Performance Study on a CSMA/CA-Based MAC Protocol for Multi-User MIMO Wireless LANs

Shanshan Wu, *Student Member, IEEE*, Wenguang Mao, and Xudong Wang, *Senior Member, IEEE*

***Abstract*—A multi-antenna access point (AP) can communicate simultaneously with multiple clients, however, this multi-user MIMO (MU-MIMO) capability is underutilized in conventional 802.11 wireless LANs (WLANs). To address this problem, researchers have recently developed a CSMA/CA-based MAC protocol to support concurrent transmissions from different clients. In this paper, we propose an analytical model to characterize the saturation throughput and mean access delay of this CSMA/CA-based MAC protocol operating in an MU-MIMO WLAN. We also consider and model a distributed opportunistic transmission scheme, where clients are able to contend for the concurrent transmission opportunities only when their concurrent rates exceed a threshold. Comparisons with simulation results show that our analytical model provides a close estimation of the network performance. By means of the developed model, we evaluate the throughput and delay performance with respect to different network parameters, including the backoff window sizes, the number of AP's antennas, the network size, and the threshold of the opportunistic transmission scheme. Performance optimization over key parameters is also conducted for the transmission schemes.**

***Index Terms*—Multi-user MIMO, wireless LAN, saturation throughput, mean access delay, opportunistic transmission.**

I. Introduction

A MULTI-USER MIMO (MU-MIMO) wireless LAN (WLAN) contains a multi-antenna access point (AP) and multiple clients. Those clients usually have small physical sizes and limited power. Hence, each client is normally equipped with a single transmit antenna. Multiple clients can communicate concurrently with the AP in both the uplink (many-to-one) and downlink (one-to-many) [1]. With spatial multiplexing and antenna diversity, an MU-MIMO system offers a high network throughput that increases with the number of antennas at the AP. However, the distributed coordination function (DCF) in current WLANs only allows one client to transmit at a time, and hence underutilizes the MU-MIMO capability in the uplink. Moreover, a random access-based MAC protocol is highly preferred in a WLAN because it allows users to access the medium in a simple manner. Therefore, how to enable multiple clients to transmit concurrently while keeping the random access property becomes a hot topic recently [2]–[4]. In [2], Tan *et al.* develop a CSMA/CA-based MAC protocol in an MU-MIMO WLAN, which enables multiple clients to communicate with the AP concurrently. This transmission scheme is then improved in [4], where the optimal bit rate is picked for each client by considering the interference from ongoing transmissions. In [3], a CSMA/CA-based MAC protocol is developed in an MIMO network, in which nodes with more antennas can join the ongoing transmissions without interfering them.

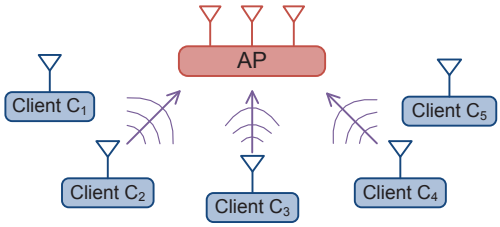


Fig. 1. A scenario of the MU-MIMO network: one three-antenna AP and five single-antenna clients.

In this paper, we summarize the key features of the CSMA/CA-based MU-MIMO WLAN proposed by [2], and then analyze its saturation throughput and mean access delay. The contributions of this paper are as follows.

First, a theoretical model is developed to characterize the saturation throughput and mean access delay of the uplink channel in a CSMA/CA-based MU-MIMO WLAN. Our derivation is based on Bianchi's Markov chain model [5], but is different from Bianchi's model in three aspects: the derivation of conditional collision probability, the formulation of saturation throughput and mean access delay. These three aspects essentially capture the difference between a conventional 802.11 MAC protocol and a CSMA/CA-based MAC in an MU-MIMO WLAN. The concurrent transmission rates are also formulated by assuming that clients experience i.i.d. time-varying Rayleigh fading.

Second, a simple distributed opportunistic scheme is considered, which allows the clients to contend for the concurrent transmission opportunity only when their concurrent data rates exceed a threshold. We also model the saturation throughput and mean access delay of this opportunistic transmission scheme by considering both MAC and PHY layer influences.

Third, comparisons between analytical and simulation results are conducted to verify our model. Numerical examples are presented to show that our analytical model provides a close estimation of the network throughput and mean access delay. The accuracy of our model is high especially when all the degrees of freedom at the AP are occupied by the

Manuscript received August 1, 2013; revised January 9, 2014; accepted March 3, 2014. The associate editor coordinating the review of this paper and approving it for publication was L. Libman.

This paper has been presented in part at the IEEE Global Communications Conference (GLOBECOM), Atlanta, USA, Dec. 2013.

The research work is supported by National Natural Science Foundation of China (NSFC) No. 61172066 and Oriental Scholar Program of Shanghai Municipal Education Commission. The authors would like to thank these sponsors for their generous support.

The authors are with the UM-SJTU Joint Institute, Shanghai Jiao Tong University, Shanghai, China. The corresponding author is Xudong Wang, (e-mail: wxudong@ieee.org).

Digital Object Identifier 10.1109/TWC.2014.131407

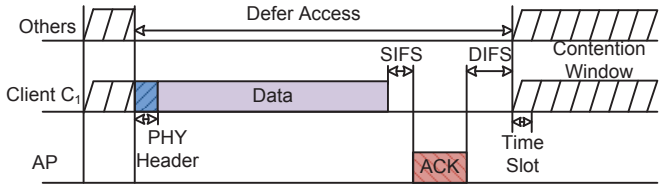


Fig. 2. Illustration of the standard 802.11 DCF access scheme.

concurrent streams.

Fourth, performance evaluation using the developed analytical model provides in-depth understanding and insights of CSMA/CA-based MU-MIMO WLAN. Specifically, the throughput and delay performance are analyzed with respect to four parameters: the transmission probability, the number of antennas at the AP, the number of clients, and the threshold in the opportunistic scheme. We show that optimal backoff window sizes can be derived to achieve the maximum throughput and the minimum access delay. Besides, we find that the throughput gain resulted from adding antennas to the AP is prominent when the total number of AP's antennas is small and the frame transmission time is long. Furthermore, for the opportunistic transmission scheme, an optimal threshold can be determined, which balances between the costs of reduced concurrent rates and increased collision probability, and the costs of decreased concurrent transmission time.

This paper is organized as follows. In Section II we describe the main features of a CSMA/CA-based MU-MIMO WLAN. In Section III a theoretical model is derived to compute the saturation throughput and mean access delay of the uplink channel. A simple distributed opportunistic scheme is introduced and modeled. In Section IV numerical analysis is carried out to validate the theoretical model. After that, variation of the model accuracy with respect to different parameters are discussed. In Section V network performance is evaluated by means of the developed model. Related work is presented in Section VI. This paper is concluded in Section VII.

II. CSMA/CA-Based MU-MIMO WLAN

In this section we summarize the main characteristics of a CSMA/CA-based MU-MIMO WLAN (see [2] for detailed descriptions). Note that we focus on the uplink throughput performance throughout this paper, because the downlink channel has been analyzed before, e.g., in [1].

In the standard 802.11 DCF access scheme, as shown in Fig. 2, only one client is allowed to transmit at a time [7]. Clients who want to transmit data enter the contention period: their *backoff counter*s are reduced each time the channel is sensed idle for a *time slot*. The client that wins the contention, i.e., has zero backoff counter, transmits data, while other clients defer their access to the channel (and stop reducing their backoff counters) until they find the medium is idle for an interval of *distributed interframe space* (DIFS). After that, a new contention period starts.

Unlike the standard 802.11 DCF access scheme, a CSMA/CA-based MU-MIMO WLAN allows multiple clients to transmit concurrently to the AP. For ease of description, let us consider a simple network with one three-antenna AP

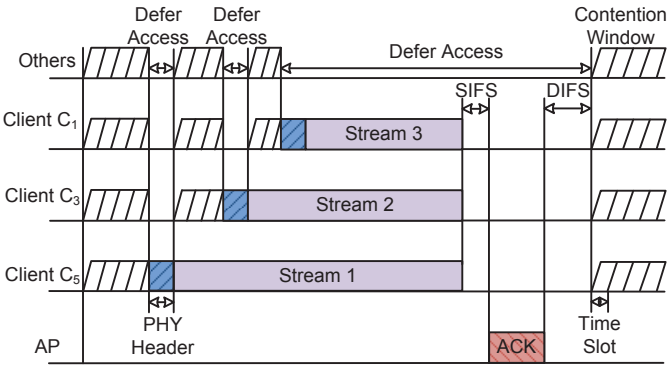


Fig. 3. Illustration of the CSMA/CA-based MU-MIMO WLAN.

and five single-antenna clients (Fig. 1). Since the AP has three antennas, the maximum number of concurrent clients cannot exceed three (assume that all the clients know this threshold through the AP's beacons). To achieve this goal, every client maintains a *transmission counter* that counts the current number of the concurrent streams by detecting their preambles[1]. If the transmission counter is smaller than the threshold (which in this case is three since the AP has three antennas), all the rest clients will continue to contend for concurrent transmission opportunities. For example, as shown in Fig. 3, when Client C_5 wins the contention and begins transmission, each of the rest clients monitors the channel and detects Client C_5's preamble[2], and then increases the transmission counter from zero to one. Since there is only one client transmitting, which is smaller than three, the rest four clients will continue to contend for the second transmission opportunity. When Client C_3 wins the second contention and transmits, the rest clients behave similarly to the previous case when Client C_5 wins the channel[3] except that the transmission counters are increased from one to two. When the transmission counter is updated to three, i.e., no concurrent transmission opportunity remains, the rest clients will defer their access to the channel until the channel is idle for an interval longer than DIFS. Assuming that the three concurrent transmissions end at the same time (which can be realized by packet fragmentation and aggregation [3] [4]), the AP will then send an ACK-to-All message to the three clients in the downlink channel through transmit beamforming [1].

A collision happens when two clients win the contention at the same time slot. As an example, consider the beginning period in Fig. 7, where both Client C_4 and Client C_5 win the second concurrent transmission opportunity. Although the rest clients can detect the preamble, they do not know that the preamble is actually two overlapping preambles. As a result,

[1]Preamble detection can be realized by correlating the received signals with the known preamble.

[2]Sometimes clients also need to decode MAC header of the first contention winner [2] before they start to compete for the concurrent transmission opportunities. However, it will not affect the following derivation process of our analytical model. Therefore, in this paper we assume that the contention period starts when previous contention winner finishes transmitting its PHY header (i.e., PLCP preamble and PLCP header), as shown in Fig. 3.

[3]Note that the rest clients will stop backoff when they detect a new preamble. Since preamble detection can be done within a slot time, they will not reduce their backoff counters once a client wins the channel and starts to transmit.

Client C_3 wins the third contention and becomes the fourth concurrent transmitter. For successful decoding, the AP needs to estimate each transmitter's channel parameters using their preambles[4]. Since the two frames of Client C_4 and Client C_5 overlap together, it is hard for the AP to nullify the interference of the two overlapped frames to extract Client C_3's frame. Besides, because the AP fails to decode the two overlapped frames, it cannot perform successive interference cancellation to extract the first contention winner's (i.e., Client C_1's) frame. In sum, the AP encounters a decoding failure when a collision happens [2]. In the case of decoding failure, no ACK message is sent to the concurrent transmitters, as shown in Fig. 7. Besides, each of the concurrent clients will select a random backoff time interval to prevent future collisions. Here we apply the binary exponential backoff mechanism which is also used in the conventional 802.11 WLANs. This backoff mechanism works as follows. Each client selects a backoff time interval from the uniform distribution over [0, CW]. CW means *contention window* and is set as $2^k - 1$, where k is a positive integer (e.g., $CW = 15$, $CW = 31$). At first, $k = k_{\min}$, $CW = CW_{\min}$, k is increased by one when a client is involved in a collision, until CW reaches $CW_{\max}$. CW is reset to $CW_{\min}$ when the client successfully transmits a packet.

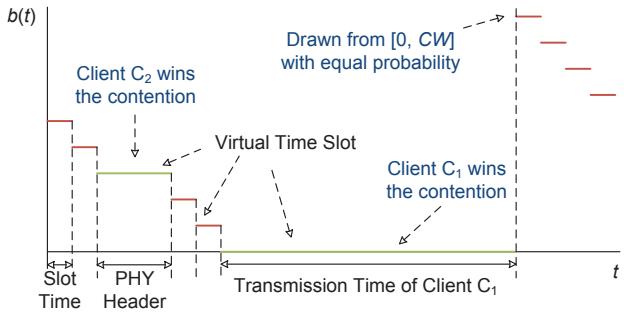


Fig. 4. Example of the stochastic process of Client C_1's backoff counter.

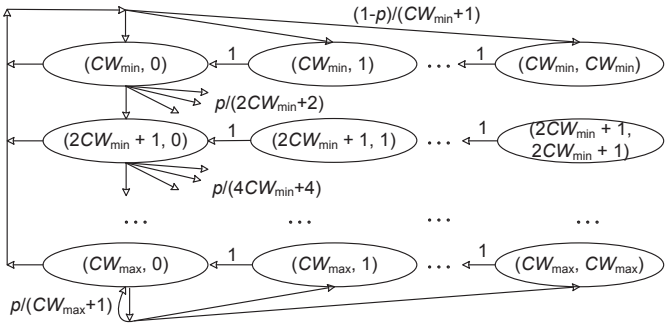


Fig. 5. Markov Chain model for the backoff counter.

III. Modeling the Uplink Channel of a CSMA/CA-Based MU-MIMO WLAN

In this section we propose an analytical model to compute the *saturation throughput* and the *mean access delay* of the uplink channel in a CSMA/CA-based MU-MIMO WLAN, under the saturation condition that each client has data to send all the time. For simplicity, only single-antenna clients are considered. Generalization to multi-antenna clients is our future work. As mentioned in the previous section, we focus on the uplink channel with no downlink transmission from AP, except ACKs.

Our analysis is based on Bianchi's Markov chain model [5]. However, Bianchi's analysis is proposed for the standard 802.11 DCF scheme, so we need to tailor it to accommodate the CSMA/CA-based MAC protocol that allows concurrent transmissions. This section consists of six subsections. In the first subsection we apply the discrete-time Markov chain model to compute the transmission probability τ of each client, which is derived as a function of the conditional collision probability p. The variable p is assumed to be constant for all the clients, and is computed in the second subsection. In the third subsection we calculate the transmission rates of the concurrent streams. The saturation throughput and mean access delay are formulated as functions of τ in the fourth and fifth subsections. An opportunistic transmission scheme is considered and modeled in the last subsection.

A. Transmission Probability

We first focus on the backoff behavior of a single client, say, Client C_1. Let $b(t)$ be the value of its backoff counter at time t, then $b(t)$ follows a stochastic process. An example of this process is shown in Fig. 4. During the contention period, $b(t)$ is reduced by one every slot time. When some client wins the contention, $b(t)$ stays in its value for a certain time interval. The length of this interval depends on who the contention winner is and whether the concurrent transmission opportunity remains. For example, if Client C_2 wins the channel and there is one more chance for concurrent transmissions, then $b(t)$ remains unchanged when Client C_2 transmits its PHY header and continues to decrease after that. Each time $b(t)$ reduces to zero, Client C_1 wins the contention and starts to transmit. After finishing transmission, Client C_1 draws a value from the uniform distribution over [0, CW] and assigns it to $b(t)$.

By regarding (CW, $b(t)$) as the states of a backoff counter, we can describe the change of (CW, $b(t)$) as a bidimensional discrete-time Markov chain. The state transition probabilities are shown in Fig. 5, where p is the conditional probability that Client C_1 encounters a failed transmission when it has won the channel. Although the MAC protocol in a CSMA/CA-based MU-MIMO WLAN is different from that in a conventional 802.11 WLAN, the state transition behavior, i.e., the Markov chain, of a client's backoff counter in the two WLANs are identical. This is because the backoff mechanism used in the two WLANs are the same. In the Markov chain, let τ be the probability that $b(t) = 0$, i.e., Client C_1 is in transmission state, then Bianchi's result can be applied here (see [5] for details):

$$\tau = \frac{2(1-2p)}{(1-2p)(W+1) + pW(1-(2p)^m)}, \qquad (1)$$

where $W = CW_{\min}+1$ and $2^m W = CW_{\max}+1$. Since clients are assumed to have packets to transmit at all times (i.e., the saturation condition), in the long term, all the clients share

[4]Readers who are interested in the detailed decoding process can refer to [2], [4], and Section 8.3 in [1].

the same transmission probability τ and conditional collision probability p. Therefore, Eq. (1) holds true not only for Client C_1 but also for all the clients in the network.

Note that Eq. (1) is derived based on the transition process of the Markov chain without considering the real time elapse. Actually, how long a client stays in its current state before jumping to the next state is different for different states, as shown in Fig. 4. Here, we give it a general name, i.e., *virtual time slot*, meaning the time interval between two consecutive states (see Fig. 4). Although the Markov models in a CSMA/CA-based MU-MIMO WLAN and a conventional 802.11 WLAN are identical, the length of virtual time slots in the two WLANs are different, which is a reason why Bianchi's method is not applicable to the derivation of conditional collision probability and saturation throughput, as illustrated in the next two subsections.

B. Conditional Collision Probability

As defined in the previous subsection, the conditional collision probability p is the probability that a client encounters transmission failure[5] given that it has won the channel. Let N be the total number of clients in a WLAN. In Bianchi's analysis, $p = 1-(1-\tau)^{N-1}$, corresponding to the probability that, in a virtual time slot, when a client (say, Client C_1) transmits, at least one of the remaining $N-1$ clients transmits at the same time. However, in a CSMA/CA-based MU-MIMO WLAN, this result does not hold true, for the following two reasons. First, when Client C_1 transmits in a virtual time slot, the number of the remaining clients that can transmit at the same virtual time slot is unknown, since some of the clients (e.g., Client C_3 and Client C_5 in Fig. 3) may have already been involved in the ongoing transmissions. Second, when Client C_1 starts to transmit, collisions may happen not only at Client C_1 but also at clients that transmit concurrently with it. As shown in Fig. 7, although Client C_1 does not encounter collisions when it first wins the channel, it still fails to transmit its data because Client C_4 and Client C_5 collide with each other. Since Bianchi's result is not applicable, we propose a new approach to compute p.

Define M as the maximum number of clients that can transmit concurrently and successfully in a CSMA/CA-based MU-MIMO WLAN with only N single-antenna clients, then

$$M = \min\{\text{the number of antennas at AP}, N\}. \quad (2)$$

We use the term *round* to denote the time interval spent by a transmission with M (or more than M if collision happens) concurrent clients[6]. A transmission round can *succeed* (or *fail*), corresponding to whether the AP can perform successful decoding in that round (see Fig. 7). According to the definition of the conditional collision probability, p can be represented as

$$\begin{aligned} p &= P(\text{Client C}_i \text{ fails}|\text{Client C}_i \text{ transmits}) \\ &= P(r \text{ fails}|\text{Client C}_i \text{ transmits in round } r), \end{aligned} \quad (3)$$

where $i \in \{1, 2, \ldots, N\}$ denotes the client that we are interested in, and r is a randomly chosen round. Let $\mathcal{R}_s$ and $\mathcal{R}_f$ denote the sets of successful and failed rounds, respectively. Then $r \in \mathcal{R}_s$ (or $r \in \mathcal{R}_f$) means that r succeeds (or fails). Accordingly, $1-p$ can be calculated as

$$\begin{aligned} 1-p &= P(r \in \mathcal{R}_s|\text{Client C}_i \text{ transmits in round } r) \\ &= \frac{P(\text{Client C}_i \text{ transmits in } r \text{ and } r \in \mathcal{R}_s)}{P(\text{Client C}_i \text{ transmits in round } r)} \\ &= \frac{P(\text{Client C}_i \text{ transmits in } r|r \in \mathcal{R}_s)P(r \in \mathcal{R}_s)}{1-P(\text{Client C}_i \text{ does not transmit in round } r)}. \end{aligned} \quad (4)$$

As discussed in the last paragraph of Section II, the AP encounters a decoding failure as long as a collision happens. Therefore, $P(r \in \mathcal{R}_s)$ represents the probability that no clients transmit at the same time in a round. This probability certainly depends on the number of allowed concurrent transmissions in a round and the number of clients competing for those transmission opportunities, i.e., M and N. Therefore, we use $P_s(M, N)$ instead of $P(r \in \mathcal{R}_s)$ to highlight its dependence on M and N.

In a successful round, there are exactly M concurrent transmissions and hence M contention periods, so $P_s(M, N)$ can be computed as the probability that, at the end of each contention period, only one client wins the transmission opportunity. Let A_j ($j \in \{1, 2, \ldots, M\}$) be the event that exactly one client wins the j-th contention, i.e., no collision happens in the j-th concurrent transmission. $P_s(M, N)$ can then be represented as

$$P(A_1)P(A_2|A_1)\cdots P(A_M|A_1, A_2, \ldots, A_{M-1}). \quad (5)$$

Since τ is the probability that Client C_i transmits ($i \in \{1, 2, \ldots, N\}$) in a randomly chosen virtual time slot, and a virtual time slot is the same as a conventional time slot during the contention period, we can compute Eq. (5) as

$$\begin{aligned} P_s(M, N) = &\frac{N\tau(1-\tau)^{N-1}}{1-(1-\tau)^N} \frac{(N-1)\tau(1-\tau)^{N-2}}{1-(1-\tau)^{N-1}} \\ &\cdots \frac{(N-M+1)\tau(1-\tau)^{N-M}}{1-(1-\tau)^{N-M+1}}, \end{aligned} \quad (6)$$

where there are M terms multiplying together, and each term corresponds to a contention period. For the first term, the denominator $1-(1-\tau)^N$ denotes the probability that at least one of the N clients transmits in a time slot, while the numerator $N\tau(1-\tau)^{N-1}$ denotes the probability that exactly one of the N clients transmits in a time slot. Therefore, the first term represents the probability that given a time slot where at least one client wins the contention[7], exactly one client transmits in that time slot. All the rest $M-1$ terms can be explained in the same way, except that they are computed under the collision-free condition of the previous contentions.

[5]Many factors, e.g., fading, interference, collisions, can cause transmission failure, however, in this paper we focus on investigating the effect of collisions.

[6]It is possible that less than M clients transmit in a round, when no clients win the contention before the ongoing transmission ends. Here we ignore this probability, which consequently results in a limitation of the analytical model. This limitation will be discussed in Section IV-C.

[7]This condition restricts the time slot to be at the end of a contention period.

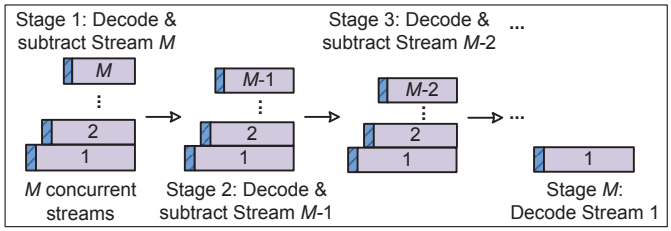


Fig. 6. The decoding procedure of ZF-SIC.

Recall that our aim is to compute the uplink throughput when every client has data to send all the time. Under this saturation condition, each client will have an equal probability to join a successful round. Consequently, the probability that Client C_i is among the M concurrent clients of a successful round is equal to the probability that i is among the M clients that are randomly picked from N clients, i.e.,

$$P(\text{Client } C_i \text{ transmits in } r | r \in \mathcal{R}_s) = \frac{\binom{N-1}{M-1}}{\binom{N}{M}} = \frac{M}{N}. \quad (7)$$

To compute $P(\text{Client } C_i \text{ does not transmit in round } r)$, note that if Client C_i is sure to be silent in a round, then the whole network will act as if Client C_i were not there, i.e., there were $N-1$ clients. Therefore, we have

$$P(r \in \mathcal{R}_s | \text{Client } C_i \text{ does not transmit in } r) = P_s(M', N-1), \quad (8)$$

where M' is defined as the maximum number of allowed concurrent transmissions in a network with the same AP but with $N-1$ single-antenna clients, i.e.,

$$M' = \min\{M, N-1\}. \quad (9)$$

Based on Bayes' Theorem, we get

$$\begin{aligned} &P(\text{Client } C_i \text{ does not transmit in round } r) \\ &= \frac{P(\text{Client } C_i \text{ does not transmit in } r | r \in \mathcal{R}_s) P(r \in \mathcal{R}_s)}{P(r \in \mathcal{R}_s | \text{Client } C_i \text{ does not transmit in round } r)} \\ &= \frac{(1 - \frac{M}{N}) P_s(M, N)}{P_s(M', N-1)}. \end{aligned} \quad (10)$$

Substituting Eqs. (7) and (10) into Eq. (4) gives

$$p = 1 - \frac{\frac{M}{N} P_s(M, N)}{1 - \frac{(1 - \frac{M}{N}) P_s(M, N)}{P_s(M', N-1)}}, \quad (11)$$

where $P_s(M, N)$ is calculated by Eq. (6). Now we have two non-linear equations of p and τ, i.e., Eqs. (1) and (11). The value of τ can be determined by solving these equations.

C. Transmission Rates

The transmission rates of concurrently transmitting clients depend on their channels and how they interact in the decoding procedure. For the CSMA/CA-based MU-MIMO WLAN, the AP uses *zero-forcing with successive interference cancellation* (ZF-SIC)[8] to decode the M independent data streams [1] [2] [4]. The decoding procedure contains M stages, as shown in Fig. 6. In the k-th stage, the AP decorrelates and decodes the $(M+1-k)$-th stream, and then subtracts off the decoded stream from the received vector so that in the $(k+1)$-th stage, there are $M-k$ remaining streams. According to this decoding procedure, when the AP decodes the k-th stream, only the interfering streams that join before the k-th stream need to be considered, since streams that join after it have already been removed. This property allows the k-th concurrent client to transmit at a rate that is determined by the channels of itself and the previous $k-1$ contention winners.

To illustrate how ZF-SIC works, let us consider a network with n-antenna AP. After decoding and removing $M-k$ streams, the remaining received vector at the AP for a symbol time can be written as

$$\mathbf{y} = \sum_{1}^{k} \mathbf{h}_i x_i + \mathbf{w}, \quad (12)$$

where $\mathbf{y}$, $\mathbf{h}_i$, and $\mathbf{w}$ are $n \times 1$ vectors. The i-th contention winner transmits a data symbol x_i through a channel $\mathbf{h}_i$. The additive white noise vector is denoted by $\mathbf{w}$, which follows a circular symmetric distribution $\mathcal{CN}(0, N_0\mathbf{I}_n)$. We assume that the data streams, the channel vectors, and the noise vectors are all independent. To decorrelate x_k, the AP projects the received $\mathbf{y}$ onto the *null space* of the matrix $[\mathbf{h}_1\mathbf{h}_2 \ldots \mathbf{h}_{k-1}]^T$, where $[\cdot]^T$ is the transpose operator. Under the assumption of independent channel vectors, the dimension of this null space is

$$d_k = n - k + 1. \quad (13)$$

We can construct a $d_k \times n$ matrix $\mathbf{Q}_k$, with its rows representing an orthogonal basis of this null space. Then the projection operation can be expressed as multiplying $\mathbf{Q}_k$ and $\mathbf{y}$, which yields

$$\mathbf{Q}_k\mathbf{y} = \mathbf{Q}_k\mathbf{h}_k x_k + \mathbf{Q}_k\mathbf{w}. \quad (14)$$

Accordingly, the k-th stream can be decoded and then removed from Eq. (12). The AP will continue to decode the $(k-1)$-th stream following a similar procedure.

To characterize the resulting rates, note that in Eq. (14), $\mathbf{Q}_k\mathbf{w}$ is still white noise, distributed as $\mathcal{CN}(0, N_0\mathbf{I}_{d_k})$. Let $P = E[||x||^2]$ be the transmission power of each client, and B be the channel bandwidth, then the maximum data rate achieved by the k-th concurrent client is

$$R_k = B\log_2(1 + P||\mathbf{Q}_k\mathbf{h}_k||^2/N_0), k = 1, \ldots, M. \quad (15)$$

We consider a time-varying i.i.d. Rayleigh fading channel model, with coherence time as a transmission round, which means that each client's channel remains unchanged during a round, but is independently variable between successive transmission rounds. Let $\mathbf{h}_i \sim \mathcal{CN}(0, \mathbf{I}_n)$, we can now calculate the average data rates of concurrent date streams. According to [1], the distribution of $\mathbf{Q}_k\mathbf{h}_k$ is $\mathcal{CN}(0, \mathbf{I}_{d_k})$ and $||\mathbf{Q}_k\mathbf{h}_k||^2$ is distributed as $\chi^2_{2d_k}$, i.e., it is Chi-squared distributed with $2d_k$ degrees of freedom. This result also holds true for the first contention winner, because when $k=1$, $||\mathbf{h}_1||^2$ follows the distribution χ^2_{2n}. Accordingly, the average transmission rate of the k-th concurrent data stream can be

[8]The zero-forcing operation is also known as decorrelator or interference nulling.

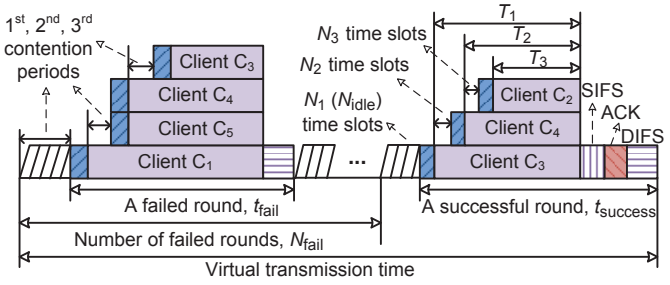


Fig. 7. Illustration of the transmission structure in a CSMA/CA-based MU-MIMO WLAN.

computed as

$$E[R_k] = \int_0^{+\infty} B\log_2(1+Px/N_0) f_{\chi^2_{2d_k}}(x)\mathrm{d}x, \quad (16)$$

where $f_{\chi^2_{2d_k}}(\cdot)$ denotes the PDF for $\chi^2_{2d_k}$ distribution with $k = 1, 2, \ldots, M$.

D. Saturation Throughput

Saturation throughput refers to the network throughput when clients always have data to transmit. To formulate it, we first introduce a concept called *virtual transmission time*, as defined in [8]. It represents the time elapse between two consecutive successful rounds (see Fig. 7). Let N_{fail} be the number of failed transmission rounds during the virtual transmission time and N_{idle} be the number of idle time slots between two consecutive rounds. Let T_j denote the transmission time of the j-th concurrent client in a round, where $j \in \{1, 2, \ldots, M\}$. Then N_{fail}, N_{idle}, and T_j are all random variables. Together with the transmission rates in Section III-C, we can express the saturation throughput as

$$\rho = \frac{\sum_{j=1}^{M} E[R_j]E[T_j]}{E[N_{\text{fail}}]t_{\text{fail}} + t_{\text{success}} + (E[N_{\text{fail}}]+1)E[N_{\text{idle}}]t_{\text{slot}}}, \quad (17)$$

where t_{fail}, t_{success}, and t_{slot} are the time elapse of a failed round, a successful round and an idle time slot, respectively.

According to the definition, N_{fail} follows a geometric distribution with parameter $1 - P_s(M, N)$, which is the probability that a randomly chosen round fails, i.e.,

$$P(N_{\text{fail}} = k) = (1 - P_s(M,N))^k P_s(N,M), \quad (18)$$

where $k = 0, 1, 2, \ldots$, then the average number of failed rounds in a virtual transmission time is

$$E[N_{\text{fail}}] = \frac{1 - P_s(M,N)}{P_s(M,N)}. \quad (19)$$

In a successful transmission round, because the transmitting clients are forced to end simultaneously, the time they spent on data transmission can be calculated recursively as

$$T_{j+1} = T_j - t_{\text{PHY}} - N_{j+1}t_{\text{slot}}, \quad (20)$$

where t_{PHY} is the time needed to transmit a PHY header, and N_j ($j \in \{1, 2, \ldots, M\}$) is the number of idle time slots elapsing in the j-th contention period (see Fig. 7). Since N_{idle} is the number of idle time slots between two consecutive rounds, it actually equals to N_1. During the first contention period, there are N clients competing for the transmission opportunity, and each client transmits in a time slot with probability τ. Thus N_1 follows a geometric distribution with the parameter $(1-\tau)^N$, i.e.,

$$P(N_1 = k) = ((1-\tau)^N)^k(1-(1-\tau)^N). \quad (21)$$

Accordingly, we have

$$E[N_{\text{idle}}] = E[N_1] = \frac{(1-\tau)^N}{1-(1-\tau)^N}. \quad (22)$$

During the j-th contention period, where $j \geq 2$, there are $N-j+1$ clients competing for the jth concurrent transmission opportunity. However, all the contending clients must have nonzero backoff counters, for otherwise they would have been involved in the ongoing transmission and can no longer join the j-th contention. In other words, no client is able to win the j-th concurrent transmission opportunity in the first time slot of the contention period, i.e., $N_j \geq 1$ for $j \geq 2$. Therefore, the distribution of N_j is

$$P(N_j = k) = ((1-\tau)^{N-j+1})^{k-1}(1-(1-\tau)^{N-j+1}), \quad (23)$$

where $k \geq 1$ and $2 \leq j \leq M$. Then its expectation becomes

$$E[N_j] = \frac{1}{1-(1-\tau)^{N-j+1}}. \quad (24)$$

Assuming that $E[T_1]$ is known, we can then use Eqs. (20) and (24) to compute the expectation of other T_js in a recursive manner, i.e.,

$$E[T_{j+1}] = E[T_j] - t_{\text{PHY}} - \frac{1}{1-(1-\tau)^{N-j}}t_{\text{slot}}. \quad (25)$$

Although each time different clients are engaged in a transmission round, how long a round lasts depends only on the data time of the first client (Fig. 7). Therefore[9],

$$t_{\text{success}} = t_{\text{PHY}} + E[T_1] + \text{SIFS} + \text{ACK} + \text{DIFS}, \quad (26)$$

$$t_{\text{fail}} = t_{\text{PHY}} + E[T_1] + \text{DIFS}. \quad (27)$$

Substituting Eqs. (6), (16), (19), (22), and (25)–(27) into Eq. (17), with values of N, M, B, P/N_0, $E[T_1]$ as well as the value of τ calculated in the last subsection, we are now able to compute the saturation throughput ρ of the uplink channel in a CSMA/CA-based MU-MIMO WLAN.

E. Access Delay

In this subsection the mean access delay is determined for each client. The access delay is defined as the time experienced by a packet, from it first becoming the head of the queue to the time it is transmitted successfully. Under the saturation condition, all the clients have the same mean access delay. Let d denote the mean access delay of a given client, say, Client C_1. Then d refers to the average time between Client C_1's consecutively transmitted packets. According to Eq. (7), we know that Client C_1 needs to wait an average of $1/P(\text{Client C}_1 \text{ transmits in } r | r \in \mathcal{R}_s)$ successful rounds to join a successful round. Based on the concept of virtual

[9]The propagation delay is normally too small compared with the total frame transmission time, so it is omitted here.

transmission time, we can then calculate the mean access delay as

$$d = \frac{E[\text{Virtual transmission time}]}{P(\text{Client C}_1 \text{ transmits in } r | r \in \mathcal{R}_s)}. \quad (28)$$

According to Eqs. (7) and (17), we can get

$$d = \frac{E[N_{\text{fail}}]t_{\text{fail}} + t_{\text{success}} + (E[N_{\text{fail}}]+1)E[N_{\text{idle}}]t_{\text{slot}}}{M/N}. \quad (29)$$

Note that Client C_1's packets are of varying sizes, which depend on the dimension Client C_1 occupies in each successful round. Therefore, d corresponds to the transmission of a packet with an average size $(\sum_{j=1}^{M} E[R_j]E[T_j])/M$.

F. Opportunistic Transmission

According to Section III-C, the k-th stream experiences interference from the previous $k-1$ concurrent clients. After projecting $\mathbf{h}_k$ onto the subspace orthogonal to the one spanned by $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_{k-1}$, the k-th stream achieves an SNR of $P||\mathbf{Q}_k\mathbf{h}_k||^2/N_0$. When the transmission power is given, the concurrent transmission rate is fully determined by $||\mathbf{Q}_k\mathbf{h}_k||$, which represents the effect of inter-stream interference. The value of $||\mathbf{Q}_k\mathbf{h}_k||$ depends on the interaction of the channels of the k concurrent streams, and is always less than or equal to $||\mathbf{h}_k||$, where equality is only achieved when $\mathbf{h}_k$ is orthogonal to the span of $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_{k-1}$.

Considering a network where the clients fade independently, we are then interested in an opportunistic transmission scheme: during the j-th contention period ($j \geq 2$), if clients with large concurrent rates, i.e., large values of $||\mathbf{Q}_j\mathbf{h}_j||$, are given high probabilities of winning the contention, then the total network throughput can be improved. In this subsection we will model the saturation throughput and mean access delay of a simple distributed opportunistic scheme for a 2-antenna AP scenario. Our modeling method can be easily generalized to scenarios with more antennas at the AP, but considering a simple two antenna case is enough to reveal the influence of the threshold value on the network performance, as indicated in Section V-D.

In Section III-C, we have shown that the maximum data rates of two concurrent streams are

$$\begin{aligned} R_1 &= B\log_2(1 + P||\mathbf{h}_1||^2/N_0), \\ R_2 &= B\log_2(1 + P||\mathbf{Q}_2\mathbf{h}_2||^2/N_0), \end{aligned} \quad (30)$$

where $\mathbf{Q}_2$ is a 1×2 unit vector that is orthogonal to $\mathbf{h}_1$. A geometric interpretation of $\mathbf{h}_1$, $\mathbf{h}_2$, and $\mathbf{Q}_2^T$ is shown in Fig. 8. Similar to Section III-C, we consider a network where the clients experience i.i.d. Rayleigh fading. The channel of each client remains unchanged during a round time, but is independently variable between successive rounds. Assuming that $\mathbf{h}_i \sim \mathcal{CN}(0, \mathbf{I}_2)$ for $i = 1, 2$, then $||\mathbf{Q}_2\mathbf{h}_2||^2$ follows a Chi-squared distribution with 2 degrees of freedom.

The opportunistic MAC protocol that we will model works as follows. In a transmission round, when a client wins the first contention, each of the rest clients would then calculate[10]

[10]A client could learn its own channel through the reverse channel. Besides, according to [4], the first winner can put its own channel information in its PLCP header so that other clients would be able to know it.

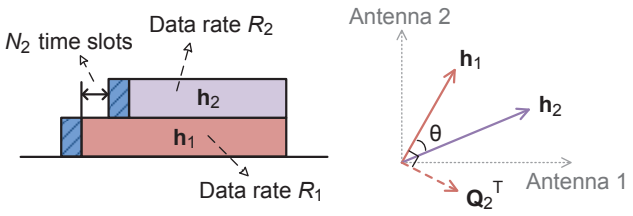


Fig. 8. A geometric interpretation of $\mathbf{h}_1$, $\mathbf{h}_2$, and $\mathbf{Q}_2^T$.

the value of $||\mathbf{Q}_2\mathbf{h}_2||^2$ by assuming that itself is the second contention winner. Only the clients satisfying

$$||\mathbf{Q}_2\mathbf{h}_2||^2 \geq \text{T} \quad (31)$$

are allowed to join the second contention, while others will defer their access until the next round. Here T acts as a threshold.

To formulate the opportunistic MAC protocol, we can follow the same procedure as developed in the previous subsections. Note that Eq. (1) still holds true here because the backoff mechanisms used in the two MAC protocols are identical.

To derive the conditional collision probability p, we need to calculate $P_s(M, N)$ first (with $M = 2$ and $N > 2$), which represents the probability that, in a network with N clients, a randomly chosen round is successful. Let N_{join} denote the number of clients that can contend the second transmission opportunity in a successful round. Then N_{join} is a random variable because of the randomness of clients' channels. Denoted by p_{join} the probability that, given the first contention winner, i.e., given $\mathbf{Q}_2$, a randomly chosen client satisfies Eq. (31), i.e.,

$$p_{\text{join}}(\text{T}) = P(||\mathbf{Q}_2\mathbf{h_2}||^2 \geq \text{T} | \mathbf{Q}_2). \quad (32)$$

Then N_{join} follows a binomial distribution as

$$P(N_{\text{join}} = k) = \binom{N-1}{k} p_{\text{join}}^k (1 - p_{\text{join}})^{N-1-k}, \quad (33)$$

where $k = 0, 1, 2..., N-1$. To compute p_{join}, let $\theta \in [0, \pi]$ be the angle between $\mathbf{h}_1$ and $\mathbf{h}_2$ in the antenna space, as shown in Fig. 8. According to [4], we have

$$||\mathbf{Q}_2\mathbf{h_2}||^2 = ||\mathbf{h_2}||^2 \sin^2(\theta). \quad (34)$$

We assume that $||\mathbf{h_2}||^2$ and θ are independent random variables. We already know that $||\mathbf{h_2}||^2 \sim \chi_4^2$. Assuming that θ is uniformly distributed in the interval $[0, \pi]$, Eq. (32) can then be computed as

$$\begin{aligned} p_{\text{join}}(\text{T}) &= 1 - P(||\mathbf{Q}_2\mathbf{h_2}||^2 < \text{T} | \mathbf{Q}_2) \\ &= 1 - P(||\mathbf{h_2}||^2 \sin^2(\theta) < \text{T}) \\ &= 1 - \int_0^\infty P(||\mathbf{h_2}||^2 = x) P(\sin^2(\theta) < \frac{\text{T}}{x}) \text{d}x \\ &= 1 - \int_0^\infty f_{\chi_4^2}(x) \frac{\arcsin(\sqrt{\text{T}/x})}{\pi/2} \text{d}x, \end{aligned} \quad (35)$$

where $f_{\chi_4^2}(\cdot)$ denotes the PDF of a Chi-squared distribution with 4 degrees of freedom. Based on Eqs. (35) and (33), we

can now calculate the probability of a successful round as the following expectation

$$P_s(2,N) = E\left[\frac{N\tau(1-\tau)^{N-1}}{1-(1-\tau)^N}\frac{N_{\text{join}}\tau(1-\tau)^{N_{\text{join}}-1}}{1-(1-\tau)^{N_{\text{join}}}}\right], \quad (36)$$

where $N > 2$ and $N_{\text{join}} = 0, 1, 2..., N-1$. In the case of $N_{\text{join}} = 0$, when a successful round contains only one stream, $P_s(M,N)$ reduces to the probability that only one client wins the first contention. Note that in Eq. (33), when p_{join} approaches 1, i.e., when less restriction is given on who can join the second contention, the value of N_{join} would be approaching $N-1$, and the above equation becomes the same as the one defined in Eq. (6).

Eq. (8) still hold true here because it is formulated without considering the detailed MAC protocol. Eq. (7) calculates the probability that Client C_i shows up in a randomly chosen successful round. Under the opportunistic transmission scheme, in each round the number of clients that can join the second contention is limited. However, since the clients' channels are assumed to be independent for different rounds, in the long run, each client has an equal probability to join a successful round and is able to transmit concurrently with any other clients. Considering the special case of $N_{\text{join}} = 0$, we denote p_0 the probability that in a randomly chosen successful round, no client contends for the second concurrent transmission opportunity, i.e.,

$$p_0 = P(N_{\text{join}} = 0 | r \in \mathcal{R}_s). \quad (37)$$

Then Eq. (7) is changed to

$$P(\text{Client C}_i \text{ transmits in } r | r \in \mathcal{R}_s) = \frac{2}{N}(1-p_0) + \frac{1}{N}p_0. \quad (38)$$

Using Eqs. (33) and (36), p_0 can be calculated as

$$p_0 = \frac{P(r \in \mathcal{R}_s \text{ and } N_{\text{join}} = 0)}{P(r \in \mathcal{R}_s)} = \frac{\frac{N\tau(1-\tau)^{N-1}}{1-(1-\tau)^N}(1-p_{\text{join}})^{N-1}}{P_s(2,N)}. \quad (39)$$

The average transmission rate of the first contention winner is the same as that of the original MAC protocol, i.e.,

$$E[R_1] = \int_0^{+\infty} B\log_2(1+Px/N_0) f_{\chi_4^2}(x)\mathrm{d}x, \quad (40)$$

where $f_{\chi_4^2}(\cdot)$ denotes the PDF of a Chi-squared distribution with 4 degrees of freedom. To compute the the second stream's average rate, note that according to the opportunistic transmission scheme, the value of $||\mathbf{Q}_2\mathbf{h}_2||^2$ falls in the range of $[\text{T}, \infty)$. Since $||\mathbf{Q}_2\mathbf{h}_2||^2 \sim \chi_2^2$, we have

$$E[R_2] = \frac{\int_{\text{T}}^{\infty} B\log_2(1+Px/N_0) f_{\chi_2^2}(x)\mathrm{d}x}{\int_{\text{T}}^{\infty} f_{\chi_2^2}(x)\mathrm{d}x}, \quad (41)$$

where $f_{\chi_2^2}(\cdot)$ denotes the PDF of a Chi-squared distribution with 2 degrees of freedom.

To compute the saturation throughput, note that since the number of clients competing for the second transmission opportunity is reduced to N_{join}, which is a random variable with distribution given by Eq. (33), we can then change Eq. (24) to

$$E[N_2] = E\left[\frac{1}{1-(1-\tau)^{N_{\text{join}}}}\right]. \quad (42)$$

TABLE I
NETWORK PARAMETERS USED TO OBTAIN NUMERICAL VALUES.

Parameter	Value
t_{slot}	9 μs
t_{PHY}	20 μs
SIFS	16 μs
DIFS	34 μs
ACK	39 μs
ACK timeout	70 μs
$E[T_1]$	2000 μs
B	20 MHz
P/N_0	10 dB
$CW_{\min}$	127
$CW_{\max}$	1023

Accordingly, the average transmission time of the second contention winner, i.e., Eq. (25), becomes

$$E[T_2] = E[T_1] - t_{\text{PHY}} - E\left[\frac{1}{1-(1-\tau)^{N_{\text{join}}}}\right] t_{\text{slot}}. \quad (43)$$

Using the above results and following the same procedure in the previous subsections of Section III, we are now able to calculate the saturation throughput and mean access delay of the opportunistic transmission scheme.

IV. Model Validation

In this section comparisons between the analytical and simulation results are presented to validate our previous analysis. It includes three subsections. In the first and second subsections examples are shown that our analytical model can closely predict the uplink throughput and mean access delay of a wireless LAN. Two MAC protocols are simulated: a CSMA/CA-based MU-MIMO MAC and its opportunistic variation. In the third subsection we discuss about the main limitations of our analytical model.

A. CSMA/CA-based MU-MIMO WLANs

We use MATLAB to simulate the uplink channel of a CSMA/CA-based MU-MIMO WLAN. Our event-driven simulation program contains all major components of the MAC protocol, e.g., contention, PHY header, ACK, ACK timeout and the interframe spaces. The program also simulates the PHY layer as described in Section III-C. Every client is assigned an $n \times 1$ channel vector, where n is the number of antennas at the AP. Each component of the channel vector is an i.i.d. $\mathcal{CN}(0,1)$ random variable. The channel vectors are generated at the beginning of every transmission round and remain unchanged during a round time. The channel vectors of a client for different rounds are independent.

The network parameters used to obtain numerical values for both analytical model and simulation program are outlined in Table I, where the upper half values are defined by 802.11 standards for OFDM PHY layer with 20 MHz channel spacing, and the lower half values are either calculated or selected according to the standards [7]. Note that our analytical model is derived using the mean frame transmission time of the first transmitter in a round, i.e., $E[T_1]$, with no restrictions on the probability distribution of T_1. For simplicity, we only consider the case of constant T_1 in the simulation program. Besides, we set ACK timeout to be 70 μs in the simulation

program, a value that is long enough to cover an SIFS and ACK transmission. Unless otherwise specified, the numerical values obtained in the rest of this paper are all based on the network parameters listed in Table I.

In Fig. 9 we plot the saturation throughput and mean access delay using both simulation (symbols) and analytical (lines) results. Different network configurations are considered: the number of antennas at the AP varies from 1 to 6 and the number of clients ranges from 2 to 50. The figure indicates that: 1) our analytical model provides a close approximation of the saturation throughput and mean access delay in a CSMA/CA-based MU-MIMO WLAN; 2) the accuracy of the saturation throughput model degrades as the number of antennas at the AP increases, the reason for which is discussed in Section IV-C.

B. Opportunistic Transmission Scheme

We modify the simulation program of the previous subsection to characterize the opportunistic transmission scheme of Section III-F. A network with a 2-antenna AP is considered. Every client has a 2×1 channel vector, with each component being an i.i.d. $\mathcal{CN}(0,1)$ random variable. The channel vectors remain unchanged for a round time and are independent between successive rounds. When a client wins the first contention, every other client calculates its concurrent transmission rate and check whether its concurrent rate is less than the threshold $B\log_2(1+P\mathrm{T}/N_0)$. If so, the corresponding client stops decreasing its backoff counter, i.e., defers its access to the channel, until the current round ends.

The saturation throughput and mean access delay are plotted in Fig. 10. Simulation (symbols) and analytical (lines) results are compared for networks with sizes varying from 5 to 50. The threshold T is set as 0.5 and 1.5. Note that the y-axis for saturation throughput ranges from 120 to 160 Mbps. Although differences exist between the simulation and analytical saturation throughput, the error percentage is less than 4%. Fig. 10 indicates that our analytical model can closely estimate the network performance of the opportunistic transmission scheme.

C. Limitations and Discussions

In this subsection we would discuss the limitations of our analytical model. Reasons are provided to explain why the accuracy of our model varies with different parameters.

As shown in Fig. 11, the accuracy of our saturation throughput model varies with respect to three parameters: frame transmission time, the number of antennas at the AP, and the number of total clients. The error percentage between simulation and analytical result is calculated as $|\rho_{\text{simulation}} - \rho_{\text{analytical}}|/\rho_{\text{simulation}}$. Comparing scenarios (a) and (b), scenarios (a) and (c) would reveal that our model is more accurate in the case of long frame transmission time, and small AP's antennas. Besides, the error percentage decreases as the number of clients grows. This fluctuation of accuracy can be explained as follows. Our analytical model does not consider the situation that there may be less than M concurrent transmissions in a round, which happens when none of contending clients win the channel before the ongoing transmission ends. During the

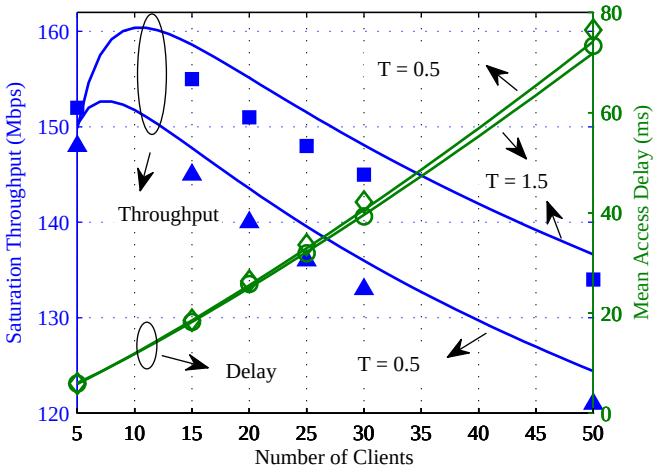


Fig. 10. Saturation throughput and mean access delay of the opportunistic transmission scheme: simulation (symbols) versus analysis (lines).

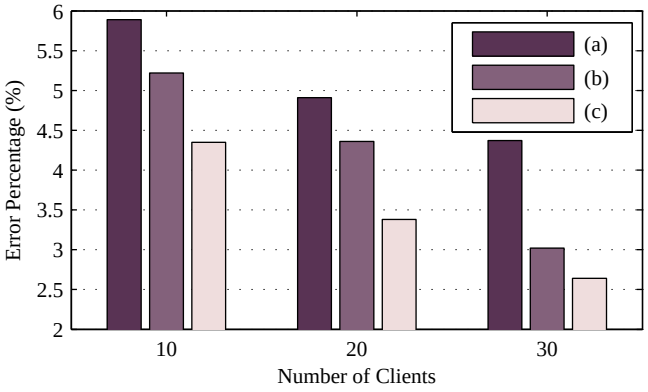


Fig. 11. Error that occurs when the analytical model is used to estimate the saturation throughput in different scenarios: (a) $M = 6$ and $E[T_1] = 2000$ μs; (b) $M = 6$ and $E[T_1] = 4000$ μs; (c) $M = 4$ and $E[T_1] = 2000$ μs.

derivation of the analytical model (e.g., Eq. (6)), we simply assume that in every round, there are M (or more than M if collision happens) concurrent transmissions. This assumption holds true with a high probability when the frame transmission time is long, the number of antennas at the AP is small, and the number of total clients is large, which explains why the accuracy of our model is high under these situations.

To make the above explanation more convincing, we change t_{slot} from 9 μs to 1 μs and simulate the CSMA/CA-based MU-MIMO WLANs[11]. The contention window sizes are set as $CW_{\min} = 511$ and $CW_{\max} = 1023$. All other network parameters are the same as those in Table I. Comparisons between the simulation (symbols) and analytical (lines) results are presented in Fig. 12, with M ranges from 7 to 20. This figure indicates that our model is extremely accurate even when the AP has a large number of antennas. In this case a successful round is ensured to have M concurrent clients. If there is a successful round with only $M-1$ concurrent clients, then all the other clients must have a backoff time longer than

[11]Please note that this change is only for illustrative purpose; it does not imply the existence of such an implementation in standard wireless networks.

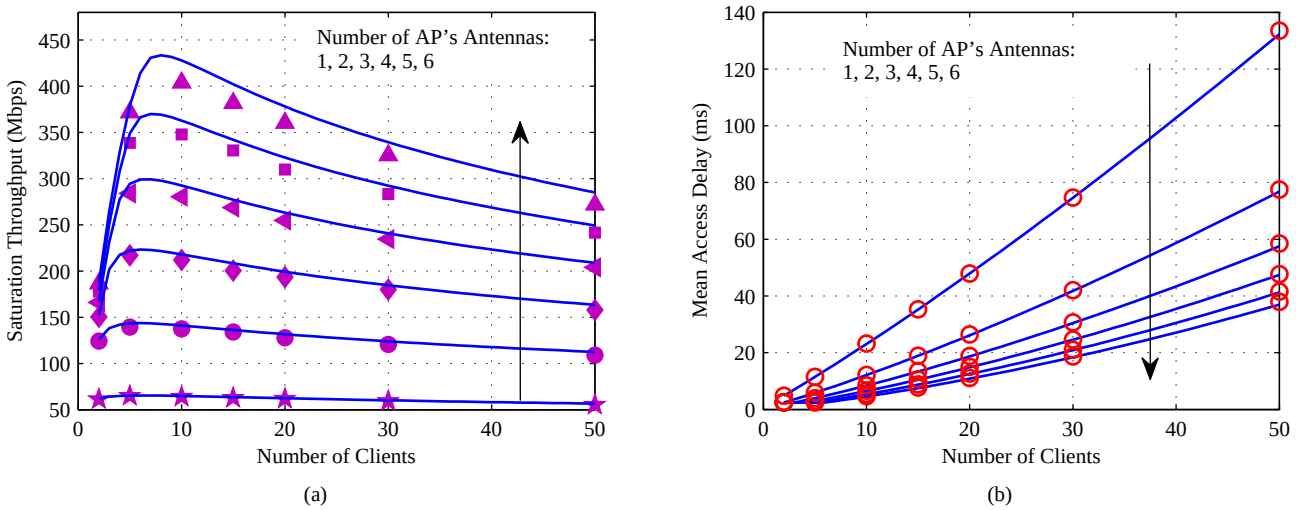


Fig. 9. Saturation throughput and mean access delay for different network configurations: simulation (symbols) versus analysis (lines).

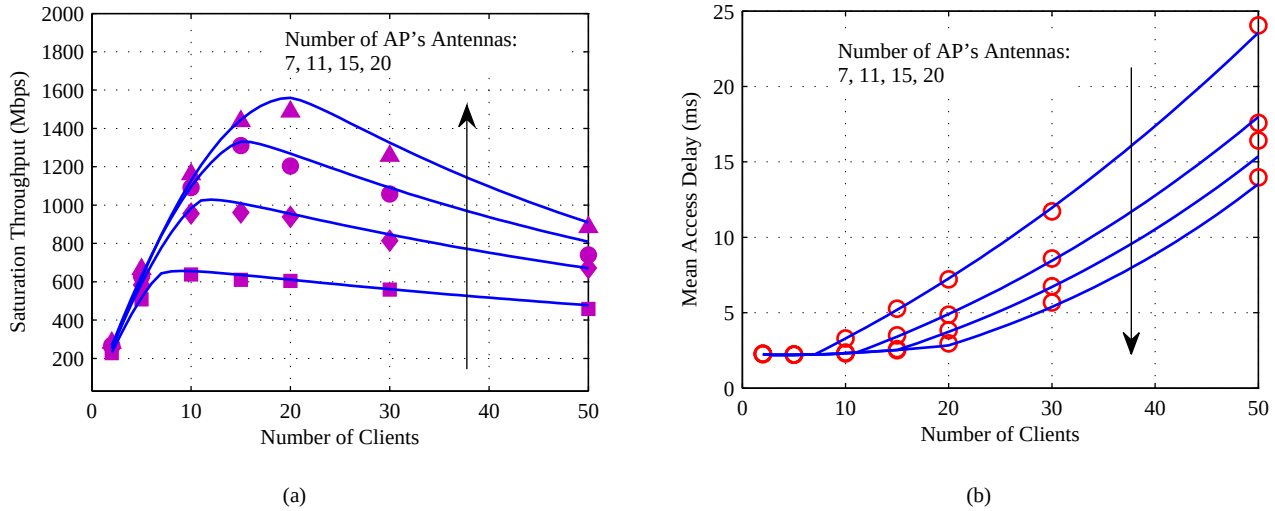


Fig. 12. Comparisons between the simulation (symbols) and analytical (lines) results by setting $t_{\text{slot}} = 1$ μs. Saturation throughput is shown in (a) while the mean access delay is shown in (b).

$E[T_1]-(M-2)t_{\text{PHY}}$. For $E[T_1] = 2000$ μs, $t_{\text{PHY}} = 20$ μs, and $M = 20$, this value is equal to 1640 μs. However, a client's backoff time is always less than $CW_{\max}t_{\text{slot}} = 1023$ μs. Since 1023 μs is smaller than 1640 μs, it is impossible for a round to have less than M concurrent transmissions.

As demonstrated by Fig. 12, our model can accurately characterize a CSMA/CA-based MU-MIMO WLAN when there are always M concurrent clients in a successful round. This corresponds to the situation when all the dimensions at the AP are utilized by the concurrent streams. When the number of concurrent streams are less than the maximum number allowed, i.e., when the AP's antennas are underutilized, the saturation throughput would reduce. In other words, our analytical model is able to characterize the maximum saturation throughput that can be achieved by the current AP in a CSMA/CA-based MU-MIMO WLAN.

V. Numerical Evaluation

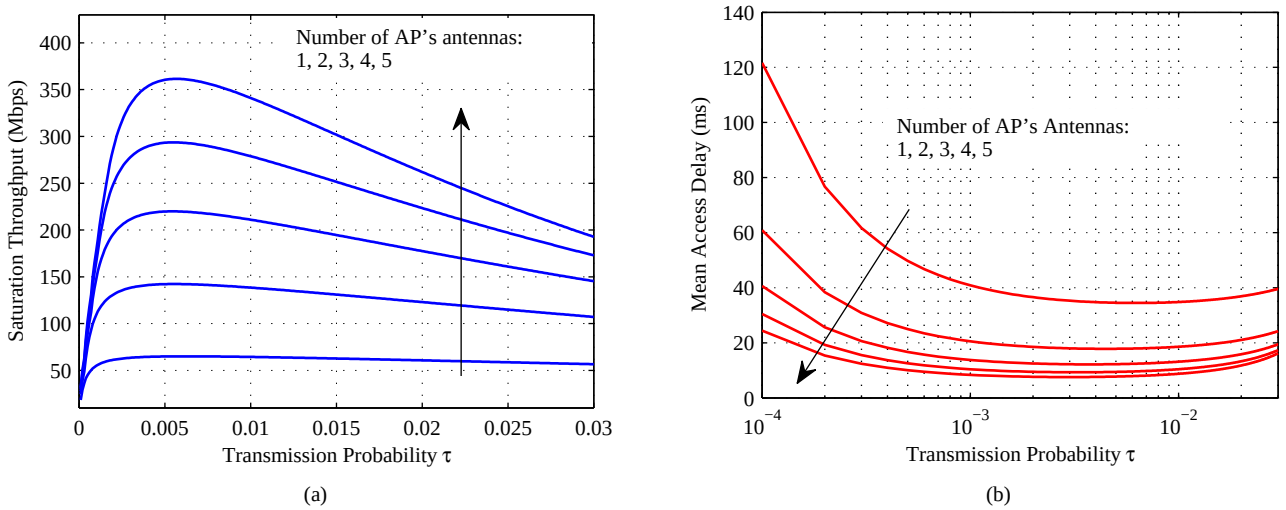
In this section the developed model is used to analyze the network performance with respect to different parameters. The CSMA/CA-based MU-MIMO WLAN is investigated in the first three subsections while performance of the opportunistic transmission scheme is analyzed in the last subsection.

As discussed in Section IV-C, our analytical model can accurately characterize the network performance when the number of concurrent streams in a successful round equals the number of antennas at the AP, i.e., when all the AP's dimensions are occupied. Therefore, performance analysis using our model can reveal the full influence of varying AP's antennas. To maintain a high accuracy when using the proposed model, in this section we will focus on networks with no more than 6 antennas at the AP.

A. Transmission Probability

In Section III-D the analytical throughput ρ is derived as a function of the transmission probability τ, and so is the mean access delay d in Section III-E. To highlight their dependence on τ, in this subsection we express ρ as $\rho(\tau)$ and d as $d(\tau)$.

In Fig. 13, we plot $\rho(\tau)$ and $d(\tau)$ for $M = 1, 2, 3, 4, 5$, $N = 15$, and τ from 0 to 0.03. As shown in the figure,



(a) (b)

Fig. 13. Saturation throughput and mean access delay versus τ for different numbers of antennas at the AP, with $N = 15$ and $E[T_1] = 2000$ μs.

TABLE II
THE MAXIMUM SATURATION THROUGHPUT AND MINIMUM MEAN ACCESS DELAY ACHIEVED IN FIG. 13.

M	$\rho_{\max}$ (Mbps)	W_ρ	$d_{\min}$ (ms)	W_d
1	65.07	[312, 327]	34.46	[302, 338]
2	142.3	[338, 384]	17.82	[407, 487]
3	219.9	[350, 384]	12.16	540
4	293.7	[356, 364]	9.296	605
5	361.5	[344, 363]	7.552	[666, 689]

$\rho(\tau)$ is maximized at a certain transmission probability τ_ρ. Similarly, $d(\tau)$ achieves its minimum when τ reaches a certain value τ_d. Note that τ is an indication of clients' willingness in transmitting during a slot time. When τ is small, few clients tend to transmit in a time slot, so a large amount of time is wasted by idle time slots. When τ is large, the probability that two or more clients transmit in the same time slot is high, so the collision probability is large. Both cases would lead to a small saturation throughput and a large mean access delay.

Since τ is the solution of two nonlinear equations (i.e., Eqs. (1) and (11)), when M, N are given, τ is fully determined by the backoff parameters, i.e., $CW_{\min}$ and $CW_{\max}$. Therefore, based on τ_ρ and τ_d, we can obtain the optimal backoff parameters, which corresponds to the maximum throughput and minimum access delay. For simplicity, consider a special backoff strategy that employs constant window size, i.e., $CW_{\min} = CW_{\max}$. Eq. (1) then becomes: $\tau = 2/(W+1)$, where $W = CW_{\min} + 1$. Using this simple relation between τ and W, optimal contention window sizes can be found. As shown in Table II, W_ρ and W_d represent the optimal backoff window sizes[12] corresponding to $\rho_{\max}$ and $d_{\min}$, respectively.

B. Number of Antennas at the AP

In this subsection the influence of AP's antennas is evaluated. For convenience, ρ and d are expressed as $\rho(M)$ and $d(M)$.

[12]Due to the precision limitation of MATLAB, in the table W_ρ (W_d) would be represented as intervals instead of a single value.

As shown in Fig. 9(b), $d(M)$ decreases as M increases from 1 to 6. This is mainly because with more concurrent transmission opportunities, a client would have a larger chance to access the channel. However, since a transmission round of M concurrent clients fails when any one of the M clients encounters a collision, the failure probability of a round increases as M grows. Therefore, $d(M)$ decreases slowly and would finally increase at a large M.

In Fig. 14(a), we plot $\rho(M)$ for $E[T_1] = 2000$ and 4000 μs with $N = 30$. Saturation throughput achieved with fixed backoff parameters ($CW_{\min} = 127$ and $CW_{\max} = 1023$) is depicted in solid lines, while the maximum saturation throughput achieved at the optimal backoff parameters is shown in dashed lines. As shown in Fig. 14(a), $\rho(M)$ increases as M increases from 1 to 6. Let $\Delta\rho(M)$ denote the the amount of increased throughput when one more antenna is added to the M-antenna AP. In Fig. 14(b) we plot $\Delta\rho(M)$ under the same scenarios as in Fig. 14(a).

Fig. 14 indicates two things. First, the throughput is high with a large $E[T_1]$, mainly because the data transmitted during virtual transmission time is large when the frame transmission time is long. Second, the throughput gain of adding more antennas at the AP decreases as M grows large. The reasons are threefold.

- The frame transmission time of the M-th concurrent client (i.e., T_M) decreases as M increases. As shown in Fig. 3, transmission time is wasted by the contention periods in a round. The number of idle time slots during the contention can be reduced by choosing suitable backoff parameters. The resulting throughput increase is indicated by the dashed lines in Fig. 14.
- The date rate achieved by the M-th stream (i.e., R_M) decreases for a large M. According to the decoding procedure (Fig. 6), the M-th stream is decoded while the previous $M-1$ streams are treated as interference. The rate reduction due to interference can be avoided if the concurrent streams have orthogonal channel gains, which can hardly happen in a pure random access MAC protocol. However, we can consider an opportunistic MAC

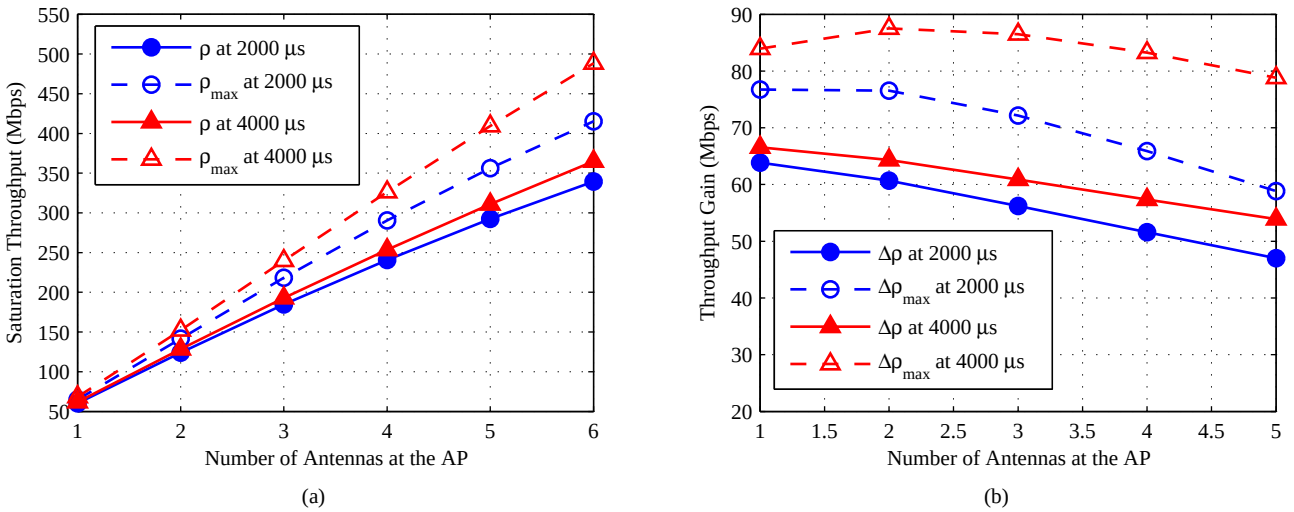


Fig. 14. (a) Saturation throughput versus the number of antennas at the AP for different $E[T_1]$s, with $N = 30$. The solid lines denote the saturation throughput when $CW_{\min} = 127$ and $CW_{\max} = 1023$. The dashed lines correspond to the optimal saturation throughput evaluated at the optimal backoff parameters. (b) The throughput gain of adding one antenna to the current AP. The solid lines and dashed lines are calculated under the same scenarios as (a).

protocol, which gives clients with larger concurrent rates higher probabilities to join the ongoing transmission. A simple opportunistic transmission scheme is modeled in Section III-F and is analyzed in Section V-D.

- A large M means a large chance for a round to fail, since any one of the M concurrent streams encounters a collision would result in transmission failure. The increased probability of a failed round is also a reason for the decrease of $\Delta\rho(M)$ as M grows.

C. Network Size

The network size refers to the total number of clients in the network, which is previously denoted as N. In this subsection we focus on $\rho(N)$ and $d(N)$. As shown in Fig. 9(a), $\rho(N)$ first increases and then decreases as N grows large. The reason is straightforward: when N is small, the number of concurrent clients in a successful round is limited by N (see Eq. (2)); when N is large, the number of clients that contend for each concurrent transmission opportunity is large, resulting in a large collision probability. How d varies with respect to N can be found in Fig. 9(b). Given the number of antennas at the AP, the mean access delay d increases with N due to increased collision probability.

D. Threshold

In this subsection we will analyze how the threshold T affects the network performance in the opportunistic transmission scheme. In Section III-F, a network with a 2-antenna AP is considered. In every transmission round only clients with concurrent rates larger than $B\log_2(1+P\mathrm{T}/N_0)$ are allowed to contend for the concurrent transmission opportunity. When $\mathrm{T} = 0$, the opportunistic scheme is just the original CSMA/CA-based MU-MIMO transmission scheme.

In Fig. 15(a) we plot $d(\mathrm{T})$ when the number of clients are 5, 15, 20, and 25. As shown in the figure, d increases slowly with T for $N = 5$ while decreases slowly for other

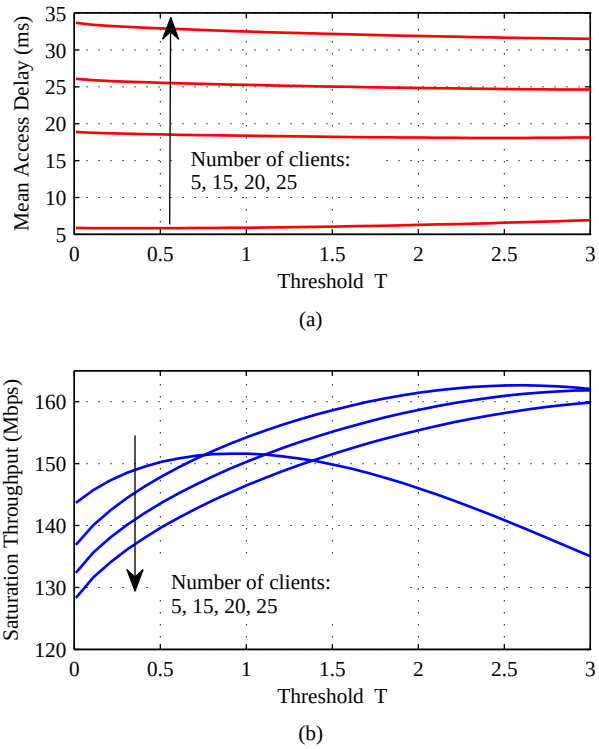


Fig. 15. Mean access delay (above) and saturation throughput (below) versus threshold for the opportunistic transmission scheme in Section III-F.

cases. In Fig. 15(b) we plot $\rho(\mathrm{T})$ for different network sizes. The saturation throughput ρ tends to first increase with T and then decrease after T reaches a certain value. This inverted U-shaped curve can be seen for $N = 5$ when T goes from 0 to 3. The relationship of d and ρ versus T can be explained by the following reasons.

- As T increases, a network would obtain two benefits. The first benefit is an increased data rate of the second stream, as indicated in Eq. (41). The second benefit is a reduced collision probability, since a large T results

in a smaller number of contending clients in the second contention period. The two benefits would drive the saturation throughput to increase with T. Besides, the reduced collision probability leads to a reduced virtual transmission time (Fig. 7), which is the cause for the decreased mean access delay.

- A large T would degrade the network performance. When T is large, the average number of clients that can contend for the concurrent transmission (i.e., $E[N_{\text{join}}]$) is small. According to Eq. (43), the average transmission time of the second stream (i.e., $E[T_2]$) would then reduce. An extreme case would be that no client contends for the concurrent transmission, i.e., $N_{\text{join}} = 0$. In that case, T_2 is equal to 0, which corresponds to the worst case since one of AP's degrees of freedom is wasted. The probability of no contending clients in the second contention period is high when the network size is small, which explains why the network performance loss is prominent when $N = 5$.

VI. Related Work

Many studies have been performed to design and analyze a wireless network that enables multiple concurrent transmissions in the uplink. In [9], Zheng *et al.* propose and analyze a RTS/CTS-based MAC protocol that supports multiple packet reception (MPR) in a WLAN. The proposed protocol is extended in [10], where adaptive resource allocation and MPR are jointly considered through a cross-layer framework. In [11] and [13], Jin *et al.* compare the network performance of single-user MIMO and MU-MIMO schemes in the uplink WLAN, where MU-MIMO transmission is enabled when multiple clients win the contention at the same time. Throughput tradeoff between downlink and uplink in an MU-MIMO based WLAN is investigated in [12]. In [14], Yoon *et al.* develop and implement a CSMA-based scheme that enables simultaneous concurrent transmissions in an ad hoc network.

Despite the many previous research efforts on the design and analysis of MU-MIMO schemes in the uplink, a key difference exists between the MU-MIMO transmission schemes that are analyzed in previous research and the one we have analyzed in this paper. In the previous schemes, concurrent streams are transmitted at the same time by different clients, i.e., their transmissions start *synchronously*. However, in the CSMA/CA-based MU-MIMO WLAN of Section II, clients are allowed to join the ongoing transmission one after another, resulting in *asynchronous* concurrent transmissions. This asynchronous characteristic results in two difficulties in performance modeling and analysis. The first difficulty is an increased complexity in performance modeling. Taking the conditional collision probability p as an example, if concurrent transmissions are synchronous, then p is simply calculated as the probability that the number of concurrent streams is larger than the maximum number allowed (see, e.g., Eq. (20) of [9], Eq. (13) of [11], and Eq. (3) of [14]). However, in the case of asynchronous concurrent transmissions, the derivation of p is more complicated, as indicated in Section III-B. The second difficulty is an increased complexity in performance analysis. As indicated in Section V, when discussing how different parameters influence the network performance, we have to jointly consider their impacts on the average transmission time of each concurrent stream, the collision probability, and the concurrent transmission rates. Modeling and analyzing the network performance by overcoming the two difficulties is the main contribution of this paper.

Over the past years, many efforts have been made to improve the throughput of MU-MIMO networks by selecting a subset of users to perform concurrent transmissions. Multi-user selection algorithms are proposed for both the downlink [15] and uplink MU-MIMO systems [16]. Joint user/antenna selection algorithms are investigated in [17]. Most of the proposed algorithms are centralized, in which a scheduler is assumed to have the clients' channel information. In this paper we consider a simple distributed opportunistic scheduling scheme, where users contend for the concurrent transmission opportunities only when their concurrent rates are large enough. We model and analyze its throughput and delay by considering both PHY and MAC layer influences.

VII. Conclusion

In this paper we modeled and investigated the saturation throughput and mean access delay of a CSMA/CA-based MU-MIMO WLAN, where clients are allowed to contend for the concurrent transmission opportunities. We also considered a simple distributed opportunistic transmission scheme, where clients are able to join the ongoing transmissions only when their concurrent rates exceed a threshold. Analytical models were developed to characterize the network performance of the transmission schemes. Comparisons between simulation and analytical results were conducted to demonstrate the validity of our analytical model. By means of the analytical model, the saturation throughput and the mean access delay were investigated with respect to four parameters. Specifically, we optimized the network performance over the backoff window sizes, the network sizes, and the threshold of the opportunistic transmission scheme. We found that the throughput gain from adding one antenna at the AP reduces as the total number of antennas grows. Performance variations with respect to different parameters were analyzed thoroughly.

Our modeling and analysis provide insights into the CSMA/CA-based MU-MIMO transmission scheme. Besides, the developed theoretical model offers a helpful tool for future study of the CSMA/CA-based MAC protocols that allow concurrent transmissions. For the opportunistic transmission scheme, future work includes developing an algorithm to determine the optimal thresholds for the concurrent rates. Another important research direction would be to build a more general model by considering the situations when the number of concurrent transmissions is less than the maximum number.

References

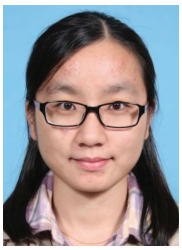
[1] D. Tse, P. Viswanath, *Fundamentals of Wireless Communication*, Cambridge, UK: Cambridge University Press, 2005.

[2] K. Tan, H. Liu, J. Fang, W. Wang, J. Zhang, M. Chen, and G. M. Voelker, "SAM: enabling practical spatial multiple access in wireless LAN," in *Proc. 2009 ACM MobiCom*, pp. 49-60.

[3] K. C.-J. Lin, S. Gollakota, and D. Katabi, "Random access heterogeneous MIMO networks," in *Proc. 2011 ACM SIGCOMM*, pp. 146-157.

[4] W. L. Shen, Y. C. Tung, K. C. Lee, K. C.-J. Lin, S. Gollakota, D. Katabi, and M. S. Chen, "Rate adaptation for 802.11 multiuser MIMO networks," in *Proc. 2012 ACM MobiCom*, pp. 29-40.

[5] 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.
[6] G. Bianchi and I. Tinnirello, "Remarks on IEEE 802.11 DCF performance analysis," *IEEE Commun. Lett.*, vol. 9, no. 8, Aug. 2005.
[7] Local and Metropolitan Area Networks–Specific Requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, *IEEE Std 802.11*, 2012.
[8] F. Calì, M. Conti, and E. Gregori, "Dynamic tuning of the IEEE 802.11 protocol to achieve a theoretical throughput limit," *IEEE/ACM Trans. Netw.*, vol. 8, no. 6, pp. 785-799, Dec. 2000.
[9] P. X. Zheng, Y. J. Zhang, and S. C. Liew, "Multipacket reception in wireless local area networks," in *Proc. 2006 IEEE International Conference on Communications*, pp. 3670-3675.
[10] W. L. Huang, K. B. Letaief, and Y. J. Zhang, "Cross-layer multi-packet reception based medium access control and resource allocation for space-time coded MIMO/OFDM," *IEEE Trans. Wireless Commun.*, vol. 7, no. 9, pp. 3372-3384, 2008.
[11] H. Jin, B. C. Jung, H. Y. Hwang, and D. K. Sung, "Performance comparison of uplink WLANs with single-user and multi-user MIMO schemes," in *Proc. 2008 IEEE Wireless Communications and Networking Conference*, pp. 1854-1859.
[12] H. Jin, B. C. Jung, H. Y. Hwang, and D. K. Sung, "A throughput balancing problem between uplink and downlink in multi-user MIMO-based WLAN systems," in *Proc. 2009 IEEE Wireless Communications and Networking Conference*.
[13] H. Jin, B. C. Jung, and D. K. Sung, "A tradeoff between single-user and multi-user MIMO schemes in multi-rate uplink WLANs," *IEEE Trans. Wireless Commun.*, vol. 10, no. 10, pp. 3332-3342, 2011.
[14] S. Yoon, I. Rhee, B. C. Jung, B. Daneshrad, and J. H. Kim, "Contrabass: concurrent transmissions without coordination for ad hoc networks," in *Proc. 2011 IEEE INFOCOM*, pp. 1134-1142.
[15] Z. Shen, R. Chen, J. G. Andrews, R. W. Heath, and B. L. Evans, "Low complexity user selection algorithms for multiuser MIMO systems with block diagonalization," *IEEE Trans. Signal Process.*, vol. 54, no. 9, pp. 3658-3663, 2006.
[16] B. Fan, W. Wang, Y. Lin, L. Huang, and K. Zheng, "Spatial multi-user pairing for uplink virtual-MIMO systems with linear receiver," in *Proc. 2009 IEEE Wireless Communications and Networking Conference*.
[17] R. Chen, Z. Shen, J. G. Andrews, and R. W. Heath, "Multimode transmission for multiuser MIMO systems with block diagonalization," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 3294-3302, 2008.
[18] S. Wu, W. Mao, and X. Wang, "Performance analysis of random access multi-user MIMO wireless LANs," in *Proc. 2013 IEEE Global Communications Conference*.



Shanshan Wu received the B.S. degree in Electrical and Computer Engineering from Shanghai Jiao Tong University (SJTU), Shanghai, China, in 2011 and the M.S. degree in Electronics Science and Technology from SJTU in 2014. She was an exchange student at the University of Hong Kong, China, in 2010. Her current research interests include rateless coding schemes, multi-user MIMO systems, two-way relaying techniques, and their applications in wireless communications and networking.

Wenguang Mao received the B.S. degree in Electrical and Computer Engineering from Shanghai Jiao Tong University (SJTU), Shanghai, China, in 2011 and the M.S. degree in Information and Communication Engineering from SJTU in 2014. His current research interests include random access MAC protocols, physical-layer cooperative coding schemes, software-defined radios, as well as mobile applications in smart phones and wearable computers.

Xudong Wang is currently with the UM-SJTU Joint Institute, Shanghai Jiao Tong University. He is a distinguished professor (Shanghai Oriental Scholar) and is the director of the Wireless and Networking (WANG) Lab. He is also an affiliate faculty member with the Electrical Engineering Department at the University of Washington. Since he received the Ph.D. degree in Electrical and Computer Engineering from Georgia Institute of Technology in August 2003, he has been working as a senior research engineer, senior network architect, and R&D manager in several companies. He has been actively involved in R&D, technology transfer, and commercialization of various wireless networking technologies. His research interests include wireless communication networks, smart grid, and cyber physical systems. He holds several patents on wireless networking technologies and most of his inventions have been successfully transferred to products. Dr. Wang is an editor for IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, Elsevier *Ad Hoc Networks*, and ACM/Kluwer *Wireless Networks*. He was also a guest editor for several journals. 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 has been a technical committee member of many international conferences and a technical reviewer for numerous international journals and conferences. Dr. Wang is a senior member of IEEE and was a voting member of IEEE 802.11 and 802.15 Standard Committees.