Backoff Design for IEEE 802.11 DCF Networks: Fundamental Tradeoff and Design Criterion

Xinghua Sun, *Member, IEEE* and Lin Dai, *Senior Member, IEEE*

***Abstract*—Binary Exponential Backoff (BEB) is a key component of the IEEE 802.11 DCF protocol. It has been shown that BEB can achieve the theoretical limit of throughput as long as the initial backoff window size is properly selected. It, however, suffers from significant delay degradation when the network becomes saturated. It is thus of special interest for us to further design backoff schemes for IEEE 802.11 DCF networks which can achieve comparable throughput as BEB, but provide better delay performance.**

This paper presents a systematic study on the effect of backoff schemes on throughput and delay performance of saturated IEEE 802.11 DCF networks. In particular, a backoff scheme is defined as a sequence of backoff window sizes $\{W_i\}$. The analysis shows that a saturated IEEE 802.11 DCF network has a single steady-state operating point as long as $\{W_i\}$ is a monotonic increasing sequence. The maximum throughput is found to be independent of $\{W_i\}$, yet the growth rate of $\{W_i\}$ determines a fundamental tradeoff between throughput and delay performance. For illustration, Polynomial Backoff is proposed, and the effect of polynomial power x on the network performance is characterized. It is demonstrated that Polynomial Backoff with a larger x is more robust against the fluctuation of the network size, but in the meanwhile suffers from a larger second moment of access delay. Quadratic Backoff (QB), i.e., Polynomial Backoff with x=2, stands out to be a favorable option as it strikes a good balance between throughput and delay performance. The comparative study between QB and BEB confirms that QB well preserves the robust nature of BEB, and achieves much better queueing performance than BEB.

***Index Terms*—IEEE 802.11 DCF networks, Binary Exponential Backoff, Quadratic Backoff, Polynomial Backoff, maximum throughput, access delay**

I. Introduction

IEEE 802.11 wireless local area networks have gained significant attention in both industry and academia. Fueled by the widespread popularity in commercial use, research activities have been intensified over the last few years, and a major focus has been put on the medium access control (MAC) layer with distributed coordination function (DCF).

As a key component of the IEEE 802.11 DCF protocol, Binary Exponential Backoff (BEB) plays a crucial role in determining the whole network performance. With BEB, the backoff window size is doubled after each unsuccessful transmission. By doing so, the transmission probability of each node can be effectively reduced, thus alleviating the contention in a distributed way. A widely adopted model of IEEE 802.11 DCF networks was proposed by Bianchi in his landmark paper [1], where it was shown that BEB can achieve quite satisfying throughput performance if the initial backoff window size W is carefully selected. A number of studies have been motivated to further focus on the dynamic tuning of the backoff window size according to the estimated number of active nodes [2]–[5] or the transmission failure rate [6].

In the meanwhile, the so-called "capture phenomenon" of BEB was also observed [7]–[9]. In particular, when the network becomes saturated, many nodes are pushed to deep states with extremely small retransmission probabilities such that the node who once succeeds can dominate the channel for a long time and produce a continuous stream of packets. In that case, the output process is no longer stationary, and nodes suffer from severe short-term unfairness [10]–[13]. A lot of effort has been made to address this issue [14]–[24]. A common belief is that unfairness is caused by the fact that the backoff window size is reset to the initial value upon successful transmission, which gives the fresh packets great advantages to succeed in the contention. Based on this observation, many schemes were proposed to slow down the decrement of the backoff window size, including Exponential Decrease [16]–[19], Linear Decrease [20]–[22], Sliding Contention Window [23], and Gentle DCF [24] in which the window size is halved after several successful transmissions.

Distinct analytical models have been established in the aforementioned studies, which makes the performance evaluation of key system parameters and various backoff schemes extremely difficult. In our recent work [25], a unified analytical framework was proposed to study the stability, throughput and delay performance of homogeneous buffered IEEE 802.11 DCF networks. In contrast to the classical Bianchi model [1], the behavior of each Head-of-Line (HOL) packet, including backoff, collision and successful transmission, is modeled as a discrete-time Markov renewal process, and two steady-state operating points are characterized by using the limiting probability of successful transmission of HOL packets given that the network is in unsaturated or saturated conditions. The analysis shows that the maximum throughput $\hat{\lambda}_{\max}$ of IEEE 802.11 DCF networks with Exponential Backoff is a function of the holding time of HOL packets in successful transmission and collision states, and BEB can achieve $\hat{\lambda}_{\max}$ as long as the initial backoff window size W is linearly adjusted according to the network size n. Nevertheless, the second moment of

Manuscript received November 22, 2012; revised August 23, 2013 and November 22, 2013. The associate editor coordinating the review of this paper and approving it for publication was G. Bianchi.

This work was fully supported by the Research Grants Council (RGC) of Hong Kong under GRF Grant CityU 112810.

X. Sun is with the Key Laboratory of Wireless Communications, Nanjing University of Posts and Telecommunications, Nanjing, China. He was with City University of Hong Kong (email: sxhgoahead@gmail.com).

L. Dai is with the Department of Electronic Engineering, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon Tong, Hong Kong, China (email: lindai@cityu.edu.hk).

access delay may grow exponentially with the cutoff phase K for a small W. In a saturated network where nodes are pushed to deep states, BEB leads to a large second moment of access delay, from which the observed capture phenomenon and the short-term unfairness stem.

The performance analysis of BEB stimulates a series of questions about the optimal backoff design criterion of IEEE 802.11 DCF networks. For instance, it was shown in [25] that the maximum throughput $\hat{\lambda}_{\max}$ is independent of backoff parameters and solely determined by which access mechanism, i.e., the basic access mechanism or the request-to-send/clear-to-send (RTS/CTS) access mechanism, is adopted. Does it suggest that all the backoff schemes achieve exactly the same maximum throughput? If so, how to achieve it? Moreover, it seems that the growth rate of the backoff window size critically determines the network performance. With BEB, the exponential growth of backoff window size leads to a huge difference of window sizes between a fresh packet and a deeply backlogged one, causing a large second moment of access delay. Is it possible for us to find a "milder" backoff scheme which achieves good throughput performance as BEB, but with less severe delay jitter?

This paper is devoted to a comprehensive study on backoff design of IEEE 802.11 DCF networks. In particular, the unified analytical framework proposed in [25] is further extended to incorporate a general backoff scheme which is defined as a sequence of backoff window sizes $\{W_i\}$.[1] It is shown that a saturated IEEE 802.11 DCF network has a single steady-state operating point, p_A, as long as $\{W_i\}$ is a monotonic increasing sequence. p_A is a function of $\{W_i\}$, indicating that the performance of a saturated IEEE 802.11 DCF network is closely dependent on the backoff scheme.

The analysis verifies that the maximum throughput $\hat{\lambda}_{\max}$ is independent of $\{W_i\}$ and solely determined by the access mechanism. To achieve $\hat{\lambda}_{\max}$, however, the backoff parameters should be carefully selected. According to whether the limiting growth rate of the backoff window size is larger than 1, the backoff schemes can be categorized into two groups, i.e., aggressive backoff with $\lim_{i\to\infty} \frac{W_{i+1}}{W_i} > 1$ and mild backoff with $\lim_{i\to\infty} \frac{W_{i+1}}{W_i} = 1$. The aggressive backoff schemes can achieve the maximum throughput even as the network size n goes to infinity. The second moment of access delay, however, may become unbounded if n exceeds a certain threshold. BEB, as a representative aggressive backoff scheme, suffers from such a delay degradation when the network size n is large.

With the mild backoff schemes, the maximum throughput and a bounded second moment of access delay can be both achieved for any finite network size. Moreover, it is found that the growth rate of the backoff window size can be further steered to strike a balance between throughput and delay performance. For demonstration, Polynomial Backoff[2] is proposed, and the effect of polynomial power x on the network performance is characterized. Intuitively, a larger x indicates a faster growth rate of the backoff window size, with which the network is better capable of absorbing the mounting contention. It, on the other hand, may lead to a more severe delay jitter due to a larger difference of backoff window sizes between a fresh packet and a deeply backlogged one. Quadratic Backoff (QB), i.e., Polynomial Backoff with x=2, stands out to be a favorable option as it provides a good tradeoff between throughput and delay performance.

The comparative study between QB and BEB further shows that both schemes achieve the same maximum throughput and are robust against the variation of network size n. Yet the queueing performance can be significantly improved with QB, which is consistent with the observation in [26] and [27]. The short-term fairness performance of QB and BEB is also evaluated, and compared with two representative slow-decrement backoff schemes, Exponential Increase Exponential Decrease (EIED) [16] and Exponential Increase Linear Decrease (EILD) [20]. The comparison corroborates that the key to improving fairness lies in the growth rate of the backoff window size. By choosing a mild growth rate, the proposed QB achieves the best short-term fairness performance.

The remainder of this paper is organized as follows. Section II presents a detailed review of backoff algorithms proposed in the literature. Section III presents the analytical model and preliminary analysis. Both throughput and delay performance of a saturated IEEE 802.11 DCF network with a general backoff scheme is characterized in Section IV. Polynomial Backoff is proposed and analyzed in Section V, and a comparative study of QB and BEB is presented in Section VI. Finally, conclusions are summarized in Section VII.

II. Overview of Backoff Design

Backoff is a key component of random-access networks. It can be characterized as a sequence of transmission probabilities $\{q_t\}$, where q_t denotes the transmission probability of the HOL packet in each node's buffer at time slot t, in a homogeneous random-access network. Early work has shown that if q_t is constant, the throughput of Aloha networks would dramatically decrease as the network size increases [28], [29]. Intuitively, q_t should be adaptively tuned to alleviate the time-varying channel contention. The ultimate aim of backoff design is then to find how to properly set the sequence of transmission probabilities $\{q_t\}$ to optimize the network performance.

A prevailing method in the literature is to adjust the transmission probability q_t according to the number of collisions that the HOL packet has experienced by time slot t, i.e., $q_t = \mathcal{Q}(i)$, where $\mathcal{Q}(i)$ is an arbitrary monotonic decreasing function[3] of the number of collisions i. The widely adopted BEB [30] is a typical example, with the backoff function $\mathcal{Q}(i) = 2^{-i}$. In addition to BEB, numerous backoff schemes have been proposed and extensively studied based on different backoff functions such as exponential function [9], [27], [31], linear function [27], [32], μ-law function [32], step function

[1]Here i denotes the number of collisions that the packet has experienced, and W_i can be an arbitrary function of i. With BEB, for instance, $W_i = W \cdot 2^i, i = 0, \ldots, K$, where W is the initial backoff window size and K is the cutoff phase (which is referred to as the maximum backoff stage in [1]).

[2]The backoff window size with Polynomial Backoff is $W_i = W \cdot (1 + i)^x, i = 0, \ldots, K$, where the polynomial power x is a non-negative integer.

[3]Intuitively, to alleviate the channel contention, nodes should reduce their transmission probabilities as they experience more collisions.

[32] and polynomial function [26], [27]. It was shown in [32] that μ-law function and step function are preferable to the exponential one in terms of the delay performance when the network size is large. Similar observations were made in [26], [27] that polynomial function could improve the delay performance compared to BEB. Due to the lack of a unified analytical framework, nevertheless, it remains largely unknown how to properly choose the backoff function to optimize the network performance.

Note that there are also backoff schemes that do not fall into the above category. For example, it was proposed in [33] that the transmission probability q_t should be tuned according to the number of backlogged nodes n_t in an Aloha network. Similar schemes could be found in CSMA networks [34], [35] and IEEE 802.11 DCF networks[4] [2]–[5]. In [16]–[24], the initial backoff window size of a fresh HOL packet is decremented according to the window sizes of its precedents, in which case the transmission probability of each HOL packet is determined by not only how many collisions it experiences, but when it enters the network.

In this paper, we limit our discussion to backoff schemes where the transmission probability q_t can be fully characterized as a function of the number of collisions i that the HOL packet has experienced by time slot t. As IEEE 802.11 DCF networks adopt the window mechanism, we focus on the window-based counterpart, where the backoff scheme is characterized by a sequence of backoff window sizes $\{W_i\}$. We will analyze the effect of $\{W_i\}$ on the throughput and delay performance of a saturated IEEE 802.11 DCF network, and reveal the design criterion to achieve the optimal network performance.

III. Preliminary Analysis

Let us first briefly review the unified analytical framework proposed for IEEE 802.11 DCF networks in [25]. Specifically, we consider an n-node homogeneous IEEE 802.11 DCF network with packet transmissions over a noiseless channel. Suppose that each node is equipped with a buffer of infinite size and the maximum number of retransmission attempts for each HOL packet is infinite. Different from the widely adopted Bianchi model [1] where the backoff process of each node was modeled as a two-dimensional Markov chain, in the following subsection, a discrete-time Markov renewal process will be established to characterize the complete behavior of each HOL packet including backoff, collision and successful transmission.

A. State Characterization of HOL Packets

Let X_j denote the state of a tagged HOL packet at the j-th transition and V_j denotes the epoch at which the j-th transition occurs. Fig. 1 shows the embedded Markov chain $\boldsymbol{X} = \{X_j\}$ of the discrete-time Markov renewal process $(\boldsymbol{X}, \boldsymbol{V}) = \{(X_j, V_j), j = 0, 1, \dots\}$.

[4]Note that a window-based backoff mechanism is adopted in IEEE 802.11 DCF networks. Specifically, instead of attempting transmission with a certain probability at each idle time slot, the node would choose a random number from a backoff window, count down when the channel is idle, and transmit the HOL packet if the counter is zero.

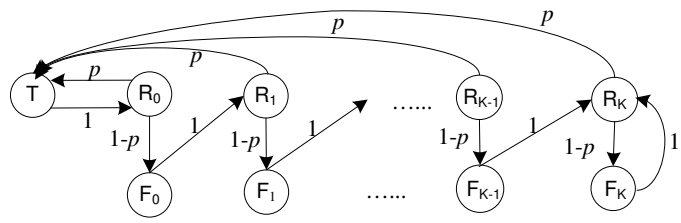


Fig. 1. Embedded Markov chain of the state transition process of an individual HOL packet in IEEE 802.11 DCF networks.

The states of $\{X_j\}$ can be divided into three categories: 1) waiting to request a transmission (State R_i, $i = 0, \dots, K$), 2) collision (State F_i, $i = 0, \dots, K$) and 3) successful transmission (State T). As Fig. 1 illustrates, a HOL packet moves from State R_i to State T if the transmission request is successful. Otherwise, it stays at State F_i until the end of the collision and then shifts to State R_{i+1}. Here i denotes the number of collisions experienced by the HOL packet and is incremented until it reaches the cutoff phase K (which is referred to as the maximum backoff stage in [1]). Note that K can be any nonnegative integer from 0 to ∞.

In IEEE 802.11 DCF networks, each HOL packet can request a transmission only if it senses the channel idle. Let p represent the steady-state probability of successful transmission of HOL packets given that the channel is idle. The stead-state probability distribution of the embedded Markov chain shown in Fig. 1 can be obtained as

$$\pi_{R_i} = \begin{cases} (1-p)^i \pi_T & i = 0, ..., K-1 \\ \frac{(1-p)^K}{p} \pi_T & i = K \end{cases} \tag{1}$$

and

$$\pi_{F_i} = \pi_{R_i} \cdot (1-p), \quad i = 0, \dots, K. \tag{2}$$

The interval between successive transitions, i.e., $V_{j+1} - V_j$, is called the holding time in State X_j, which solely depends on State X_j. In IEEE 802.11 DCF networks, the holding time τ_T in State T and the holding time τ_F in State $\mathrm{F}_i, i = 0, \dots, K$, vary under different access mechanisms. A graphic illustration of τ_T and τ_F in the basic and RTS/CTS access mechanisms can be found in Fig. 5 of [1] (corresponding to T_s and T_c, respectively, in unit of time slots), and the typical values are provided in Table V of [1].

The mean holding time τ_{R_i} in State $\mathrm{R}_i, i = 0, \dots, K$, on the other hand, is determined by the backoff protocol. In IEEE 802.11 DCF networks, when a HOL packet enters State R_i, it randomly selects a value from $\{0, \dots, W_i - 1\}$, where W_i is the backoff window size, $i = 0, \dots, K$, and then counts down at each idle time slot. It leaves State R_i and makes a transmission request when the channel is idle and the counter is zero. It is shown in [25] that the mean holding time τ_{R_i} is given by

$$\tau_{R_i} = \frac{1}{2\alpha}(W_i + 1), \tag{3}$$

$i = 0, \dots, K$, where α denotes the steady-state probability of sensing the channel idle, which is given by

$$\alpha = \frac{1}{1 + \tau_F - \tau_F p - (\tau_T - \tau_F) p \ln p}. \tag{4}$$

$$G''_{D_0}(1)=\sum_{i=0}^{K-1}(1-p)^i G''_{Y_i}(1)+\frac{(1-p)^K}{p}G''_{Y_K}(1)+2\left(p\tau_T+(1-p)\tau_F\right)\cdot\left(\sum_{i=0}^{K-1}(1-p)^i G'_{Y_i}(1)+\frac{(1-p)^K}{p}G'_{Y_K}(1)\right)$$
$$+2\sum_{i=0}^{K-1}\left(\tau_F+G'_{Y_i}(1)\right)\cdot\left(\sum_{j=i+1}^{K-1}(1-p)^j\left(p\tau_T+(1-p)\tau_F+G'_{Y_j}(1)\right)+\frac{(1-p)^K}{p}\left(p\tau_T+(1-p)\tau_F+G'_{Y_K}(1)\right)\right)$$
$$+2\frac{(1-p)^{K+1}}{p^2}\left(\tau_F+G'_{Y_K}(1)\right)\cdot\left(p\tau_T+(1-p)\tau_F+G'_{Y_K}(1)\right)+\tau_T\left(\tau_T-1\right)+\frac{1-p}{p}\tau_F\left(\tau_F-1\right). \quad (10)$$

Finally, the limiting state probabilities of the Markov renewal process (***X***, ***V***) are given by

$$\tilde{\pi}_j=\frac{\pi_j\cdot\tau_j}{\sum_{i\in S}\pi_i\cdot\tau_i}, \quad (5)$$

$j\in S$, where S is the state space of ***X***. Specifically, the probability of being in State T can be obtained as

$$\tilde{\pi}_T=\tau_T\Big/\left(\tau_T+\frac{1-p}{p}\tau_F+\frac{1+\tau_F-\tau_F p-(\tau_T-\tau_F)p\ln p}{2}\cdot\left(\sum_{i=0}^{K-1}(1-p)^i W_i+\frac{(1-p)^K}{p}W_K+\frac{1}{p}\right)\right), \quad (6)$$

by substituting (1-4) into (5). Note that $\tilde{\pi}_T$ is also the service rate of each node's queue as each queue has a successful output if and only if the HOL packet stays at State T.

B. First and Second Moments of Access Delay

The access delay performance can be also characterized based on the embedded Markov chain shown in Fig. 1. Let Y_i denote the holding time of a HOL packet in State R_i, and D_i denote the time spent from the beginning of State R_i until the service completion, $i=0,\ldots,K$. According to Fig. 1, we have

$$D_i=\begin{cases}Y_i+\tau_T & \text{with probability } p\\ Y_i+\tau_F+D_{i+1} & \text{with probability } 1-p,\end{cases} \quad (7)$$

i=0, ..., $K-1$, and

$$D_K=\begin{cases}Y_K+\tau_T & \text{with probability } p\\ Y_K+\tau_F+D_K & \text{with probability } 1-p,\end{cases} \quad (8)$$

where τ_T and τ_F are holding time in State T and States F_i, $i=0,\ldots,K$, respectively.

Note that D_0 is the service time of HOL packets (also the access delay). Let $G_{D_0}(z)$ denote its probability generating function. It is shown in [25] that

$$G'_{D_0}(1)=\tau_T+\frac{1-p}{p}\tau_F+\sum_{i=0}^{K-1}(1-p)^i G'_{Y_i}(1)+\frac{(1-p)^K}{p}G'_{Y_K}(1), \quad (9)$$

and $G''_{D_0}(1)$ is given in (10), which is shown at the top of this page, where $G'_{Y_i}(1)$ and $G''_{Y_i}(1)$ are given by

$$G'_{Y_i}(1)=\frac{1}{2\alpha}\left(W_i+1\right), \quad (11)$$

and

$$G''_{Y_i}(1)=\frac{1}{3\alpha^2}W_i^2+\frac{1-\alpha}{\alpha^2}W_i+\frac{2-3\alpha}{3\alpha^2}, \quad (12)$$

respectively, i=0, ..., K.

Finally, the mean access delay $E[D_0]$ (in the unit of time slots) is given by

$$E[D_0]=G'_{D_0}(1), \quad (13)$$

and the second moment of access delay $E[D_0^2]$ is

$$E[D_0^2]=G''_{D_0}(1)+G'_{D_0}(1), \quad (14)$$

which can be obtained by substituting (9-12) into (13) and (14), respectively.

The above analysis clearly indicates that the network performance critically depends on the steady-state probability of successful transmission of HOL packets given that the channel is idle, p. It is revealed in [25] that when the network is unsaturated, it operates at the desired stable point $p=p_L$ which is independent of backoff parameters and solely determined by the aggregate input rate $\hat{\lambda}$, the holding time τ_T in State T and the holding time τ_F in State $\mathrm{F}_i, i=0,\ldots,K$. The network throughput $\hat{\lambda}_{out}$ at the desired stable point p_L is always equal to the aggregate input rate $\hat{\lambda}$. In contrast, if all the nodes become saturated with non-empty queues, the network operating point will shift to the undesired stable point $p=p_A$, at which the network performance is closely dependent on backoff parameters such as the cutoff phase K and the backoff window size $W_i, i=0,\ldots,K$. In the following section, we will characterize the effect of backoff parameters on the throughput and delay performance of saturated networks.

IV. Throughput and Delay Performance of Saturated Networks

In this section, we consider a saturated network, in which all the nodes are busy with non-empty queues, and the network throughput $\hat{\lambda}_{out}$ falls below the aggregate input rate $\hat{\lambda}$. In that case, all the HOL packets must be in State R_i, $i=0,\ldots,K$, if the channel is idle. Moreover, for a given HOL packet, its transmission request is successful if and only if the other $n-1$ nodes are not requesting any transmission. The steady-state probability of successful transmission of HOL packets given that the channel is idle, p, can then be written as

$$p=\left\{\frac{\sum_{i=0}^{K}\tilde{\pi}_{R_i}(1-r_i)}{\sum_{i=0}^{K}\tilde{\pi}_{R_i}}\right\}^{n-1}\overset{\text{with a large } n}{\approx}\exp\left\{-n\sum_{i=0}^{K}\tilde{\pi}_{R_i}r_i\right\}, \quad (15)$$

where r_i is the conditional probability of a State-R_i HOL packet making a transmission request given that the channel is idle. It is shown in [25] that

$$r_i=\frac{2}{1+W_i}, \quad (16)$$

$i = 0, \ldots, K$. By substituting (1-3), (5) and (16) into (15), we have

$$p = \exp\left\{-n/\left(\alpha(\tau_T \cdot p + \tau_F \cdot (1-p)) + \frac{1}{2}\left(1 + \sum_{i=0}^{K-1} p(1-p)^i \cdot W_i + (1-p)^K \cdot W_K\right)\right)\right\}, \quad (17)$$

where α is the probability of sensing the channel idle which is given in (4). Intuitively, in a saturated network, there is a small probability for each node to sense the channel idle or successfully transmit due to a large number of transmission requests. With a small α and p, the first term of the denominator in the right-hand side of (17) is much smaller than the second term, and can be ignored.[5] (17) can then be written as

$$p = \exp\left\{-\frac{2n}{1 + \sum_{i=0}^{K-1} p(1-p)^i W_i + (1-p)^K W_K}\right\}. \quad (18)$$

Note that (18) is consistent to Eq. (4) in [36].[6] It was shown in Theorem 5.1 [36] that Eq. (4) has a unique fixed point if the sequence of mean backoff duration $\{b_i\}$ is non-decreasing. The following theorem states the existence and uniqueness of the root of the fixed-point equation (18).

Theorem 1. *The fixed-point equation* (18) *has one single non-zero root p_A if $\{W_i\}$ is a monotonic increasing sequence.*

Proof: See Appendix A. ■

Similar to [25], we refer to p_A as the undesired stable point. It is clear from (18) that the undesired stable point p_A is determined by the sequence of backoff window sizes $\{W_i\}$, the network size n and the cutoff phase K. Intuitively, a smaller network size n leads to a better chance of successful transmission, implying a larger p_A. Corollary 1 summarizes the monotonicity properties of p_A with regard to the cutoff phase K and the network size n.

Corollary 1. *If $\{W_i\}$ is a monotonic increasing sequence, then 1) p_A is a monotonic increasing function of the cutoff phase K;*
2) p_A is a monotonic decreasing function of the network size n.

Proof: See Appendix B. ■

According to Corollary 1, p_A decreases as the network size n increases. As n goes to infinity, it is clear from (18) that

$$\lim_{n\to\infty} p_A^{K<\infty} = 0. \quad (19)$$

The cutoff phase K determines the maximum backoff window size nodes can have. A larger cutoff phase K indicates that nodes have more room to reduce their transmission probabilities if they experience collisions, and the network is then better capable of absorbing the mounting contention as the network size n grows. With $n\to\infty$, the huge contention brought by infinite competing nodes cannot be alleviated if K is finite, thus dragging the probability of successful transmission down to zero. With an infinite cutoff phase $K = \infty$, on the other hand, nodes can always back off to deeper states no matter how large n is. Corollary 2 shows that with $K = \infty$, a positive $p_A^{K=\infty} > 0$ as $n \to \infty$ is possible if the limiting growth rate of backoff window size is larger than 1.

Corollary 2. $\lim_{n\to\infty} p_A^{K=\infty} = 1 - \lim_{i\to\infty} \frac{W_i}{W_{i+1}}$.

Proof: See Appendix C. ■

With a monotonic increasing sequence of backoff window sizes $\{W_i\}$, the limiting growth rate of backoff window size is

$$\lim_{i\to\infty} \frac{W_{i+1}}{W_i} \geq 1. \quad (20)$$

Accordingly, we can divide the backoff schemes into two categories:

1) *Aggressive Backoff*: $\lim_{i\to\infty} \frac{W_{i+1}}{W_i} > 1$;
2) *Mild Backoff*: $\lim_{i\to\infty} \frac{W_{i+1}}{W_i} = 1$.

It is clear from Corollary 2 that for the mild backoff schemes, $\lim_{n\to\infty} p_A^{K=\infty} = 0$. In this case, the growth rate of the backoff window size is not fast enough to catch up with the ever-increasing contention as the network size n goes to infinity, and thus the network will eventually collapse with no packets getting through. For the aggressive backoff schemes, in contrast, a positive $p_A^{K=\infty} > 0$ can be achieved even with an infinite number of nodes n. For instance, with Exponential Backoff [31], i.e., $W_i = W \cdot q^{-i}$ for some $q \in (0, 1)$, the limiting growth rate of the backoff window size is $\lim_{i\to\infty} \frac{W_{i+1}}{W_i} = 1/q > 1$, and we have

$$\lim_{n\to\infty} p_A^{EB,K=\infty} = 1 - q > 0, \quad (21)$$

according to Corollary 2. In this case, a non-zero network throughput can be achieved even as the network size n grows without bound. In the following subsections, we will specifically discuss the throughput and delay performance of the aggressive and mild backoff schemes.

A. Saturation Throughput

In a saturated network with non-empty queues, the throughput is usually referred to as saturation throughput $\hat{\lambda}_s$ [1], which is equal to the aggregate service rate $n\tilde{\pi}_T$. By combining (4), (6) and (17), the saturation throughput $\hat{\lambda}_s$ can be written as

$$\hat{\lambda}_s = \frac{-\tau_T p_A \ln p_A}{1 + \tau_F - \tau_F p_A - (\tau_T - \tau_F) p_A \ln p_A}. \quad (22)$$

The following theorem presents the maximum throughput $\hat{\lambda}_{\max} = \max_{p_A} \hat{\lambda}_s$ and the corresponding steady-state point p_A^*.

Theorem 2. *The maximum throughput $\hat{\lambda}_{\max}$ is given by*

$$\hat{\lambda}_{\max} = \frac{-\mathbb{W}_0\left(-\frac{1}{e(1+1/\tau_F)}\right)}{\tau_F/\tau_T - (1 - \tau_F/\tau_T)\mathbb{W}_0\left(-\frac{1}{e(1+1/\tau_F)}\right)}, \quad (23)$$

[5] Specifically, with a small p, $\alpha(\tau_T \cdot p + \tau_F \cdot (1-p)) \approx \frac{\tau_F}{1+\tau_F} < 1$, which is much smaller than the second term of the denominator.

[6] Note that different notations were used in [36]. Specifically, in Eq. (4) of [36], γ and b_i denote the probability of collision and the mean backoff duration for the i-th attempt, respectively, which can be written as $\gamma = 1 - p$, $b_i = \frac{1+W_i}{2}$, for $i \leq K-1$, and $b_i = \frac{1+W_K}{2}$, for $i \geq K$ by using our notations.

$$E[D_{0,p=p_A}^{2,K=\infty}] = \frac{3+p_A-3\alpha p_A}{6\alpha^2 p_A^2} + \left(\frac{2}{\alpha}+(2-p_A)\tau_F\right)\frac{(1-p_A)\tau_F}{p_A^2} + \left(\tau_T + 2\frac{1-p_A}{p_A}\tau_F + \frac{1}{\alpha p_A}\right)\tau_T$$
$$+ \left(\frac{1-\alpha}{2\alpha^2} + \frac{1}{2\alpha^2 p_A} + \frac{p_A\tau_T+(1-p_A)\tau_F}{\alpha p_A}\right)\sum_{i=0}^{\infty}(1-p_A)^i W_i + \frac{1}{\alpha}\left(\tau_F + \frac{1}{2\alpha}\right)\sum_{i=0}^{\infty}\left(\sum_{j=i+1}^{\infty}(1-p_A)^j W_j\right)$$
$$+ \frac{1}{3\alpha^2}\sum_{i=0}^{\infty}(1-p_A)^i W_i^2 + \frac{1}{2\alpha^2}\sum_{i=0}^{\infty} W_i\left(\sum_{j=i+1}^{\infty}(1-p_A)^j W_j\right). \quad (28)$$

where $\mathbb{W}_0$ is the principal branch of the Lambert W function [37]. $\hat{\lambda}_{\max}$ is achieved when

$$p_A = p_A^* = -(1+1/\tau_F)\mathbb{W}_0\left(-\frac{1}{e(1+1/\tau_F)}\right). \quad (24)$$

Proof: See Appendix D. ■

We can see from Theorem 2 that the maximum throughput $\hat{\lambda}_{\max}$ is equal to that of IEEE 802.11 DCF networks with Exponential Backoff which is given in Eq. (36) in [25]. It is solely determined by τ_T and τ_F, i.e., the holding time of HOL packets in the successful transmission and collision states, which indicates that the maximum throughput is the same for all the backoff schemes. To achieve $\hat{\lambda}_{\max}$, however, the backoff parameters should be carefully tuned such that $p_A = p_A^*$.

Recall that p_A declines as the network size n increases according to Corollary 1. For the mild backoff schemes, p_A approaches zero as n goes to infinity, implying a zero throughput. With the aggressive backoff schemes, in contrast, the maximum throughput $\hat{\lambda}_{\max}$ is achievable even with an infinite number of nodes. For instance, with Exponential Backoff, we can choose the retransmission factor $q = 1-p_A^*$, such that $\lim_{n\to\infty} p_A^{EB,K=\infty} = 1-q = p_A^*$. The maximum throughput $\hat{\lambda}_{\max}$ can then be achieved as the network size $n \to \infty$ according to Theorem 2. Corollary 3 presents the necessary and sufficient condition to achieve $\hat{\lambda}_{\max}$ as $n \to \infty$.

Corollary 3. *$\hat{\lambda}_{\max}$ is achievable as $n \to \infty$ if and only if $K = \infty$ and $\lim_{i\to\infty}\frac{W_{i+1}}{W_i} > 1$.*

Proof: See Appendix E. ■

B. Access Delay

The expressions of first and second moments of access delay have been shown in Section III. At the undesired stable point p_A, the mean access delay $E[D_{0,p=p_A}]$ can be obtained as

$$E[D_{0,p=p_A}] = \tau_T + \frac{1-p_A}{p_A}\tau_F + \frac{1+\tau_F-\tau_F p_A-(\tau_T-\tau_F)p_A\ln p_A}{2}$$
$$\cdot\left(\sum_{i=0}^{K-1}(1-p_A)^i W_i + \frac{(1-p_A)^K}{p_A}W_K + \frac{1}{p_A}\right), \quad (25)$$

by combining (4), (9), (11) and (13). We can see from (25) and (6) that the mean access delay $E[D_{0,p=p_A}]$ is inversely proportional to the saturation throughput $\hat{\lambda}_s$, i.e.,

$$E[D_{0,p=p_A}] = n\tau_T/\hat{\lambda}_s. \quad (26)$$

Theorem 2 shows that $\hat{\lambda}_s$ is maximized when $p_A = p_A^*$; therefore, $E[D_{0,p=p_A}]$ is minimized at

$$\min_{p_A} E[D_{0,p=p_A}] = n\left(\tau_T - \tau_F\left(1+1/\mathbb{W}_0\left(-\frac{1}{e(1+1/\tau_F)}\right)\right)\right), \quad (27)$$

when $p_A = p_A^*$. It is clear from (27) that all the backoff schemes can achieve the same minimum mean access delay, which is solely determined by the network size n and the holding time in the successful transmission and collision states, τ_T and τ_F.

The second moment of access delay at the undesired stable point p_A can be obtained by combining (9-12) and (14). With an infinite cutoff phase $K = \infty$, an explicit expression of the second moment of access delay $E[D_{0,p=p_A}^{2,K=\infty}]$ can be obtained as (28), which is shown at the top of the page. Theorem 3 presents the convergence property of $E[D_{0,p=p_A}^{2,K=\infty}]$.

Theorem 3. *If $\lim_{i\to\infty}\frac{W_{i+1}}{W_i} = 1$, $E[D_{0,p=p_A}^{2,K=\infty}] < \infty$ for any finite $n < \infty$; otherwise, if $\lim_{i\to\infty}\frac{W_{i+1}}{W_i} > 1$, there exists $n' < \infty$ such that $E[D_{0,p=p_A}^{2,K=\infty}] = \infty$ when $n > n'$.*

Proof: See Appendix F. ■

Theorem 3 indicates that for the mild backoff schemes, the second moment of access delay is always finite for any network size $n < \infty$. For the aggressive backoff schemes, in contrast, the second moment of access delay may become unbounded if the network size n exceeds a certain threshold n'. With Exponential Backoff, for instance, it is shown in [25] that its second moment of access delay is infinite if $n > n' = \frac{-(1+q)\ln(1-q^2)}{2}\cdot W$.

In saturated networks, HOL packets are normally pushed to deep states due to intensive contention. With the aggressive backoff schemes, the backoff window size W_i is constantly enlarged as the number of collisions i grows. As a result, nodes can always effectively reduce their transmission probabilities by backing off to deeper states to alleviate the contention, such that a non-zero throughput is achievable even as the network size n grows without bound. The cost is, nevertheless, a vast difference of window sizes between fresh packets and deeply backlogged ones. In that case, the node who once succeeds has a much higher transmission probability than those with deeply backlogged HOL packets. It can then capture the channel and produce a continuous stream of packets while other nodes have to wait for a long time. The access delay performance across the network becomes extremely unbalanced, indicating a huge second moment of access delay. This irregular behavior has long been observed in networks with Exponential Backoff,

and is referred to as the "capture phenomenon" [7]–[9]. Here Theorem 3 suggests that all the aggressive backoff schemes suffer from large delay jitter as well as severe short-term unfairness once the network grows to a certain size.

C. Tradeoff between Throughput and Delay

So far we have shown that the throughput and delay performance of saturated IEEE 802.11 DCF networks critically depends on the sequence of backoff window sizes $\{W_i\}$. According to whether the limiting growth rate of backoff window size exceeds one, the backoff schemes can be categorized into two groups: mild backoff and aggressive backoff. With the aggressive backoff schemes, the maximum throughput can be achieved even as the network size n goes to infinity. The delay performance, nevertheless, may significantly deteriorate due to an unbounded second moment of access delay when the network size becomes large. In contrast, the mild backoff schemes can reach a better balance: the maximum throughput and a bounded second moment of access delay can be both achieved as long as the network size is finite.

For the mild backoff schemes, the growth rate of backoff window size can be further tuned to trade off between the throughput and delay performance. For demonstration, let $W_i = W\cdot\omega(i)$, where W is the initial backoff window size and $\omega(i)$ is an arbitrary monotonic increasing function of i with $\omega(0) = 1$ and $\lim_{i\to\infty}\frac{\omega(i+1)}{\omega(i)} = 1$. According to Theorem 2 and (26), the optimal initial backoff window size to achieve the maximum throughput $\hat{\lambda}_{\max}$ and the minimum mean access delay can be obtained from (18) as

$$W_m=\frac{-\frac{2n}{\ln p_A^*}-1}{\sum_{i=0}^{K-1}p_A^*(1-p_A^*)^i\cdot\omega(i)+(1-p_A^*)^K\cdot\omega(K)}. \quad (29)$$

(29) shows that the optimal initial backoff window size W_m linearly increases with the network size n, and the slope is determined by $\omega(i)$. With a higher growth rate of $\omega(i)$, W_m becomes less sensitive against n, implying that a smaller throughput degradation is caused if the initial backoff window size is not updated with n in time.

In particular, suppose that the initial backoff window size W is set to be the optimal value $W_m^{n=n_0}$ when the network size n is equal to n_0. As n increases from n_0, if the initial backoff window size is not enlarged accordingly, the network stead-state point will deviate from p_A^*, causing a degradation of network throughput. With $K{=}\infty$, the derivative of the saturation throughput $\hat{\lambda}_s$ with regard to the network size n can be obtained from (18) and (22) as

$$\left|\frac{d\hat{\lambda}_s}{dn}\right|=\left|\frac{d\hat{\lambda}_s}{dp_A}\cdot\frac{dp_A}{dn}\right|=\frac{|-\tau_T(1{+}\tau_F)(1{+}\ln p_A){+}\tau_T\tau_F p_A|}{(1{+}\tau_F{-}\tau_F p_A{-}(\tau_T{-}\tau_F)p_A\ln p_A)^2}\cdot \frac{1}{(\frac{n_0}{-\ln p_A^*}-\frac{1}{2})\cdot\frac{\sum_{i=0}^{\infty}(1-p_A)^{i-1}\omega(i)((1+\ln p_A)(1-p_A)-ip_A\ln p_A)}{\sum_{i=0}^{\infty}p_A^*(1-p_A^*)^i\omega(i)}+\frac{1}{2p_A}}. \quad (30)$$

We can clearly see from (30) that the derivative decreases as the growth rate of $\omega(i)$ increases, indicating that the network throughput becomes more robust against the variation of the network size n.

On the other hand, the second moment of access delay with $W{=}W_m$ and $K{=}\infty$ can be explicitly written as

$$E[D_{0,p=p_A^*}^{2,K=\infty}] = C+\frac{4n^2\tau_T^2}{3\hat{\lambda}_{\max}^2}\cdot\frac{\sum_{i=0}^{\infty}(1-p_A^*)^i\omega^2(i)}{\left(\sum_{i=0}^{\infty}(1-p_A^*)^i\omega(i)\right)^2}+\frac{2n^2\tau_T^2}{\hat{\lambda}_{\max}^2}\cdot\frac{\sum_{i=0}^{\infty}\sum_{j=i+1}^{\infty}(1-p_A^*)^j\omega(i)\omega(j)}{\left(\sum_{i=0}^{\infty}(1-p_A^*)^i\omega(i)\right)^2}+\frac{2n\tau_T}{\hat{\lambda}_{\max}}\left(\tau_F-\frac{\tau_T p_A^*\ln p_A^*}{2\hat{\lambda}_{\max}}\right)\cdot\frac{\sum_{i=1}^{\infty}i(1-p_A^*)^i\omega(i)}{1+\sum_{i=1}^{\infty}(1-p_A^*)^i\omega(i)}, \quad (31)$$

where C is a function of τ_T and τ_F, which is independent of $\omega(i)$. In contrast to (30), (31) indicates that $E[D_{0,p=p_A^*}^{2,K=\infty}]$ sharply increases with the growth rate of $\omega(i)$. Intuitively, a higher growth rate of backoff window size leads to a larger difference of window sizes between fresh packets and backlogged ones. The network is, on the other hand, also better capable of alleviating the contention and becomes more robust. In the next section, we will take the example of Polynomial Backoff to demonstrate how to steer the growth rate of backoff window size to strike a balance between the throughput and delay performance.

V. Polynomial Backoff

In this section, we focus on Polynomial Backoff, where the backoff window size W_i of a State-R$_i$ HOL packet can be written as

$$W_i = W\cdot(1+i)^x, \quad (32)$$

$i = 0,\ldots,K$. The polynomial power x is a non-negative integer, which determines the growth rate of the backoff window size. It is clear from (32) that Polynomial Backoff belongs to the group of mild backoff as $\lim_{i\to\infty}\frac{W_{i+1}}{W_i} = 1$. In the following subsections, we will apply the analysis in Section IV to Polynomial Backoff, and illustrate the effect of polynomial power x on the network performance.

A. Undesired Stable Point p_A

It has been shown in Section IV that a saturated IEEE 802.11 DCF network operates at the undesired stable point p_A. With Polynomial Backoff, the undesired stable point p_A can be obtained by combining (18) and (32). For instance, when $x = 0$ and $K = \infty$, it can be explicitly written as

$$p_A^{K=\infty,x=0} = \exp\left\{-\frac{2n}{1+W}\right\}. \quad (33)$$

With $x = 1$ and $K = \infty$, we have

$$p_A^{K=\infty,x=1} \approx \frac{W}{2n}\mathbb{W}_0\left(\frac{2n}{W}\right). \quad (34)$$

Both $p_A^{K=\infty,x=0}$ and $p_A^{K=\infty,x=1}$ decline as the network size n increases.

Fig. 2 illustrates the undesired stable point p_A of a 50-node IEEE 802.11 DCF network with Polynomial Backoff. It can be clearly observed from Fig. 2 that p_A increases as the network size n decreases or the cutoff phase K increases, which corroborates Corollary 1. Moreover, p_A is improved with a larger polynomial power x owing to a faster growth rate of backoff window size.

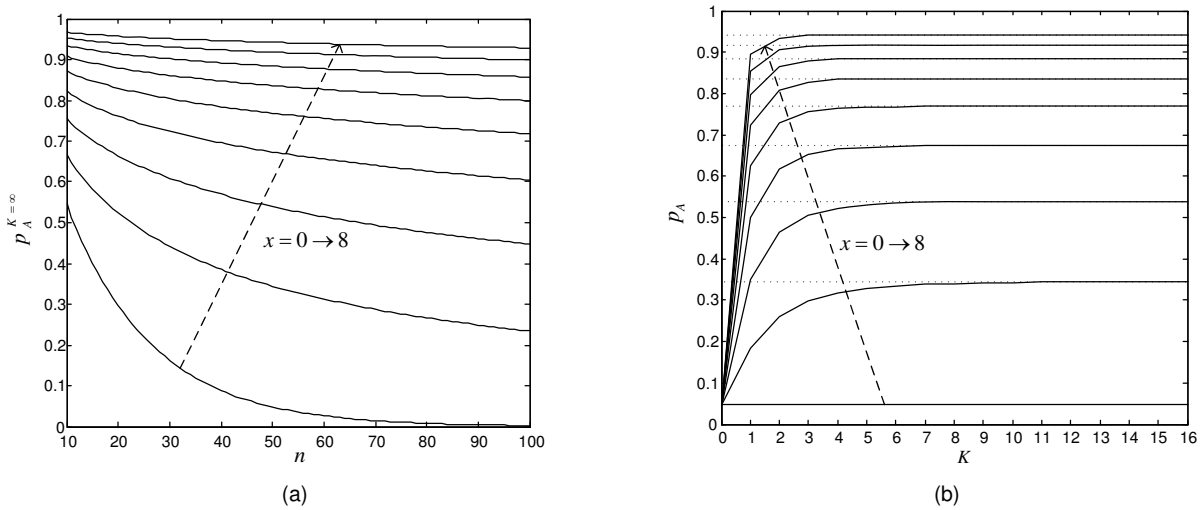


Fig. 2. Undesired stable point p_A of saturated IEEE 802.11 DCF networks with Polynomial Backoff. $W = 32$. (a) $p_A^{K=\infty}$ versus network size n. $K = \infty$; (b) p_A versus cutoff phase K. $n = 50$.

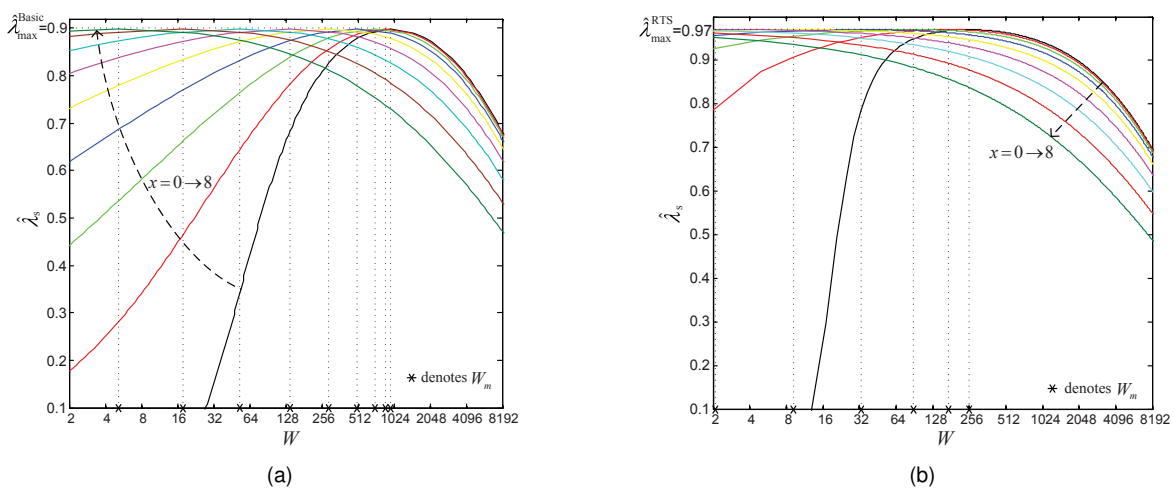


Fig. 3. Saturation throughput $\hat{\lambda}_s$ versus initial backoff window size W in IEEE 802.11 DCF networks with Polynomial Backoff. $n = 50$ and $K = \infty$. (a) Basic access mechanism; (b) RTS/CTS access mechanism.

B. Saturation Throughput

With Polynomial Backoff, the saturation throughput $\hat{\lambda}_s$ can be obtained by combining (18), (22) and (32). The curves of $\hat{\lambda}_s$ versus the initial backoff window size W in the basic and RTS/CTS access mechanisms are presented in Fig. 3a and Fig. 3b, respectively. Here we use the typical values of τ_T and τ_F provided in Table V of [1] to demonstrate the numerical results. That is, $\tau_T^{Basic} = 180$ time slots and $\tau_F^{Basic} = 175$ time slots for the basic access mechanism, and $\tau_T^{RTS} = 192$ time slots and $\tau_F^{RTS} = 9$ time slots for the RTS/CTS access mechanism. We can clearly see from Fig. 3 that the maximum throughput $\hat{\lambda}_{\max}$ is independent of the polynomial power x, and solely determined by τ_T and τ_F. The optimal initial backoff window size to achieve $\hat{\lambda}_{\max}$ can be obtained by combining (29) and (32) as

$$W_m = \frac{-\frac{2n}{\ln p_A^*} - 1}{\sum_{i=0}^{K-1} p_A^*(1-p_A^*)^i \cdot (1+i)^x + (1-p_A^*)^K \cdot (1+K)^x}, \quad (35)$$

where p_A^* is given by (24). We can see from (35) that the optimal initial backoff window size W_m decreases as the network size n decreases or the polynomial power x increases.

Section IV-C also shows that the network becomes more robust with a larger growth rate of the backoff window size. As Fig. 4 illustrates, the saturation throughput $\hat{\lambda}_s$ starts to decline from $\hat{\lambda}_{\max}$ as the network size n increases from 10, if the initial backoff window size W is fixed to be $W = W_m^{K=\infty,n=10}$. The throughput degradation is, nevertheless, much less significant with a larger polynomial power x. With the RTS/CTS access mechanism, the throughput barely varies with n when $x \geq 2$. We can conclude that the network is increasingly robust against the variation of the network size n as the polynomial power x grows.

C. Access Delay

With Polynomial Backoff, the mean access delay at the undesired stable point $E[D_{0,p=p_A}]$ can be obtained by combining

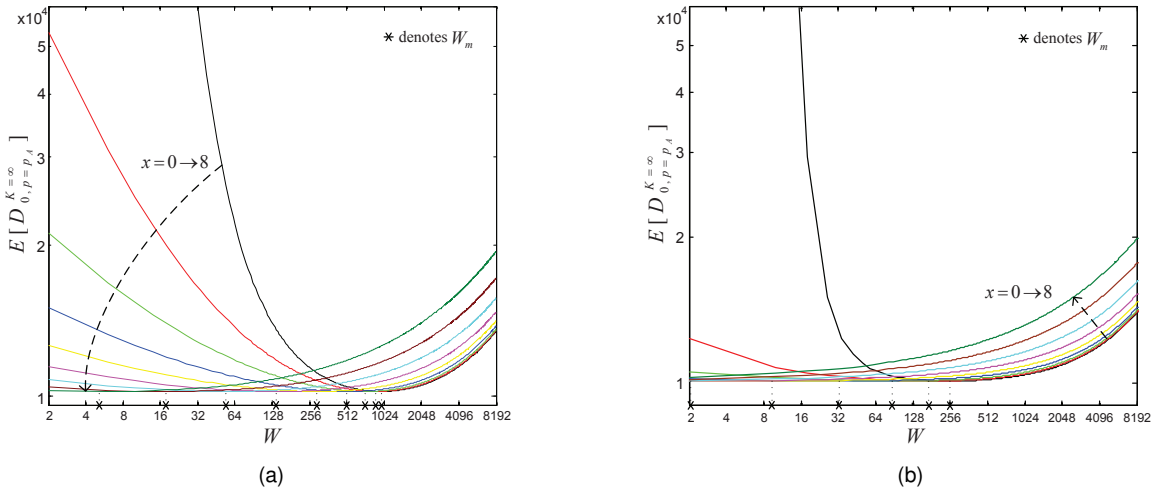


Fig. 5. Mean access delay $E[D^{K=\infty}_{0,p=p_A}]$ versus initial backoff window size W in saturated IEEE 802.11 DCF networks with Polynomial Backoff. $n = 50$ and $K = \infty$. (a) Basic access mechanism; (b) RTS/CTS access mechanism.

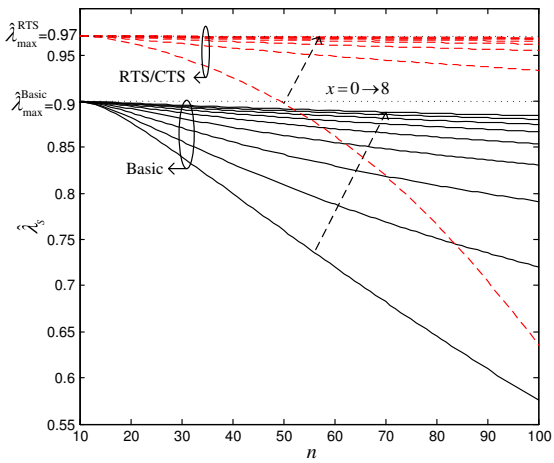


Fig. 4. Saturation throughput $\hat{\lambda}_s$ versus network size n in IEEE 802.11 DCF networks with Polynomial Backoff. $K = \infty$ and $W = W_m^{K=\infty,n=10}$.

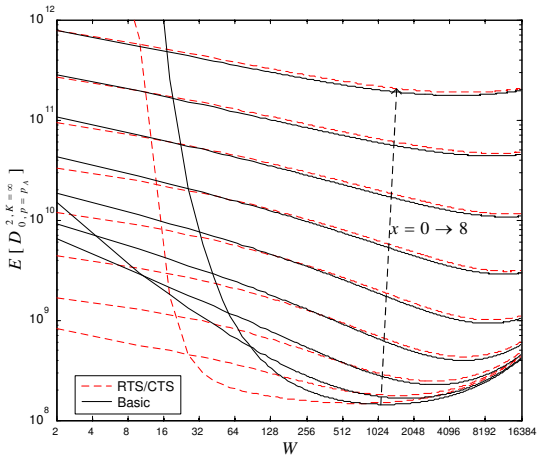


Fig. 6. Second moment of access delay $E[D^{2,K=\infty}_{0,p=p_A}]$ versus initial backoff window size W in saturated IEEE 802.11 DCF networks with Polynomial Backoff. $n = 50$ and $K = \infty$.

(18), (25) and (32). As we can see from Fig. 5, the minimum mean access delay is independent of the polynomial power x, and is achieved when the initial backoff window size W is tuned to be W_m according to (35).

The second moment of access delay at the undesired stable point $E[D^2_{0,p=p_A}]$ can be obtained by combining (9-14), (18) and (32). Fig. 6 presents the curves of $E[D^{2,K=\infty}_{0,p=p_A}]$ versus the initial backoff window size W under various values of polynomial power x. We can see from Fig. 6 that similar to the mean access delay, $E[D^{2,K=\infty}_{0,p=p_A}]$ is also sensitive to the initial backoff window size W when x is small. It is minimized when W is carefully tuned, and the minimum second moment of access delay increases with the polynomial power x.

As we have pointed out in Section IV-C, a faster growth rate of backoff window size leads to a larger difference of window sizes between fresh packets and backlogged ones, and thus incurs a higher second moment of access delay. This can be clearly observed from Fig. 7, where the second moment of access delay $E[D^{2,K=\infty}_{0,p=p_A^*}]$ with $W = W_m$ is plotted against

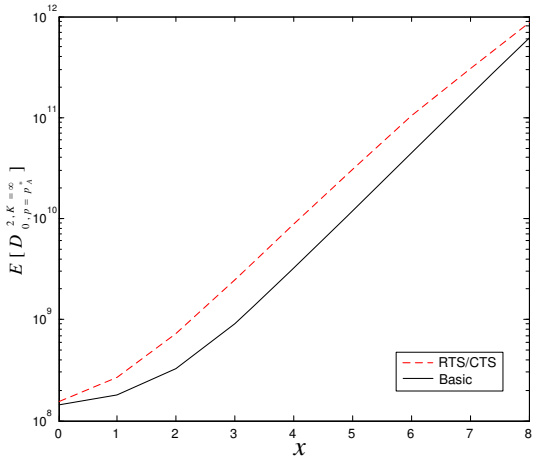


Fig. 7. Second moment of access delay $E[D^{2,K=\infty}_{0,p=p_A^*}]$ versus polynomial power x in saturated IEEE 802.11 DCF networks with Polynomial Backoff. $K = \infty$, $W = W_m$ and $n = 50$.

the polynomial power x.

So far we have shown that Polynomial Backoff with various polynomial powers can achieve the same maximum throughput and the same minimum mean access delay when the initial backoff window size W is carefully tuned. The polynomial power x determines the growth rate of the backoff window size, and also the tradeoff between the throughput and delay performance. With a larger x, the network becomes more robust, while the delay performance deteriorates. To balance between throughput and delay, Quadratic Backoff (QB), i.e., Polynomial Backoff with $x = 2$, stands out as a favorable option. In the following section, we will further compare the performance of QB with the default backoff scheme of IEEE 802.11 DCF networks, Binary Exponential Backoff (BEB).

VI. Performance Comparison of Quadratic Backoff and Binary Exponential Backoff

With BEB, the backoff window size W_i of a State-R$_i$ HOL packet is given by

$$W_i = W \cdot 2^i, \tag{36}$$

$i = 0, \ldots, K$. BEB belongs to the group of aggressive backoff, as $\lim_{i\to\infty} \frac{W_{i+1}}{W_i}$=2>1. According to Theorem 3, the second moment of access delay of BEB diverges if the network size n is too large. With QB, the backoff window size is given by

$$W_i = W \cdot (1+i)^2, \tag{37}$$

$i = 0, \ldots, K$. As a mild backoff scheme, QB can always achieve a bounded second moment of access delay.

The backoff window sizes with both QB and BEB are presented in Fig. 8. As we can see from Fig. 8, when i is small, the backoff window size with QB grows faster than that with BEB. A smaller increasing rate of W_i, however, is observed with QB when i is large. Intuitively, when the network becomes saturated, a fast increase of the backoff window size in the first few states, i.e., with a small i, is helpful for effective alleviation of the channel contention. After nodes are pushed to deep states, a small backoff window size would be desirable to reduce the delay jitter. In the following subsections, we will compare the throughput and delay performance of QB with that of BEB.

A. Undesired Stable Point p_A

The undesired stable points of BEB and QB can be obtained by combining (18) and (36-37), respectively. With an infinite cutoff phase K=∞, the undesired stable point of BEB can be explicitly written as

$$p_A^{BEB,K=\infty} \approx \frac{2n/W}{\mathbb{W}_0\left(2n/W \cdot \exp\left(4n/W\right)\right)}. \tag{38}$$

Similarly, $p_A^{QB,K=\infty}$ can be numerically obtained by solving

$$p = \exp\left\{-\frac{2n}{1+W(2-p)/p^2}\right\}. \tag{39}$$

(38-39) is verified by the simulation results presented in Fig. 9. In this section, all the simulations are conducted using the *ns*-2 simulator, and the values of system parameters are

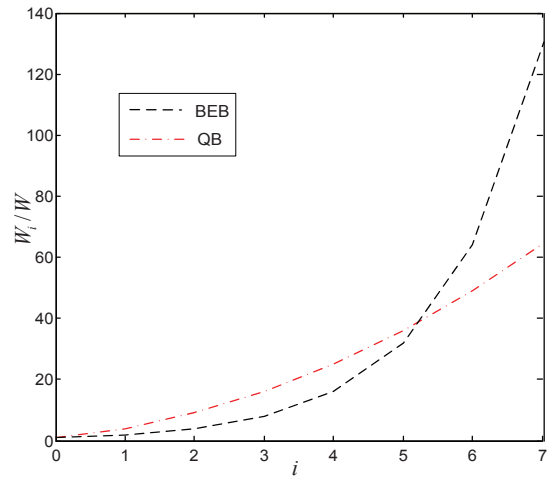


Fig. 8. Backoff window sizes of BEB and QB.

in accordance with [1] (which were summarized in Table II of [1]). As we can see from Fig. 9a, with an infinite cutoff phase $K = \infty$, $p_A^{QB,K=\infty}$ is comparable to $p_A^{BEB,K=\infty}$ if the initial backoff window size W is not too small. With a finite $K < \infty$, since the backoff window sizes with QB in the first few states are larger than those with BEB, p_A^{QB} converges faster to $p_A^{QB,K=\infty}$ than p_A^{BEB} does, as shown in Fig. 9b. A perfect match between the theoretical and simulation results can be observed from Fig. 9.

B. Saturation Throughput

The saturation throughput of BEB and QB can be obtained by combining (18), (22) and (36-37). It can be clearly seen from Fig. 10 that both QB and BEB can achieve the same maximum throughput $\hat{\lambda}_{\max}$ when the initial backoff window size W is properly tuned. The optimal initial backoff window sizes of BEB and QB to achieve the maximum throughput $\hat{\lambda}_{\max}$ can be obtained as

$$W_m^{BEB,K=\infty}=\frac{-2n\left(2\frac{1+\tau_F}{\tau_F}\mathbb{W}_0\left(-\frac{\tau_F}{e(1+\tau_F)}\right)+1\right)}{\frac{1+\tau_F}{\tau_F}\mathbb{W}_0\left(-\frac{\tau_F}{e(1+\tau_F)}\right)\ln\left(-\frac{1+\tau_F}{\tau_F}\mathbb{W}_0\left(-\frac{\tau_F}{e(1+\tau_F)}\right)\right)} \tag{40}$$

and

$$W_m^{QB,K=\infty}=\frac{-2n\left(-\frac{1+\tau_F}{\tau_F}\mathbb{W}_0\left(-\frac{\tau_F}{e(1+\tau_F)}\right)\right)^2}{\left(2+\frac{1+\tau_F}{\tau_F}\mathbb{W}_0\left(-\frac{\tau_F}{e(1+\tau_F)}\right)\right)\ln\left(-\frac{1+\tau_F}{\tau_F}\mathbb{W}_0\left(-\frac{\tau_F}{e(1+\tau_F)}\right)\right)} \tag{41}$$

by combining (24), (29) and (36-37), respectively.

We can see from (40-41) that to achieve the maximum throughput $\hat{\lambda}_{\max}$, the initial backoff window size W_m should be carefully tuned according to the network size n. Otherwise, the saturation throughput may decline as n grows. Fig. 11 presents the throughput performance of QB and BEB with W fixed to be $W_m^{K=\infty,n=10}$. A small throughput degradation can be observed in both cases, indicating that both QB and BEB are robust against the variation of network size.

The above results are obtained when the cutoff phase K=∞. With a finite cutoff phase K<∞, the saturation

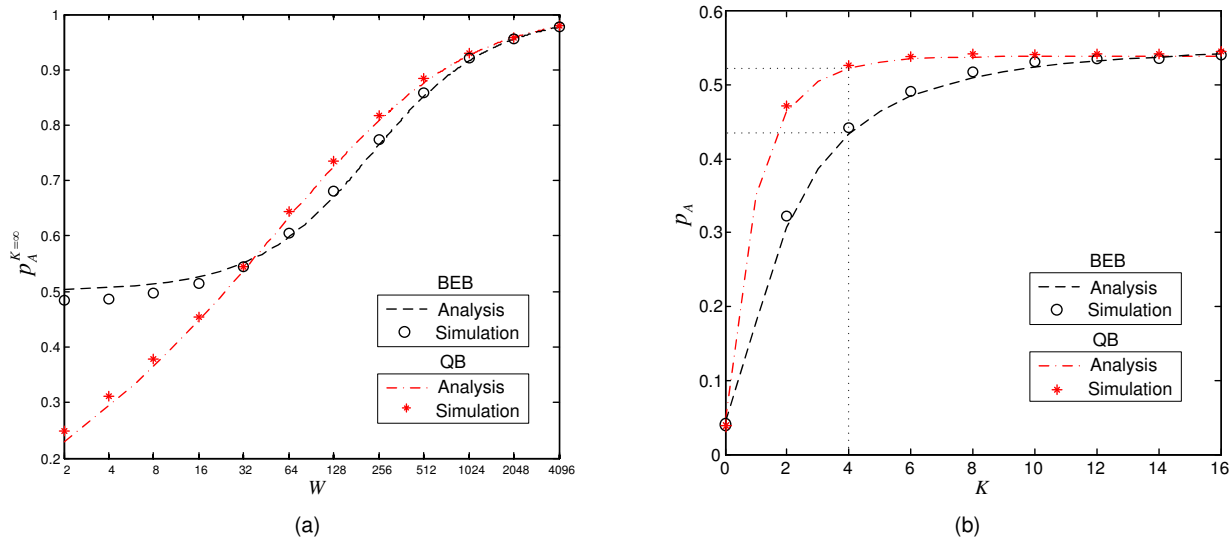


Fig. 9. Undesired stable point p_A in saturated IEEE 802.11 DCF networks with BEB and QB. (a) $p_A^{K=\infty}$ versus initial backoff window size W. $n = 50$ and $K = \infty$; (b) p_A versus cutoff phase K. $W = 32$ and $n = 50$.

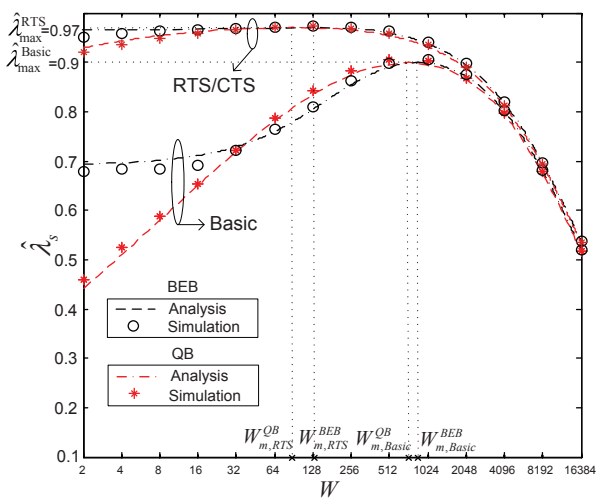


Fig. 10. Saturation throughput $\hat{\lambda}_s$ versus initial backoff window size W in IEEE 802.11 DCF networks with BEB and QB. $n = 50$ and $K = \infty$.

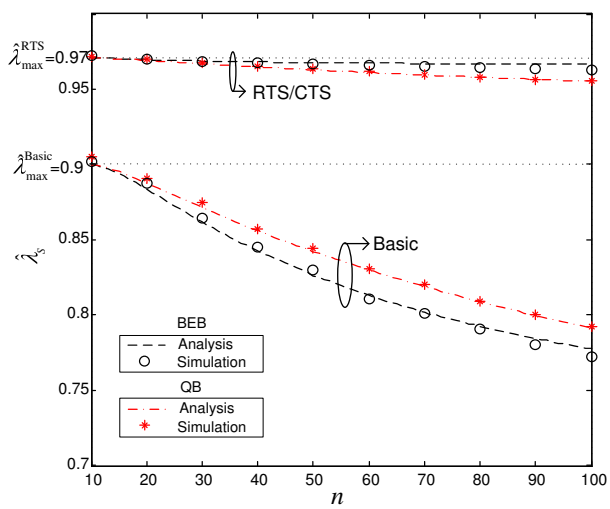


Fig. 11. Saturation throughput $\hat{\lambda}_s$ versus network size n in IEEE 802.11 DCF networks with BEB and QB. $K = \infty$ and $W = W_m^{K=\infty,n=10}$.

throughput may be significantly lower than that with $K{=}\infty$ due to a smaller p_A. It has been shown in Fig. 9b that p_A^{QB} converges to $p_A^{QB,K=\infty}$ at a faster rate. Fig. 12 further illustrates that the saturation throughput of QB also converges faster to the limiting value than that of BEB. In particular, with QB, $K{=}4$ is good enough to approach the limiting throughput with $K{=}\infty$ for both the basic and RTS/CTS access mechanisms, which is much smaller than that with BEB, i.e., $K{=}16$ for the basic access mechanism and $K{=}6$ for the RTS/CTS access mechanism. Fig. 12 validates that with enlarged backoff window sizes in the first few states, QB can better alleviate the channel contention, resulting in a larger throughput. Nodes with BEB, in contrast, have to back off to much deeper states to achieve a comparable throughput. A larger cutoff phase K is therefore required to approach the limiting throughput.

C. Access Delay

The mean access delay at the undesired stable point $E[D_{0,p=p_A}]$ with BEB and QB can be obtained by combining (18), (25) and (36-37). Both BEB and QB achieve the same minimum mean access delay if their initial backoff window sizes are tuned to be $W_m^{BEB,K=\infty}$ and $W_m^{QB,K=\infty}$ according to (40) and (41), respectively. Fig. 13 shows that both schemes have similar mean access delay performance as the network size n increases.

The second moments of access delay at the undesired stable point $E[D^2_{0,p=p_A}]$ with BEB and QB can be obtained by combining (9-14), (18) and (36-37). With $K = \infty$, they can

be explicitly written as

$$E[D_{0,p=p_A}^{2,BEB,K=\infty}]=\frac{3+p_A-3\alpha p_A}{6\alpha^2 p_A^2}+\left(\frac{2}{\alpha}+(2-p_A)\tau_F\right)\frac{(1-p_A)\tau_F}{p_A^2}$$
$$+\left(\tau_T+2\frac{1-p_A}{p_A}\tau_F+\frac{1}{\alpha p_A}\right)\tau_T+\left(\tau_T+\frac{1-p_A}{p_A}\tau_F+\frac{2(1-p_A)}{2p_A-1}\right.$$
$$\left.\cdot\left(\tau_F+\frac{1}{2\alpha}\right)-\frac{1}{2}+\frac{1+p_A}{2\alpha p_A}\right)\cdot\frac{W}{\alpha\,(2p_A-1)}+\left(\frac{1}{3}+\frac{1-p_A}{2p_A-1}\right)$$
$$\cdot\sum_{i=0}^{\infty}(4-4p_A)^i\cdot\frac{W^2}{\alpha^2}, \tag{42}$$

and

$$E[D_{0,p=p_A}^{2,QB,K=\infty}]=\frac{3+p_A-3\alpha p_A}{6\alpha^2 p_A^2}+\left(\frac{2}{\alpha}+(2-p_A)\tau_F\right)\frac{(1-p_A)\tau_F}{p_A^2}$$
$$+\left(\tau_T+2\frac{1-p_A}{p_A}\tau_F+\frac{1}{\alpha p_A}\right)\tau_T+\left(\frac{3}{2}p_A(2-p_A)\tau_T+(1-p_A)\right.$$
$$\left.\cdot(8-3p_A)\tau_F+\frac{\alpha(1+\tau_T)(p_A-2)p_A+p_A^2-7p_A+8}{2\alpha}\right)\cdot\frac{W}{\alpha p_A^4}$$
$$+(p_A^4-26p_A^3+135p_A^2-228p_A+120)\cdot\frac{W^2}{6\alpha^2 p_A^6}, \tag{43}$$

respectively, where α is given by (4). As Theorem 3 indicates, the second moment of access delay for mild backoff is always finite. With aggressive backoff, however, it diverges if the network size n is too large. Here we can clearly see from (42) that $E[D_{0,p=p_A}^{2,BEB,K=\infty}]$ becomes infinite if

$$p_A^{BEB}<3/4, \tag{44}$$

which is equivalent to

$$n>\frac{3}{4}\left(\ln\frac{4}{3}\right)W, \tag{45}$$

according to (38).

The second moments of access delay $E[D_{0,p=p_A}^2]$ with BEB and QB with $W=W_m^{K=\infty,n=10}$ are illustrated in Fig. 13. $W_m^{K=\infty,n=10}$ with BEB in the basic access and the RTS/CTS access modes can be obtained from (40) as 173 and 26, respectively. According to (45), the second moment of access delay with BEB $E[D_{0,p=p_A}^{2,BEB,K=\infty}]$ is infinite if the network size $n>37$ in the basic access mode, and $n>5$ in the RTS/CTS access mode. As Fig. 13 shows, $E[D_{0,p=p_A}^{2,BEB,K=\infty}]$ sharply grows as the network size n increases, and eventually becomes unbounded. In contrast, a finite second moment of access delay can be always achieved with QB.

For a 50-node IEEE 802.11 network with BEB, (45) indicates that the initial backoff window size W should be larger than 232. Otherwise, the second moment of access delay $E[D_{0,p=p_A}^{2,BEB}]$ will sharply grow with the cutoff phase K, and become infinite as $K\to\infty$. This fact can be clearly observed from Fig. 14. Also note that a large cutoff phase K is required for BEB to achieve the limiting throughput with $K=\infty$. With $W=32$, for instance, Fig. 12 shows that K should be at least 16 for BEB to approach the limiting throughput when the basic access mechanism is adopted. In contrast, $K\geq 4$ is sufficient

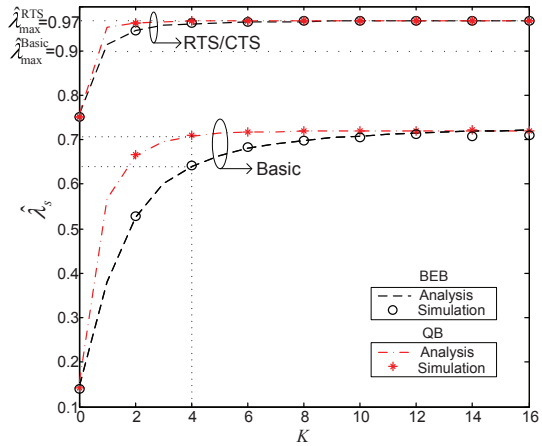


Fig. 12. Saturation throughput $\hat{\lambda}_s$ versus cutoff phase K in IEEE 802.11 DCF networks with BEB and QB. $W=32$ and $n=50$.

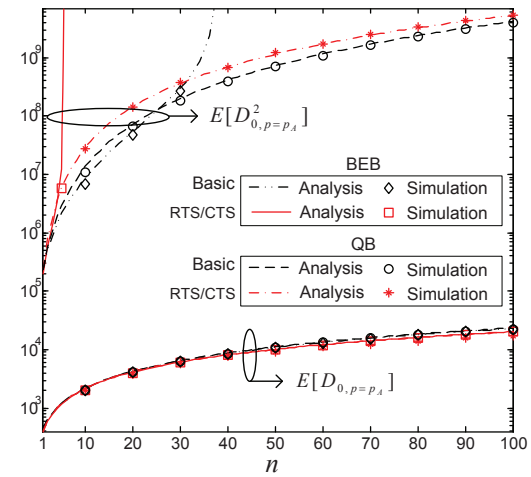


Fig. 13. Mean access delay $E[D_{0,p=p_A}]$ (in unit of time slots) and second moment of access delay $E[D_{0,p=p_A}^2]$ versus network size n in saturated IEEE 802.11 DCF networks with BEB and QB. $K=\infty$ and $W=W_m^{K=\infty,n=10}$.

for QB, and the corresponding second moment of access delay is much smaller than that with BEB.

Recall that a huge second moment of access delay indicates that the access delay performance across the network is fairly disproportionate: some node may produce a continuous stream of packets, while others have to wait for a long time. It is the inherent reason behind the capture phenomenon and the poor delay performance of BEB. The above results clearly show that QB can dramatically reduce the second moment of access delay, and therefore greatly improve the queueing performance.

D. Short-term Unfairness

Note that the capture phenomenon also indicates serious short-term unfairness: the node who captures the channel has a much higher throughput than those with deeply backlogged HOL packets in a certain time period.[7] To better understand

[7] In the long run, nodes would capture the channel in turn and achieve similar throughput performance.

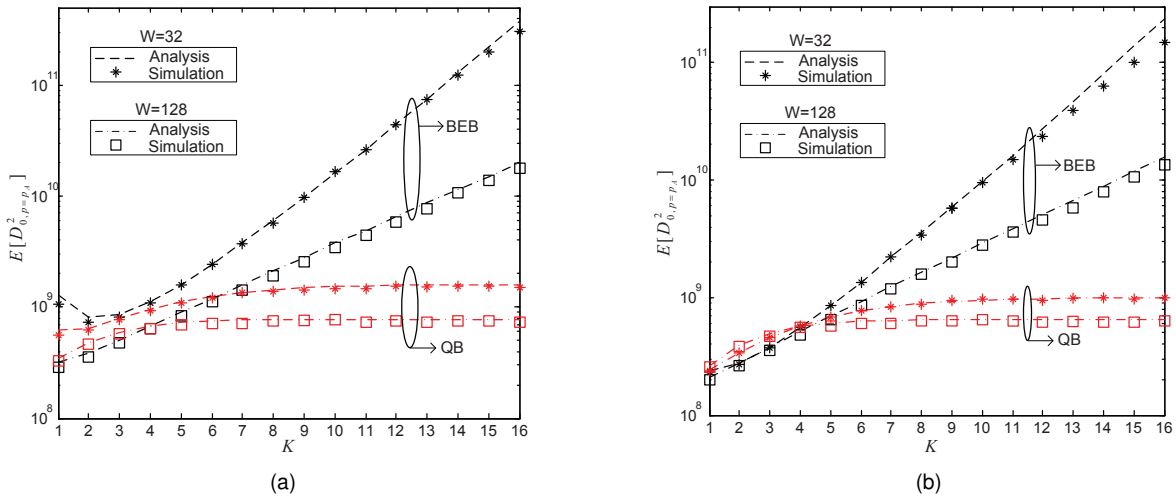


Fig. 14. Second moment of access delay $E[D^2_{0,p=p_A}]$ versus cutoff phase K in saturated IEEE 802.11 DCF networks with BEB and QB. $n = 50$. (a) Basic access mechanism; (b) RTS/CTS access mechanism.

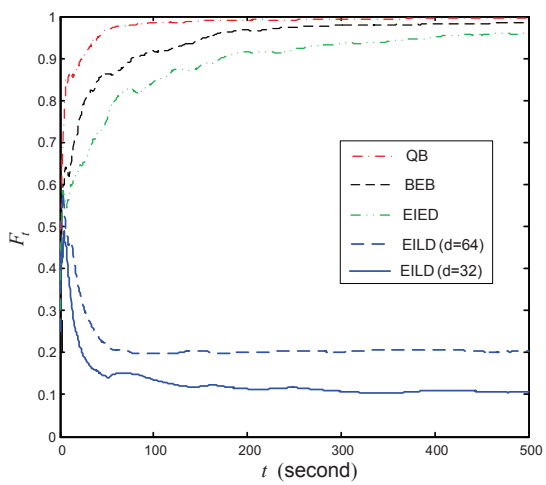


Fig. 15. Jain's Fairness Index F_t versus time t in saturated IEEE 802.11 DCF networks with BEB, QB, EIED and EILD. $W = 32$, $K = 6$ and $n = 50$. With EIED, the backoff window size is doubled upon collision and halved upon success. With EILD, the backoff window size is doubled upon collision and linearly decreased by some value d upon success.

the short-term fairness performance, let us introduce Jain's Fairness Index [38]:

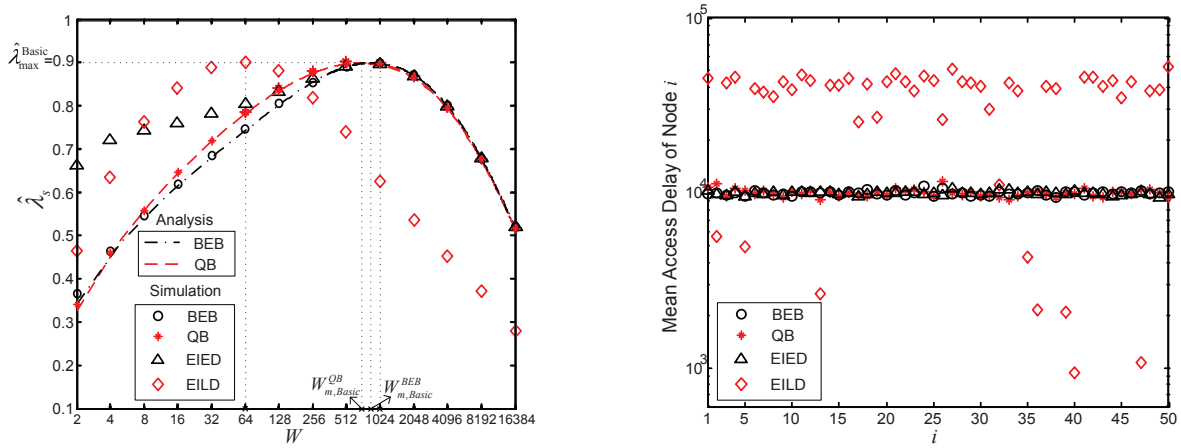
$$F_t = \frac{\left(\sum_{i=0}^{n} \lambda_t^i\right)^2}{n \sum_{i=0}^{n} (\lambda_t^i)^2}, \tag{46}$$

where λ_t^i is the throughput of node i measured in the time interval $(0, t)$. The index characterizes the difference of the throughput performance of nodes. It is clear from (46) that a higher F_t indicates better short-term fairness. It approaches unity when each node has an equal throughput. Fig. 15 shows the fairness indexes of BEB and QB. It can be clearly observed that the fairness index of QB converges much faster to unity than that of BEB, which validates that QB can effectively mitigate the capture phenomenon and significantly improve the short-term fairness performance.

As mentioned in Section I, it is widely believed that the short-term unfairness of BEB is rooted in the setting of the initial backoff window size: by setting the window size of fresh HOL packets to a small value, they have a much higher transmission probability than those deeply backlogged ones. Various schemes were then proposed to decrement the backoff window sizes of fresh HOL packets from their preceding ones. Fig. 15 illustrates the short-term fairness performance of two representative schemes: Exponential Increase Exponential Decrease (EIED) [16] and Exponential Increase Linear Decrease (EILD) [20]. On the contrary to the common belief, the fairness indexes of both schemes are found to be significantly lower than that of BEB.

The reason lies in the difference of backoff window sizes of nodes. With a slow decrement of the window size upon successful transmission, nodes that ever have deeply backlogged HOL packets would stay with a large backoff window size for a long time. They can hardly access the channel due to a small transmission probability, and thus have much worse throughput performance. As the window size decreases more slowly, it takes more time for them to return to a small backoff window size. The channel is then captured by other nodes for longer time, leading to more severe short-term unfairness. As we can see from Fig. 15, compared to EIED where the window size is exponentially decreased upon successful transmission, the fairness performance of EILD is much worse due to a linear decrement. Although the fairness index of EILD can be improved by choosing a larger decrement d of the window size, it is far below the other backoff schemes.

Similar to the second moment of access delay, unfairness originates from a large difference of backoff window sizes of nodes. By decrementing the backoff window sizes of fresh HOL packets from their preceding ones, the window size difference of packets from the same node is reduced, but the difference among nodes is enlarged. That is why the short-term unfairness is worsened rather than improved by using EILD/EIED. Here we can conclude that the key to improving fairness lies in the growth rate of backoff window size. With a slower growth rate, the difference of backoff window size



(a) (b)

Fig. 16. Network throughput and mean access delay with BEB, QB, EIED and EILD in saturated IEEE 802.11 DCF networks with the basic access mechanism. $n = 50$. $K = 6$. $d = 64$ for EILD. (a) Saturation throughput $\hat{\lambda}_s$ versus initial backoff window size W. (b) Mean access delay (in unit of time slots) of each node with $W = W_m$. The x-axis i denotes the index of a node.

between nodes with fresh packets and those with deeply backlogged ones becomes smaller, and thus better fairness performance can be achieved by QB.

For the sake of comparison, the throughput and delay performance of EIED and EILD is also presented in Fig. 16. It can be observed from Fig. 16a that EIED and EILD achieve the same maximum network throughput as BEB and QB, though with different initial backoff window sizes, i.e., $W_m = 1024$ for EIED and $W_m = 64$ for EILD. Fig. 16b illustrates the mean access delay performance of each node where the initial backoff window size W is set to be the optimal value W_m according to Fig. 16a. It can be clearly seen from Fig. 16b that although the minimum mean access delay across the network, which is inversely proportional to the maximum network throughput, is equal for all the backoff schemes, the delay performance with EILD significantly varies from node to node. As we have observed from Fig. 15, with EILD, only a small subset of nodes could capture the channel and have superior performance. The rest of them still suffer from a long delay as they can rarely access the channel, thus leading to serious unfairness. Each node with EIED, on the other hand, achieves comparable mean access delay performance as that with QB and BEB. Fig. 17 further shows the second moment of access delay with EIED. We can see from Fig. 17 that although it is lower than that of BEB when the initial backoff window size is small, i.e., $W = 32$, the gap is diminished as W increases. Similar to BEB, the second moment of access delay with EIED increases with the cutoff phase K unboundedly, and becomes much higher than that with QB when K is large.

VII. Conclusion

In this paper, the effect of backoff schemes on saturated IEEE 802.11 DCF networks is characterized based on a unified analytical framework. According to whether the limiting growth rate of the backoff window size is larger than one, the backoff schemes are categorized into two groups: aggressive

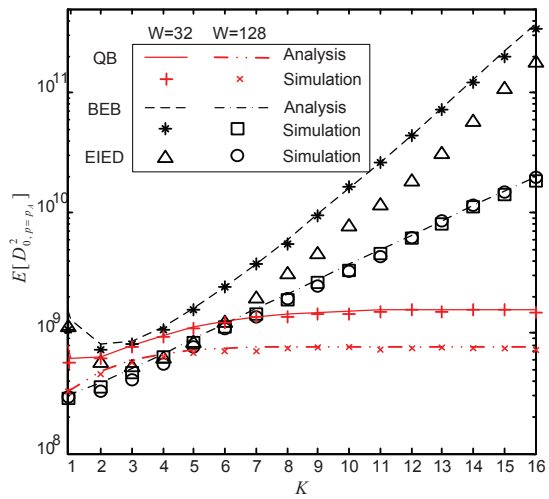


Fig. 17. Second moment of access delay $E[D^2_{0,p=p_A}]$ with BEB, QB and EIED versus cutoff phase K in saturated IEEE 802.11 DCF networks with the basic access mechanism. $n = 50$.

backoff and mild backoff. It is shown that both groups have the same maximum throughput $\hat{\lambda}_{\max}$ which is solely determined by the access mechanism, i.e., the basic access or the RTS/CTS access. Yet with the aggressive backoff schemes, an infinite second moment of access delay may be incurred if the network size exceeds a certain threshold. The default backoff scheme of IEEE 802.11 DCF networks, BEB, belongs to the group of aggressive backoff, and therefore suffers from such a delay degradation when the network size is large.

To further demonstrate how to trade off between throughput and delay by tuning the growth rate of the backoff window size for mild backoff schemes, the performance of Polynomial Backoff with various polynomial powers is evaluated, and QB is selected as a competitive candidate for IEEE 802.11 DCF networks. The comparative study of QB and BEB shows that both schemes achieve the same maximum throughput and are robust against the variation of network size. With BEB,

however, the second moment of access delay rapidly grows as the cutoff phase K increases, indicating deteriorating queueing performance as well as severe unfairness. With QB, in contrast, the second moment of access delay is insensitive to the value of K. Much better queueing and fairness performance is therefore achieved, which makes QB a favorable option for IEEE 802.11 DCF networks.

Appendix A
Proof of Theorem 1

Proof: Denote

$$g_K(p) = \sum_{i=0}^{K-1} p(1-p)^i W_i + (1-p)^K W_K. \tag{47}$$

(18) can be written as

$$p = \exp\left\{-\frac{2n}{1+g_K(p)}\right\}. \tag{48}$$

Let us first prove the monotonicity of $g_K(p)$ with regard to p.

Lemma 1. *$g_K(p)$ is a monotonic non-increasing function of $p \in (0,1)$ if $\{W_i\}$ is a monotonic increasing sequence.*

Proof: Define $\{\tilde{W}_i\}$ as an infinite sequence with $\tilde{W}_i = W_i$ for $0 \le i \le K-1$ and $\tilde{W}_i = W_K$ for $i \ge K$. $g_K(p)$ can then be written as

$$g_K(p) = \sum_{i=0}^{\infty} p(1-p)^i \tilde{W}_i = E_X[\tilde{W}_X], \tag{49}$$

where X is a geometric random variable with parameter p, and $E_X[\cdot]$ denotes expectation with respect to X.

Suppose that $0<p_1<p_2<1$. Let X_1 and X_2 denote geometric random variables with parameters p_1 and p_2, respectively. Clearly we have $X_1 \ge_{st} X_2$ [39]. Further note that $\{\tilde{W}_i\}$ is a monotonic non-decreasing sequence. As a result, $\tilde{W}_{X_1} \ge_{st} \tilde{W}_{X_2}$, and we can then conclude from (49) that $g(p_1) \ge g(p_2)$. ■

According to Lemma 1, $\exp\left\{-\frac{2n}{1+g_K(p)}\right\} - p$ is a monotonic decreasing function of $p \in (0,1)$. Moreover,

$$\lim_{p\to 0} \exp\left\{-\frac{2n}{1+g_K(p)}\right\} - p = \exp\left\{-\frac{2n}{1+W_K}\right\} > 0, \tag{50}$$

and

$$\lim_{p\to 1} \exp\left\{-\frac{2n}{1+g_K(p)}\right\} - p = \exp\left\{-\frac{2n}{1+W_0}\right\} - 1 < 0. \tag{51}$$

Therefore, (18) has one single non-zero root. ■

Appendix B
Proof of Corollary 1

The proof is divided into two parts. In the first part, we will prove the monotonicity of p_A with regard to K; in the second part, we will prove the monotonicity of p_A with regard to n.

1) If $0 \le K_1 < K_2$, then $p_A^{K_1} < p_A^{K_2}$.

Proof: As p_A is a function of $g_K(p_A)$ which is shown in (48), let us first prove the monotonicity of $g_K(p)$ with regard to K.

Lemma 2. *$g_K(p)$ is a monotonic increasing function of $K \ge 0$ if $\{W_i\}$ is a monotonic increasing sequence.*

Proof: Suppose that $0 \le K_1 < K_2$. Define $\{\tilde{W}_i^{K_j}\}$, $j=1,2$, as two infinite sequences with $\tilde{W}_i^{K_j} = W_i$ for $0 \le i \le K_j - 1$ and $\tilde{W}_i^{K_j} = W_{K_j}$ for $i \ge K_j$. We have

$$\tilde{W}_i^{K_1} = \tilde{W}_i^{K_2} \text{ if } i \le K_1, \text{ and } \tilde{W}_i^{K_1} < \tilde{W}_i^{K_2} \text{ if } i > K_1. \tag{52}$$

Further define X as a geometric random variable with parameter p. $g_{K_j}(p)$, $j=1,2$, can then be written as

$$g_{K_j}(p) = \sum_{i=0}^{\infty} p(1-p)^i \cdot \tilde{W}_i^{K_j} = E_X\left[\tilde{W}_X^{K_j}\right]. \tag{53}$$

By combining (52) and (53), we can easily see that $g_{K_1}(p) < g_{K_2}(p)$. ■

According to (48), we have

$$-\ln p_A^{K_1} \cdot \left(1 + g_{K_1}\left(p_A^{K_1}\right)\right) = -\ln p_A^{K_2} \cdot \left(1 + g_{K_2}\left(p_A^{K_2}\right)\right) = 2n. \tag{54}$$

According to Lemma 2,

$$-\ln p_A^{K_1} \cdot \left(1 + g_{K_1}\left(p_A^{K_1}\right)\right) < -\ln p_A^{K_1} \cdot \left(1 + g_{K_2}\left(p_A^{K_1}\right)\right), \tag{55}$$

if $0 \le K_1 < K_2$. By combining (54) and (55), we can obtain that

$$-\ln p_A^{K_2} \cdot \left(1 + g_{K_2}\left(p_A^{K_2}\right)\right) < -\ln p_A^{K_1} \cdot \left(1 + g_{K_2}\left(p_A^{K_1}\right)\right). \tag{56}$$

Lemma 1 shows that $g_K(p)$ is a monotonic non-increasing function of $p \in (0,1)$. As a result, $-\ln p \cdot (1 + g_{K_2}(p))$ is a monotonic decreasing function of $p \in (0,1)$, and we can then conclude from (56) that $p_A^{K_1} < p_A^{K_2}$. ■

2) If $0 < n_1 < n_2$, then $p_A^{n_1} > p_A^{n_2}$.

Proof: With $0 < n_1 < n_2$, we have

$$2n_1 = -\ln p_A^{n_1} \cdot (1 + g_K(p_A^{n_1})) < 2n_2 = -\ln p_A^{n_2} \cdot (1 + g_K(p_A^{n_2})), \tag{57}$$

according to (48). As $-\ln p \cdot (1 + g_K(p))$ is a monotonic decreasing function of $p \in (0,1)$, we can then conclude from (57) that $p_A^{n_1} > p_A^{n_2}$. ■

Appendix C
Proof of Corollary 2

Proof: It is clear from (18) that $\sum_{i=0}^{\infty} p_A^{K=\infty}\left(1 - p_A^{K=\infty}\right)^i W_i < \infty$ for any finite $n < \infty$. We then have

$$\lim_{n\to\infty} p_A^{K=\infty} \ge 1 - \lim_{i\to\infty} \frac{W_i}{W_{i+1}}. \tag{58}$$

Now suppose that $\lim_{n\to\infty} p_A^{K=\infty} > 1 - \lim_{i\to\infty} \frac{W_i}{W_{i+1}}$. As $\lim_{i\to\infty} \frac{W_i}{W_{i+1}} \le 1$ according to (20), we have $\lim_{n\to\infty} p_A^{K=\infty} > 0$. On the other hand, if $\lim_{n\to\infty} p_A^{K=\infty} > 1 - \lim_{i\to\infty} \frac{W_i}{W_{i+1}}$, the series $\sum_{i=0}^{\infty} p_A^{K=\infty}\left(1 - p_A^{K=\infty}\right)^i W_i$ would converge as $n \to \infty$.

We can then obtain from (18) that $\lim_{n\to\infty} p_A^{K=\infty} = 0$, which results in a contradiction. Therefore, the postulate $\lim_{n\to\infty} p_A^{K=\infty} > 1 - \lim_{i\to\infty} \frac{W_i}{W_{i+1}}$ is not true, and we can conclude from (58) that $\lim_{n\to\infty} p_A^{K=\infty} = 1 - \lim_{i\to\infty} \frac{W_i}{W_{i+1}}$. ■

Appendix D
Proof of Theorem 2

Proof: According to (22), the derivative of $\hat{\lambda}_s$ is given by

$$\frac{d\hat{\lambda}_s}{dp_A} = \frac{-\tau_T(1+\tau_F)(1+\ln p_A) + \tau_T\tau_F p_A}{(1+\tau_F-\tau_F p_A-(\tau_T-\tau_F)p_A \ln p_A)^2} = \frac{f(p_A)}{(1+\tau_F-\tau_F p_A-(\tau_T-\tau_F)p_A \ln p_A)^2}, \quad (59)$$

where $f(p_A) = -\tau_T(1+\tau_F)(1+\ln p_A) + \tau_T\tau_F p_A$. It can be easily obtained that $p_A^* = -(1+1/\tau_F)\mathbb{W}_0\left(-\frac{1}{e(1+1/\tau_F)}\right)$ is the root of $f(p_A) = 0$. Moreover, the derivative of $f(p_A)$ is given by

$$\frac{df(p_A)}{dp_A} = \tau_T\tau_F - \tau_T(1+\tau_F)/p_A = \tau_T\tau_F(1-1/p_A) - \tau_T/p_A < 0 \quad (60)$$

for $p_A \in (0,1)$. Therefore, $f(p_A) > 0$ if $p_A \in (0, p_A^*)$, and $f(p_A) < 0$ if $p_A \in (p_A^*, 1)$. We can then conclude from (59) that $\hat{\lambda}_s$ is monotonically increasing with p_A if $p_A \in (0, p_A^*)$, and decreasing with p_A if $p_A \in (p_A^*, 1)$. It reaches the maximum value with $p_A = p_A^*$. (23) can be obtained by substituting (24) into (22). ■

Appendix E
Proof of Corollary 3

Proof: 1) If: if $K = \infty$ and $\lim_{i\to\infty} \frac{W_{i+1}}{W_i} > 1$, we have $\lim_{n\to\infty} p_A^{K=\infty} = 1 - \lim_{i\to\infty} \frac{W_i}{W_{i+1}} > 0$, according to Corollary 2. We can choose $\lim_{i\to\infty} \frac{W_{i+1}}{W_i} = \frac{1}{1-p_A^*}$ such that $\lim_{n\to\infty} p_A^{K=\infty} = p_A^*$, and $\hat{\lambda}_{\max}$ can be achieved according to Theorem 2.

2) Only if: if $K < \infty$, we have $\lim_{n\to\infty} p_A^{K<\infty} = 0$ according to (19). If $\lim_{i\to\infty} \frac{W_{i+1}}{W_i} = 1$, we have $\lim_{n\to\infty} p_A^{K=\infty} = 0$ according to Corollary 2. In both cases, the saturation throughput $\hat{\lambda}_s$ approaches 0 as $n\to\infty$ according to (22). ■

Appendix F
Proof of Theorem 3

Proof: 1) Let us first consider the case when $\lim_{i\to\infty} \frac{W_{i+1}}{W_i} = 1$.

Lemma 3. *If* $\lim_{i\to\infty} \frac{W_{i+1}}{W_i} = 1$, $\sum_{i=0}^{\infty}(1-p_A)^i W_i^2 < \infty$ *and* $\sum_{i=0}^{\infty} W_i \left(\sum_{j=i+1}^{\infty}(1-p_A)^j W_j\right) < \infty$ *for any finite* $n < \infty$.

Proof: For the series $\sum_{i=0}^{\infty}(1-p_A)^i W_i^2$, we have

$$\lim_{i\to\infty} \frac{(1-p_A)^{i+1}W_{i+1}^2}{(1-p_A)^i W_i^2} = (1-p_A) \lim_{i\to\infty} \left(\frac{W_{i+1}}{W_i}\right)^2 = 1-p_A. \quad (61)$$

For any finite n, $p_A \geq \exp\{-2n\} > 0$ according to (18). Therefore, $\lim_{i\to\infty} \frac{(1-p_A)^{i+1}W_{i+1}^2}{(1-p_A)^i W_i^2} < 1$, and we have $\sum_{i=0}^{\infty}(1-p_A)^i W_i^2 < \infty$.

Similarly, for the series $\sum_{i=0}^{\infty} W_i \left(\sum_{j=i+1}^{\infty}(1-p_A)^j W_j\right)$, because

$$\lim_{i\to\infty} \frac{W_{i+1}\left(\sum_{j=i+2}^{\infty}(1-p_A)^j W_j\right)}{W_i\left(\sum_{j=i+1}^{\infty}(1-p_A)^j W_j\right)} = (1-p_A) \cdot \lim_{i\to\infty} \frac{\sum_{j=i+1}^{\infty}(1-p_A)^j W_{j+1}}{\sum_{j=i+1}^{\infty}(1-p_A)^j W_j} = 1-p_A < 1, \quad (62)$$

we have $\sum_{i=0}^{\infty} W_i \left(\sum_{j=i+1}^{\infty}(1-p_A)^j W_j\right) < \infty$. ■

If both $\sum_{i=0}^{\infty} W_i \left(\sum_{j=i+1}^{\infty}(1-p_A)^j W_j\right)$ and $\sum_{i=0}^{\infty}(1-p_A)^i W_i^2$ converge, then $\sum_{i=0}^{\infty}(1-p_A)^i W_i$ and $\sum_{i=0}^{\infty}\left(\sum_{j=i+1}^{\infty}(1-p_A)^j W_j\right)$ converge. As a result, the second moment of access delay $E[D_{0,p=p_A}^{2,K=\infty}]$ converges according to (28).

2) When $\lim_{i\to\infty} \frac{W_{i+1}}{W_i} > 1$, we have

$$\lim_{n\to\infty} p_A = 1 - \lim_{i\to\infty} \frac{W_i}{W_{i+1}} < 1 - \lim_{i\to\infty} \left(\frac{W_i}{W_{i+1}}\right)^2 < 1 = \lim_{n\to 0} p_A. \quad (63)$$

Corollary 1 shows that p_A is a monotonic decreasing function of n. Therefore, there must exist $n' < \infty$, such that when $n > n'$, $p_A < 1 - \lim_{i\to\infty}\left(\frac{W_i}{W_{i+1}}\right)^2$. In this case,

$$\lim_{i\to\infty}(1-p_A)\left(\frac{W_{i+1}}{W_i}\right)^2 > 1. \quad (64)$$

We can then conclude that the series $\sum_{i=0}^{\infty}(1-p_A)^i W_i^2$ diverges, leading to the divergence of $E[D_{0,p=p_A}^{2,K=\infty}]$ when $n > n'$. ■

Acknowledgment

The authors would like to thank an anonymous reviewer for the insightful comments on the proof of Corollary 2.

References

[1] 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.

[2] F. Cali, M. Conti, and E. Gregori, "IEEE 802.11 protocol: design and performance evaluation of an adaptive backoff mechanism," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 9, pp. 1774–1786, Sept. 2000.

[3] ——, "Dynamic tuning of the IEEE 802.11 protocol to achieve a theoretical throughput limit," *IEEE/ACM Trans. Networking*, vol. 8, no. 6, pp. 785–799, Dec. 2000.

[4] G. Bianchi and I. Tinnirello, "Kalman filter estimation of the number of competing terminals in an IEEE 802.11 network," in *Proc. INFOCOM*, vol. 2, Mar. 2003, pp. 844–852.

[5] A. L. Toledo, T. Vercauteren, and X. Wang, "Adaptive Optimization of IEEE 802.11 DCF Based on Bayesian Estimation of the Number of Competing Terminals," *IEEE Trans. Mobile Computing*, vol. 5, pp. 1283–1296, Sept. 2006.

[6] D. Deng, C. Ke, H. Chen, and Y. Huang, "Contention window optimization for IEEE 802.11 DCF access control," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 5129–5135, Dec. 2008.

[7] R. Rom and M. Sidi, *Multiple Access Protocols: Performance and Analysis*. Springer-Verlag New York Inc., 1990.

[8] K. Ramakrishnan and H. Yang, "The Ethernet capture effect: analysis and solution," in *Proc. 19th Local Computer Networks Conf.*, Oct. 1994, pp. 228–240.

[9] K. Medepalli and F. Tobagi, "On Optimization of CSMA/CA based Wireless LANs: Part I - Impact of Exponential Backoff," in *Proc. ICC*, vol. 5, June 2006, pp. 2089–2094.

[10] C. Koksal, H. Kassab, and H. Balakrishnan, "An analysis of short-term fairness in wireless media access protocols," *SIGMETRICS Perform. Eval. Rev.*, vol. 28, pp. 118–119, June 2000.

[11] N. Vaidya, A. Dugar, S. Gupta, and P. Bahl, "Distributed fair scheduling in a wireless LAN," *IEEE Trans. Mobile Computing*, vol. 4, no. 6, pp. 616–629, Nov./Dec. 2005.

[12] S. Ye and Y. Tseng, "A Multichain Backoff Mechanism for IEEE 802.11 WLANs," *IEEE Trans. Veh. Technol.*, vol. 55, no. 5, pp. 1613–1620, 2006.

[13] J. Choi, J. Yoo, and C. Kim, "A Distributed Fair Scheduling Scheme With a New Analysis Model in IEEE 802.11 Wireless LANs," *IEEE Trans. Veh. Technol.*, vol. 57, no. 5, pp. 3083–3093, 2008.

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

[15] J. Choi, J. Yoo, S. Choi, and C. Kim, "EBA: an enhancement of the IEEE 802.11 DCF via distributed reservation," *IEEE Trans. Mobile Computing*, vol. 4, no. 4, pp. 378–390, July/Aug. 2005.

[16] N.-O. Song, B.-J. Kwak, J. Song, and M. Miller, "Enhancement of IEEE 802.11 Distributed Coordination Function with Exponential Increase Exponential Decrease Backoff Algorithm," in *Proc. VTC 2003-Spring*, vol. 4, 2003, pp. 2775–2778.

[17] Q. Ni, I. Aad, C. Barakat, and T. Turletti, "Modeling and Analysis of Slow CW Decrease IEEE 802.11 WLAN," in *Proc. PIMRC*, vol. 2, Sept. 2003, pp. 1717–1721.

[18] Z. Haas and J. Deng, "On optimizing the backoff interval for random access schemes," *IEEE Trans. Commun.*, vol. 51, no. 12, pp. 2081–2090, Dec. 2003.

[19] I. Aad, Q. Ni, C. Barakat, and T. Turletti, "Enhancing IEEE 802.11 MAC in congested environments," *Elsevier Comput. Commun. J.*, vol. 28, no. 14, pp. 1605–1617, Sept. 2005.

[20] V. Bharghavan, A. Demers, S. Shenker, and L. Zhang, "MACAW: a media access protocol for wireless LAN's," *SIGCOMM Comput. Commun. Rev.*, vol. 24, pp. 212–225, Oct. 1994.

[21] V. Vitsas, "Throughput Analysis of Linear Backoff Scheme in Wireless LANs," *Electronics Letters*, vol. 39, no. 1, pp. 99–100, Jan. 2003.

[22] J. Deng, P. K. Varshney, and Z. Haas, "A New Backoff Algorithm for the IEEE 802.11 Distributed Coordination Function," in *Proc. CNDS*, 2004, pp. 215–225.

[23] A. Nafaa, A. Ksentini, A. Mehaoua, B. lshibashi, Y. Iraqi, and R. Boutaba, "Sliding contention window (SCW): towards backoff range-based service differentiation over IEEE 802.11 wireless LAN networks," *IEEE Network*, vol. 19, no. 4, pp. 45–51, July-Aug. 2005.

[24] C. G. Wang, B. Li, and L. Li, "A new collision resolution mechanism to enhance the performance of IEEE 802.11 DCF," *IEEE Trans. Veh. Technol.*, vol. 53, no. 4, pp. 1235–1246, July 2004.

[25] L. Dai and X. Sun, "A Unified Analysis of IEEE 802.11 DCF Networks: Stability, Throughput and Delay," *IEEE Trans. Mobile Computing*, vol. 12, no. 8, pp. 1558–1572, 2013.

[26] J. Hastad, T. Leighton, and B. Rogoff, " Analysis of Backoff Protocols for Mulitiple Access Channels," *SIAM Journal on Computing*, vol. 25, pp. 740–774, Aug. 1996.

[27] D. Xu, T. Sakurai, and H. Vu, "An Analysis of Different Backoff Functions for an IEEE 802.11 WLAN," in *Proc. VTC 2008-Fall*, 2008, pp. 1–5.

[28] L. Kleinrock and S. Lam, "Packet Switching in a Multiaccess Broadcast Channel: Performance Evaluation," *IEEE Trans. Commun.*, vol. 23, no. 4, pp. 410–423, 1975.

[29] G. Fayolle, E. Gelenbe, and J. Labetoulle, "Stability and Optimal Control of the Packet Switching Broadcast Channel," *J. Assoc. Comput. Machinery*, vol. 24, no. 3, pp. 375–386, Jul. 1977.

[30] R. M. Metcalfe and D. R. Boggs, "Ethernet: distributed packet switching for local computer networks," *Commun. ACM*, vol. 19, no. 7, pp. 395–404, Jul. 1976.

[31] B.-J. Kwak, N.-O. Song, and L. E. Miller, "Performance analysis of Exponential Backoff," *IEEE/ACM Trans. Networking*, vol. 13, pp. 343–355, Apr. 2005.

[32] D. Raychaudhuri and K. Joseph, "Performance evaluation of slotted ALOHA with generalized retransmission backoff," *IEEE Trans. Commun.*, vol. 38, no. 1, pp. 117–122, 1990.

[33] S. Lam and L. Kleinrock, "Packet Switching in a Multiaccess Broadcast Channel: Dynamic Control Procedures," *IEEE Trans. Commun.*, vol. 23, no. 9, pp. 891–904, 1975.

[34] F. Tobagi and L. Kleinrock, "Packet Switching in Radio Channels: Part IV–Stability Considerations and Dynamic Control in Carrier Sense Multiple Access," *IEEE Trans. Comm.*, vol. 25, no. 10, pp. 1103–1119, 1977.

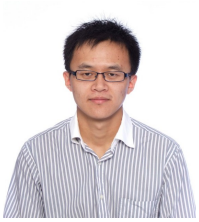
[35] J. Meditch and C.-T. Lea, "Stability and Optimization of the CSMA and CSMA/CD Channels," *IEEE Trans. Comm.*, vol. 31, no. 6, pp. 763–774, 1983.

[36] A. Kumar, E. Altman, D. Miorandi, and M. Goyal, "New insights from a fixed-point analysis of single cell IEEE 802.11 WLANs," *IEEE/ACM Trans. Networking*, vol. 15, pp. 588–601, June 2007.

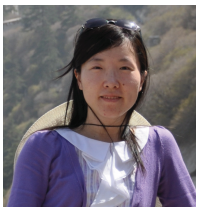
[37] R. M. Corlessa, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, "On the Lambert W function," *Adv. Comput. Math.*, vol. 5, pp. 329–359, 1996.

[38] R. K. Jain, D.-M. W. Chiu, and W. R. Hawe, "A Quantitative Measure Of Fairness And Discrimination For Resource Allocation In Shared Computer Systems," DEC-TR-301, Digital Equipment Corporation, Tech. Rep., Sept. 1984.

[39] M. Shaked and J. Shanthikumar, *Stochastic Orders*. Springer Series in Statistics, 2007.



Xinghua Sun (M'13) received the BS degree from Nanjing University of Posts and Telecommunications (NUPT) in 2008 and the PhD degree from City University of Hong Kong (CityU) in 2013. From March 2010 to July 2010, he was a visiting student with INRIA, France. From March 2013 to August 2013, he was a postdoctoral fellow at CityU. He is now an assistant professor in NUPT. His research interests are in the area of wireless communications and networking.



Lin Dai (S'00-M'03-SM'13) received the B.S. degree from Huazhong University of Science and Technology, Wuhan, China, in 1998, and the M.S. and Ph.D. degrees from Tsinghua University, Beijing, China, in 2003, all in electronic engineering. She was a postdoctoral fellow at the Hong Kong University of Science and Technology and University of Delaware. Since 2007, she has been with City University of Hong Kong, where she is an associate professor. She has broad interest in communications and networking theory, with special interest in wireless communications. She received the Best Paper Award at IEEE Wireless Communications and Networking Conference (WCNC) 2007 and the IEEE Marconi Prize Paper Award in 2009.