Multi-Dimensional Busy-Tone Arbitration for OFDMA Random Access in IEEE 802.11ax

Dianhan Xie, *Student Member, IEEE*, Jiawei Zhang, Aimin Tang, *Member, IEEE*, and Xudong Wang, *Fellow, IEEE*

***Abstract*— IEEE 802.11ax has adopted orthogonal frequency division multiple access (OFDMA) to support multi-user (MU) transmissions. There exist two uplink MU OFDMA access methods in IEEE 802.11ax. The first one is uplink OFDMA random access (UORA) in which stations randomly select resource units (RUs) to send physical protocol data units (PPDUs), so its access efficiency is low. The second one is uplink OFDMA nonrandom access (UONRA) in which the access point (AP) schedules MU transmissions based on buffer status reports (BSR) from stations. However, stations usually rely on UORA to send BSRs, so the low efficiency of UORA can be a bottleneck of UONRA. Thus, UORA needs to be renovated to improve the performance of itself and UONRA. To this end, a multi-dimensional busy-tone arbitration (MBTA) mechanism is developed in this paper to reduce collisions among stations contending the same RU. Since there lacks an algorithm in IEEE 802.11ax to support coexistence of UORA and UONRA, a dynamic access-method selection (DAMS) algorithm is designed for the AP and stations to choose an optimal access method. Both MBTA and DAMS are analyzed rigorously and are further validated via simulations. The analytical and simulation results show that: 1) The MBTA dramatically improves the access efficiency of UORA; 2) DAMS always achieves a higher throughput than both UORA and UONRA.**

***Index Terms*— Medium access control, uplink OFDMA random access, uplink OFDMA nonrandom access, IEEE 802.11ax.**

I. Introduction

THE traffic of the Internet increases rapidly in recent years. According to [1], the monthly global Wi-Fi traffic will exceed 80 exabytes by 2021. However, the traditional single-user (SU) transmission in wireless local area networks (WLAN) suffers low efficiency, especially in dense scenarios [2], [3]. To this end, the IEEE 802.11ax task group tries to improve channel efficiency by proposing the multi-user (MU) OFDMA access technology [4], [5]. In IEEE 802.11ax, there are two categories of medium access methods. The first category is uplink OFDMA random access (UORA). As shown in Fig. 1, the channel is divided into several resource units (RUs), each of which contains a subset of sub-carriers.

Manuscript received May 19, 2019; revised October 20, 2019 and February 3, 2020; accepted March 3, 2020. Date of publication March 18, 2020; date of current version June 10, 2020. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 61771312. The associate editor coordinating the review of this article and approving it for publication was G. Fodor. *(Corresponding author: Xudong Wang.)*

The authors are with the University of Michigan-Shanghai Jiao Tong University Joint Institute, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: wxudong@ieee.org).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2020.2979852

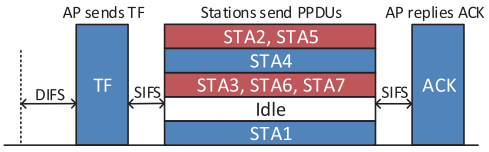


Fig. 1. Uplink OFDMA random access in IEEE 802.11ax.

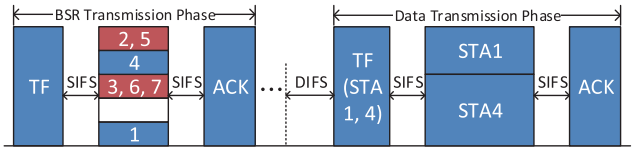


Fig. 2. Uplink OFDMA non-random access in IEEE 802.11ax.

UORA is started by a trigger frame (TF) that carries the number of the RUs for random access (called available RU). The TF also serves as a synchronization reference for the stations. When receiving the TF, the station whose OFDMA backoff (OBO) counter decreases to zero randomly chooses one available RU to transmit its physical protocol data unit (PPDU). Since the stations access the channel without coordination, RUs are not utilized efficiently. The performance of UORA has been analyzed in [6]–[8]. The channel access efficiency of UORA is the same as that of the slotted Aloha scheme. Thus, UORA is not suitable for the transmission of long frames.

The second category is the uplink OFDMA non-random access (UONRA), which is shown in Fig. 2. The AP requests stations to transmit their buffer status reports (BSR) that carry the information about the traffic size in their buffers. After that, the AP allocates RUs to stations based on their BSRs and channel conditions. Since the uplink transmissions in UONRA are scheduled by the AP, the data frame collisions can be avoided. Thus, the bottleneck of UONRA is mainly in the BSR transmission phase. To this end, the AP can send a BSR poll (BSRP) to poll stations for BSRs. However, in the scenario of a dense network, the polling scheme results in low efficiency in collecting BSRs. Besides the polling scheme, stations can use UORA to reply their BSR. Thus, UORA plays an important role in sending BSR for UONRA. However, the low efficiency of UORA causes a large delay in the BSR transmission phase. For user datagram protocol (UDP) traffic, the large delay degrades the quality of service (QoS) of the network. For transmission control protocol (TCP) traffic, it increases the round trip time (RTT) of a TCP link,

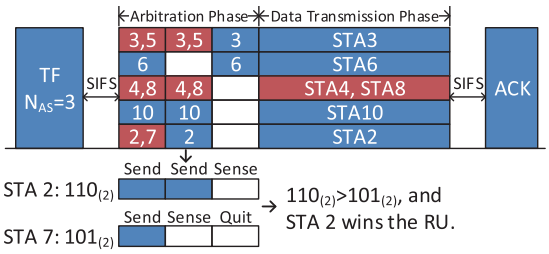


Fig. 3. The multi-dimensional busy-tone arbitration mechanism.

which impacts throughput performance, as the TCP congestion control adjusts the congestion window per RTT. To solve these problems, the efficiency of UORA needs to be improved.

Before IEEE 802.11ax is being standardized, multiple schemes have been developed to achieve the MU OFDMA access in WLAN. In [9]–[11], the carrier-sense medium access (CSMA) mechanism is conducted in each OFDMA sub-channel. However, in IEEE 802.11ax, the header of the PPDU contains a legacy preamble that occupies the entire 20 MHz channel. The legacy preamble is crucial for backward compatibility with legacy stations, as it can be demodulated by legacy stations. If the CSMA mechanisms in [9]–[11] are applied to UORA, the PPDUs in different RUs are not aligned, and the legacy header can damage the concurrent transmissions in different RUs. In [12], [13], OFDMA transmissions are scheduled by the AP, which is similar to UONRA in IEEE 802.11ax. However, neither of these two schemes improves the efficiency of random access.

In this paper, a multi-dimensional busy-tone arbitration (MBTA) mechanism is developed to significantly improve the efficiency of UORA. As shown in Fig. 3, an arbitration phase is added before the stations transmit PPDUs on the RUs. Each station randomly selects an RU and an arbitration number. In the arbitration phase, the stations selecting the same RU arbitrate their arbitration numbers by sending busy-tone signals in their selected RU. The station with the maximum arbitration number wins the RU, and the other stations quit the contention and proceed with the retransmission procedure. As a result, collisions occur only when more than one station chooses both the same RU and the same arbitration number. In the above mechanism, stations contend the access opportunity in both the frequency domain (i.e., RUs) and the arbitration number domain, and thus it is called a multi-dimensional busy-tone arbitration (MBTA) mechanism in this paper.

The bitwise arbitration mechanisms in [14], [15] look similar to MBTA, but they are not applicable to IEEE 802.11ax. In [14], arbitration continues until collisions among stations are resolved. If this scheme is applied to UORA, the arbitrations in different RUs are not aligned, but PPDUs in UORA need to be aligned to support the legacy preamble. In [15], a pseudo-noise (PN) sequence is used to conduct the pipeline channel contention during packet transmission. The PN sequence is much longer than a busy tone signal in MBTA, and it is not feasible in IEEE 802.11ax. Furthermore, the pipeline contention scheme in [15] arbitrates contentions in the frequency domain and the antenna domain, which does not resolve collisions in an RU in the time domain.

Besides the MBTA mechanism, a dynamic access method selection (DAMS) algorithm is designed in this paper for the AP and its associated stations to choose the optimal access strategy. Since UORA is efficient for short frames but UONRA is efficient for long frames, there exists an optimal boundary of frame duration for using UORA or UONRA. In the DAMS algorithm, the AP estimates this optimal boundary and then send this value to stations where it is used as the threshold for selecting an access method between UORA and UONRA. To the best of our knowledge, it is the first dynamic access method selection algorithm for IEEE 802.11ax.

The analytical model for the MBTA mechanism is formulated, through which the network throughput and the access delay of stations are rigorously derived. Furthermore, simulations are carried out to validate the analytical model and also evaluate the performance of the MBTA mechanism and the DAMS algorithm. Results show that the analytical model is consistent with the actual behavior of the MBTA mechanism. Moreover, it is shown that MBTA significantly improves channel access efficiency. Simulation results also illustrate that the DAMS algorithm achieves the highest throughput, compared with the scenarios where stations choose either UORA or UONRA only for channel access.

The main contributions of this paper are listed as follows:

- The MBTA mechanism is developed to significantly improve random access efficiency of IEEE 802.11ax;
- The DAMS algorithm is designed for the AP and stations to choose an access method in an optimal way;
- The analytical models for the MBTA mechanism and the DAMS algorithm are formulated, and the performance of MBTA and DAMS are evaluated extensively via simulations.

The rest of the paper is organized as follows. In Section II, the MBTA mechanism is developed, and its performance is analyzed in Section III. In Section IV, the optimal frame duration boundary between UORA and UONRA is derived, and then the DAMS algorithm is designed accordingly. The simulation setup and performance results are presented in Section V and Section VI, respectively. The paper is concluded in Section VII.

II. Multi-Dimensional Busy-Tone Arbitration

In this section, the MBTA mechanism is designed to improve the random access efficiency in IEEE 802.11ax. In addition, the arbitration slot is illustrated, and the feasibility of the MBTA mechanism is clarified.

A. Details of the MBTA Mechanism

To reduce the collisions among stations, an arbitration phase is added for the stations to coordinate with each other. The detail of the MBTA mechanism is designed as follows.

As shown in Fig. 3, the AP transmits a TF that contains the number of arbitration slots (denoted by N_{AS}), and the number of available RUs (denoted by N_{RU}). The stations with data to transmit decrease their OBO counters by one for each available RU. When the OBO counter of a station decreases to zero, the station randomly selects an available RU and a N_{AS}-bit

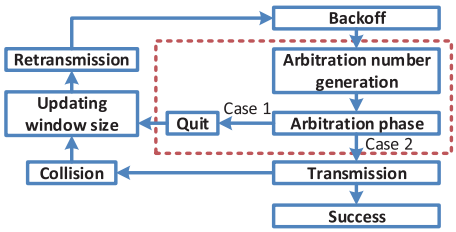


Fig. 4. The protocol flow of the stations in the MBTA mechanism.

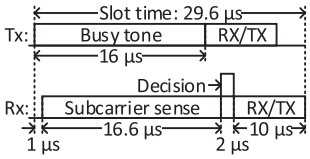


Fig. 5. The details of the arbitration slot.

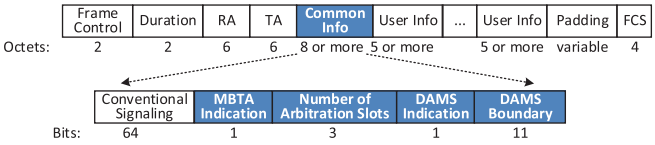


Fig. 6. The newly designed TF format.

binary number (from 0 to $2^{N_{AS}}-1$) as the arbitration number. In the arbitration phase, the station evaluates each bit (in the order from the most-significant one to the least-significant one) and then reacts as follows. If the bit is one, the station sends a busy-tone signal in the selected RU; otherwise, it senses the selected RU. Two cases need to be considered in the arbitration phase, as shown in Fig. 4.

In the first case, the station detects a busy-tone signal in the selected RU, which means there exists at least one station with a larger arbitration number in the RU. The station then quits the contention and conducts retransmission.

In the second case, the RU is idle in all the slots when the station senses the RU. This case occurs when the station has the maximum arbitration number. Thus, the station wins the selected RU and then sends its frame in this RU. However, it is possible that more than one station wins the RU, which leads to collisions. The collided stations then follow the binary exponential backoff algorithm to conduct retransmission.

For example, as shown in Fig. 3, station 2 and station 7 select the same RU, and they randomly choose a 3-bit binary number ($N_{AS}=3$), more specifically 110_2 for station 2 and 101_2 for station 7. In the first slot (the most-significant bit), both stations send a busy-tone signal in the selected RU. In the second slot, station 2 transmits a busy-tone signal while station 7 is sensing the selected RU, so station 7 can detect the busy-tone signal transmitted by station 2, stop contending, and enter the retransmission state. In the third slot, station 2 senses the RU and finds it idle. Finally, station 2 wins the RU, and the collision between station 2 and station 7 is avoided.

The conventional UORA has low efficiency as most RUs experience collisions. In contrast, by using MBTA, the collisions among the stations that choose the same RU is greatly reduced. It should be noted that the arbitration phase causes extra overhead, so the arbitration slots need to be carefully designed.

B. Design of the Arbitration Slot

To improve the efficiency of MBTA, the arbitration slot needs to be designed as short as possible. In addition, the functions required by MBTA need to be properly supported.

The design of the arbitration slot is shown in Fig. 5. The maximum propagation delay is assumed to be 1 μs, which is reasonable for WLAN. The busy-tone signal needs to contain one OFDM symbol (12.8 μs in IEEE 802.11ax) so that it can cover an RU. According to the IEEE 802.11ax standard, there are three cyclic prefix (CP) durations, which are 0.8 μs, 1.6 μs, and 3.2 μs. A large CP duration lengthens the busy-tone signal, which induces a large overhead. In this paper, the worst case of our design is considered. The CP duration is set to 3.2 μs. Thus, the duration of the busy-tone signal is 16 μs. To support the sub-carrier sensing function, FFT operation is needed by the sensing function of a station. Since the FFT size in IEEE 802.11ax is 256, the FFT operation can be done within 0.6 μs [16]. Based on the sub-carrier sensing result, the station decides whether to quit the arbitration or not, which can be fulfilled via a low-complex logical circuit. The processing time is within 2 μs. Moreover, the RX/TX switching time is 10 μs [17]. Hence, the total duration of an arbitration slot is 29.6 μs, and the functions required by MBTA can be properly supported.

C. Format of Trigger Frame

In MBTA, the information about the number of arbitration slots is sent by an AP via a TF. In DAMS, which will be introduced in Section IV, an AP broadcasts a frame duration boundary to stations via a TF. Hence, the TF format needs to be designed to fulfill these functions.

The newly designed TF frame format is shown in Fig. 6. According to the specification in the IEEE 802.11ax standard, a TF contains a common field that can be extended. Hence, 2 bytes are added to the common field to indicate the number of arbitration slots in MBTA and the frame duration boundary in DAMS. In the newly added bytes, the first bit is used to indicate whether MBTA is selected or not. The following three bits indicate the number of arbitration slots, which varies from 0 to 7. The 5-th bit is to indicate whether DAMS is selected or not. The last eleven bits report the frame duration boundary with a granularity of 2 μs. Thus, this field can represent a frame duration boundary of up to 4096 μs (Based on the analysis in Section IV, the frame duration boundary is within 2760 μs approximately).

D. Coexistence With Legacy Stations

Legacy stations may interrupt the arbitration phase. To prevent interruptions of legacy stations, the TF sent by the AP needs to carry a network allocation vector (NAV) to indicate

the duration of the arbitration phase and the data transmission phase. In this way, the legacy stations will keep silent in the arbitration phase. Considering the legacy stations that just wake up and miss the NAV, The legacy stations follow carrier-sense multiple access with collision avoidance (CSMA/CA) protocol to access the channel. The legacy stations can access a channel when the channel is idle for a distributed interframe space (DIFS) duration (34 μs). In an arbitration phase, as a large number of stations randomly send busy-tone signals, the probability that the whole channel is idle for a DIFS duration is very low.

E. Feasibility of MBTA

For the feasibility of MBTA, a few issues need to be considered. First, the synchronization among stations needs to be explained. The synchronization among the stations is achieved via a training field in the TF from an AP, which has been standardized in IEEE 802.11ax. Hence, the stations can conduct MBTA synchronously.

Second, the busy-tone signals sent by a station may not be detected by another station, which leads to the hidden node problem. The hidden node problem can be alleviated by using larger power for busy-tone signals. In this paper, the transmit power of busy tone signals is set to 12 dBm (In contrast, the transmit power of payload signals in each RU is 5.5 dBm). The minimum sensing threshold is -93 dBm. The maximum sensing range of each station can achieve 100 m based on the path loss model in Section V-A. For a network with a radius of 50 m, all stations in the network can hear the busy tone signals. The typical applications include hotspot areas such as train stations, airports, conference centers, and stadiums.

In practical systems, the hidden nodes are hard to be totally avoided, due to the randomness of wireless channels. Hence, the hidden node problem is investigated via extensive simulations in Section VI-B.

To evaluate the MBTA mechanism more deeply, it is analyzed in the next section.

III. Theoretical Analysis

In this section, theoretical analysis is carried out to validate that the MBTA mechanism can significantly improve saturation throughput of UORA (channel access efficiency) and reduce stations' access delay.

Similar to [18], N_{STA} stations are assumed to work in saturation conditions, i.e, the transmission queue of each station is always nonempty. Hidden terminals and capture effect are not considered. The analysis consists of three parts. Firstly, the behavior of a single station is studied with a Markov model, and then the transmission probability τ of the station when receiving a TF for random access is derived. Secondly, the saturation throughput of an RU is expressed as a function of τ. Finally, the access delay of a station is expressed as a function of τ.

A. Transmission Probability

The derivation of transmission probability takes three steps. First, the backoff procedure is reviewed with the Markov chain model in Fig. 7. Second, based on the Markov chain, stations' transmission probability τ is derived as a function of stations' failure probability (denoted by p), which is the probability that a station collides in transmission or quits in the arbitration phase. Third, p is derived as a function of τ by analyzing the channel access procedure. Finally, τ is obtained by solving these two nonlinear equations.

TABLE I
DEFINITIONS OF VARIABLES

Variables	Definition
N_{STA}	the number of stations
N_{RU}	the number of RUs
N_{AS}	the number of arbitration slots
τ	the transmission probability of a station in a UORA
p	the probability that a station fails in a UORA
W_0	the minimum OFDMA contention window
W_m	the maximum OFDMA contention window
m	the maximum backoff stage
i	the backoff stage
k	the value of backoff counter
N	the number of stations that select a certain RU
N_{win}	the number of stations that win their selected RUs
N_s	the expected number of successful stations in an RU
S_j	the expected payload transmitted by the j-th station
S	the expected overall payload transmitted by stations
$R_{tx}(j)$	the PHY rate of the j-th station
$\overline{R}_{tx}$	the average PHY rate over all stations
T_{pay}	the duration of the payload in a UORA
T_{DIFS}	the duration of DIFS
T_{TF}	the duration of TF
T_{SIFS}	the duration of SIFS
T_{AS}	the duration of an arbitration slot
T_{PH}	the duration of a PHY header
T_{ACK}	the duration of an acknowledgement (ACK)
T_{all}	the total duration of a UORA
η	the saturated network throughput
T_I	the interval between two adjacent TFs for random access
D_i	the backoff duration in the i-th backoff stage
α	the payload duration in UORA
L_{max}	the maximum frame length
L_{min}	the minimum frame length
$f(x)$	the frame length distribution in a network
T_{RA}	the required time duration in UORA
T_{NRA}	the required time duration in UONRA

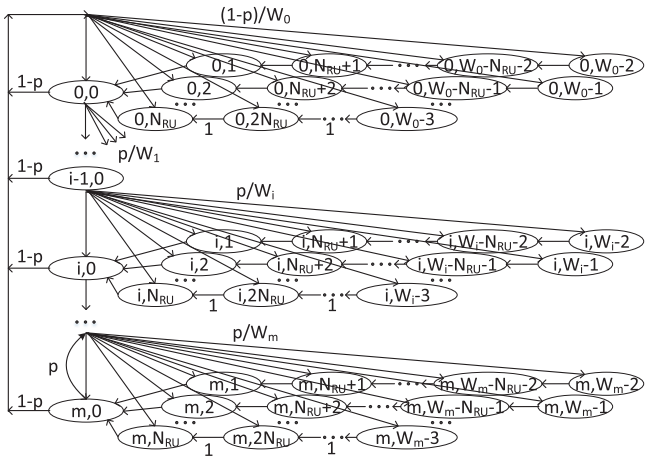


Fig. 7. Markov chain model for backoff window size in IEEE 802.11ax.

1) Backoff Procedure: In the first backoff stage, the station randomly selects a number from $[0, W_0 - 1]$ as the backoff counter, where W_0 is the minimum backoff window. When receiving a TF for random access, the station decreases its backoff counter by the number of RUs for random access (N_{RU}). When the backoff counter decreases to zero, the station will access the channel. If the transmission is collided or the station quits in the arbitration phase, the station will double the backoff window, and conduct retransmission. i.e., $W_i = 2^i W_0$, $0 \le i \le m$, where W_i is the backoff window of the i-th backoff stage, and m is the maximum backoff stage. If the transmission is successful, the station will set the backoff window as the minimum value W_0, and conduct the next transmission. Once the backoff stage reaches the maximum value m, it does not increase further.

2) Transmission Probability τ From Markov Chain: The stationary states of the Markov chain in Fig. 7 are denoted by $\{i, k\}$, where i is the backoff stage, and k is the value of backoff counter. In this Markov chain, the transition probabilities among the stationary states are

$$\begin{cases} P\{i, k - N_{RU} | i, k\} = 1 & k \in [N_{RU}, W_i - 1], i \in [0, m] \\ P\{i, 0 | i, k\} = 1 & k \in [1, N_{RU} - 1], i \in [0, m] \\ P\{0, k | i, 0\} = (1-p)W_0 & k \in [0, W_i - 1], i \in [0, m] \\ P\{i, k | i-1, 0\} = p/W_i & k \in [0, W_i - 1], i \in [1, m] \\ P\{m, k | m, 0\} = p/W_m & k \in [0, W_m - 1]. \end{cases} \tag{1}$$

The first equation in Eq. (1) models the fact that, when receiving a TF for random access, the backoff counter decreases by N_{RU}. The second equation models the fact that, when a station receives a TF and its backoff counter is less than N_{RU}, the backoff counter directly decreases to zero. The third equation models the fact that, once the station successfully transmits a packet, the station will start with backoff stage 0. The fourth equation accounts for the fact that, if the station has a collided transmission or quits in the arbitration phase, the backoff window will enter the next stage. The fifth equation account for the fact that, if the backoff stage reaches the maximum value m, it does not increase further.

The closed-form solution for this Markov chain is obtained based on Eq. (1). First, the relationship between $P\{0, 0\}$ and $P\{i, 0\}$ is

$$\begin{aligned} P\{i-1, 0\} \cdot p &= P\{i, 0\}, \\ P\{i, 0\} &= p^i P\{0, 0\}, \quad 0 < i < m. \end{aligned} \tag{2}$$

For $i = m$, we have

$$\begin{aligned} P\{m-1, 0\} \cdot p &= (1-p) P\{m, 0\}, \\ P\{m, 0\} &= \frac{p^m P\{0, 0\}}{1-p}. \end{aligned} \tag{3}$$

Based on chain regularities and the relations in Eq. (2) and Eq. (3), the relation between $P\{i, k\}$ and $P\{i, 0\}$ is

$$P\{i, k\} = \frac{\left\lceil \frac{W_i - k}{N_{RU}} \right\rceil}{W_i} P\{i, 0\}, \quad i \in [0, m], \ k \in [1, W_i - 1], \tag{4}$$

where $\lceil \cdot \rceil$ is the ceiling function. Based on the relations in Eq. (2), Eq. (3), and Eq. (4), all the values of $P\{i, k\}$ can be expressed as functions of $P\{0, 0\}$, and the failure probability p. By the normalization condition, $P\{0, 0\}$ is obtained as

$$\begin{aligned} 1 &= \sum_{i=0}^{m} \sum_{k=0}^{W_i - 1} P\{i, k\} \\ &= \sum_{i=0}^{m-1} p^i P\{0, 0\} \left(\sum_{k=1}^{W_i - 1} \frac{\left\lceil \frac{W_i - k}{N_{RU}} \right\rceil}{W_i} + 1 \right) \\ &\quad + \frac{p^m}{1-p} P\{0, 0\} \left(\sum_{k=1}^{W_m - 1} \frac{\left\lceil \frac{W_m - k}{N_{RU}} \right\rceil}{W_m} + 1 \right), \end{aligned} \tag{5}$$

$$\begin{aligned} P\{0, 0\} = \Bigg[&\sum_{i=0}^{m-1} p^i \left(\sum_{k=1}^{W_i - 1} \frac{\left\lceil \frac{W_i - k}{N_{RU}} \right\rceil}{W_i} + 1 \right) \\ &+ \frac{p^m}{1-p} \left(\sum_{k=1}^{W_m - 1} \frac{\left\lceil \frac{W_m - k}{N_{RU}} \right\rceil}{W_m} + 1 \right) \Bigg]^{-1}. \end{aligned} \tag{6}$$

The transmission probability τ is given as

$$\tau = \sum_{i=0}^{m} P\{i, 0\} = \frac{P\{0, 0\}}{1-p}. \tag{7}$$

However, the failure probability p is still unknown. To obtain p, the channel access procedure is analyzed.

3) Failure Probability From Channel Access Procedure: The stations whose backoff counters are zero randomly select an RU and enter the arbitration phase. For a given RU, the probability that it is selected by a generic station is $\frac{\tau}{N_{RU}}$. The number of stations that select the RU is denoted by N. The distribution of N is

$$P\{N = n\} = \binom{N_{STA}}{n} \left(\frac{\tau}{N_{RU}} \right)^n \left(1 - \frac{\tau}{N_{RU}} \right)^{N_{STA} - n}. \tag{8}$$

In the arbitration phase, the stations randomly choose an arbitration number from $[0, 2^{N_{AS}} - 1]$. The result of the arbitration phase is that the stations with the maximal arbitration number win the RU. N_{win} denotes the number of stations that win the RU. The event $\{N_{win} = 1\}$ denotes that the RU successfully supports a transmission, since only one station wins the RU. The probability of $\{N_{win} = 1\}$ with the condition of $\{N = n\}$ is given as

$$P\{N_{win} = 1 | N = n\} = \frac{\sum_{l=0}^{2^{N_{AS}} - 1} \binom{n}{1} l^{n-1}}{(2^{N_{AS}})^n}, \tag{9}$$

where l is the maximum number among these arbitration numbers, $\binom{n}{1} l^{n-1}$ denotes the number of events that only one station chooses l. As l traverses from 0 to $2^{N_{AS}} - 1$, the number of the events that only one station wins the RU is $\sum_{l=0}^{2^{N_{AS}} - 1} \binom{n}{1} l^{n-1}$.

Based on Eq. (8) and Eq. (9), the probability that the RU successfully supports a transmission is given as

$$P\{N_{win} = 1\} = \sum_{n=1}^{N_{STA}} P\{N_{win} = 1 | N = n\} P\{N = n\}. \tag{10}$$

Hence, the expectation of the number of successful stations on the RU (denoted by N_s) is expressed as

$$\mathbb{E}(N_s) = 1 \cdot P\{N_{win} = 1\}. \tag{11}$$

In addition, the expectation of the number of stations that select the RU is $\mathbb{E}(N) = \frac{\tau N_{STA}}{N_{RU}}$. Therefore, the success probability of a station is

$$\begin{aligned} 1-p &= \frac{\mathbb{E}(N_s)}{\mathbb{E}(N)} \\ &= \frac{N_{RU}}{\tau N_{STA}} \sum_{n=1}^{N_{STA}} \left[\frac{n \sum_{l=0}^{2^{N_{AS}}-1} l^{n-1}}{(2^{N_{AS}})^n} \right. \\ &\quad \left. \cdot \binom{N_{STA}}{n} \left(\frac{\tau}{N_{RU}}\right)^n \left(1-\frac{\tau}{N_{RU}}\right)^{N_{STA}-n} \right]. \end{aligned} \tag{12}$$

The nonlinear system represented by Eq. (7) and Eq. (12) has two unknowns τ and p, and can be solved by numerical techniques. Thus, the transmission probability τ is obtained.

B. Saturated Throughput

For channel efficiency analysis, the saturated throughput of the channel is considered. In generic UORA procedure, the transmission probability of the j-th station is τ, and its successful transmission probability is $\frac{P\{N_{win}=1\}N_{RU}}{\tau N_{STA}}$. Hence, the expected payload transmitted by the j-th station is

$$S_j = \tau \frac{P\{N_{win}=1\}N_{RU}}{\tau N_{STA}} T_{pay} R_{tx}(j), \tag{13}$$

where T_{pay} is the duration of the payload and $R_{tx}(j)$ is the physical layer (PHY) rate of the j-th station. The expected payload transmitted by all stations in a UORA transmission duration is expressed as

$$\begin{aligned} S &= \sum_{j=1}^{N_{STA}} S_j \\ &= \frac{P\{N_{win}=1\}N_{RU} T_{pay} \overline{R}_{tx}}{N_{STA}}, \end{aligned} \tag{14}$$

where $\overline{R}_{tx}$ is the average PHY rate over all the stations. The total duration of a UORA transmission (denoted by T_{all}) is given as

$$\begin{aligned} T_{all} = T_{DIFS} + T_{TF} + 2T_{SIFS} + N_{AS} \cdot T_{AS} \\ + T_{PH} + T_{pay} + T_{ACK}, \end{aligned} \tag{15}$$

where T_{DIFS} is the duration of DIFS, T_{TF} is the duration of the TF, T_{SIFS} is the duration of short interframe space (SIFS), T_{AS} is the duration of one arbitration slot, T_{PH} is the duration of PHY header, and T_{ACK} is the duration of ACK. Finally, the saturated throughput (denoted by η) is given as

$$\begin{aligned} \eta &= \frac{P\{N_{win}=1\}T_{pay}N_{RU}\overline{R}_{tx}}{T_{all}} \\ &= \frac{T_{pay}N_{RU}\overline{R}_{tx}}{T_{all}} \sum_{n=1}^{N_{STA}} \left[\frac{n \sum_{l=0}^{2^{N_{AS}}-1} l^{n-1}}{(2^{N_{AS}})^n} \right. \\ &\quad \left. \cdot \binom{N_{STA}}{n} \left(\frac{\tau}{N_{RU}}\right)^n \left(1-\frac{\tau}{N_{RU}}\right)^{N_{STA}-n} \right]. \end{aligned} \tag{16}$$

It is shown that the saturated throughput is a function of τ, where τ can be obtained via Eq. (7) and Eq. (12).

Moreover, if the number of arbitration slots is zero ($N_{AS} = 0$), the random access scheme converts to the conventional UORA. The saturated throughput in the conventional UORA (denoted by η_c) is expressed as

$$\begin{aligned} \eta_c &= \frac{P\{N=1\}T_{pay}N_{RU}\overline{R}_{tx}}{T_{all}} \\ &= \frac{T_{pay}N_{RU}\overline{R}_{tx}N_{STA}}{T_{all}} \left(\frac{\tau}{N_{RU}}\right)\left(1-\frac{\tau}{N_{RU}}\right)^{N_{STA}-1}. \end{aligned} \tag{17}$$

C. Access Delay

In addition to the throughput, the access delay of stations is also important to evaluate network performance. In this subsection, the access delay of stations is derived. Formally, the *access delay* is defined as the duration from the time when a station has data to send to the time when it successfully accesses the channel. The derivation of the access delay contains two steps. Firstly, the expected backoff duration for each backoff stage is derived. Secondly, the expected access delay is obtained via the weighted summation of the expected backoff durations in a generic backoff procedure.

1) The Backoff Duration for Each Backoff Stage: In a backoff procedure, once receiving a TF for random access, the stations decrease their backoff counters by N_{RU}. The backoff duration in the i-th backoff stage is denoted by D_i, which is given as

$$D_i = \left\lceil \frac{k_0}{N_{RU}} \right\rceil T_I, \quad k_0 \in [0, W_i - 1], \quad i \in [0, m], \tag{18}$$

where k_0 is the initial backoff counter when a station enters the i-th backoff stage, and T_I is the time interval between two adjacent TFs for random access.

Since k_0 is uniformly distributed on $[0, W_i - 1]$, the expectation of the backoff duration in the i-th stage is

$$\mathbb{E}\{D_i\} = \sum_{k_0=0}^{W_i-1} \left\lceil \frac{k_0}{N_{RU}} \right\rceil \frac{T_I}{W_i}, \quad i \in [0, m]. \tag{19}$$

2) The Access Delay in a Generic Backoff Procedure: Assume that to successfully transmit a frame, a station fails (quits or collides) for f times. The average delay of the station (denoted by $D(f)$) is given as

$$\mathbb{E}\{D(f)\} = \begin{cases} \sum_{i=0}^{f} \mathbb{E}\{D_i\}, & f \in [0, m] \\ \sum_{i=0}^{m} \mathbb{E}\{D_i\} + (f-m)\mathbb{E}\{D_m\}, & f > m. \end{cases} \tag{20}$$

In addition, the probability of the event that the station fails for f times is $(1-p)p^f$. Thus, the expectation of access delay

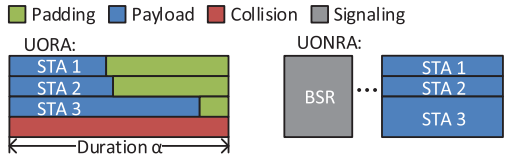


Fig. 8. The padding and data frame collisions degrade UORA. The BSR phase causes an extra overhead in UONRA.

(denoted by D_a) is obtained as

$$\begin{aligned}\mathbb{E}\{D_a\} &= \sum_{f=0}^{\infty}(1-p)p^f\mathbb{E}\{D(f)\}\\ &= \sum_{f=0}^{m}(1-p)p^f\sum_{i=0}^{f}\sum_{k_0=0}^{W_i-1}\left\lceil\frac{k_0}{N_{RU}}\right\rceil\frac{T_I}{W_i}\\ &\quad+\sum_{f=m+1}^{\infty}(1-p)p^f\sum_{i=0}^{m}\sum_{k_0=0}^{W_i-1}\left\lceil\frac{k_0}{N_{RU}}\right\rceil\frac{T_I}{W_i}\\ &\quad+\sum_{f=m+1}^{\infty}(1-p)p^f\sum_{k_0=0}^{W_m-1}\left\lceil\frac{k_0}{N_{RU}}\right\rceil\frac{T_I}{W_m}. \quad (21)\end{aligned}$$

In conclusion, the theoretical results about the saturated network throughput and the access delay can be obtained by Eq. (16) and Eq. (21), respectively. In addition, the theoretical analysis will be validated via simulations in Section VI.

IV. Dynamic Selection Between UORA and UONRA

In UORA, data frame collisions cannot be completely avoided. Hence, UORA is only effective for frame transmissions with a relatively short length. In addition, the padding problem also degrades the performance of UORA. As shown in Fig. 8, the transmission duration (denoted by α) in UORA is determined by the AP [4]. If the frame of a station is shorter than α, the station pads '0' in the frame to make its duration reach α. However, the concurrently transmitted frames are not necessarily identical, the padding bits cannot be avoided. The overhead caused by padding depends on the frame length distribution in the network (including an AP and all stations). The larger the variance of frame lengths is, the larger the overhead is induced by padding. Therefore, the padding problem and frame collisions degrade the performance of UORA. In UONRA, data frame collisions can be avoided, and the overhead caused by padding can be minimized by the frame scheduling at the AP. However, the BSR transmission phase leads to overhead. Thus, for short frame transmissions, UONRA is not effective. Instead, it prefers long frame transmissions to compensate for the overhead due to the BSR transmission phase.

As a result, there exists a frame duration boundary between UORA and UONRA (denoted by α_0). If the frame duration is less than α_0, UORA is more efficient than UONRA. Otherwise, UONRA is more efficient than UORA. To this end, a dynamic access-method selection (DAMS) algorithm is developed in this section.

In the DAMS algorithm, the AP estimates the optimal boundary to maximize the network throughput and then broadcasts it to the stations. A station dynamically chooses an access method based on the duration of the frame to be sent and the optimal boundary. If the frame duration is shorter than the boundary, the station chooses UORA; Otherwise, it chooses UONRA. In what follows, the optimal boundary is derived, and then the DAMS algorithm is designed.

A. The Optimal Boundary Between UORA and UONRA

The derivation of the optimal boundary comprises two steps. First, the relation between the frame duration boundary and the average network throughput (both UORA and UONRA) is derived. After that, the optimal boundary is obtained by maximizing the average network throughput.

Assume the payload transmitted by a station is a random variable x. The range of x is $[L_{min}, L_{max}]$, where L_{min} is the minimum frame length, and L_{max} is the maximum frame length in the network. The probability density function (PDF) of x is $f(x)$. Given a certain frame duration boundary α_0, if the frame length of the j-th station is less than $\alpha_0 R_{tx}(j)$, it chooses UORA. Otherwise, it chooses UONRA. Consider the scenario where there are N_{STA} active stations in the network. To send out all the frames in the network, the required duration in UORA and that in UONRA are T_{RA} and T_{NRA}, respectively.

In UORA, the expected number of stations that choose UORA is $\sum_{j=1}^{N_{STA}} P\{x < \alpha_0 R_{tx}(j)\}$. The expected number of successful transmissions in a UORA transmission is $N_{RU}P\{N_{win} = 1\}$. Moreover, the duration of a UORA is $(T_{DIFS} + T_{TF} + 2T_{SIFS} + N_{AS}T_{AS} + T_{PH} + \alpha_0 + T_{ACK})$. Hence, to send all the frames that select UORA, the required duration T_{RA} is derived as

$$\begin{aligned}T_{RA} = &\frac{\sum_{j=1}^{N_{STA}} P\{x < \alpha_0 R_{tx}(j)\}}{N_{RU}P\{N_{win}=1\}}(T_{DIFS} + T_{TF}\\ &+2T_{SIFS} + N_{AS}T_{AS} + T_{PH} + \alpha_0 + T_{ACK}). \quad (22)\end{aligned}$$

In UONRA, a station firstly transmits its BSR, and afterwards sends its frame in the following scheduled resources. Since various high efficient resource allocation schemes in wireless networks have been developed, such as [19], [20], the padding part of the PPDU in UONRA is not considered in this paper. The expected number of stations that choose UONRA is $\sum_{j=1}^{N_{STA}} P\{x > \alpha_0 R_{tx}(j)\}$. The expected number of stations that successfully send their BSRs is $N_{RU}P\{N_{win} = 1\}$. To send all the frames that select UONRA, the required duration T_{NRA} is expressed as

$$\begin{aligned}T_{NRA} = &(T_{DIFS} + T_{TF} + 2T_{SIFS} + T_{BSR} + N_{AS}T_{AS}\\ &+T_{PH} + T_{ACK})\frac{\sum_{j=1}^{N_{STA}} P\{x > \alpha_0 R_{tx}(j)\}}{N_{RU}P\{N_{win}=1\}}\\ &+\sum_{j=1}^{N_{STA}}\frac{\int_{\alpha_0 R_{tx}(j)}^{L_{max}} xf(x)\mathrm{d}x}{N_{RU}R_{tx}(j)}, \quad (23)\end{aligned}$$

where T_{BSR} is the duration of the BSR transmission phase. In Eq. (23), $\int_{\alpha_0 R_{tx}(j)}^{L_{max}} xf(x)\mathrm{d}x$ is the expectation of the frame length for the j-th station (that chooses UONRA). As shown in Fig. 8, the AP can allocate all the RUs to

the stations. The duration of the frame transmission phase is $\sum_{j=1}^{N_{STA}} \frac{\int_{\alpha_0 R_{tx}(j)}^{L_{max}} x f(x)\mathrm{d}x}{N_{RU} R_{tx}(j)}$.

The average throughput of the network (denoted by S_0) is given as

$$S_0 = \frac{N_{STA}\mathbb{E}\{x\}}{T_{RA} + T_{NRA}}. \tag{24}$$

Since the numerator term in Eq. (24) is independent of α_0, maximizing S_0 is equivalent to minimizing the denominator term (denoted by T_0). Hence, the optimal frame duration boundary (denoted by α_0^*) is obtained as

$$\alpha_0^* = \arg\min T_0. \tag{25}$$

Based on Eq. (22) and Eq. (23), T_0 is expressed as

$$\begin{aligned} T_0 &= T_{RA} + T_{NRA} \\ &= \frac{\sum_{j=1}^{N_{STA}} (\alpha_0 P\{x < \alpha_0 R_{tx}(j)\} + T_{BSR} P\{x > \alpha_0 R_{tx}(j)\})}{N_{RU} P\{N_{win} = 1\}} \\ &\quad + \frac{T_{DIFS} + T_{TF} + 2T_{SIFS} + N_{AS}T_{AS} + T_{PH} + T_{ACK}}{N_{RU} P\{N_{win} = 1\}} \\ &\quad + \sum_{j=1}^{N_{STA}} \frac{\int_{\alpha_0 R_{tx}(j)}^{L_{max}} x f(x)\mathrm{d}x}{N_{RU} R_{tx}(j)}. \end{aligned} \tag{26}$$

To determine the range of α_0^*, the derivative of T_0 with respect to α_0 is given as

$$\begin{aligned} &\frac{\partial T_0}{\partial \alpha_0}(\alpha_0) \\ &= \frac{\sum_{j=1}^{N_{STA}} [\alpha_0 (1 - P\{N_{win} = 1\}) - T_{BSR}] f(\alpha_0 R_{tx}(j))}{N_{RU} P\{N_{win} = 1\}} \\ &\quad + \frac{\sum_{j=1}^{N_{STA}} \left\{ \int_{\alpha_0 R_{tx}(j)}^{L_{max}} f(x)\mathrm{d}x \right\}}{N_{RU} P\{N_{win} = 1\}}. \end{aligned} \tag{27}$$

Since α_0^* is a minimum point of T_0, $\frac{\partial T_0}{\partial \alpha_0}(\alpha_0^*) = 0$. As the second term on the right side of Eq. (27) is greater than zero, we have

$$\alpha_0^*(1 - P\{N_{win} = 1\}) - T_{BSR} < 0.$$

In other words, the range of α_0^* lies in the following set:

$$\zeta = \left\{ \alpha_0^* : 0 < \alpha_0^* < \frac{T_{BSR}}{1 - P\{N_{win} = 1\}} \right\}. \tag{28}$$

Hence,

$$\alpha_0^* = \arg \min_{\alpha_0 \in \zeta} \{T_0(\alpha_0)\}. \tag{29}$$

α_0^* can be calculated based on Eq. (29). In an IEEE 802.11ax network, the duration of BSR phase T_{BSR} is 276 μs, and $P\{N_{win} = 1\}$ is typically within 0.9 (with four arbitration slots), so α_0^* is within 2760 μs.

B. Dynamic Access-Method Selection Algorithm

To obtain α_0, the AP needs to estimate the success transmission probability $P\{N_{win} = 1\}$, the frame length distribution in the network $f(x)$, and the PHY rate of each station $R_{tx}(j), j \in [1, N_{STA}]$. $P\{N_{win} = 1\}$ can be calculated by Eq. (10) in Section III-A, and it can also be obtained by recording the number of successful transmissions N_{st} and the total number of transmissions N_{total} in each RU as

$$\hat{P}\{N_{win} = 1\} = \frac{N_{st}}{N_{total}}. \tag{30}$$

To estimate $f(x)$, the AP can record the length of the received frames and obtain the statistical estimation $\hat{f}(x)$. The PHY rate of each station can be obtained by recording the PHY rate of the station in the previous transmissions.

With the knowledge of $\hat{P}\{N_{win} = 1\}$, $\hat{f}(x)$, and $R_{tx}(j)$, $j \in [1, N_{STA}]$, the AP can obtain α_0^* based on Eq. (29), and broadcast α_0^* to the stations. After that, the AP triggers UORA with probability of $\frac{\sum_{j=1}^{N_{STA}} \int_{L_{min}}^{\alpha_0^* R_{tx}(j)} \hat{f}(x)\mathrm{d}x}{N_{STA}}$, and triggers UONRA with probability of $\frac{\sum_{j=1}^{N_{STA}} \int_{\alpha_0^* R_{tx}(j)}^{L_{max}} \hat{f}(x)\mathrm{d}x}{N_{STA}}$. For a station, if its frame duration is shorter than α_0^*, the station chooses UORA. Otherwise, the station chooses UONRA.

Note that there is another option for stations in uplink: If a station has frames accumulated in the buffer, the station can piggyback its BSR in its uplink data frames and afterwards sends its following data frames via UONRA. In this case, it is unnecessary to use DAMS.

MBTA and DAMS are independent mechanisms. They can work together to improve network performance. MBTA improves the efficiency of UORA by reducing the collisions among stations, while DAMS guides stations to choose the optimal access strategy.

V. Simulation Setup

To validate theoretical analysis and evaluate the performance of the MBTA mechanism and the DAMS algorithm, three scenarios of simulations are carried out: 1) UORA only; 2) UONRA only (but BSRs are sent using UORA); 3) Dynamic selection between UORA and UONRA. The simulations are conducted in MATLAB platform.

A. PHY Channel Model

The radius of the simulated network is set to 50 m. An AP is located in the center point of the network and 50 to 200 stations are uniformly distributed in the network. The channel between each station and the AP is modeled by the channel model in a large indoor environment [21]. In this model, the path loss between a transmitter and a receiver (denoted by $PL(d)$) is given as

$$PL(d) = \begin{cases} 32.4 + 20\log_{10} f_{GHz} + 20\log_{10} d, & d \le 5\mathrm{m} \\ 46.4 + 20\log_{10} f_{GHz} + 35\log_{10}\left(\frac{d}{5}\right), & d > 5\mathrm{m}, \end{cases} \tag{31}$$

where d is the distance between the transmitter and the receiver in meter, and f_{GHz} is the carrier frequency in GHz. In this

TABLE II
PHY RATE ON A 26-TONE RU

Index	Modulation	Coding rate	Sensitivity	PHY rate
0	BPSK	1/2	-91 dBm	0.8 Mbps
1	QPSK	1/2	-88 dBm	1.5 Mbps
2	QPSK	3/4	-86 dBm	2.3 Mbps
3	16-QAM	1/2	-83 dBm	3.0 Mbps
4	16-QAM	3/4	-79 dBm	4.5 Mbps
5	64-QAM	2/3	-75 dBm	6.0 Mbps
6	64-QAM	3/4	-74 dBm	6.8 Mbps
7	64-QAM	5/6	-73 dBm	7.5 Mbps
8	256-QAM	3/4	-68 dBm	9.0 Mbps
9	256-QAM	5/6	-66 dBm	10.0 Mbps

TABLE III
SYSTEM PARAMETERS

Parameters	Value	Parameters	Value
N_{STA}	50 - 200	W_0	16
Number of APs	1	W_m	1024
Network radius	50 m	T_{AS}	29.6 μs
Carrier frequency	5.2 GHz	N_{AS}	0 - 4
TX power of payload	5.5 dBm	T_{PH}	56 μs
TX power of busy-tone	12 dBm	T_I	10 ms
Sensing sensitivity	-93 dBm	T_{SIFS}	16 μs
N_{RU} in UORA	18 (40 MHz)	T_{DIFS}	34 μs
N_{RU} in UONRA	37 (80 MHz)	T_{TF}	70 μs
Frame duration in UORA	1 ms	T_{ACK}	60 μs
TCP layer MSS	576 bytes	T_{BSR}	80 μs
Backoff slot in CSMA/CA	9 μs		

paper, the carrier frequency is set to 5.2 GHz. The network bandwidth is 40 MHz. The transmit power of each station is 18 dBm. As a 40 MHz channel contains 18 RUs, the transmit power of payload signals in an RU is 5.5 dBm. To extend the sensing range of the stations in MBTA, the transmit power of busy-tone signals is set to 12 dBm, and the minimum sensing threshold of each station is set to -93 dBm. Based on the channel model in Eq. (31), the received power at the AP can be obtained. The PHY rate of each station can be determined via the map between the receiving power and the PHY rate, as shown in Table II. In each simulation scenario, 50 rounds of simulations are carried out. In each round, the stations are uniformly distributed in the network. The average throughput and the average access delay over all rounds of simulations are collected.

B. General Setup in Medium Access Control (MAC) Layer

In the MAC layer, only uplink transmission is simulated, as the downlink transmissions have no direct impact on either UORA or UONRA. The simulated network is always saturated to evaluate the network capability, i.e., the packet generation rate is larger than the packet transmission rate for the overall network. The general system parameters are summarized in Table III. The number of stations varies from 50 to 200. All the stations are served by one AP, except the simulation for overlapping basic service set (OBSS) scenarios in Section V-C. The number of arbitration slots N_{AS} varies from 0 to 4, where $N_{AS} = 0$ denotes the conventional access scheme. The maximum contention window and the minimum contention window are set to 1024 and 16, respectively. The arbitration slot is set to 29.6 μs based on the design in Section II-B. According to the specification in IEEE 802.11ax standard [4], TF is set to 70 μs, ACK is set to 60 μs, BSR frame is set to 80 μs, and PHY header is set to 56 μs. Moreover, detailed setups for various simulation scenarios are elaborated as follows.

C. UORA

In the UORA simulation system, the number of RUs for random access is 18 (40 MHz). The frame duration is 1 ms, and the frame generation interval follows the exponential distribution. The average packet generation rate is 600 frames per second. The network throughput and access delay of the stations are collected.

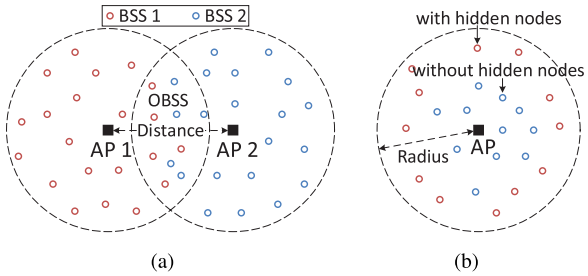


Fig. 9. (a) An example of network with OBSS; (b) An example of network with hidden nodes.

In addition, the simulations in many meaningful OBSS scenarios are carried out. As shown in Fig. 9(a), there are two basic service sets (BSS) with 400 stations in the network. An overlapping area is between these two BSSs, i,e, OBSS. In each BSS, an AP is located in the center of the BSS area. In the simulations, the distance between these two APs varies from 50 m to 100 m to represent the different sizes of OBSS areas. APs follow the CSMA/CA protocol to transmit TFs.

Moreover, many scenarios with hidden nodes are also simulated, as shown in Fig. 9. 200 stations are uniformly distributed in the network. The network radius varies from 50 m to 70 m to represent various hidden-nodes scenarios.

Besides, the scenarios where stations have a small size of data in their buffers are simulated. Such scenarios correspond to the non-saturated traffic load in an IEEE 802.11ax network, and are suited to UORA.

D. UONRA

In the UONRA simulation system, the TF for random access is sent every 10 ms to solicit the BSR of the stations in a 40 MHz channel (18 RUs) as suggested in [17]. The AP allocates an 80 MHz channel resource (37 RUs) to the stations based on their BSRs. Since various high efficient resource allocation schemes in wireless networks have been developed, the padding bits are not considered. In a communication system, the transport layer exchanging packets with the MAC layer can be based on a UDP or TCP protocol. Both UDP

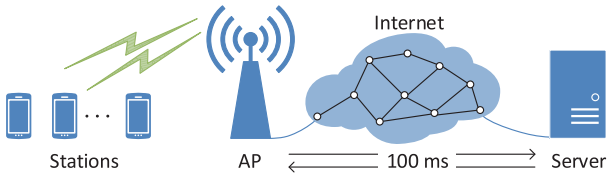


Fig. 10. The simulation scenario for TCP traffic.

traffic and the TCP traffic are considered when evaluating UONRA.

1) The UDP Traffic: In the UDP traffic scenario, the frame generation interval follows the exponential distribution. The frame length is 1500 bytes, which is a suitable choice for UONRA according to the DAMS algorithm. The access delay of the stations is collected.

2) The TCP Traffic: The simulation scenario of TCP traffic is shown in Fig. 10. The AP connects with the Internet server via a wired network. Normally, the round trip time between the AP and Internet server is less than 100 ms. Thus, the round trip time between the AP and Internet server is set to 100 ms in the simulation. The TCP layer delivers the data segments to the MAC layer based on the congestion control algorithm TCP Tahoe [22]. The maximum segment size (MSS) is set to 576 bytes. The buffer of each station is large enough so that its transmission window is only limited by its congestion window.

E. Dynamic Selection Between UORA and UONRA

To evaluate the performance of the DAMS algorithm, three scenarios (i.e., UORA only, UONRA only, and DAMS algorithm) are compared. The simulation system works in a 40 MHz channel. In addition, the number of arbitration slots is set to 0 and 4 to investigate the impact of MBTA.

In the UORA only scenario, the stations only choose UORA. The AP only triggers UORA and sets the PPDU duration as the maximum duration of the frames from the stations. In the UONRA only scenario, the stations only choose UONRA to send data. The AP only triggers UONRA (the BSRs are sent using UORA, the data frames are sent using UONRA). In the DAMS algorithm scenario, the station chooses UORA or UONRA based on the optimal boundary that is estimated by AP. The AP schedules UORA and UONRA based on the DAMS algorithm in Section IV-B.

As the optimal boundary also depends on frame length distributions, five typical frame length distributions are simulated:

- Distribution 1: 40% of the frames are 40 bytes, and 40% of the frames are 1500 bytes. The remaining 20% frames are uniformly distributed in $[40, 1500]$.
- Distribution 2: The frame length is fixed at 40 bytes.
- Distribution 3: The frame length follows the truncated normal distribution $N(500, 300^2)$. The range of the frame length is between 40 bytes and 1500 bytes.
- Distribution 4: The frame length is fixed at 500 bytes.
- Distribution 5: The frame length is fixed at 1500 bytes.

The reasons for applying the above distributions are as follows. Distribution 1 is a typical packet length distribution in the Internet [23]. Hence, distribution 1 is carried out to evaluate the performance of DAMS for Internet traffic. From distribution 2 to distribution 5, these distributions are to model the typical traffic distributions that can influence the performance of UORA and UONRA in a different way. For UORA, we have two conclusions: 1) it is suitable for short frames; 2) it is degraded by the paddings caused by unequal-length frames. To validate the first conclusion, distribution 2 is carried out to model burst traffic with short frames. To verify the second conclusion, distribution 3 and distribution 4 are carried out. The means of both distribution 3 and distribution 4 are 500 bytes. The variance of distribution 3 is 300 bytes while the variation of distribution 4 is 0. For UONRA, it is suitable for long frames. To verify this, distribution 5 is carried out to model burst traffic with long frames.

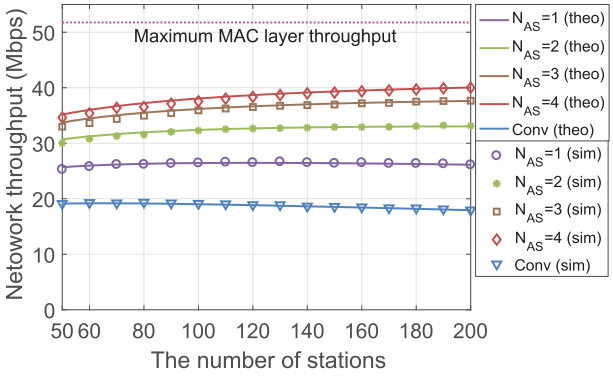


Fig. 11. The average network throughput.

VI. Performance Results

A. Validation of Theoretical Analysis

1) Saturated Throughput: The saturated network throughput with a different number of arbitration slots is shown in Fig. 11. The theoretical results of the saturated throughput are indicated by solid lines, while simulation results are represented by marks of different shapes. It can be observed that the theoretical results match the simulation results really well, which confirms the validity of the theoretical analysis.

2) Access Delay: The average access delays of the stations with different N_{AS} in the UORA are shown in Fig. 12, where the theoretical results from Eq. (21) are presented by solid lines and the simulation results are indicated by marks of different shapes. Comparisons illustrate that theoretical analysis for access delay has properly captured the real behaviour of the protocol.

B. Uplink OFDMA Random Access

In UORA, the performance of the proposed mechanism is evaluated in two aspects: the saturated network throughput and the access delay of the stations.

1) The Saturated Throughput: The saturated throughput is shown in Fig. 11. It is shown that the proposed mechanism improves the saturated throughput of the RU by 110%, compared with the conventional access scheme. The reason is that

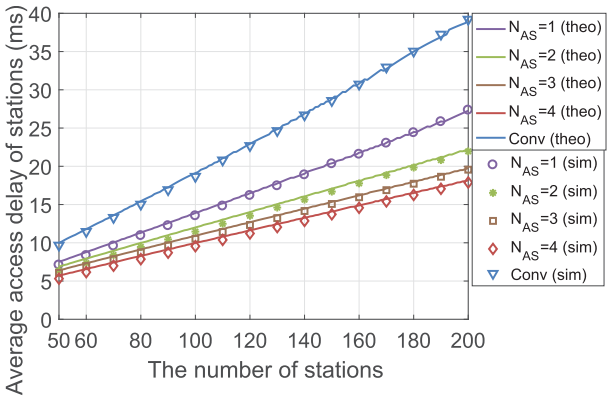


Fig. 12. The average access delay of the stations in UORA.

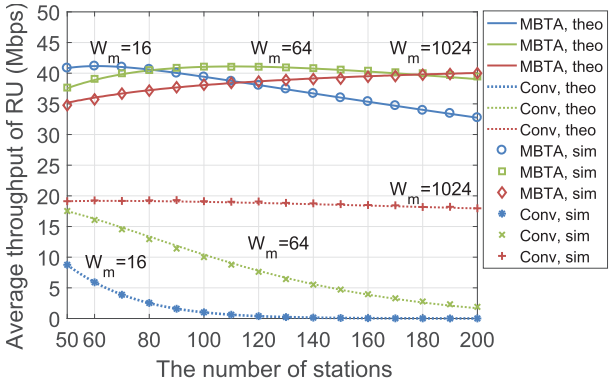


Fig. 13. The average network throughput with different W_m values.

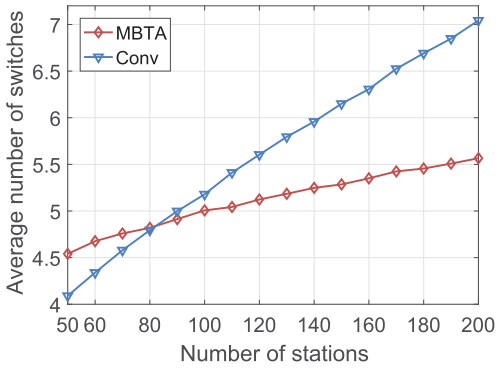


Fig. 14. The average number of switches needed for a successful transmission in different schemes.

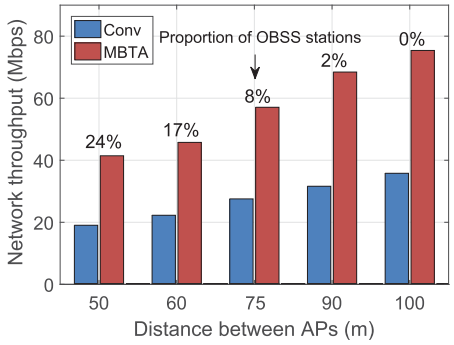


Fig. 15. The average network throughput in different OBSS scenarios.

the MBTA mechanism can significantly reduce the number of collisions.

In addition, once N_{AS} reaches 4, the saturated throughput increases slowly with the increase of N_{AS}. The reason is that, in this case, the number of collisions in the network is already low, and the overhead caused by the arbitration phase increases with the increase of N_{AS}. Hence, there is an optimal N_{AS} for this system, which is $N_{AS} = 4$.

Moreover, the maximum MAC layer throughput is shown by the dashed line in Fig. 11. The *maximum MAC layer throughput* is the throughput in the ideal case where there is neither collision nor extra overhead caused by MBTA. It is shown that the saturated throughput in MBTA is actually lower than the maximum MAC layer throughput, due to the extra overhead in the arbitration phase.

2) The Access Delay of the Stations: The access delay of stations is shown in Fig. 12. Compared with the conventional access scheme, MBTA reduces the access delay of the stations by approximately 50%. The reason is that MBTA can greatly reduce the collision probability of the stations, and thus reduces the number of retransmissions of the stations.

In addition, if N_{AS} reaches 4, the collision probability of stations is already low. At the same time, the extra overhead caused by the arbitration phase increases as N_{AS} increases. Hence, the access delay cannot be effectively reduced further, once N_{AS} reaches 4.

3) The Influence of OFDMA Contention Window: The influence of the OFDMA contention window size is investigated. The maximum OFDMA contention window size W_m is set to 16, 64, and 1024. The results are shown in Fig. 13. It can be observed that the gain in the scenario where there are 50 stations with $W_m = 16$ is as significant as that in the scenario where there are 200 stations with $W_m = 1024$.

4) The Complexity of MBTA: We have collected the average number of TX/RX switch operations for each successful transmission in simulations. The result is shown in Fig. 14. It can be observed that MBTA can reduce the average number of RX/TX switches in dense scenarios. The reason is that the number of transmission failures is greatly reduced by MBTA.

5) The OBSS Scenarios: The simulation results for OBSS scenarios are shown in Fig. 15. It can be observed that the network throughput decreases as the proportion of OBSS stations increases. The reason is that the OBSS stations cause interference to the AP, and the AP needs to wait for a clear channel to access. It should be mentioned that MBTA outperforms conventional UORA in all OBSS scenarios.

6) The Influence of Hidden Nodes: The simulation results for the scenarios with hidden nodes are shown in Fig. 16. First, it can be observed that network throughput decreases as the network radius increases. The reason is that the average PHY rate of the stations decreases as the network radius increases. Second, it is shown that the performance of MBTA degrades slightly as the number of hidden nodes increases (compared with the conventional scheme). The reason is that if two stations are hidden from each other, one station cannot hear the busy tone from the other one. Hence, these two stations

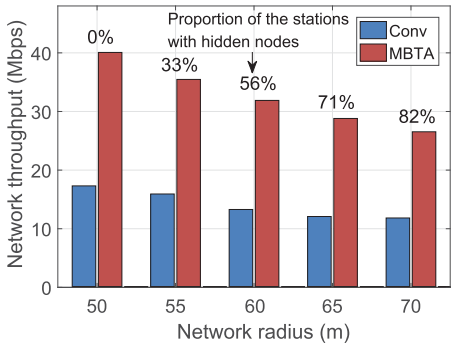


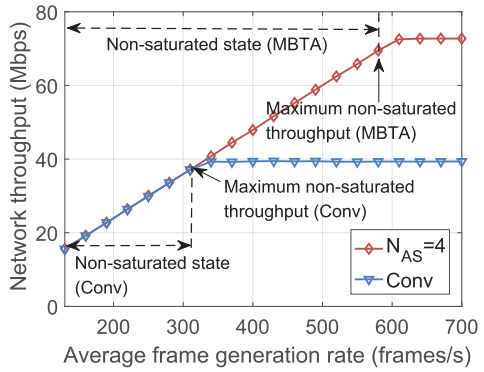
Fig. 16. The average network throughput with different network radii.

cannot resolve collisions among them via MBTA. Note that MBTA still outperforms the conventional UORA scheme when 82% stations have hidden nodes.

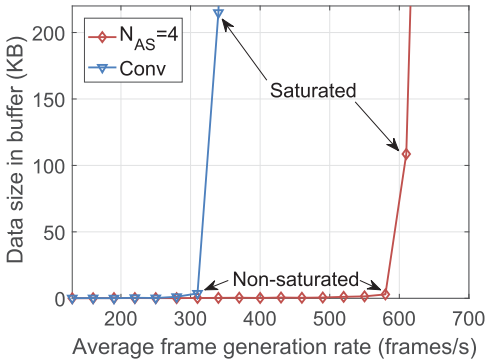
7) Non-Saturated Traffic Load: UORA is suited to the scenarios where STAs have a small size of data in their buffers. Such scenarios correspond to the non-saturated traffic load in an IEEE 802.11ax network. To evaluate UORA under this situation, simulations are carried out with the following system settings. The number of stations varies from 5 to 120. The radius of the network is 10 m, and an AP is located in the center of the network. The transmit power of a station is set to 5.5 dBm. The non-saturated traffic load is generated as follows. The number of frames generated per second at each station follows a Poisson distribution, while the frame length is fixed at 500 bytes, which is a suitable choice for UORA according to the DAMS algorithm.[1] When the frame generation rate becomes greater, the traffic load of stations also increases and eventually transitions from a non-saturated state to a saturated state. During this process, the maximum network throughput under which stations stay in the non-saturated state can be collected, and such a throughput is called *the maximum non-saturated throughput* in this paper.

Considering an example where 30 stations are served by an AP, the frame generation rate increases from 130 frames/s to 700 frames/s. The network throughput is shown in Fig. 17(a). From the results of the conventional UORA, it can be observed that, when the frame generation rate is lower than 310 frames/s, the network throughput increases linearly with the frame generation rate. The reason is that all the generated frames at the stations can be successfully delivered to the AP, and so the generated frames do not accumulate in the buffers of stations. As shown in Fig. 17(b), the average data size in a buffer is within a small value. If the traffic load further increases, then the network enters the saturated state, and the generated frames accumulate quickly in the buffers of stations, as shown in Fig. 17(b). Thus, the maximum throughput before the network enters the saturated state can be determined from the simulation results (i.e., the maximum non-saturated throughput is 37.1 Mbps corresponding to a

[1]Note that this simulation is different from the simulation of frame length distribution 4 in Section V-E, where the radius of network is 50 m and thus the frame length of 500 bytes is not suitable for UORA according to the DAMS algorithm.



(a)



(b)

Fig. 17. (a) The network throughput versus the frame generation rate; (b) The average data size in the buffer of a station.

frame generation rate of 310 frames/s). If the frame generation rate further increases, the network becomes saturated, and then the conventional UORA is not suitable for the network. In comparison, for MBTA ($N_{AS} = 4$), the network is still non-saturated even when the frame generation rate exceeds 580 frames/s (the maximum non-saturated throughput is 69.5 Mbps). Hence, the maximum non-saturated throughput is improved significantly by MBTA. It should be noted that the maximum non-saturated throughput corresponds to the maximum bursty range of data traffic, i.e., if the maximum non-saturated throughput is higher, then the network can support traffic types with larger burstiness. This property is definitely preferred by a data network. It should also be noted that, when the 802.11ax network enters saturation (i.e., data accumulates in the buffer), the BSR information can be piggybacked to the AP for scheduled transmissions. In this way, the throughput of the network can be further improved. However, a piggyback-based reservation scheme is only effective when the traffic load is maintained at a high level. Otherwise, due to the lack of data in the buffer, the BSR information cannot be piggybacked to the AP in a timely manner. In other words, a piggyback-based reservation scheme is inefficient to handle data traffic with a large range of burstiness.

Besides the above case with a fixed number of stations, the scenarios with a different number of stations (varying from 5 to 120) are simulated. The maximum non-saturated throughput is shown in Fig. 17. It can be observed that the improvement by MBTA under a small number of stations is

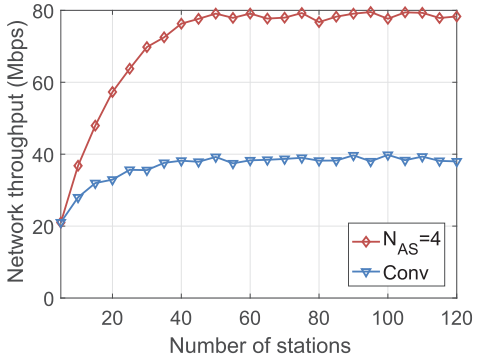


Fig. 18. The maximum non-saturated throughput versus the number of stations.

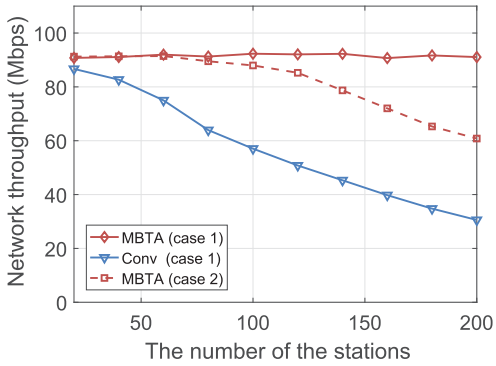


Fig. 20. The average network throughput with TCP traffic in different access schemes.

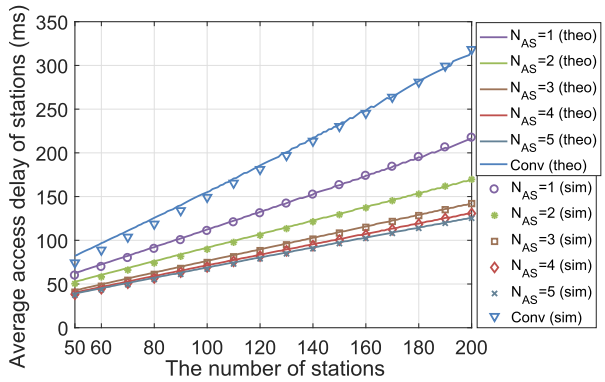


Fig. 19. The average access delay of the stations in UONRA.

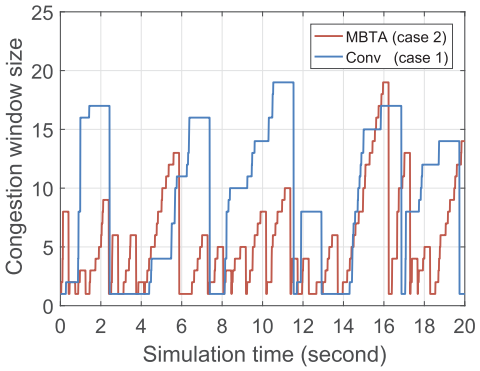


Fig. 21. The TCP congestion window of a station with different access schemes. Case 1: The same timeout threshold case; Case 2: The same timeout probability case.

not as significant as that under a large number of stations. The reason is that, given a small number of stations, the collisions among the stations are not severe.

C. Uplink OFDMA Non-Random Access

To evaluate the proposed mechanism in UONRA, the access delay of the stations and the throughput of the network with TCP traffic are analyzed.

1) The Access Delay of the Stations in UONRA: The access delay of stations in UONRA is shown in Fig. 19. It can be observed that the MBTA mechanism reduces the access delay of the stations by 60% approximately.

Moreover, the optimal N_{AS} in UONRA is larger than the optimal N_{AS} in UORA. Since the UONRA is suitable for long frame transmission, the TF interval is much larger than that in UORA. As a result, the overhead caused by the arbitration phase has little impact on access delay. Hence, the access delay is further reduced, when N_{AS} exceeds 4, as shown in Fig. 19.

2) The Throughput of the Network With TCP Traffic: In the TCP traffic simulation system, two cases are simulated. In each case, 50 geographical distributions (of stations' locations) are carried out. The average network throughput over all the 50 geographical distributions is collected. In case 1, the time-out thresholds in the networks with different access schemes are set to the same value 1.1 s. The saturated throughput of the network is shown by the solid lines in Fig. 20. These results show that the saturated throughput of the network is dramatically improved by the MBTA mechanism, especially station density increases. This feature is attributed to the reduced timeout probability by using the MBTA mechanism to significantly reduce access delay in the IEEE 802.11ax network. Thus, the average size of the congestion window of stations is greatly improved.

Moreover, the proposed scheme also improves the saturated throughput by reducing the RTT. As the RTT is reduced, the frame transmission rate is improved. To validate this, case 2 is simulated. In case 2, to remove the impact of the timeout, the timeout probability in the network with MBTA is set to the same value as that in the network with conventional UORA. To this end, the timeout threshold in the network with MBTA needs to be adjusted to 0.36 s. The saturated throughput of the networks in shown by the dashed lines in Fig. 20. Even if their timeout probabilities are the same, MBTA still achieves a higher throughput than the conventional scheme.

To a further look into the TCP behaviour, the congestion windows of a station in different schemes are shown in Fig. 21, where 200 stations exist in the network. It can be observed that the station with MBTA adjusts its congestion window much faster than that with the conventional access scheme.

D. Dynamic Access-Method Selection Algorithm

The saturated throughput of the network with different access algorithms is shown in Fig. 22. In distribution 1, the UORA is mainly degraded by padding, and the UONRA is

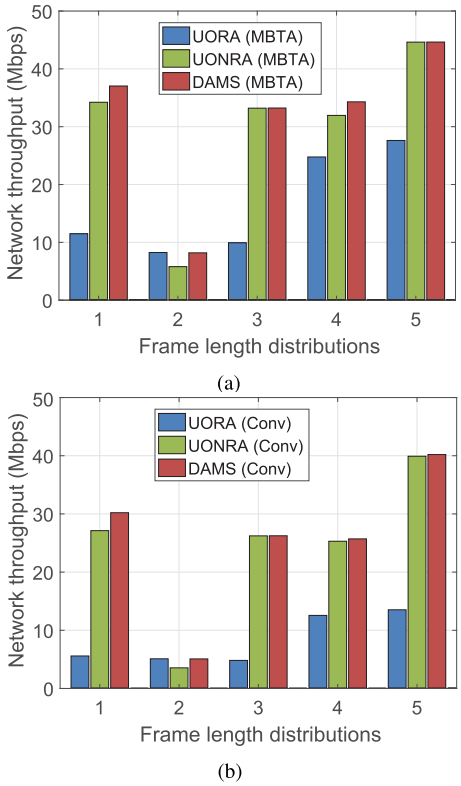


Fig. 22. Throughput of the network with different access algorithms in various frame distributions. (a) The MBTA mechanism is applied ($N_{AS} = 4$); (b) The conventional random access scheme is applied ($N_{AS} = 0$). Distribution 1: Frame length follows the bimodal distribution; Distribution 2: Frame length is fixed at 40 bytes; Distribution 3: Frame length follows truncated normal distribution $N(500, 300^2)$; Distribution 4: Frame length is fixed at 500 bytes; Distribution 5: Frame length is fixed at 1500 bytes.

degraded by the extra overhead caused by the BSR transmission. The DAMS algorithm can jointly minimize the overhead caused by padding, collision, and BSR transmission. Hence, the DAMS algorithm achieves the highest throughput.

The results in distribution 2 demonstrate that UORA is more suitable for short frames. The results in distribution 5 show that UONRA is more efficient for long frames. For both two distributions, the DAMS algorithm automatically selects the optimal access method.

The performance of UORA is degraded in distribution 3, compared with distribution 4. The reason is that the padding problem in distribution 3 is worse than that in distribution 4. In distribution 4, although all frames are 500 bytes, UORA is also degraded by padding problem. The reason is that the PHY rates of concurrent transmitting stations are different, and the transmission durations of the stations are not identical. Hence, UONRA outperforms UORA in distribution 4. For these two distributions, the DAMS algorithm achieves the highest throughput, compared with the UORA and the UONRA.

By comparing Fig. 22(a) with Fig. 22(b), it can be observed that the DAMS algorithm with the MBTA mechanism outperforms that without the MBTA mechanism.

VII. Conclusion

Study in this paper showed that the bottleneck of the existing medium access scheme in IEEE 802.11ax is rooted in the low efficiency of UORA. To solve this problem, the MBTA mechanism was developed. In this mechanism, the stations contend a resource unit in multiple arbitration slots to improve the efficiency of UORA. In this paper, a dynamic access-method selection algorithm called DAMS was also designed for the AP and stations to choose an optimal access method based on the optimal frame duration boundary between UORA and UONRA. The MBTA mechanism and the DAMS algorithm were analyzed rigorously and were also evaluated via simulations. Both theoretical and simulation results showed that the MBTA mechanism significantly improves network throughput for UORA and reduces the access delay of stations. In the case of UONRA, the reduced access delay by the MBTA mechanism dramatically improved the throughput in TCP traffic. Moreover, under various distributions of frame lengths, the DAMS algorithm was proved to be always effective to achieve the highest throughput, as compared to the case of using either UORA or UONRA only.

Acknowledgment

The authors would like to thank the editor and the anonymous reviewers for their insightful comments.

References

[1] "Cisco visual networking index: Global mobile data traffic forecast update, 2016–2021," Cisco, San Jose, CA, USA, White Paper C11-738429-00, Feb. 2017, pp. 18–22. [Online]. Available: https://www.ramonmillan.com/documentos/bibliografia/VisualNetworkingIndexGlobalMobileDataTrafficForecastUpdate2016_Cisco.pdf

[2] H. A. Omar, K. Abboud, N. Cheng, K. R. Malekshan, A. T. Gamage, and W. Zhuang, "A survey on high efficiency wireless local area networks: Next generation WiFi," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 4, pp. 2315–2344, Apr. 2016.

[3] D.-J. Deng, S.-Y. Lien, J. Lee, and K.-C. Chen, "On quality-of-service provisioning in IEEE 802.11ax WLANs," *IEEE Access*, vol. 4, pp. 6086–6104, 2016.

[4] H. Z. Robert Stacey, R. Porat, and A. Chen, *IEEE 802.11 HEW SG Proposed TGax Draft Specification (IEEE 802.11-16/0024r1)*, IEEE Standard 802.11, Mar. 2016.

[5] D.-J. Deng *et al.*, "IEEE 802.11ax: Highly efficient WLANs for intelligent information infrastructure," *IEEE Commun. Mag.*, vol. 55, no. 12, pp. 52–59, Dec. 2017.

[6] L. Lanante, H. O. T. Uwai, Y. Nagao, M. Kurosaki, and C. Ghosh, "Performance analysis of the 802.11ax UL OFDMA random access protocol in dense networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.

[7] H. Yang, D.-J. Deng, and K.-C. Chen, "Performance analysis of IEEE 802.11ax UL OFDMA-based random access mechanism," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2017, pp. 1–6.

[8] G. Naik, S. Bhattarai, and J.-M. Park, "Performance analysis of uplink multi-user OFDMA in IEEE 802.11ax," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–6.

[9] H. Kwon, H. Seo, S. Kim, and B. Gi Lee, "Generalized CSMA/CA for OFDMA systems: Protocol design, throughput analysis, and implementation issues," *IEEE Trans. Wireless Commun.*, vol. 8, no. 8, pp. 4176–4187, Aug. 2009.

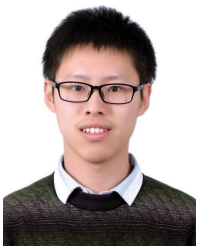
[10] X. Wang and H. Wang, "A novel random access mechanism for OFDMA wireless networks," in *Proc. IEEE Global Telecommun. Conf. GLOBECOM*, Dec. 2010, pp. 1–5.

[11] Q. Qu, B. Li, M. Yang, and Z. Yan, "An OFDMA based concurrent multiuser MAC for upcoming IEEE 802.11ax," in *Proc. IEEE Wireless Commun. Netw. Conf. Workshops (WCNCW)*, Mar. 2015, pp. 136–141.

[12] Y. P. Fallah, S. Khan, P. Nasiopoulos, and H. Alnuweiri, "Hybrid OFDMA/CSMA based medium access control for next-generation wireless LANs," in *Proc. IEEE Int. Conf. Commun.*, May 2008, pp. 2762–2768.

[13] J. Lee and C. Kim, "An efficient multiple access coordination scheme for OFDMA WLAN," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 596–599, Mar. 2017.
[14] P. Huang, X. Yang, and L. Xiao, "WiFi-BA: Choosing arbitration over backoff in high speed multicarrier wireless networks," in *Proc. IEEE INFOCOM*, Apr. 2013, pp. 1375–1383.
[15] S. Lv *et al.*, "3D pipeline contention: Asymmetric full duplex in wireless networks," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr. 2014, pp. 790–798.
[16] *FFT Megacore Function User Guide*, Altera Corp., San Jose, CA, USA, Dec. 2010, pp. 7–10.
[17] R. S. Simone Merlin, L. Cariou, and R. Porat, *IEEE P802.11 Wireless LANs: TGax Simulation Scenarios (IEEE 802.11-14/0980r16)*, Standard P802.11, Jul. 2015, pp. 24–25.
[18] 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.
[19] S.-B. Lee, I. Pefkianakis, A. Meyerson, S. Xu, and S. Lu, "Proportional fair frequency-domain packet scheduling for 3GPP LTE uplink," in *Proc. IEEE 28th Conf. Comput. Commun. (INFOCOM)*, Apr. 2009, pp. 2611–2615.
[20] D. Bankov, A. Didenko, E. Khorov, and A. Lyakhov, "OFDMA uplink scheduling in IEEE 802.11 ax networks," in *Proc. IEEE ICC*, May 2018, pp. 1–6.
[21] J. Liu *et al.*, *IEEE 802.11 ax Channel Model Document*, documents IEEE 802.11ax, Task Group Offical, 2014, pp. 5–8.
[22] R. Braden, *Requirements for Internet Hosts-Communication Layers*, document RFC 1122, Internet Engineering Task Force, 1989.
[23] R. Sinha, C. Papadopoulos, and J. Heidemann, "Internet packet size distributions: Some observations," USC/Inf. Sci. Inst., Los Angeles, CA, USA, Tech. Rep. ISI-TR-2007-643, 2007.

Dianhan Xie (Student Member, IEEE) received the B.S. degree in optoelectronic information engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2016. He is currently pursuing the Ph.D. degree with the Wireless Networking and Artificial Intelligence Laboratory, Shanghai Jiao Tong University. His current research interests include physical layer security, next-generation WiFi, and wireless sensing.



Jiawei Zhang received the B.S. degree in communication engineering from the Harbin Institute of Technology, Harbin, China, in 2016. He is currently pursuing the Ph.D. degree with the Wireless Networking and Artificial Intelligence Laboratory, Shanghai Jiao Tong University. His current research interests include edge computing and the Internet of Things.

Aimin Tang (Member, IEEE) received the B.S. and Ph.D. degrees in information and communication engineering from Shanghai Jiao Tong University, Shanghai, China, in 2013 and 2018, respectively. He was a Visiting Scholar with the University of Washington, Seattle, WA, USA, from January 2016 to November 2016. He is currently a Research Assistant Professor with the University of Michigan-Shanghai Jiao Tong University (UM-SJTU) Joint Institute, Shanghai Jiao Tong University. His current research interests include 5G networks, full-duplex communications, rateless coding, coded caching, and smart connected systems.

Xudong Wang (Fellow, IEEE) received the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology in 2003. He is currently a tenured Professor with the UM-SJTU Joint Institute, Shanghai Jiao Tong University. He is also an affiliate faculty member with the Electrical Engineering Department, University of Washington. He has been working as a Senior Research Engineer, a Senior Network Architect, and an Research and Development Manager in several companies. He has been actively involved in Research and Development, technology transfer, and commercialization of various wireless networking technologies. He holds a number of patents on wireless networking technologies and most of his inventions have been successfully transferred to products. His research interests include wireless communication networks (5G and beyond), smart connected systems, and machine learning. He is a Fellow of the IEEE Communications Society and was a Voting Member of IEEE 802.11 and 802.15 Standard Committees. He was a General Co-Chair of the 2017 IEEE 5G Summit in Shanghai and a TPC Co-Chair of the 32nd International Conference on Information Networking. He was the demo Co-Chair of the ACM International Symposium on Mobile Ad Hoc Networking and Computing (ACM MOBIHOC 2006), a Technical Program Co-Chair of Wireless Internet Conference (WICON) 2007, and a General Co-Chair of WICON 2008. He was also a Guest Editor for several international journals. He is also an editor for the IEEE TRANSACTIONS ON MOBILE COMPUTING, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, *Ad Hoc Networks* (Elsevier), and *China Communications*.