IEEE Transactions on Wireless Communications Volume: 18 , Issue: 4 , April 2019

QoS-Aware User Association and Resource Allocation in LAA-LTE/WiFi Coexistence Systems

Junjie Tan, *Student Member, IEEE*, Sa Xiao, Shiying Han, *Member, IEEE*, Ying-Chang Liang, *Fellow, IEEE*, and Victor C. M. Leung, *Fellow, IEEE*

***Abstract*—The licensed-assisted access based long term evolution (LAA-LTE) is a promising solution to provide enhanced LTE services by sharing unlicensed bands with WiFi systems. However, the intense contention with the incumbent WiFi system makes it challenging for the LAA-LTE system to support guaranteed quality-of-service (QoS) for the users. This paper is interested in the QoS-aware LAA-LTE/WiFi coexistence system. We first propose a flexible coexistence framework using the listen-before-talk mechanism, based on which the QoS metrics of LAA-LTE and WiFi systems are quantified. Then, a joint user association and resource allocation problem is formulated, which aims to maximize the number of QoS-preferred users supported by LAA-LTE, while protecting the WiFi users. The considered optimization problem is equivalently decomposed into two subproblems, the sum-power minimization problem and the user association problem. For the first subproblem, the deep-cut ellipsoid method is adopted to optimize the LAA-LTE transmission time, subcarrier assignment and power allocation. For the latter one, an efficient algorithm called successive user removal is proposed. Simulation results have demonstrated the effectiveness of the proposed scheme, based on which the tradeoff among different QoS metrics in the coexistence system is observed.**

***Index Terms*—Licensed-assisted access (LAA), long term evolution (LTE), unlicensed band, coexistence system, Quality-of-Service (QoS).**

I. Introduction

The global mobile traffic is expected to increase exponentially because of the explosive growth of mobile devices and emerging mobile applications [1]. In contrast, it is hard to further improve the capacity of existing cellular networks, due to the lack of dedicated spectra. A promising solution is to harvest additional spectra through *cognitive radio* (CR) technology, which supports spectrum sharing between multiple systems using opportunistic or concurrent access [2]–[8]. *Licensed-assisted access based long term evolution* (LAA-LTE) has thus been proposed to enhance the cellular network capacity by sharing the unlicensed bands with WiFi systems [9].

Since the WiFi system in unlicensed bands employs contention-based *media access control* (MAC) protocols [10], e.g., *Carrier Sense Multiple Access with Collision Avoidance* (CSMA/CA), its performance could be severely degraded if the LAA-LTE system adopts aggressive spectrum sharing strategies [11]–[13]. It is thus crucial to design fair and efficient coexistence mechanisms for the two systems when operating on the same unlicensed band [14]. So far, duty cycle and *listen-before-talk* (LBT) mechanisms have been proposed to support the coexistence between LAA-LTE and WiFi systems. For duty cycle mechanism, such as *carrier sense adaptive transmission* (CSAT) and *almost blank subframe* (ABS) schemes [15], [16], the LAA-LTE system is allowed to occupy the unlicensed bands by proactively conserving partial time of the WiFi transmissions. Though this mechanism requires minor modifications to existing LTE protocols and renders high throughput for LAA-LTE system, it may cause more collisions to the WiFi system due to the lack of *clear channel assessment* (CCA) [17]. On the contrary, supported by CCA, LBT brings more friendliness to the WiFi system. In addition, LBT is compulsory in the regulations for some countries such as Japan, making it indispensable in the global standardization of LAA-LTE [18].

There has been much literature exploring the LBT mechanism to support the LAA-LTE and WiFi coexistence. The achievable throughput for the LAA-LTE system or the overall channel throughput is of concern in [19]–[22]. In [19], the normalized throughput of unlicensed bands is maximized by designing a novel LBT-based MAC protocol, which is further extended in [20] to the scenario of dynamic network settings and solved by a learning-based method. In [21], a cross-layer optimization method is proposed to maximize the overall expected throughput of the unlicensed band. A user offloading method to achieve throughput improvement to both LAA-LTE and WiFi systems is proposed in [22]. Besides throughput, the energy efficiency is also optimized for LAA-LTE/WiFi coexistence systems in [23] and [24]. The aforementioned studies, however, mainly focuses on the overall system performance, without considering the individual *quality-of-service* (QoS) of the users. In [25], the rate requirement of each LAA-LTE user is guaranteed by aggregating licensed and unlicensed bands with a *contention window* (CW) optimization method. In [26], the maximum queue delay of users is guaranteed,

This work was supported in part by the National Natural Science Foundation of China under Grant 61631005, Grant U1801261, Grant 61571100, Grant 61601247, China Postdoctoral Science Foundation under Grant 2018M631075, Tianjin Natural Science Foundation under Grant 17JCQNJC01500, and 111 international collaboration project.

J. Tan and S. Xiao are with the National Key Laboratory on Communications, and the Center for Intelligent Networking and Communications (CINC), University of Electronic Science and Technology of China (UESTC), Chengdu, 611731, China.

S. Han is with the College of Electronic Information and Optical Engineering, Nankai University, Tianjin, 300071, China.

Y.-C. Liang is with the Center for Intelligent Networking and Communications (CINC), University of Electronic Science and Technology of China (UESTC), Chengdu, 611731, China. (Email: liangyc@ieee.org)

V. C. M. Leung is with the Department of Electrical and Computer Engineering, the University of British Columbia, Vancouver, BC, V6T1Z4, Canada (e-mail: vleung@ece.ubc.ca).

but the individual rates are not considered. It is pointed out that, in future wireless networks, many applications involve multiple usage scenarios, including, e.g., *enhanced mobile broadband* (eMBB) and *ultra-reliable low-latency communication* (URLLC). In these cases, multiple QoS requirements should be considered simultaneously. Thus, novel schemes that can precisely guarantee multiple QoS metrics for each user are highly desirable for the LAA-LTE system when coexisting with the WiFi system.

Several challenges remain to be dealt with when designing QoS-aware LAA-LTE/WiFi coexistence mechanisms. First, the protocol of the coexistence framework needs to be well designed, as it determines the flexibility for meeting the QoS requirements. Such flexibility can be achieved by many techniques, including adjustable transmission time and CW in LBT, subcarrier assignment and power allocation in *orthogonal frequency division multiple access* (OFDMA), and time slot and power allocation in *time division multiple access* (TDMA). However, it is challenging to combine them organically into a practical and harmonious coexistence mechanism. Moreover, in order to optimize the LAA-LTE system with protection to the WiFi system, the performance of both systems should be quantified exactly. Nevertheless, the heterogeneity of the two systems makes the existing quantification methods difficult to apply directly. Finally, effective solutions should be developed to optimize the coupling parameters for the efficient and fair coexistence.

In this paper, we propose a QoS-aware scheme for the LAA-LTE/WiFi coexistence system, where the QoS metrics, namely throughput and transmission delay, are of interest. The target of the scheme is to maximize the number of users admitted to the LAA-LTE network with their preferred QoS, while at the same time maintaining a certain level of QoS for the WiFi users. To solve the challenges mentioned before, we first adopt an LBT mechanism with adjustable transmission time and employ the OFDMA technique in the LAA-LTE *base station* (BS). The adopted LBT mechanism belongs to Cat.3 LBT as defined in the standards [18], which has a fixed CW. Thus, LAA-LTE BS does not require extra procedures or signalings to tune the CW. It is worth noting that both CW and transmission time reflect the aggressiveness of the LAA-LTE in sharing the channel, though the principles behind them are different. In addition, the adopted OFDMA technique meets the LTE standards, which makes it more practical. Then, we extend the Bianchi model in [10] and the existing QoS quantification methods to the coexisting LAA-LTE and WiFi systems. With the QoS metrics, the constraints are well designed to take the multiple QoS requirements and fair coexistence into consideration. Next, we formulate the proposed QoS-aware scheme into a joint user association and resource allocation problem. To make the problem tractable, we decompose it into two subproblems: the sum-power minimization problem and the user association problem. For the first subproblem, the LAA-LTE transmission time, subcarrier assignment and power allocation are optimized by the deep-cut ellipsoid method [27] to minimize the required sum-power for a given user association strategy. For the latter one, we develop an efficient algorithm called *successive user removal* (SUR) to exploit the available transmission power to admit as many users as possible to the LAA-LTE network, where their QoS requirements are satisfied.

The contributions of this work are summarized as follows.

- User association, LAA-LTE transmission time, subcarrier assignment and power allocation are jointly optimized when the QoS-aware scheme is designed for LAA-LTE to provide QoS-guaranteed services and to protect the WiFi system. As far as we know, this is the first work on designing the QoS-aware coexistence schemes for LAA-LTE and WiFi systems.
- The throughput and delay of the LAA-LTE/WiFi coexistence system are extensively analyzed and quantified into QoS metrics. With these metrics, the constraints of the QoS guarantees of the LAA-LTE users and the protection to the WiFi users can be formulated.
- Simulation results have demonstrated the effectiveness and efficiency of the proposed scheme, and also revealed the fundamental tradeoff of the QoS metrics in the coexistence system.

Notice that, the proposed novel QoS-aware coexistence scheme is more advantageous than the conventional ones from two aspects. For operators, it can utilize unlicensed bands to provide QoS-guaranteed services similar to licensed LTE services. Therefore, the operators can increase revenue by admitting more users without extra investment for additional licensed spectra. For users, compared with the best-effort services of WiFi systems and the high expense of LTE systems, it can provide them with QoS-guaranteed connections at much lower prices.

The remainder of the paper is organized as follows. In Section II, the system model is described, followed by the QoS analysis in Section III. The problem of the design of the QoS-aware coexistence scheme is formulated in Section IV. To efficiently solve the problem, we decompose it into two subproblems, which are solved in Section V-A and Section V-B, respectively. Then, simulation results are presented in Section VI. Finally, the conclusions are drawn in Section VII.

The main notations in this paper are listed as follows. The boldface lowercase, $\mathbf{a}$, and boldface uppercase letter, $\mathbf{A}$, denote a vector and a matrix, respectively. The calligraphic uppercase letter, $\mathcal{A}$, stands for a set. $|\mathcal{A}|$ denotes the cardinality of the set $\mathcal{A}$. The operators $\cup$ and $\cap$ denote the union and intersection of two sets, respective. The symbols $\emptyset$ and $\mathbb{E}\left[\cdot\right]$ denote the empty set and the expectation operator, respectively.

II. System Model

A. *LAA-LTE/WiFi Coexistence System*

Fig. 1 shows the model of the LAA-LTE/WiFi coexistence system considered in this paper. The LTE BS operates on a licensed band, while the WiFi AP and LAA-LTE BS share the same unlicensed band. We consider the *supplemental downlink* (SDL) deployment of LAA-LTE [28], where the LAA-LTE BS only utilizes the unlicensed band for *downlink* (DL) operation, while the control and *uplink* (UL) signals are transmitted by the LTE BS via licensed bands. Thus, the LAA-LTE BS and LTE BS are connected via optical fiber networks.

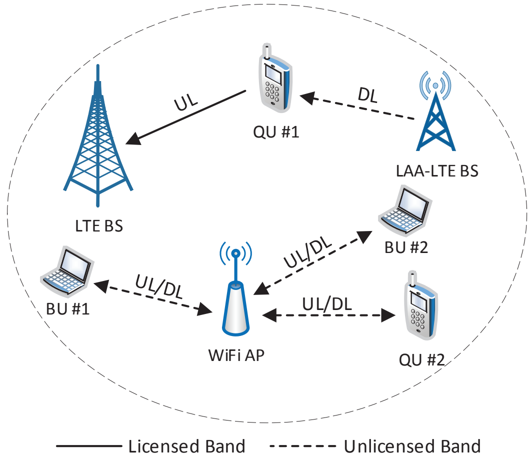


Fig. 1. The LAA-LTE/WiFi coexistence system.

There are two types of users in the coexistence system, i.e., *best-effort users* (BUs) and *QoS-preferred users* (QUs). BUs do not have QoS requirements and only join the WiFi network for best effort services because of the device capability or price-sensitivity. In contrast, QUs have QoS preferences, and they are willing to pay for the QoS guarantees by joining the LAA-LTE network. If the LAA-LTE network cannot admit them, they join the WiFi network for best effort services. Denote by $\mathcal{N}_0$ and $\mathcal{N}$ the sets of BUs and QUs, respectively. Let $\mathcal{N}_1$ and $\mathcal{N}_2$ be the sets of the QUs accessing the WiFi network and the LAA-LTE network, respectively. Note that we have $\mathcal{N}_1 \cup \mathcal{N}_2 = \mathcal{N}$, $\mathcal{N}_1 \cap \mathcal{N}_2 = \emptyset$, $N = |\mathcal{N}|$, and $N_i = |\mathcal{N}_i|, \forall i = 0, 1, 2$. Similar to the standard [18], the LAA-LTE BS adopts a simplified CSMA/CA protocol as its LBT mechanism, with a fixed CW and adjustable transmission time. Within the transmission time, OFDMA is employed to support the LAA-LTE users simultaneously. The design objective for the coexistence system is to maximize the number of the QUs accessing the LAA-LTE network, and to maintain the acceptable level of QoS for the WiFi services.

B. CSMA/CA Protocol

In this part, we analyze the contention among the WiFi stations and LAA-LTE BS in the channel, and derive their stationary transmission and collision probabilities[1] for given system parameters and user association. For WiFi stations, the original CSMA/CA protocol is adopted, which is based on the exponential backoff rules. It works as follows: 1) each competing station randomly chooses an integer from $[0, W_0 - 1]$ (where W_0 is the initial CW of WiFi stations) as the counter and sets the backoff stage, i, as 0; 2) the counter counts down when the channel is sensed idle, otherwise, the counter freezes; 3) transmission will be triggered when the counter becomes zero; 4) when the transmitted packet suffers a collision, retransmission will be triggered, i will be increased by 1 until it reaches the maximum backoff stage, m, and the counter will be randomly chosen from $[0, 2^i W_0 - 1]$. Under the assumption of saturation[2], i.e., the buffer of any station is full, the above protocol can be modelled as a Markov chain. The probability, τ_W, of a WiFi station to transmit is given by [10]

$$\tau_W = \frac{2(1-2p_1)}{(1-2p_1)(W_0+1)+p_1 W_0 (1-(2p_1)^m)}, \quad (1)$$

where p_1 denotes the collision or retransmission probability of a transmitted packet from a WiFi station.

As for the LAA-LTE BS, a simplified CSMA/CA protocol is adopted, which has a fixed CW, W_L, without backoff stage. From [25], we can let m be 0 in (1) and write the transmission probability, τ_L, of the LAA-LTE BS as

$$\tau_L = \frac{2(1-2p_2)}{(1-2p_2)(W_L+1)} = \frac{2}{(W_L+1)}, \quad (2)$$

where p_2 denotes the collision probability of an attempted transmission of the LAA-LTE BS.

According to (2), τ_L can be determined once W_L is chosen, but τ_W, p_1, and p_2 cannot be obtained directly. Since a transmitted WiFi packet or an attempted LAA-LTE transmission encounters a collision if there is any other contender transmitting, their collision probabilities, p_1 and p_2 can be written as the functions of transmission probabilities and the number of contenders. As both UL and DL of the WiFi system operate on the same channel, the WiFi AP is also a contender. The number of competing WiFi stations is N_0+N_1+1. Hence, we obtain

$$p_1 = 1-(1-\tau_L)(1-\tau_W)^{N_0+N_1}, \quad (3)$$
$$p_2 = 1-(1-\tau_W)^{N_0+N_1+1}, \quad (4)$$

similar to [10] and [25]. Terms τ_W and p_1 are two variables to be solved for in the two nonlinear equations (1) and (3). We first rewrite τ_W in (3) as

$$\tau_W(p_1) = 1-(1-\tau_L)^{-1/(N_0+N_1)}(1-p_1)^{1/(N_0+N_1)}, \quad (5)$$

which shows $\tau_W(p_1)$ is a monotonically increasing function of p_1. Meanwhile, [10] has shown that the right-hand side (RHS) of (1) is monotonically decreasing with respect to p_1. Hence, a bisection search can be applied to solve the simultaneous equations (1) and (5), and the unique p_1 and τ_W can be obtained. Then, the corresponding p_2 can also be obtained according to (4). With the help of τ_W, τ_L, p_1 and p_2, the QoS metrics concerned can be derived in the next section.

III. QoS Analysis

To enable QoS-awareness, we derive the QoS metrics of both the pure WiFi system and the LAA-LTE/WiFi coexistence system, where the performance of the pure WiFi system serves as the benchmark to protect the WiFi system in the coexistence scenario. The concerned QoS metrics consist of throughput and delay, of which the analyses are presented in the following two subsections, respectively.

[1] Probability stands for stationary probability in the rest of the paper.

[2] Saturation is also assumed in this paper for the convenience of analysis, but the results can be extended to the unsaturated situation by following [29].

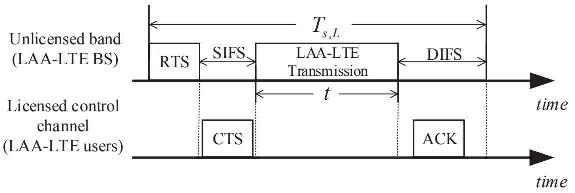


Fig. 2. Channel busy time of a successful transmission of the LAA-LTE BS, $T_{s,L}$.

A. Throughput Analysis

We first review the throughput analysis of the pure WiFi system, where all the BUs and QUs access the WiFi network without the LAA-LTE BS. Thus, the total number of competing WiFi stations is $N_0 + N + 1$. The total throughput of the pure WiFi system is given by [10]

$$R_0 = \frac{P_{tr} P_s \mathbb{E}\left[P\right]}{(1-P_{tr})\theta + P_{tr} P_s T_{s,W} + P_{tr}(1-P_s)T_c}, \quad (6)$$

where $\mathbb{E}\left[P\right]$ is the average packet payload size; P_{tr} and P_s denote the busy probability of the channel and the success probability of a transmission on the channel, respectively; θ denotes the time duration of an empty slot; $T_{s,W}$ represents the average channel busy time for a packet that is transmitted successfully and T_c represents the channel busy time for a collision. According to Eq.(10) and Eq.(11) of [10], P_{tr} and P_s can be written as

$$P_{tr} = 1 - (1-\tau_P)^{N_0+N+1}, \quad (7)$$

$$P_s = \frac{(N_0+N+1)\,\tau_P (1-\tau_P)^{N_0+N}}{P_{tr}}, \quad (8)$$

respectively, where τ_P is the transmission probability of a station in the pure WiFi system. Term τ_P can be obtained by replacing p_1 in (1) with p_3, where p_3 stands for the collision probability of a transmitted packet in the pure WiFi system, and it can be given by

$$p_3 = 1 - (1-\tau_P)^{N_0+N}. \quad (9)$$

Assuming that the mechanism of *request to send* (RTS) and *clear to send* (CTS) is adopted for avoiding the hidden terminal problem, therefore, T_s and T_c are defined as

$$\begin{aligned} T_{s,W} &= \left(RTS + CTS + H + \mathbb{E}\left[P\right] + ACK\right)/C \\ &\quad + 3SIFS + DIFS + 4\delta, \quad (10) \\ T_c &= \left(RTS + DIFS\right)/C, \quad (11) \end{aligned}$$

respectively, where RTS, CTS, ACK and H stand for the bit lengths of the RTS frame, the CTS frame, the acknowledgement (ACK) frame, and the MAC header, respectively; C is the channel bit rate; $SIFS$, $DIFS$, and δ are the time durations of the short interframe space, the distributed interframe space and the propagation delay, respectively. Note that, similar to [10], [13], [17], [23], and [25], C is assumed to be a constant, for which perfect power control is adopted by the WiFi system to compensate the fading effects and to avoid the outage.

Next, we analyze the throughput of the WiFi system and the LAA-LTE system in the coexistence scenario. It is noticed that there is only a pair of transmitter and receiver in each transmission for the WiFi system, and thus the feedback packets, i.e., CTS and ACK, can be transmitted on the unlicensed band without collisions. In contrast, there are multiple receivers, i.e., LAA-LTE users, when the LAA-LTE BS transmits, which causes collisions if they feedback as WiFi stations do. Hence, we propose to let LAA-LTE users transmit CTS and ACK via the licensed control channel with advanced multiple access techniques. As such, the channel busy time, $T_{s,L}$, for a successful transmission of the LAA-LTE BS is different from $T_{s,W}$. Let t be the LAA-LTE transmission time. As Fig. 2 shows, $T_{s,L}$ is given by

$$T_{s,L}\left(t\right) = RTS + \delta + SIFS + t + \delta + DIFS. \quad (12)$$

With (1) and (2), we can obtain P_{tr}^{Co} (the busy probability of the channel), $P_{s,W}$ (the probability that a transmission on the channel belongs to the WiFi stations and succeeds), and $P_{s,L}$ (the probability that a transmission on the channel belongs to the LAA-LTE BS and succeeds) by reformulating (7) and (8) in the coexistence scenario, given by

$$P_{tr}^{Co} = 1 - (1-\tau_L)(1-\tau_W)^{N_0+N_1+1}, \quad (13)$$

$$P_{s,W} = \frac{(N_0+N_1+1)\tau_W(1-\tau_W)^{N_0+N_1}(1-\tau_L)}{P_{tr}^{Co}}, \quad (14)$$

$$P_{s,L} = \frac{\tau_L(1-\tau_W)^{N_0+N_1+1}}{P_{tr}^{Co}}. \quad (15)$$

Therefore, the total WiFi throughput, $R_W(N_1,t)$, and the LAA-LTE efficiency, $f_L(N_1,t)$, can be given by (16) and (17) on the top of next page, respectively, where LAA-LTE efficiency presents the time ratio of successful LAA-LTE transmissions. Within the LAA-LTE transmission time, OFDMA is adopted to support the DL of the QUs in $\mathcal{N}_2$. Assume that QU k is in the LAA-LTE network, i.e., $k \in \mathcal{N}_2$. Denote by $\mathcal{S}_k$ the set of the subcarriers allocated to QU k. Let $h_{k,i}$ be the channel gain from the LAA-LTE BS to QU k, and $q_{k,i}$ the power allocated to QU k, on subcarrier i. The individual rate of QU k can be given by

$$R_k(N_1,t) = f_L(N_1,t)B_0 \sum_{i\in\mathcal{S}_k} r_{k,i}, \quad (18)$$

where

$$r_{k,i} = \log_2\left(1 + \frac{q_{k,i}h_{k,i}}{\sigma^2}\right). \quad (19)$$

In (18) and (19), B_0 and σ^2 denote the bandwidth and the power of *additive white Gaussian noise* (AWGN) of a subcarrier, respectively.

B. Delay Analysis

Similarly, we first go over the delay analysis of the pure WiFi system. For the WiFi system, its delay is defined as the average elapsed time for transmitting a packet successfully since it is put into service. Several methods have been presented in [30]–[32] to evaluate the delay of the pure WiFi system, while [32] increases the accuracy by considering the N-WAIT packets. N-WAIT packets are the packets transmitted without waiting, which follow a successful transmission and

$$R_W(N_1,t)=\frac{P_{tr}^{Co}P_{s,W}\mathbb{E}[P]}{(1-P_{tr}^{Co})\theta+P_{tr}^{Co}P_{s,W}T_{s,W}+P_{tr}^{Co}P_{s,L}T_{s,L}(t)+P_{tr}^{Co}[1-P_{s,W}-P_{s,L}]T_c}, \tag{16}$$

$$f_L(N_1,t)=\frac{P_{tr}^{Co}P_{s,L}t}{(1-P_{tr}^{Co})\theta+P_{tr}^{Co}P_{s,W}T_{s,W}+P_{tr}^{Co}P_{s,L}T_{s,L}(t)+P_{tr}^{Co}[1-P_{s,W}-P_{s,L}]T_c}. \tag{17}$$

$$T_M^{Pure}=(1-\tau_P)^{N_0+N}\theta+(N_0+N)\tau_P(1-\tau_P)^{N_0+N-1}(T_{s,W}+\theta)+\Big[1-(1-\tau_P)^{N_0+N}-(N_0+N)\tau_P(1-\tau_P)^{N_0+N-1}\Big](T_c+\theta). \tag{24}$$

their counters happen to be 0. On the other hand, the packets that need waiting are WAIT packets. For completeness, we summarize the results in [32] as the following lemma.

Lemma 1: Let J be the retransmission limit. The delay, $T_{Delay}^{Pure,J}$, of the pure WiFi system is given by

$$T_{Delay}^{Pure,J}=\frac{T_{Wait}^{Pure,J}+\varphi_W T_{s,W}}{1+\varphi_W}. \tag{20}$$

where $\varphi_W=1/\left(W_0-1\right)$, and $T_{Wait}^{Pure,J}$ is given by

$$T_{Wait}^{Pure,J}=\theta+D_2^{Pure,J}+D_3^{Pure,J}, \tag{21}$$

where

$$D_2^{Pure,J}=\frac{\left[\sum\limits_{j=0}^{J}\left(p_3^j-p_3^{J+1}\right)\left(2^jW_0-1\right)\right]T_M^{Pure}}{2\left(1-p_3^{J+1}\right)}, \tag{22}$$

$$D_3^{Pure,J}=\frac{\sum\limits_{j=0}^{J}\left(1-p_3\right)p_3^j\left(T_{s,W}+jT_c\right)}{1-p_3^{J+1}}, \tag{23}$$

and T_M^{Pure} in (22) is shown on the top of this page as (24).

Proof: The result follows from [32] by modifying the counter range from $[0,2^m]$ to $[0,2^m-1]$ for the backoff stage m, in order to be consistent with the standard [10]. ■

Then, we propose the following **Theorem 1** to consider the asymptotic property of (22) and (23) by letting J approach infinity.

Theorem 1: Under the assumption of infinite retransmission limit, i.e., $J\to\infty$, $D_2^{Pure,J}$ and $D_3^{Pure,J}$ converge respectively to D_2^{Pure} and D_3^{Pure}, as shown below

$$D_2^{Pure}=\frac{1}{2}\left[\sum_{j=0}^{m}p_3^j2^jW_0+\frac{p_3^{m+1}}{1-p_3}2^mW_0-\frac{1}{1-p_3}\right]T_M^{Pure}, \tag{25}$$

$$D_3^{Pure}=T_{s,W}+\frac{p_3}{1-p_3}T_c. \tag{26}$$

Proof: Please refer to Appendix A. ■

With **Theorem 1**, $T_{Wait}^{Pure,J}$ and $T_{Delay}^{Pure,J}$ can be asymptotically expressed as

$$T_{Wait}^{Pure}=\theta+D_2^{Pure}+D_3^{Pure}, \tag{27}$$

$$T_{Delay}^{Pure}=\frac{T_{Wait}^{Pure}+\varphi_W T_{s,W}}{1+\varphi_W}, \tag{28}$$

respectively. When the retransmission limit is large, the asymptotic delay T_{Delay}^{Pure} can be used to replace $T_{Delay}^{Pure,J}$. In the sequel, we use the asymptotic delay to simplify the derivations.

Next, we extend the above results to the LAA-LTE/WiFi coexistence system. The delay of the LAA-LTE system is defined as the average time consumed for a successful transmission after the previous one. We propose the following **Theorem 2** and **Theorem 3** to quantify the delays of the coexisting WiFi and LAA-LTE systems, respectively.

Theorem 2: The delay of the WiFi system in the coexistence system can be asymptotically given by

$$T_{Delay}^{W}(N_1,t)=\frac{T_{Wait}^{W}+\varphi_W T_{s,W}}{1+\varphi_W}, \tag{29}$$

where

$$T_{Wait}^{W}=\theta+D_2^{W}+D_3^{W}. \tag{30}$$

In (30), D_2^W can be obtained from D_2^{Pure} in (25) if we replace p_3 with p_1 and T_M with T_M^W, and D_3^W can be obtained from D_3^{Pure} in (26) if we replace p_3 with p_1. Term T_M^W is given by (31) on the top of next page.

Proof: Please refer to Appendix B. ■

Theorem 3: The delay of the LAA-LTE system in the coexistence system can be asymptotically given by

$$T_{Delay}^{L}(N_1,t)=\frac{T_{Wait}^{L}+\varphi_L T_{s,L}(t)}{1+\varphi_L}, \tag{32}$$

where

$$T_{Wait}^{L}=\theta+D_2^{L}+D_3^{L}, \tag{33}$$

and $\varphi_L=1/\left(W_L-1\right)$. Terms D_2^L and D_3^L in (33) are written as

$$D_2^L=\frac{1}{2}\left(W_L-1\right)T_M^L, \tag{34}$$

$$D_3^L=T_{s,L}(t)+\frac{p_2}{1-p_2}T_c, \tag{35}$$

respectively, and T_M^L is given by (36) on the top of next page.

Proof: Please refer to Appendix C. ■

IV. Problem Formulation

In this section, the problem of designing the QoS-aware coexistence scheme is formulated, which aims to maximize the number of the QUs admitted to the LAA-LTE network while preserving a certain level of QoS for the WiFi system. The following constraints are imposed.

Constraint 1 (Throughput guarantee for QUs): For fulfilling the QoS guarantees, the QUs in the LAA-LTE network should be provided with their desired rates. Let $\bar{R}_k$ denote the rate requirement of QU k. The throughput guarantee for the QUs can be expressed as

$$R_k(N_1,t)\geq\bar{R}_k,\forall k\in\mathcal{N}_2. \tag{37}$$

$$T_M^W = (1-\tau_W)^{N_0+N_1}(1-\tau_L)\,\theta + (N_0+N_1)\tau_W(1-\tau_W)^{N_0+N_1-1}(1-\tau_L)(T_{s,W}+\theta) + \tau_L(1-\tau_W)^{N_0+N_1}(T_{s,L}(t)+\theta) + \left[1-(1-\tau_W)^{N_0+N_1}(1-\tau_L) - (N_0+N_1)\tau_W(1-\tau_W)^{N_0+N_1-1}(1-\tau_L) - \tau_L(1-\tau_W)^{N_0+N_1}\right](T_c+\theta)\,. \tag{31}$$

$$T_M^L = (1-\tau_W)^{N_0+N_1+1}\theta + (N_0+N_1+1)\,\tau_W\,(1-\tau_W)^{N_0+N_1}(T_{s,W}+\theta) + \left[1-(1-\tau_W)^{N_0+N_1} - (N_0+N_1+1)\,\tau_W(1-\tau_W)^{N_0+N_1}\right](T_c+\theta)\,. \tag{36}$$

Constraint 2 (Throughput protection for WiFi users): For fairness, the individual throughput of the pure WiFi system should be guaranteed for the WiFi system when coexisting with the LAA-LTE system. Thus, we require

$$\frac{R_W(N_1,t)}{N_0+N_1} \geq \frac{R_0}{N_0+N}. \tag{38}$$

Constraint 3 (Delay guarantee for QUs): To provide similar QoS of licensed LTE systems, the delay of the QUs in the LAA-LTE network should also be guaranteed. Denote by T_{Max_Delay} the maximum tolerable delay of the LAA-LTE system. This constraint can be written as

$$T^L_{Delay}(N_1,t) \leq T_{Max\_Delay}. \tag{39}$$

Constraint 4 (Delay protection for WiFi users): The delay performance of the pure WiFi system should be guaranteed for the WiFi users when coexisting with the LAA-LTE system. This leads to the constraint

$$T^W_{Delay}(N_1,t) \leq T^{Pure}_{Delay}. \tag{40}$$

With the above constraints, the QoS-aware user association and resource allocation problem can be mathematically formulated as

Problem 1:

$$\max_{t,\{q_{k,i}\},\{\mathrm{I}_k\}} \sum_{k\in\mathcal{N}} \mathrm{I}_k$$

$$\text{s.t.} \quad R_k\Big(\sum\nolimits_{k\in\mathcal{N}} \bar{\mathrm{I}}_k, t\Big) \geq \bar{R}_k, \forall k\in\mathcal{N} \text{ and } \mathrm{I}_k=1, \tag{P1.C1}$$

$$\frac{R_W\big(\sum_{k\in\mathcal{N}} \bar{\mathrm{I}}_k, t\big)}{N_0+\sum_{k\in\mathcal{N}} \bar{\mathrm{I}}_k} \geq \frac{R_0}{N_0+N}, \tag{P1.C2}$$

$$T^L_{Delay}\Big(\sum\nolimits_{k\in\mathcal{N}} \bar{\mathrm{I}}_k, t\Big) \leq T_{Max\_Delay}, \tag{P1.C3}$$

$$T^W_{Delay}\Big(\sum\nolimits_{k\in\mathcal{N}} \bar{\mathrm{I}}_k, t\Big) \leq T^{Pure}_{Delay}, \tag{P1.C4}$$

$$\mathrm{I}_k \in \{0,1\}, \forall k \in \mathcal{N}, \tag{P1.C5}$$

$$\mathcal{S}_k = \emptyset, \forall k \in \mathcal{N} \text{ and } \mathrm{I}_k = 0, \tag{P1.C6}$$

$$\bigcup_{k\in\mathcal{N}} \mathcal{S}_k \subseteq \mathcal{S}, \tag{P1.C7}$$

$$\bigcap_{k\in\mathcal{N}} \mathcal{S}_k = \emptyset, \tag{P1.C8}$$

$$q_{k,i} \geq 0, \forall k \in \mathcal{N} \text{ and } i \in \mathcal{S}, \tag{P1.C9}$$

$$\sum_{k\in\mathcal{N}} \sum_{i\in S_k} q_{k,i} \leq \bar{Q}. \tag{P1.C10}$$

In **Problem 1**, I_k indicates whether QU k is in the LAA-LTE network. If yes, $\mathrm{I}_k = 1$; otherwise, $\mathrm{I}_k = 0$. Moreover, we define $\bar{\mathrm{I}}_k = 1-\mathrm{I}_k$. The objective of the problem is to maximize the number of the QUs served by the LAA-LTE network with QoS guarantees, by optimizing the user association of the QUs, LAA-LTE transmission time, subcarrier assignment, and power allocation. (P1.C1) - (P1.C4) reflect the constraints defined in (37) - (40), while (P1.C6) - (P1.C8) are the exclusive subcarrier assignment constraints of the OFDMA system. $\mathcal{S}$ is the set of subcarriers, and $|\mathcal{S}| = S$. (P1.C9) indicates non-negative power allocation. (P1.C10) restricts the total transmission power of the LAA-LTE BS, due to regulations or physical limitations.

Problem 1 is a combinatorial and nonconvex problem, due to the user association, the subcarrier assignment, and the nonlinearity brought by the CSMA/CA mechanism. To solve this problem efficiently, we propose a QoS-aware joint user allocation and resource allocation algorithm in the next section.

V. QoS-aware Joint User Association and Resource Allocation

By investigating **Problem 1**, we can observe that except (P1.C10), all constraints can always be satisfied for any $\{\mathrm{I}_k\}$ with $t > 0$ and $|\mathcal{S}_k| \geq 1$, $\forall k \in \mathcal{N}_2$ by increasing the transmission power. Therefore, the maximum transmission power, $\bar{Q}$, plays a key role in the user association. According to this feature, we first decompose **Problem 1** into the sum-power minimization subproblem and the user association subproblem. Then, we develop the QoS-aware joint user allocation and resource allocation algorithm by solving the two subproblems.

A. Sum-Power Minimization

We first consider the sum-power minimization problem under a given user association strategy, $\{\mathrm{I}_k\}$. Our target is to minimize the sum-power while maintaining the constraints in **Problem 1** except for (P1.C5) and (P1.C10). The problem can be written as

Problem 2:

$$\min_{t,\{q_{k,i}\}} \sum_{k\in\mathcal{N}_2} \sum_{i\in\mathcal{S}_k} q_{k,i}$$

$$\text{s.t.} \quad f_L(N_1,t)B_0 \sum_{i\in\mathcal{S}_k} r_{k,i} \geq \bar{R}_k, \forall k \in \mathcal{N}_2, \tag{P2.C1}$$

$$\frac{R_W(N_1,t)}{N_0+N_1} \geq \frac{R_0}{N_0+N}, \tag{P2.C2}$$

$$T^L_{Delay}(N_1,t) \leq T_{Max\_Delay}, \tag{P2.C3}$$

$$T^W_{Delay}(N_1,t) \leq T^{Pure}_{Delay}, \tag{P2.C4}$$

$$\mathcal{S}_k = \emptyset, \forall k \in \mathcal{N}_2, \tag{P2.C5}$$

$$\bigcup_{k\in\mathcal{N}_2} \mathcal{S}_k \subseteq \mathcal{S}, \tag{P2.C6}$$

$$\bigcap_{k\in\mathcal{N}_2} \mathcal{S}_k = \emptyset, \tag{P2.C7}$$

$$q_{k,i} \geq 0, \forall k \in \mathcal{N}_2 \text{ and } i \in \mathcal{S}. \tag{P2.C8}$$

Since $\mathcal{N}_1$, $\mathcal{N}_2$, N_1, and N_2 are fixed for the given $\{\mathrm{I}_k\}$, the RHS of (P2.C1) - (P2.C4) are all determined. However, the LAA-LTE transmission time, subcarrier assignment, and power allocation are coupled, making **Problem 2** still intractable. To solve the problem, we first investigate the properties of the functions in (P2.C1) - (P2.C4).

Theorem 4: With a fixed N_1, $f_L(N_1, t)$ is a monotonically increasing function of t.

Proof: Please refer to Appendix D. ■

Theorem 5: With a fixed N_1, both $T^L_{Delay}(N_1, t)$ and $T^W_{Delay}(N_1, t)$ are monotonically increasing functions of t, while $R_W(N_1, t)$ is monotonically decreasing with respect to t.

Proof: With a given N_1, we can easily show from (29) that $\frac{\partial T^L_{Delay}(N_1,t)}{\partial t} > 0$. Thus, the monotonicity of $T^L_{Delay}(N_1, t)$ with respect to t is proved. Similarly, from (32) and (16), we have $\frac{\partial T^W_{Delay}(N_1,t)}{\partial t} > 0$, $\frac{\partial R_W(N_1,t)}{\partial t} < 0, \forall t$. The theorem is proved. ■

By rewriting (P2.C1) as

$$\sum_{i\in\mathcal{S}_k} r_{k,i} \geq \frac{\bar{R}_k}{B_0 f_L(N_1, t)}, \forall k \in \mathcal{N}_2, \tag{P2.C1T}$$

we can conclude that the required sum-power can be reduced if the RHS of (P2.C1T) is decreased. In other words, the minimum required sum-power can be reached with the maximum $f_L(N_1, t)$. According to **Theorem 4**, $f_L(N_1, t)$ can be maximized with the largest feasible t, which is denoted by t^*. Then, according to **Theorem 5**, (P2.C2) - (P2.C4) can be transformed into

$$t \leq t_1(N_1), \tag{P2.C2T}$$

$$t \leq t_2(N_1), \tag{P2.C3T}$$

$$t \leq t_3(N_1), \tag{P2.C4T}$$

respectively, where $t_1(N_1)$, $t_2(N_1)$, and $t_3(N_1)$ are attained when (P2.C2) - (P2.C4) are respectively satisfied with equality. Hence, to simultaneously satisfy (P2.C2T) - (P2.C4T), the maximum value of t, i.e., $t^*(N_1)$, is given by $\min\{t_1(N_1), t_2(N_1), t_3(N_1)\}$. Note that $t^*(N_1)$ may be negative and then leads to an invalid solution. A proper LAA-LTE CW can avoid this problem, which will be discussed in Section VI. Thus, **Problem 2** becomes

Problem 3:

$$\min_{\{q_{k,i}\}} \sum_{k\in\mathcal{N}_2} \sum_{i\in\mathcal{S}_k} q_{k,i}$$

$$\text{s.t.} \quad \sum_{i\in\mathcal{S}_k} r_{k,i} \geq \frac{\bar{R}_k}{B_0 f_L(N_1, t^*(N_1))}, \forall k \in \mathcal{N}_2, \tag{P3.C1}$$

(P2.C5), (P2.C6), (P2.C7), (P2.C8)

Though **Problem 3** is still nonconvex, it can be solved by the dual decomposition method without duality gap [33], [34]. We define

$$\bar{r}_k = \frac{\bar{R}_k}{B_0 f_L(N_1, t^*(N_1))}. \tag{41}$$

With relaxation variables $\{\lambda_k\}, \forall k \in \mathcal{N}_2$, the Lagrangian of **Problem 3** is formulated as

$$\mathcal{L}(\{q_{k,i}\}, \{r_{k,i}\}, \{\lambda_k\}) = \sum_{k\in\mathcal{N}_2} \sum_{i=1}^{S} q_{k,i} - \sum_{k\in\mathcal{N}_2} \lambda_k \Big(\sum_{i=1}^{S} r_{k,i} - \bar{r}_k\Big), \tag{42}$$

where $q_{k,i}$ is positive only if subcarrier i is allocated to QU k, otherwise, it equals zero. Because of the exclusive assignment, no more than one QU can get positive power on each subcarrier. We define $\boldsymbol{\lambda} = \{\lambda_k\}, \forall k \in \mathcal{N}_2$. The Lagrangian dual function can be written as follows.

$$\begin{aligned} \mathcal{G}(\boldsymbol{\lambda}) &= \min_{\{q_{k,i}\},\{r_{k,i}\}} \mathcal{L}(\{q_{k,i}\}, \{r_{k,i}\}, \boldsymbol{\lambda}) \\ &= \min_{\{q_{k,i}\},\{r_{k,i}\}} \sum_{k\in\mathcal{N}_2} \sum_{i=1}^{S} (q_{k,i} - \lambda_k r_{k,i}) + \sum_{k\in\mathcal{N}_2} \lambda_k \bar{r}_k. \end{aligned} \tag{43}$$

Assuming QU k is allocated with subcarrier i, $\mathcal{L}(\{q_{k,i}\}, \{r_{k,i}\}, \{\lambda_k\})$ in (42) can be minimized by taking the derivative with respect to $q_{k,i}$ and setting it to be zero, i.e.,

$$\frac{\partial \mathcal{L}(\{q_{k,i}\}, \{r_{k,i}\}, \boldsymbol{\lambda})}{\partial q_{k,i}} = 0, \tag{44}$$

yielding

$$q_{k,i} = \left[\frac{\lambda_k}{\ln 2} - \frac{\sigma^2}{h_{k,i}}\right]^+. \tag{45}$$

We can observe that (43) can be decomposed into

$$\mathcal{G}(\boldsymbol{\lambda}) = \sum_{i=1}^{S} \mathcal{G}'_i(\lambda_k) + \sum_{k\in\mathcal{N}_2} \lambda_k \bar{r}_k, \tag{46}$$

in which

$$\mathcal{G}'_i(\lambda_k) = \min_{\{q_{k,i}\},\{r_{k,i}\}} \sum_{k\in\mathcal{N}_2} q_{k,i} - \lambda_k r_{k,i}. \tag{47}$$

Therefore, by using (45), $\mathcal{G}(\boldsymbol{\lambda})$ can be obtained by assigning subcarrier i to the QU that minimizes the RHS of (47). The primal problem is now transformed to the dual problem:

$$\begin{aligned} \max_{\boldsymbol{\lambda}} \quad & \mathcal{G}(\boldsymbol{\lambda}) \\ \text{s.t.} \quad & \boldsymbol{\lambda} > 0, \end{aligned}$$

in which the subgradients are chosen to be

$$d_k = \bar{r}_k - \sum_{i=1}^{S} r_{k,i}, \forall k \in \mathcal{N}_2. \tag{48}$$

Both ellipsoid and subgradient methods can be applied to iteratively solve the dual problem and obtain the optimal relaxation variables, $\boldsymbol{\lambda}^*$. The ellipsoid method, specifically the deep-cut ellipsoid method, is adopted here, for its stability and fast convergence [27], [35]. The basic idea is to iteratively produce a sequence of smaller ellipsoids in volume from an initial ellipsoid $\mathcal{E}^{(0)}$ containing $\boldsymbol{\lambda}^*$. Each new ellipsoid $\mathcal{E}^{(u)}$ is generated by keeping half of the previous ellipsoid $\mathcal{E}^{(u-1)}$ that contains $\boldsymbol{\lambda}^*$. Since an ellipsoid $\mathcal{E}$ can be described as $\mathcal{E} = \{\mathbf{z}|(\mathbf{z}-\mathbf{x})^{\mathbf{T}}\mathbf{A}^{-1}(\mathbf{x}-\mathbf{z}) \leq 1\}$, a cubic region can be chosen as $\mathcal{E}^{(0)}$, and the corresponding parameters $\mathbf{A}^{(\mathbf{0})}, \mathbf{z}^{(\mathbf{0})}$ are given by

$$\mathbf{A}^{(\mathbf{0})} = N_2 \text{diag}\left[(\frac{\lambda_{k_1,\max}}{2})^2, ..., (\frac{\lambda_{k_{N_2},\max}}{2})^2 \right], \tag{49}$$

$$\mathbf{z}^{(\mathbf{0})} = \left[\frac{\lambda_{k_1,\max}}{2}, ..., \frac{\lambda_{k_{N_2},\max}}{2} \right]^T, \tag{50}$$

where k_1 and k_{N_2} represent the smallest and largest indexes of the QUs in $\mathcal{N}_2$ respectively, and $\lambda_{k_j,\max}$ denotes the upper bound of the optimal relaxation variable for the QU with the j-th largest index in $\mathcal{N}_2$. The following proposition proposes one of the upper bounds of $\boldsymbol{\lambda}^*$.

Proposition: The optimal relaxation variables $\boldsymbol{\lambda}^*$ satisfy

$$\lambda_k^* \leq \ln 2 \left[Q_{\max} + \frac{\sigma^2}{\min\limits_{i \in \mathcal{S}}\{\, h_{k,i}\}} \right], \forall k \in \mathcal{N}_2, \tag{51}$$

where $Q_{\max}$ is one of the upper bounds of the optimal sum-power, which can be obtained by the following steps:

1) Assign the subcarriers to the QUs in $\mathcal{N}_2$ uniformly.
2) Perform bisection search over λ_k in (45) to equalize (P3.C1), for each QU in $\mathcal{N}_2$.
3) $Q_{\max}$ is the summation of the power allocated on all subcarriers obtained in last step.

Proof: According to KKT conditons, $\boldsymbol{\lambda}^*$ must satisfy (44), which leads to

$$\lambda_k^* = \ln 2 \left[q_{k,i}^* + \frac{\sigma^2}{h_{k,i}} \right], \tag{52}$$

where $q_{k,i}^*$ is the optimal power allocated to QU k on subcarrier i, under the optimal subcarrier allocation. Obviously, $q_{k,i}^*$ is less than the sum-power, $Q_{\max}$, of any suboptimal power and subcarrier allocation. Term $Q_{\max}$ can be obtained by water-filling over any subcarrier allocation to meet (P3.C1). In details, it can be summarized as the steps above. Since $1/h_{k,i} \leq 1/\min\limits_{i \in \mathcal{S}}\{\, h_{k,i}\}$, (51) is thus proved. ■

To improve the convergence speed of the basic ellipsoid method and to reduce the fluctuations during the iterations, the idea introduced into the deep-cut ellipsoid method is to further cut down the ellipsoids by excluding the part of the region where the objective function is not monotonic.

Denote $\boldsymbol{\lambda}^{(u)}$ and $d_k^{(u)}$ as the values of $\boldsymbol{\lambda}$ and d_k obtained in the u-th iteration, respectively. The objective function in the u-th iteration and the maximum objective value over the past u iterations are correspondingly denoted by $\mathcal{G}^{(u)}(\boldsymbol{\lambda}^{(u)})$ and $\mathcal{G}_{\max}^{(u)}$. We define $\mathbf{d}^{(u)} = [d_{k_1}^{(u)}, ..., d_{k_{N_2}}^{(u)}]^T$. The update procedure is described as:

$$1) \quad \alpha = \frac{\mathcal{G}_{\max}^{(u)} - \mathcal{G}^{(u)}(\boldsymbol{\lambda}^{(u)})}{\sqrt{\mathbf{d}^{(u)T}\mathbf{A}^{(u)}\mathbf{d}^{(u)}}}, \tag{53}$$

$$2) \quad \tilde{\mathbf{d}}^{(u)} = \frac{\mathbf{d}^{(u)}}{\sqrt{\mathbf{d}^{(u)T}\mathbf{A}^{(u)}\mathbf{d}^{(u)}}}, \tag{54}$$

$$3) \quad \mathbf{z}^{(u+1)} = \mathbf{z}^{(u)} + \frac{1 + N_2\alpha}{N_2 + 1}\mathbf{A}^{(u)}\tilde{\mathbf{d}}^{(u)}, \tag{55}$$

Algorithm 1 Sum-power minimization algorithm

```
Obtain $N_1, N_2, \mathcal{N}_1$, and $\mathcal{N}_2$ according to $\{\mathrm{I}_k\}$.
Obtain $t^*$, $f_L(N_1, t^*(N_1))$, and $\{\bar{r}_k\}$ according to (P2.C2T) - (P2.C4T), (17), and (41), respectively.
Obtain $\mathbf{A}^{(\mathbf{0})}$ and $\mathbf{z}^{(\mathbf{0})}$ according to (49) and (50), respectively.
Initialize $u = 0$, $\mathcal{G}_{\max}^{(0)} = -\infty$.
repeat
    Obtain $\mathcal{G}^{(u)}(\mathbf{z}^{(u)})$ according to (43), where $\{q_{k,i}\}$ and $\{r_{k,i}\}$ are obtained by (45) and (19).
    if $\mathcal{G}_{\max}^{(u)} \leq \mathcal{G}^{(u)}(\mathbf{z}^{(u)})$ then
        $\mathcal{G}_{\max}^{(u)} = \mathcal{G}^{(u)}(\mathbf{z}^{(u)})$.
    end if
    Obtain the subgradient $\mathbf{d}^{(u)}$ according to (48).
    Obtain $\alpha$, $\tilde{\mathbf{d}}^{(u)}$, $\mathbf{z}^{(u+1)}$, and $\mathbf{A}^{(u+1)}$ according to (53) - (56), respectively.
    $u = u + 1$.
until $\left|\mathbf{d}^{(u+1)}\right| \leq \varepsilon$
Obtain $\{q_{k,i}^*\}$ according to (45) with $\boldsymbol{\lambda} = \mathbf{z}^{(u)}$.
return $\{q_{k,i}^*\}$
```

$$4) \quad \mathbf{A}^{(u+1)} = \frac{N_2(1-\alpha^2)}{N_2^2 - 1} \times \left(\mathbf{A}^{(u)} - \frac{2(1+N_2\alpha)}{(N_2+1)(1+\alpha)} \mathbf{A}^{(u)}\tilde{\mathbf{d}}^{(u)}\tilde{\mathbf{d}}^{(u)T}\mathbf{A}^{(u)} \right). \tag{56}$$

From the above, the solution to the sum-power minimization problem for a given user association strategy $\{\mathrm{I}_k\}$ can be summarized as **Algorithm 1**, where ε is a small constant. Due to the matrix multiplication in (56), the computational complexity of **Algorithm 1** is $\mathcal{O}(VN_2^6)$, where V is the number of iterations that **Algorithm 1** needs to converge. When the algorithm converges, the power allocation $\{q_{k,i}^*\}$ that achieves minimum required sum-power are returned. With $\{q_{k,i}^*\}$, the minimum required individual sum-power, $\sum_{i \in \mathcal{S}} q_{k,i}^*, \forall k \in \mathcal{N}_2$, can be calculated, which helps to further optimize $\{\mathrm{I}_k\}$. For brevity, the term "minimum required" is omitted in the following.

B. User Association

With **Algorithm 1**, **Problem 1** becomes the problem of finding the user association strategy to maximize N_2 while keeping the sum-power within the maximum transmission power $\bar{Q}$. Particularly, for a special case, if $\bar{Q}$ is larger or equal to the sum-power obtained by **Algorithm 1** with $\mathrm{I}_k = 1, \forall k \in \mathcal{N}$, all QUs can be admitted to the LAA-LTE network. Otherwise, exhaustive search is needed to find the optimal solution from all possible $\{\mathrm{I}_k\}$, which requires to run **Algorithm 1** for 2^N times in the worst case. To reduce the complexity, we develop a suboptimal user association method, which is inspired by an observation of (P2.C2).

Observation: With (P2.C2) rewritten as $R_W(N_1, t) \geq \frac{N_0+N_1}{N_0+N} R_0$, it can be seen that $t_1(N_1)$ decreases as N_1 grows, making $\{\bar{r}_k\}$ non-decreasing, and thus more power is needed to meet the rate requirement. However, the subcarriers released by moving more QUs to the WiFi network have the converse impact.

Algorithm 2 QoS-aware joint user association and resource allocation algorithm

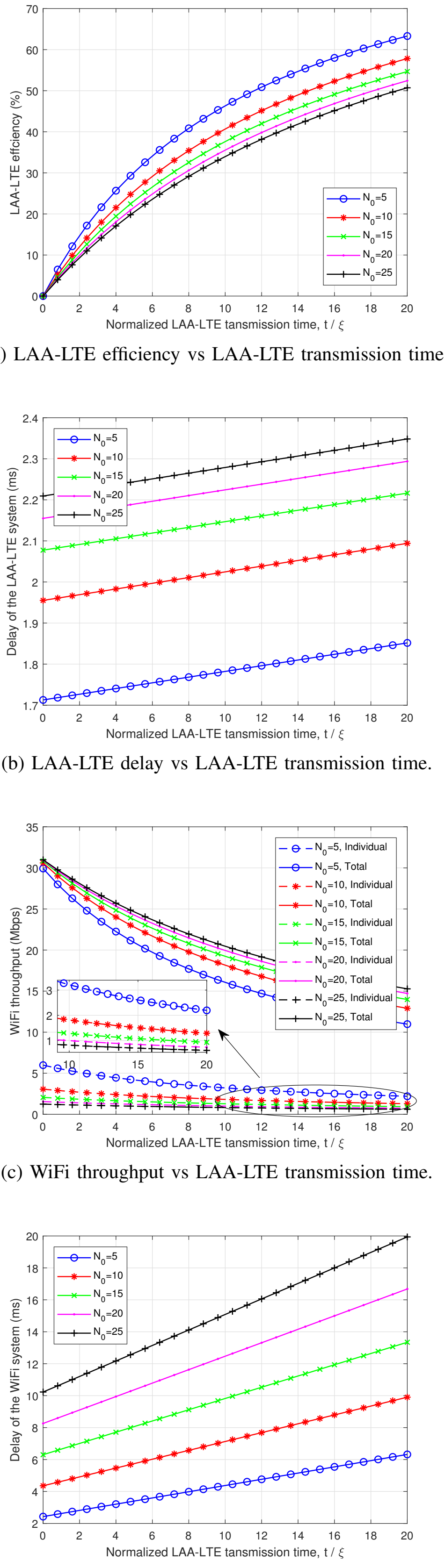
```
Initialize $N_2 = N$, $\mathcal{N}_2 = \mathcal{N}$, and $\mathrm{I}_k = 1, \forall k \in \mathcal{N}$.
repeat
    Run Algorithm 1 for the given $\{\mathrm{I}_k\}$, and obtain $\sum_{i \in \mathcal{S}} q^*_{k,i}, \forall k \in \mathcal{N}_2$ and $\sum_{k \in \mathcal{N}_2} \sum_{i \in \mathcal{S}} q^*_{k,i}$.
    if $\sum_{k \in \mathcal{N}_2} \sum_{i \in \mathcal{S}} q^*_{k,i} > \bar{Q}$ then
        Sort $\sum_{i \in \mathcal{S}} q^*_{k,i}, \forall k \in \mathcal{N}_2$ in ascending order, select the first $N_2 - 1$ user index as the new $\mathcal{N}_2$.
        Set $N_2 = N_2 - 1$ and $\mathrm{I}_k = 0, \forall k \notin \mathcal{N}_2$.
    else
        Break the loop;
    end if
until $N_2 = 0$
return $\{\mathrm{I}_k\}$ and $\{q^*_{k,i}\}$
```

From the observation, the optimal N_1 and N_2 cannot be obtained directly. Instead of searching $\mathcal{N}_1$ and $\mathcal{N}_2$ exhaustively, we propose to enumerate the possible N_1 and N_2 at much lower cost. At the beginning, all the QUs are assumed to access the LAA-LTE network, i.e., $\mathcal{N}_2 = \mathcal{N}$. Then, the QU with the largest individual power is removed from $\mathcal{N}_2$ if the sum-power exceeds $\bar{Q}$. By repeating the previous step, the maximum N_2 that satisfies the power constraint can be found. The proposed method obtains the maximum N_2 and the corresponding $\mathcal{N}_2$ by successively removing the QUs from the LAA-LTE network, and therefore we call it the *successive user removal* (SUR) algorithm. The details of the SUR algorithm are included in the proposed QoS-aware joint user association and resource allocation algorithm, which is shown in **Algorithm 2**.

Notice that **Algorithm 2** cannot guarantee a global optimal solution, because for different N_2, the subcarriers released by removing users from the LAA-LTE network may change the order of the individual powers of the QUs. However, the numerical results indicate that the SUR algorithm approaches the performance of the exhaustive search algorithm. Furthermore, compared with the exhaustive search, the SUR algorithm is much more efficient, since it only requires to run **Algorithm 1** for N times in the worst case, instead of 2^N. Hence, the overall computational complexity of the proposed joint user association and resource allocation algorithm is $\mathcal{O}(VN^7)$.

VI. Simulation Results

In this section, the proposed QoS-aware coexistence scheme is evaluated by numerical simulations. First, we focus on the CSMA/CA protocol performance to reveal the tradeoff among the QoS metrics, without considering the particular channels of the QUs. Then, by taking the particular channels into consideration, we examine the performance of the proposed algorithms to demonstrate their effectiveness and efficiency. The common system parameters used in the simulations are listed in Table I.



(a) LAA-LTE efficiency vs LAA-LTE transmission time.

(b) LAA-LTE delay vs LAA-LTE transmission time.

(c) WiFi throughput vs LAA-LTE transmission time.

(d) WiFi delay vs LAA-LTE transmission time.

Fig. 3. QoS metrics vs LAA-LTE transmission time, for the number of BUs $N_0 = 5, 10, 15, 20,$ and 25.

TABLE I
SYSTEM PARAMETERS

Parameters	Value
WiFi packet payload, $\mathbb{E}[P]$	12000 bits
MAC header, H	192 bits
PHY header	224 bits
Bit length of ACK frame, ACK	112 bits + PHY header
Bit length of CTS frame, CTS	112 bits + PHY header
Bit length of RTS frame, RTS	160 bits + PHY header
WiFi channel bit rate, C	54 Mbps
Prorogation Delay, δ	1 μs
Slot time, θ	20 μs
$DIFS$	34 μs
$SIFS$	16 μs
WiFi initial CW, W_0	16
WiFi maximum backoff stage, m	6
Unlicensed bandwidth	20 MHz
Number of subcarriers of LAA-LTE, S	512
AWGN power on each subcarrier, σ^2	-174 dBm/Hz * 20 MHz / 512
Path loss model (dB)	$-15.3 - 50\log_{10}(d(m))$

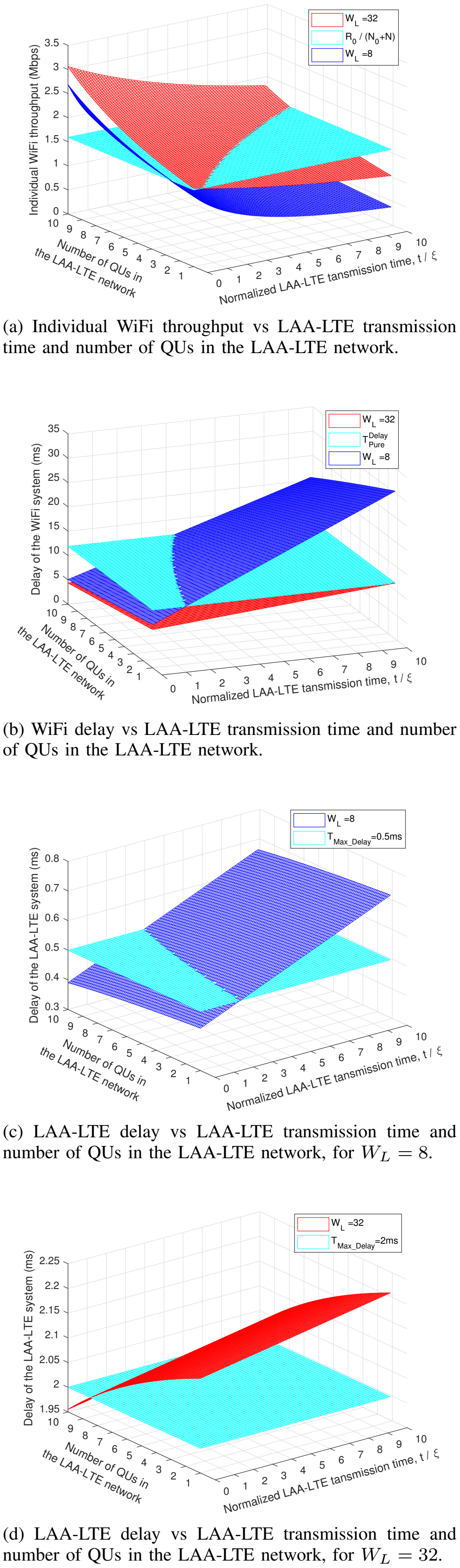
A. Effects of the LAA-LTE Transmission Time and the Number of BUs

In Fig. 3, different QoS metrics are evaluated with the varying LAA-LTE transmission time, t, for different numbers of BUs, N_0, where t is normalized by the WiFi payload transmission time, $\xi = \mathbb{E}(P)/C$. The number of QUs, N, and that of the QUs in the LAA-LTE network, N_2, are both set to be 10, i.e., $N = N_2 = 10$. Besides, we let the LAA-LTE CW, W_L, be 32. As Fig. 3a shows, for any fixed N_0, the LAA-LTE efficiency increases as t increases, resulting in more effective transmissions and less required transmission power for the LAA-LTE system. However, the throughput and delay of the WiFi system, and the delay of the LAA-LTE system are degraded as t increases, as depicted in Fig. 3b - 3d.

In addition, Fig. 3 demonstrates that the growth of N_0 causes entire performance loss to both networks, except the total throughput of the WiFi system shown in Fig. 3c. This exception also contradicts with the results of the pure WiFi system [10]. Intuitively, since the whole WiFi system contends with the LAA-LTE BS in the coexistence scenario, more WiFi stations can increase the chance of the WiFi system to transmit, and thus higher total throughput is achieved. Nevertheless, the individual throughput of concern is still decreased as N_0 increases, as Fig. 3c shows. It is implied that transferring WiFi users to the LAA-LTE network is beneficial to both networks.

B. Effects of the LAA-LTE Contention Window

As mentioned in Section V-A, the constraints (P2.C2) - (P2.C4) may result in negative t and make **Problem 1** infeasible, which can be solved by choosing an appropriate LAA-LTE CW, W_L. To describe the impacts of W_L on the feasibility of t, we depict the left-hand-sides (LHS), i.e., the QoS metrics that are constrained, and RHS, i.e., the target values, of (P2.C2) - (P2.C4) as the change of N_2 and t in Fig. 4, for $W_L = 8$ and $W_L = 32$. Besides, we choose $N_0 = 10$ and $N = 10$. Note that N_2 are integers in practice, but we use decimal values in Fig. 4 for smoother display. In Fig. 4a - Fig. 4d, the blue and red planes represent the QoS metrics



(a) Individual WiFi throughput vs LAA-LTE transmission time and number of QUs in the LAA-LTE network.

(b) WiFi delay vs LAA-LTE transmission time and number of QUs in the LAA-LTE network.

(c) LAA-LTE delay vs LAA-LTE transmission time and number of QUs in the LAA-LTE network, for $W_L = 8$.

(d) LAA-LTE delay vs LAA-LTE transmission time and number of QUs in the LAA-LTE network, for $W_L = 32$.

Fig. 4. QoS metrics vs LAA-LTE transmission time and number of QUs in the LAA-LTE network, for different LAA-LTE CW.

obtained with $W_L = 8$ and $W_L = 32$, respectively, while the cyan planes stand for the target values.

Fig. 4a depicts the variations of the individual WiFi throughput, which is constrained by (P2.C2). Hence, the RHS of (P2.C2), $\frac{R_0}{N_0+N}$, is the cyan plane. To satisfy (P2.C2), the individual WiFi throughput of the red or blue plane should be higher or equal to that of the cyan plane. From the figure, for $W_L = 8$, there does not exist positive t to satisfy the constraint if $N_2 < 3$, that is, **Problem 1** is infeasible. In contrast, for $W_L = 32$, positive t always exists for any N_2.

Fig. 4b describes the changes of the WiFi delay, which should not be larger than T_{Max_Delay} according to (P2.C3). Thus, T_{Max_Delay} is the benchmark cyan plane. Accordingly, the WiFi delay of the red or blue plane should be lower or equal to that of the cyan plane. From the cross points of the planes, there always exist positive t for both $W_L = 8$ and $W_L = 32$. However, the maximum t achieved with $W_L = 8$ is much smaller than that with $W_L = 32$ for any N_2.

Different from (P2.C2) and (P2.C3), the RHS of (P2.C4), i.e., T_{Max_Delay}, is determined manually. Fig. 4c and Fig. 4d portray the changes of the LAA-LTE delay for $W_L = 8$ and 32, respectively, where we let $T_{Max\_Delay} = 0.5$ in Fig. 4c and $T_{Max\_Delay} = 2$ in Fig. 4d for better comparison. According to (P2.C4), the LAA-LTE delay of the red or blue plane should be lower or equal to that of the cyan plane. It shows that there is an intrinsic minimum delay at $N_2 = N$ and $t = 0$ for a given W_L, which should match the designed T_{Max_Delay} to avoid negative t.

The aforementioned observation reveals a fundamental tradeoff among the three QoS metrics. A smaller W_L leads to a more aggressive contention strategy of the LAA-LTE BS, which degrades the throughput and delay performance of the WiFi system but improves the delay performance of the LAA-LTE system. W_L should be chosen appropriately to balance the QoS metrics and satisfy the constraints so that the feasibility of **Problem 1** can be achieved. Although an appropriate W_L cannot be analytically derived due to the coupling variables in the complex nonlinear system, it can be determined empirically based on experiments.

C. Performance of the Proposed Algorithms

In this part, the performance of the proposed algorithms is evaluated with user channels. Besides path loss, Rayleigh fading with 8 multipaths is also considered in the channels from the LAA-LTE BS to the QUs. We set N, T_{Max_Delay}, and W_L to be 10, 4ms, and 55, respectively. The QUs are uniformly located between 1m and 55m from the LAA-LTE BS.

Fig. 5 compares the evolution of the deep-cut ellipsoid method in **Algorithm 1** with that of the basic ellipsoid method and the subgradient method, for $N_0 = 10$, $N_2 = 5$, and $\{\bar{R}_k\} = 5$Mbps. As Fig. 5 shows, the subgradient method has slower convergence even if its initial values and step sizes are fine-tuned. On the contrary, both the ellipsoid method and its deep-cut version converge more quickly without tuning any parameter. In contrast to its design objective, the deep-cut ellipsoid method still fluctuates here due to the nonconvexity of **Problem 3**. However, it indeed reduces the duality gap and the fluctuation of the basic ellipsoid method.

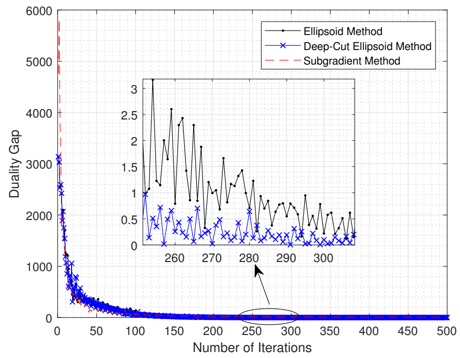
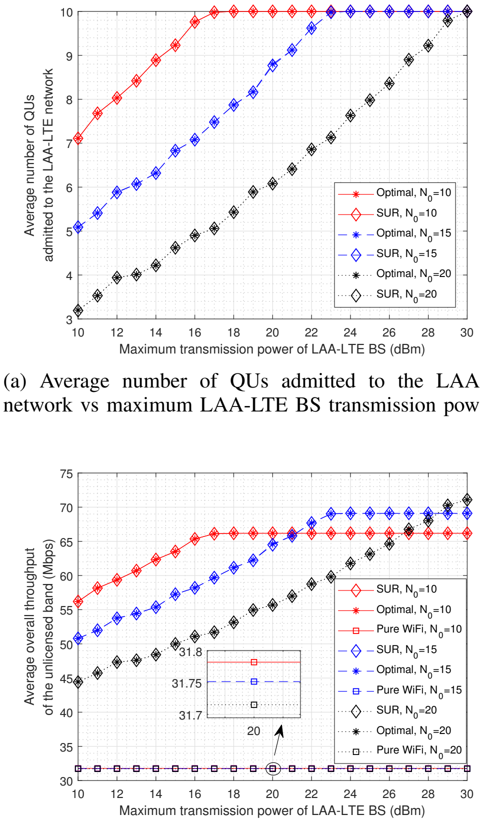


Fig. 5. Evolution curves of the three algorithms.

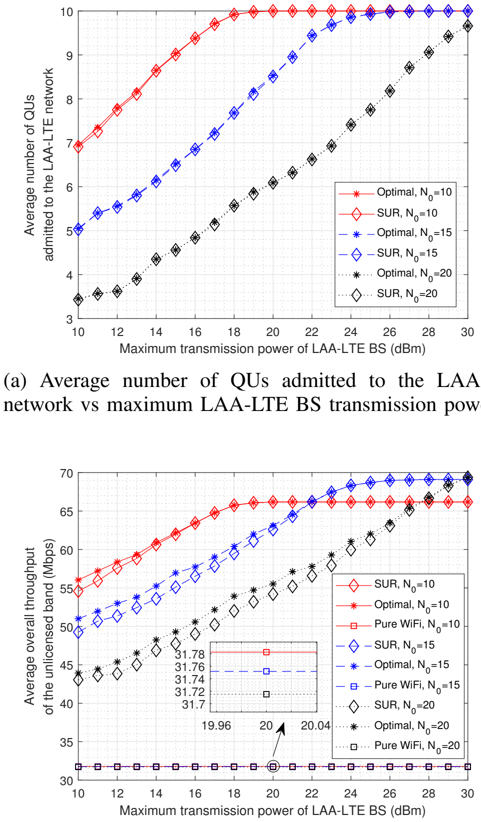


(a) Average number of QUs admitted to the LAA-LTE network vs maximum LAA-LTE BS transmission power.

(b) Average overall throughput of the unlicensed band vs maximum LAA-LTE BS transmission power.

Fig. 6. The performance of Algorithm 2. QUs have same rate requirements.

Next, we evaluate the performance of the proposed QoS-aware joint user association and resource allocation algorithm, i.e., **Algorithm 2**. Fig. 6a and Fig. 7a depict the average number of QUs admitted to the LAA-LTE network obtained by **Algorithm 2** and that by exhaustive search, respectively, for different maximum LAA-LTE transmission power, $\bar{Q}$, and N_0. Fig. 6b and Fig. 7b illustrate the average overall throughput of the unlicensed band corresponding to Fig. 6a and Fig. 7a, respectively. It is noticed that the solutions obtained by exhaustive search may have multiple user association strategies



(a) Average number of QUs admitted to the LAA-LTE network vs maximum LAA-LTE BS transmission power.

(b) Average overall throughput of the unlicensed band vs maximum LAA-LTE BS transmission power.

Fig. 7. The performance of Algorithm 2. QUs have different rate requirements.

and lead to multiple throughputs. Hence, we choose the largest one as the throughput of the optimal solution. Both Fig. 6 and Fig. 7 are obtained by averaging the results of 100 repeated experiments with fixed locations and random small fading, yet considering different scenarios. Fig. 6 considers the QUs of homogeneous QoS preferences, where all QUs have the rate requirement of 5Mbps. In contrast, Fig. 7 considers the heterogeneous scenario, where half of the QUs have the rate requirement of 3Mbps and the other 7Mbps. Particularly, the rate requirements are selected randomly in each experiment. As Fig. 6a and Fig. 7a show, **Algorithm 2** admits a similar number of QUs to the LAA-LTE network as the optimal solution for both scenarios. However, this conclusion does not hold for the average overall throughput of the unlicensed band according to Fig. 6b and Fig. 7b. This phenomenon is reasonable, because the basic idea of the SUR algorithm included in **Algorithm 2** is to remove the QU of the highest individual power successively, regardless of rate requirements. Intuitively, such operation gives up the most power-consuming QUs and leaves the power to admit more QUs, which is exactly the design target of the coexistence system, i.e., the objective function of **Problem 1**. As explained in Section V-B, the optimality of the SUR algorithm is not guaranteed, because subcarriers are released after a QU is removed from the LAA-LTE network, which may change the order of the individual power of QUs. As such, the QU of the highest individual power may not be the highest one if another QU is removed, which means the optimal solution may keep that QU. However, it happens rarely in practice, making **Algorithm 2** approach the number of QUs admitted to the LAA-LTE network achieved by exhaustive search. Although the proposed scheme only focuses on the user number, Fig. 6b and Fig. 7b demonstrate that the coexistence system can achieve much higher spectrum efficiency than the pure WiFi system.

Then, we take Fig. 6 for instance and discuss about the effects of N_0. As Fig. 6a shows, when N_0 grows, a larger $\bar{Q}$ is needed to admit the same number of the QUs to the LAA-LTE network, because a larger N_0 tightens the protection to the WiFi system according to (P2.C2) and (P2.C3). Nevertheless, it is interestingly shown in Fig. 6b that a large N_0 can result in higher overall throughput of the unlicensed band when $\bar{Q}$ is large enough. We first focus on a fixed N_0 in Fig. 6b. It can be seen that $\bar{Q}$ can increase the throughput when it is small, because more QUs are admitted to the LAA-LTE network with their desired rates. Then, when $\bar{Q}$ gets even larger, it gradually becomes saturated when all the QUs have been associated with the LAA-LTE network, where the throughputs of both networks become constants. The final throughputs of the LAA-LTE network are the same for different N_0, which is exactly the summation of the rate requirements of all QUs, but those of the WiFi network differ for different N_0, because more throughput is preserved for the WiFi system with a larger N_0, according to (P2.C2). Thus, the throughput achieved with a larger N_0 can finally surpass that with a small one, as Fig. 6b shows.

VII. Conclusions

In this paper, we have investigated the QoS-aware coexistence between the LAA-LTE system and the WiFi system. To enable the QoS-awareness, we have quantified the QoS metrics, i.e., throughput and delay, of the two coexisting systems. Then, the problem of designing the QoS-aware coexistence scheme has been formulated, which aims to support as many QoS-preferred users as possible in the LAA-LTE network, while maintaining fair QoS guarantees for the WiFi system. This is achieved by jointly optimizing the LAA-LTE transmission time, subcarrier assignment, power allocation, and user association. We have developed two efficient algorithms by decomposing the nonlinear and nonconvex problem into two subproblems. Through simulations, the effectiveness of the proposed algorithms have been demonstrated, and the fundamental tradeoff of the QoS metrics has been revealed. While this paper considers equal channel bit rates for all WiFi users, it is pointed out that the proposed methodology for coexistence optimization can also be applied to the scenario in which the WiFi users have different channel bit rates, and this will be one of the future research directions.

Appendix A: Proof of Theorem 1

A. *Proof of D_2^{Pure}*

We first let the retransmission limit go to infinity, i.e., $J \to \infty$. Note that the CW stays at $2^m W_0 - 1$ when the retransmission time is larger than m because of the maximum backoff stage m. Besides, we have $p_3 \in (0,1)$ according to (9). Hence, we can rewrite $D_2^{Pure,J}$ in (22) as

$$\begin{aligned} D_2^{Pure} &= \lim_{J\to\infty} D_2^{Pure,J} \\ &= \frac{1}{2}\left[\sum_{j=0}^{m} p_3^j 2^j W_0 + \sum_{j=m+1}^{\infty} p_3^j 2^m W_0 - \sum_{j=0}^{\infty} p_3^j\right] T_M^{Pure} \\ &\overset{(a)}{=} \frac{1}{2}\left[\sum_{j=0}^{m} p_3^j 2^j W_0 + \frac{p_3^{m+1}}{1-p_3} 2^m W_0 - \frac{1}{1-p_3}\right] T_M^{Pure}, \end{aligned} \tag{57}$$

where (a) is due to $\sum\limits_{j=m+1}^{\infty} p_3^j = \frac{p_3^{m+1}}{1-p_3}$ and $\sum\limits_{j=0}^{\infty} p_3^j = \frac{1}{1-p_3}$ according to the property of the geometric progression. Therefore, (25) is proved.

B. Proof of D_3^{Pure}

With $J \to \infty$, we can extract the expression in (23) and reformulate $D_3^{Pure,J}$ as

$$\begin{aligned} D_3^{Pure} &= \lim_{J\to\infty} \frac{(1-p_3)}{1-p_3^{J+1}} \sum_{j=0}^{J} p_3^j T_{s,W} + j p_3^j T_c \\ &= (1-p_3)\left[T_{s,W}\sum_{j=0}^{\infty} p_3^j + T_c \sum_{j=0}^{\infty} j p_3^j\right]. \end{aligned} \tag{58}$$

To derive (26) from (58), we introduce the following **Lemma 2**.

Lemma 2: For $x \in (0,1)$, $\sum\limits_{j=0}^{\infty} j x^j = \frac{x}{(1-x)^2}$.

With $p_3 \in (0,1)$, the property of geometric progression, and **Lemma 2**, we can rewrite (58) as

$$\begin{aligned} D_3^{Pure} &= (1-p_3)\left[T_{s,W}\sum_{j=0}^{\infty} p_3^j + T_c \sum_{j=0}^{\infty} j p_3^j\right] \\ &= (1-p_3)\left[T_{s,W}\frac{1}{1-p_3} + T_c \frac{p_3}{(1-p_3)^2}\right] \\ &= T_{s,W} + \frac{p_3}{1-p_3} T_c. \end{aligned} \tag{59}$$

Therefore, (26) is proved.

From the above, **Theorem 1** is thus proved.

Appendix B: Proof of Theorem 3

In this section, we prove the theorem on the WiFi packet delay in the LAA-LTE/WiFi coexistence system. The proof follows the same line of thought in [32], and infinite retransmission limit is also assumed for the asymptote. We first derive the average delay, T_{Wait}^W, for the WAIT packets. T_{Wait}^W consists of three parts: an idle slot, θ, as the start of the backoff process, the expected WiFi countdown time, D_2^W, and the expected WiFi packet transmission time, D_3^W. Due to the freezing mechanism, each countdown slot contains the channel idle time, the successful transmission time, and the collision time of other WiFi stations and the LAA-LTE BS. D_2^W can be modeled as $D_2^W = M_{CD}^W T_M^W$, where M_{CD}^W is the expected countdown slots and T_M^W is the expected time for a countdown slot. One of the WiFi stations is taken as the reference and it is supposed to be in the countdown procedure. Let p_I^W, p_W^W and p_L^W be the probabilities of the idle channel (no packet on the channel), successful transmission of other WiFi stations (only one WiFi station transmits) and successful transmission of the LAA-LTE BS (only the LAA-LTE BS transmits), respectively. We have

$$p_I^W = (1-\tau_W)^{N_0+N_1}(1-\tau_L), \tag{60}$$

$$p_W^W = (N_0+N_1)\tau_W(1-\tau_W)^{N_0+N_1-1}(1-\tau_L,) \tag{61}$$

$$p_L^W = \tau_L(1-\tau_W)^{N_0+N_1}, \tag{62}$$

The probability, p_C^W, of collision can be denoted by $p_C^W = 1-(p_I^W + p_W^W + p_L^W)$. Hence, by multiplying p_I^W,p_W^W,p_L^W and p_C^W with their corresponding time durations, we can obtain T_M^W in (31). With the retransmission probability, p_1, the CW will be doubled for each retransmission until m is reached. Since the counter is uniformly chosen within the CW, the expected countdown slots M_{CD}^W can be written as $M_{CD}^W = \frac{1}{2}[\sum_{j=0}^{m} p_1^j(2^j W_0 - 1) + \sum_{j=m+1}^{\infty} p_1^j (2^m W_0 - 1)]$. Hence, similar to Appendix A, D_2^W can be simplified as

$$D_2^W = \frac{1}{2}\left[\sum_{j=0}^{m} p_1^j 2^j W_0 + \frac{p_1^{m+1}}{1-p_1} 2^m W_0 - \frac{1}{1-p_1}\right] T_M^W. \tag{63}$$

D_3^W consists of the successful transmission time and the collision time. For instance, if a packet is successfully transmitted on the j-th retransmission, the actual transmission time is $T_{s,W} + jT_c$. Hence, $D_3^W = \sum_{j=0}^{\infty}(1-p_1)\, p_1^j\,(T_{s,W} + jT_c)$. With **Lemma 2**, D_3^W can be written as

$$D_3^W = T_{s,W} + \frac{p_1}{1-p_1} T_c. \tag{64}$$

Thus, T_{Wait}^W can be denoted by $T_{Wait}^W = \theta + D_2^W + D_3^W$.

Then, for N-WAIT packets, they occur when they are following a previous successful transmission and their new counter is 0, of which the probability is $\kappa_W = 1/W_0$. Therefore, for each WAIT packet that is transmitted successfully, $\varphi_W = \sum_{j=0}^{\infty} \kappa_W = 1/(W_0 - 1)$ N-WAIT packets can be consecutively transmitted in average. From the above, (29) can be obtained.

Appendix C: Proof of Theorem 4

The delay of the LAA-LTE system can be derived following the proof of **Theorem 3**, where infinite retransmission limit is also assumed. Similar to the N-WAIT packet of the WiFi system, the transmission without waiting of the LAA-LTE BS is called N-WAIT transmission. WAIT transmission is defined contrarily. The average delay, T_{Wait}^L, for the WAIT transmissions also contains an idle slot, θ, the expected LAA-LTE countdown time, D_2^L, and the expected LAA-LTE transmission time, D_3^L. Term D_2^L can be decomposed as $D_2^L = M_{CD}^L T_M^L$, where M_{CD}^L and T_M^L stand for the expected countdown slots and the expected time for a countdown slot, respectively. Suppose that the LAA-LTE BS is in the countdown procedure. The probability, p_I^L, of idle channel (no packet on the channel),

$$\frac{\partial f_L(N_1,t)}{\partial t} = \frac{\tau_L \bar{\tau}_W^{\mathrm{n}+1}\left\{\bar{\tau}_W^{\mathrm{n}}\left[\bar{\tau}_W\bar{\tau}_L\theta + (\mathrm{n}+1)\,\tau_W\bar{\tau}_L T_{s,W} + \bar{\tau}_W\tau_L V\right] + \left(1-\bar{\tau}_W^{\mathrm{n}} - n\bar{\tau}_W^{n}\tau_W\bar{\tau}_L + \bar{\tau}_W^{\mathrm{n}}\tau_W\tau_L\right)T_c\right\}}{\left\{\bar{\tau}_W^{\mathrm{n}}\left[\bar{\tau}_W\bar{\tau}_L\theta + (\mathrm{n}+1)\,\tau_W\bar{\tau}_L T_{s,W} + \tau_L\bar{\tau}_W\,(V+t)\right] + \left(1-\bar{\tau}_W^{\mathrm{n}} - n\bar{\tau}_W^{n}\tau_W\bar{\tau}_L + \bar{\tau}_W^{\mathrm{n}}\tau_W\tau_L\right)T_c\right\}^2}. \tag{67}$$

the probability, p_W^L, of the successful transmission of WiFi stations (only one WiFi station transmits) can be expressed by

$$p_I^L = (1-\tau_W)^{N_0+N_1+1}, \tag{65}$$

$$p_W^L = (N_0+N_1+1)\,\tau_W(1-\tau_W)^{N_0+N_1}, \tag{66}$$

respectively. The probability, p_C^L, of a collision can be denoted by $p_C^L = 1-(p_I^L+p_W^L)$. Term T_M^L is thus obtained from the multiplication of p_I^L, p_W^L, p_C^L by their corresponding time durations. Because of the zero stage of LAA-LTE, the expected countdown slots can be simply written as $M_{CD}^L = \frac{1}{2}(W_2-1)$. Hence, (34) is obtained, and T_{Wait}^L is denoted by $T_{Wait}^L = \theta + D_2^L + D_3^L$.

After a successful WAIT transmission, $\varphi_L = \sum_{j=0}^{\infty}\kappa_L = 1/(W_L-1)$ N-WAIT transmissions can be consecutively completed in average. Therefore, (32) can be derived.

Appendix D: Proof of Theorem 5

For notational simplicity, we make some definitions: $n = N_0+N_1$, $\bar{\tau}_W = 1-\tau_W$, $\bar{\tau}_L = 1-\tau_L$, and $V = T_{s,L}(t)-t$. With a given N_1, the derivative of $f_L(N_1,t)$, i.e., $\frac{\partial f_L(N_1,t)}{\partial t}$, is shown on the top of this page as (67).

Surprisingly, t does not affect the sign symbol of (67), because t only appears in the denominator and is squared. The parts in the numerator are all positive except $(1-\bar{\tau}_W^{\mathrm{n}} - n\bar{\tau}_W^{n}\tau_W\bar{\tau}_L + \bar{\tau}_W^{\mathrm{n}}\tau_W\tau_L)$, which can be rewritten as $1-\bar{\tau}_W^{\mathrm{n}+1} - (n+1)\,\bar{\tau}_W^{n}\tau_W\bar{\tau}_L$. Intuitively, $\bar{\tau}_W^{\mathrm{n}+1}$ is the probability that no WiFi stations attempt to transmit, while $(n+1)\,\bar{\tau}_W^{n}\tau_W\bar{\tau}_L$ is the probability that a WiFi station attempts to transmit and succeeds. Since the summation of the two probabilities is less than 1, we have $\frac{\partial f_L(N_1,t)}{\partial t} > 0, \forall t$. Thus, $f_L(N_1,t)$ is monotonically increasing with t.

References

[1] *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016—2021 White Paper*, Mar. 2017.

[2] S.-Y. Lien, K.-C. Chen, Y.-C. Liang, and Y. Lin, "Cognitive radio resource management for future cellular networks," *IEEE Wireless Commun.*, vol. 21, no. 1, pp. 70–79, Feb. 2014.

[3] Y.-C. Liang, K. C. Chen, G. Y. Li, and P. Mahonen, "Cognitive radio networking and communications: an overview," *IEEE Trans. Vehi. Technol.*, vol. 60, no. 7, pp. 3386–3407, Sept. 2011.

[4] L. Zhang, Y.-C. Liang, and M. Xiao, "Spectrum sharing for internet of things: A survey," *IEEE Wireless Commun.*, 2019, *Early Access*, doi:10.1109/MWC.2018.1800259.

[5] Q. Zhang, H. Guo, Y.-C. Liang, and X. Yuan, "Constellation learning-based signal detection for ambient backscatter communication systems," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 2, pp. 452–463, Feb. 2019.

[6] Q. Zhang, L. Zhang, Y.-C. Liang, and P.-Y. Kam, "Backscatter-NOMA: A symbiotic system of cellular and Internet-of-Things networks," *IEEE Access*, vol. 7, pp. 20 000–20 013, Feb. 2019.

[7] G. Yang, Y.-C. Liang, R. Zhang, and Y. Pei, "Modulation in the air: Backscatter communication over ambient OFDM carrier," *IEEE Trans. Commun.*, vol. 66, no. 3, pp. 1219–1233, Mar. 2018.

[8] G. Yang, Q. Zhang, and Y.-C. Liang, "Cooperative ambient backscatter communications for green Internet-of-Things," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 1116–1130, Apr. 2018.

[9] L. Zhang, M. Xiao, G. Wu, M. Alam, Y.-C. Liang, and S. Li, "A survey of advanced techniques for spectrum sharing in 5G networks," *IEEE Wireless Commun.*, vol. 24, no. 5, pp. 44–51, Oct. 2017.

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

[11] N. Rupasinghe and İ. Güvenç, "Licensed-assisted access for WiFi-LTE coexistence in the unlicensed spectrum," in *IEEE GLOBECOM Wkshps*, 2014, pp. 894–899.

[12] L. Berlemann, C. Hoymann, G. Hiertz, and B. Walke, "Unlicensed operation of IEEE 802.16: Coexistence with 802.11(A) in shared frequency bands," in *Proc. IEEE PIMRC*, 2006, pp. 1–5.

[13] R. Yin, G. Yu, A. Maaref, and G. Y. Li, "A framework for co-channel interference and collision probability tradeoff in LTE licensed-assisted access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6078–6090, Jun. 2016.

[14] Y. Huang, J. Tan, and Y.-C. Liang, "Wireless big data: transforming heterogeneous networks to smart networks," *J. Commun. Inf. Netw.*, vol. 2, no. 1, pp. 19–32, Mar. 2017.

[15] X. Wang, T. Q. S. Quek, M. Sheng, and J. Li, "Throughput and fairness analysis of Wi-Fi and LTE-U in unlicensed band," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 1, pp. 63–78, Jan. 2017.

[16] H. Zhang, X. Chu, W. Guo, and S. Wang, "Coexistence of Wi-Fi and heterogeneous small cell networks sharing unlicensed spectrum," *IEEE Commun. Mag.*, vol. 53, no. 3, pp. 158–164, Mar. 2015.

[17] C. Cano and D. J. Leith, "Unlicensed LTE/WiFi coexistence: Is LBT inherently fairer than CSAT?" in *Proc. IEEE ICC*, 2015, pp. 1–6.

[18] K. Havish and M. David, "Feasibility study on licensed-assisted access to unlicensed spectrum," 3GPP, Tech. Rep., 2015, https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2579.

[19] S. Han, Y.-C. Liang, Q. Chen, and B. H. Soong, "Licensed-assisted access for LTE in unlicensed spectrum: A MAC protocol design," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 10, pp. 2550–2561, Oct. 2016.

[20] J. Tan, S. Xiao, S. Han, and Y.-C. Liang, "A learning-based coexistence mechanism for LAA-LTE based HetNets," in *Proc. IEEE ICC*, 2018, pp. 1–6.

[21] H. He, H. Shan, A. Huang, L. X. Cai, and T. Q. S. Quek, "Proportional fairness-based resource allocation for LTE-U coexisting with Wi-Fi," *IEEE Access*, vol. 5, pp. 4720–4731, Sept. 2017.

[22] Q. Chen, G. Yu, A. Maaref, G. Y. Li, and A. Huang, "Rethinking mobile data offloading for LTE in unlicensed spectrum," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4987–5000, Jul. 2016.

[23] Q. Chen, G. Yu, R. Yin, A. Maaref, G. Y. Li, and A. Huang, "Energy efficiency optimization in licensed-assisted access," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 723–734, Apr. 2016.

[24] A. Galanopoulos, F. Foukalas, and T. A. Tsiftsis, "Efficient coexistence of LTE with WiFi in the licensed and unlicensed spectrum aggregation," *IEEE Trans. on Cogn. Commun. Netw.*, vol. 2, no. 2, pp. 129–140, Jun. 2016.

[25] R. Yin, G. Yu, A. Maaref, and G. Y. Li, "LBT-based adaptive channel access for LTE-U systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 6585–6597, Oct. 2016.

[26] Q. Cui, Y. Gu, W. Ni, and R. P. Liu, "Effective capacity of licensed-assisted access in unlicensed spectrum for 5G: from theory to application," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 8, pp. 1754–1767, Aug. 2017.

[27] J. B. G. Frenk, J. Gromicho, and S. Zhang, "A deep cut ellipsoid algorithm for convex programming: Theory and applications," *Math. Program.*, vol. 63, no. 1, pp. 83–108, Jan. 1994.

[28] 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, Jul. 2017.

[29] O. Tickoo and B. Sikdar, "Modeling queueing and channel access delay in unsaturated IEEE 802.11 random access MAC based wireless networks," *IEEE/ACM Trans. Netw.*, vol. 16, no. 4, pp. 878–891, Aug. 2008.

[30] P. Raptis, V. Vitsas, and K. Paparrizos, "Packet delay metrics for IEEE 802.11 distributed coordination function," *Mobile Netw. Appl.*, vol. 14, no. 6, pp. 772–781, Dec. 2009.

IEEE Transactions on Wireless Communications Volume: 18 , Issue: 4 , April 2019

[31] J. W. Tantra, F. Chuan Heng, and A. B. Mnaouer, “Throughput and delay analysis of the IEEE 802.11e EDCA saturation,” in *Proc. IEEE ICC*, 2005.

[32] K. Kang, X. Lin, and H. Hu, “An accurate MAC delay model for IEEE 802.11 DCF,” in *Proc. IEEE ICT-MICC*, 2007, pp. 654–657.

[33] W. Yu and R. Lui, “Dual methods for nonconvex spectrum optimization of multicarrier systems,” *IEEE Trans. Commun.*, vol. 54, no. 7, pp. 1310–1322, Jul. 2006.

[34] K. Seong, M. Mohseni, and J. M. Cioffi, “Optimal resource allocation for OFDMA downlink systems,” in *Proc. IEEE ISIT*, 2006, pp. 1394–1398.

[35] S. Boyd and C. Barratt, *Linear controller design: limits of performance*. Prentice-Hall, Englewood Cliffs, N.J., 1991.