cognitive radio network under congestion constraints: A network-level study," in *Proc. 9th Int. Conf. Cogn. Radio Oriented Wireless Netw. Commun. (CROWNCOM)*, Oulu, Finland, Jun. 2014, pp. 162–166.

[25] A. E. Ewaisha and C. Tepedelenlioglu, "Throughput optimization in multi-channel cognitive radios with hard-deadline constraints," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 2355–2638, Apr. 2016.

[26] S. Singh, M. Geraseminko, S.-P. Yeh, N. Himayat, and S. Talwar, "Proportional fair traffic splitting and aggregation in heterogeneous wireless networks," *IEEE Commun. Lett.*, vol. 20, no. 5, pp. 1010–1013, Mar. 2016.

[27] F. Mehmeti and T. Spyropoulos, "Performance analysis of mobile data offloading in heterogeneous networks," *IEEE Trans. Mobile Comput.*, vol. 16, no. 2, pp. 482–497, Feb. 2017.

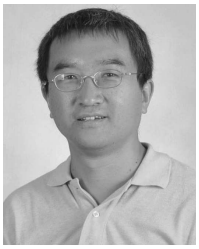
[28] C. Hua, H. Yu, R. Zheng, J. Li, and R. Ni, "Online packet dispatching for delay optimal concurrent transmissions in heterogeneous multi-RAT networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 5076–5085, Jul. 2016.
[29] Z. Zhou, D. Guo, and M. L. Honig, "Licensed and unlicensed spectrum allocation in heterogeneous networks," *IEEE Trans. Commun.*, vol. 65, no. 4, pp. 1815–1827, Apr. 2017.
[30] A. Ephremides and B. Hajek, "Information theory and communication networks: An unconsummated union," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2416–2434, Oct. 1998.
[31] N. Pappas, M. Kountouris, A. Ephremides, and V. Angelakis, "Stable throughput region of the two-user broadcast channel," *IEEE Trans. Commun.*, vol. 66, no. 10, pp. 4611–4621, Oct. 2018.
[32] W. Szpankowski, "Stability conditions for some distributed systems: Buffered random access systems," *Adv. Appl. Probab.*, vol. 26, no. 2, pp. 498–515, Jun. 1994.
[33] R. M. Loynes, "The stability of a queue with non-independent inter-arrival and service times," in *Math. Proc. Cambridge Philos. Soc.*, vol. 58, no. 3, pp. 497–520, Jul. 1962.
[34] Z. Chen, N. Pappas, M. Kountouris, and V. Angelakis, "Throughput with delay constraints in a shared access network with priorities," *IEEE Trans. Wireless Commun.*, vol. 17, no. 9, pp. 5885–5899, Sep. 2018.

BOLIN CHEN received the B.Eng. degree in photo-electronic information engineering from the Changchun University of Science and Technology, in 2014. He is currently pursuing the Ph.D. degree with the Department of Electronic and Electrical Engineering, The University of Sheffield. Since 2016, he has been a Visiting Researcher with Linköping University, Sweden. His research interests include 5G heterogeneous networks, network planning, cognitive femtocells, and cellular traffic offloading through Wi-Fi networks.

NIKOLAOS PAPPAS (S'07–M'13) received the B.Sc. degree in computer science, the B.Sc. degree in mathematics, the M.Sc. degree in computer science, and the Ph.D. degree in computer science from the University of Crete, Greece, in 2005, 2012, 2007, and 2012, respectively. From 2005 to 2012, he was a Graduate Research Assistant with the Telecommunications and Networks Laboratory, Institute of Computer Science, Foundation for Research and Technology, Hellas, and a Visiting Scholar with the Institute of Systems Research, University of Maryland at College Park, College Park, MD, USA. From 2012 to 2014, he was a Postdoctoral Researcher with the Department of Telecommunications, Supélec, France. Since 2014, he has been with Linköping University, as a Marie Curie Fellow. He is currently an Associate Professor in mobile telecommunications with the Department of Science and Technology, Linköping University, Norrköping, Sweden. His current research interests include wireless communication networks with emphasis on the stability analysis, energy harvesting networks, network-level cooperation, age-of-information, network coding, and stochastic geometry. From 2013 to 2018, he was an Editor of the IEEE COMMUNICATIONS LETTERS. He is currently an Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS and the IEEE/KICS JOURNAL OF COMMUNICATIONS AND NETWORKS.

ZHENG CHEN (S'14–M'17) received the B.S. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2011, and the M.S. and Ph.D. degrees from CentraleSupélec, Gif-sur-Yvette, France, in 2013 and 2016, respectively. In 2015, she was a Visiting Scholar with the Singapore University of Technology and Design, Singapore. Since 2017, she has been a Postdoctoral Researcher with Linköping University, Linköping, Sweden. Her research interests include stochastic geometry, queuing analysis, stochastic optimization, device-to-device communication, and wireless caching networks. She was selected as an Exemplary Reviewer for the IEEE COMMUNICATIONS LETTERS, in 2016, and the Best Reviewer for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, in 2017.

DI YUAN (M'03–SM'15) received the M.Sc. degree in computer science and engineering and the Ph.D. degree in optimization from the Linkoping Institute of Technology, in 1996 and 2001, respectively. He was a Guest Professor with the Technical University of Milan, Italy, in 2008, and a Senior Visiting Scientist with Ranplan Wireless Network Design Ltd., U.K., in 2009 and 2012. In 2011 and 2013, he has been part time with Ericsson Research, Sweden. In 2014 and 2015, he was a Visiting Professor with the University of Maryland at College Park, College Park, MD, USA. He is currently a Full Professor in telecommunications with the Department of Science and Technology, Linköping University, Sweden. His current research interests include network optimization of 4G and 5G systems, and capacity optimization of wireless networks. He was a co-recipient of the IEEE ICC12 Best Paper Award and a Supervisor of the Best Student Journal Paper Award by the IEEE Sweden Joint VT-COM-IT Chapter, in 2014. He is an Area Editor of the *Computer Networks* journal. He has been in the management committee of four European Cooperation in Scientific and Technical Research (COST) actions, an Invited Lecturer of European Network of Excellence EuroNF, and Principal Investigator of several European FP7 and Horizon 2020 projects.



JIE ZHANG received the M.Eng. and Ph.D. degrees from the Department of Automatic Control and Electronic Engineering, East China University of Science and Technology, Shanghai, China. He became a Lecturer, a Reader, and a Professor, in 2002, 2005, and 2006, respectively. He has been a Full Professor and the Chair in wireless systems with the Department of Electronic and Electrical Engineering, The University of Sheffield, since 2011. He is currently a Visiting Professor with the Chongqing University of Posts and Telecommunications and the East China University of Science and Technology. He and his students/colleagues have pioneered research in femto/small cell and HetNets and published some of the earliest and/or most cited publications in these topics. Since 2005, he received more than 20 grants by the EPSRC, the EC FP6/FP7/H2020 and industry, including some of the world's earliest research projects on femtocell/HetNets. He co-founded RANPLAN Wireless Network Design Ltd., which produces a suite of world leading in-building DAS, indoor–outdoor small cell/HetNet network design, and optimization tools iBuildNet that have been used by Ericsson, Huawei, and Cisco.