

Performance Limits and Analysis of Contention-based IEEE 802.11 MAC

Shao-Cheng Wang and Ahmed Helmy

Department of Electrical Engineering, University of Southern California
(shaochew, helmy}@usc.edu

Abstract— Recent advance in IEEE 802.11 based standard has pushed the wireless bandwidth up to 600Mbps while keeping the same wireless medium access control (MAC) schemes for full backward compatibility. However, it has been shown that the inefficient protocol overhead casts a theoretical throughput upper limit and delay lower limit for the IEEE 802.11 based protocols, even the wireless data rate goes to infinitely high. Such limits are important to understand the bottleneck of the current technology and develop insight for protocol performance improvements.

This paper uses a queuing system approach to extend the discussions of IEEE 802.11 protocol throughput and delay limits to the situation that arbitrary non-saturated background traffic is present in the network. We derive analytical models to quantify the limits for Distributed Coordination Function (DCF) of legacy 802.11a/b/g and Enhanced Distributed Coordination Access (EDCA) of IEEE 802.11e. We find such limits are functions of the underlying MAC layer backoff parameters and algorithms, and are highly dependent on the load that background traffic injects into the network. Surprisingly, depending on the rate of arrival traffic, the packet delay limit may become unbounded such that no delay sensitive services can be operated under such condition. Moreover, we also discuss the effects of different frame aggregation schemes on the performance limits when data rate is infinite. The developed model and analysis provide a comprehensive understanding of the performance limitations for IEEE 802.11 MAC, and are useful in gauging the expected QoS for the purposes such as admission control.

Keywords—IEEE 802.11, MAC, DCF, EDCA, Performance Analysis, Non-saturation, Throughput, Delay.

I. INTRODUCTION

In recent years, the IEEE 802.11-based [1] wireless local area networks (WLANs), namely IEEE 802.11b [2], 802.11g [3], and 802.11a [4], have been increasingly popular in providing low-cost high-bandwidth (up to 54Mbps) wireless connections. With the growing demands of higher bandwidth for applications such as high-definition video streaming, network storage, and online gaming, the industry has been working to seek higher data rate (HDR) extensions [5]-[7] to the family of IEEE 802.11 specifications. Earlier this year, IEEE Working Group meeting approved the first draft of IEEE 802.11n [8], in which the data rate is expected to be as high as 600Mbps. Moreover, the 802.11n specification adopts the same medium access control (MAC) schemes to ensure backward compatibility with existing IEEE 802.11 specifications.

The industry also seeks advancement in providing better Quality-of-Service (QoS) at the MAC layer. A QoS amendment of IEEE 802.11 MAC, IEEE 802.11e [9], aims to provide service differentiations to different traffic types. In

particular, the Enhanced Distributed Channel Access (EDCA) contention-based medium access improves the legacy IEEE 802.11 Distributed Coordination Function (DCF) by providing differentiated medium contention opportunities to high priority traffic.

Despite the efforts on advancing data rate and QoS of IEEE 802.11, an analysis of theoretical throughput and delay limit was first discussed in [10] by Xiao and Rosdahl. The paper emphasized on the 802.11 MAC overhead effectiveness and proved the existence of theoretical throughput and delay limits for IEEE 802.11 DCF protocol. The authors concluded that, given that the PHY data rate has advanced to infinitely high and only one station transmits in the ideal channel condition, the minimum time required for completing one packet transmission task is bounded by PHY and MAC headers as well as MAC layer backoff waiting time, and consequently bounds the maximum achievable throughput and minimum achievable packet delay. In [11], the authors extended the derivation of packet transmission time to consider collisions and backoff freeze in the case that multiple stations transmit in saturation mode.

However, the results in [10] and [11] only represent the throughput and delay limit in aforementioned special cases but are unsuitable to real-world scenarios, which typically consist of multiple wireless stations operating in non-saturation mode. Besides, the delay analysis presented in [11] only considers the “medium access delay” and fails to address the “queuing delay” for the waiting time packets spent when backlogged. On the other hand, as the models used in [10] and [11] are only applicable to legacy IEEE 802.11 DCF, it is also important to expand the explorations of theoretical limits to the QoS enhanced IEEE 802.11e specification. In particular, it is essential to answer the following questions: will the similar performance boundaries exist in the EDCA MAC protocol? If so, how to quantify such boundaries in different prioritized traffic categories and what are the subsequent impacts in fulfilling the QoS requirements promised by IEEE 802.11e EDCA?

Therefore, this paper aims to provide a comprehensive understanding of the performance limitations on throughput and total system delay of both IEEE 802.11 DCF and EDCA MAC protocols with arbitrary amount of non-saturated competing traffic. Such analysis is critical in pinpointing the performance bottleneck of state-of-the-art IEEE 802.11 technologies and in developing insight for future protocol performance improvements. We propose a queuing system point of view to directly analyze the access dynamics of 802.11 contention-based MAC. The proposed model works with any saturated or non-saturated underlying competing traffic patterns. Each wireless station is modeled as a queuing

system with the packet generation process as the ‘arrival process’ and the variable amount of time a packet spends on MAC layer medium contention as the packet ‘service process’. The packet throughput and delay bound is then derived with infinitely high operating data rate. The results are validated through extensive simulations under various network loading and operation conditions. The challenges to such analysis are in modeling the dynamic interactions between the arrival pattern of the considered node and the significantly variable amount of network delay incurred by the backoff, collision, and re-transmission procedures under different background traffic load level.

Our paper makes the following contributions:

- We construct a lightweight mathematical model for characterizing the throughput and delay limits and performance of contention-based IEEE 802.11 MAC. The proposed model enables us to systematically explore the effects of backoff settings, arrival processes, competing traffic characteristics, and frame aggregation schemes on theoretical throughput and delay limit of different versions of IEEE 802.11 MAC protocols, including legacy DCF and QoS enhanced EDCA.
- We discover a performance bottleneck of the 802.11 DCF and EDCA under the presence of background traffic: there is a turning point when packets arrive faster than the packet ‘service’ rate, packet delay becomes unboundedly high beyond such network condition.

The rest of the paper is outlined as follows. Section II provides background information for IEEE802.11 MAC standard. Section III describes the queuing system based mathematical model for packet throughput and delay of IEEE 802.11 MAC. Evaluations and simulation comparisons of the proposed model is presented in Section IV. Section V concludes and provides future work directions.

II. BACKGROUND

In this section, we briefly review the legacy IEEE 802.11 DCF MAC protocol and the enhanced IEEE 802.11e EDCA. We also describe the differences among 802.11b, 802.11g, and 802.11a. We highlight the different backoff settings, including the special protection feature for 802.11g devices to interoperate with 802.11b devices, which affect the derivation of the proposed model.

A. DCF and EDCA of 802.11 Standard

The DCF of IEEE 802.11 is a “listen-before-talk” medium access scheme based on the Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) protocol. Before initiating any packet delivery, the station detects the wireless medium to be idle for a minimum duration called DCF Interframe Space (DIFS). The station randomly select the backoff timer interval from $[0, CW_{min}]$ number of slot_time, where slot_time is a parameter depends on the underlying physical layer (PHY), and then enters the backoff process. During the count-down of backoff timer, if the station senses the medium busy, it stops decrementing the timer and does not reactivate the paused value until the channel is sensed idle again for more than a DIFS. At the timer expiration, the station is free to access the medium for packet transmission. Upon receiving an acknowledgement frame, the transmission

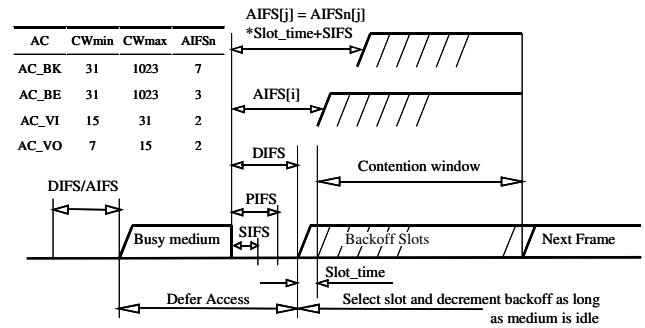


Figure 1. IEEE 802.11e EDCA

TABLE I. TIMING PARAMETERS OF 802.11A, 802.11G, AND 802.11B STANDARD

	802.11a	802.11g (pure/hybrid)	802.11b
SlotTime	9 μ s	9 μ s/20 μ s	20 μ s
SIFS	16 μ s	16 μ s/10 μ s	10 μ s
DIFS	34 μ s	34 μ s/50 μ s	50 μ s
T_p	16 μ s	16 μ s/72 μ s*	72 μ s*
T_{PHY}	4 μ s	4 μ s/24 μ s	24 μ s
CW_{min}	15	15	31
Supported Data Bit Rate (Mbps)	54, 48, 36, 24, 18, 12, 9, 6	54, 48, 36, 24, 18, 12, 9, 6/ 11, 5.5, 2, 1	11, 5.5, 2, 1

T_p : transmission time of physical preambles

T_{PHY} : transmission time of PHY header

*short preamble mechanism

is considered successful; the CW is reset to minimum CW_{min} and the station stands-by for the next packet arrival. The transmission is considered failed if no acknowledgement is received within a specified timeout; the station repeats the backoff process with CW selection range doubled up to maximum contention window, CW_{max} . If the transmission has been re-tried for up to RetryLimit times, the packet will be discarded and the CW is reset to CW_{min} .

The EDCF is a variant of DCF and provides prioritized Quality-of-Service (QoS) support among different traffic types. Each QoS-enhanced station (QSTA) maps the packets arriving at MAC layer into four different access categories (ACs) and assigns a set of backoff parameters, namely Arbitration IFS (AIFS), CW_{min} , and CW_{max} to each AC. As illustrated in Figure 1, each AC uses its own backoff parameters to contend for the wireless medium by the same backoff rules as legacy DCF stations in the previous paragraph. The $AIFS[AC]$, determined by $AIFS[AC] = AIFSn[AC] \cdot slot_time + aSIFSTime$, replaces the fixed DIFS in DCF. Shorter $AIFS[AC]$ in higher priority AC provides high priority traffic earlier timing to unfreeze the paused timer after each busy wait period. On the other hand, smaller CW sizes probabilistically provide shorter backoff stages to high priority traffic. More detailed description of DCF and EDCF can be found in [2] and [9] respectively.

B. IEEE 802.11b, 802.11g, and 802.11a

The IEEE 802.11b, 802.11a, and 802.11g are higher-speed physical layer (PHY) extensions of the IEEE 802.11 standard. They all use the same DCF medium access mechanism described in previous section. Note that, on the other hand,

IEEE 802.11e is the MAC QoS enhancement amendment to the IEEE 802.11 standard and can be incorporated with any of the three higher-speed PHY extensions.

The detailed operational parameter settings of the three versions of standard are summarized in Table I. Compared to IEEE 802.11b, IEEE 802.11a offers higher data rate, shorter PHY header, MAC slot time, and lower minimum contention window. On the other hand, IEEE 802.11g operates at the same band as 802.11b and supports data rates and PHY/MAC parameters of both 802.11b and 802.11a. When there are no 802.11b stations present in the network, all 802.11g stations utilize the same PHY/MAC settings specified in 802.11a. In the case when 802.11g stations co-exist with 802.11b stations in the network, 802.11g-capable stations shall switch to longer 20ms slot time in order to be synchronized with the timing of 802.11b stations. In addition, whenever 802.11g-capable stations use OFDM modulated high rate to transmit DATA and ACK frames, a special protection RTS-CTS or CTS-to-self exchange formed in 802.11b decodable control frames must precede the data frame in order to ensure interoperability. As we will show later, these parameters have substantial effects in theoretical protocol performance limits.

III. ANALYSIS MODEL

In this section, we derive the theoretical throughput and delay limit with non-saturation background traffic for IEEE 802.11 DCF and EDCA contention-based wireless medium access methods. We consider the infrastructure Basic Service Set (BSS) scenario, which consists of multiple wireless nodes and a base station connected with wired networks. Following the “best-case scenario” philosophy in [10] and [11], we make the following assumptions:

- 1) The wireless channel is ideal without errors.
- 2) All nodes are within “carrier sensing range” of each other.
- 3) All nodes use the basic access operation (no RTS/CTS) for shorter transmission cycles.

The key idea to our analysis is to model the MAC layer timing dynamics, from packet arriving into the sending station until the packet received by the intended node, as a G/G/1 queuing system. The theoretical throughput and delay limit are thus derived with infinite data rate. We will show that, even with infinitely high data rate, the overhead of background packets causing non-negligible amount of time in the backoff stages is the dominant factor that bounds the MAC layer throughput and delay limit. In the extended version [18], we show that our model can also be applied to quantify the throughput and delay performance in practical scenarios such as finite data rate and non-ideal wireless channel.

A. Packet Arrivals

Depending on the application layer, the traffic arriving at each wireless station can be characterized with different probabilistic models. In our proposed model, we treat the packet arrivals as the ‘general’ arrival process of G/G/1 queue. For special case arrival process such as Voice over IP (VoIP) with deterministic arrival rate, it can be treated as D/G/1 queue in our model.

B. MAC Layer Service Time

The packet service time of the proposed model is defined as MAC layer service time: the time duration from the instant

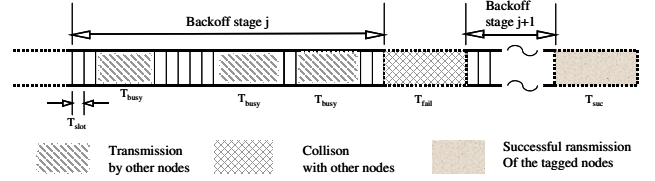


Figure 2. Packet transmission and collision events during MAC backoff

that a packet becomes the head of the transmission queue and starts MAC layer contention backoff process, to the instant that the packet is successfully received or being dropped because of maximum retry limit has reached. As shown in Figure 2, we model MAC layer service time by analyzing the duration and occurring probability of different events take place at backoff stages.

1) When the backoff timer decrements, either no transmission is sensed in the time slot and T_{slot} (the length of one time slot) elapses, or the slot is sensed busy with T_{busy} taken. Here T_{busy} is the average time interval the wireless medium being occupied by background traffic transmissions. We define P_{busy} be the probability that, at a given time slot, the backoff timer is frozen due to busy medium in carrier sensing. The occurring probability of idle slot is simply $1 - P_{busy}$.

2) When the backoff timer expires (i.e. decrements to zero), the attempt of packet transmission might either fail or succeed. In the failure case, which occurs with probability P_{fail} , T_{fail} is taken. Note that, in the case of ideal wireless channel, P_{fail} equals P_{busy} because transmission attempt collides only when the slot is supposed to be busy.

3) In the success case, it takes T_{succ} for the packet transmission process and then the packet is considered ‘served’ and therefore departs the queue. The probability of transmission attempt succeeds is simply $1 - P_{fail}$.

Note that we assume that P_{busy} (and P_{fail}) is constant in steady-state and independent of the backoff stages of the node under consideration (i.e. the tagged node).¹ As a result, nodes can obtain P_{busy} (and P_{fail}) by monitoring the channel activity and gathering the long-term statistics of the ratio that medium is busy over all time slots [12]. Likewise, T_{busy} and T_{fail} can also be obtained by channel activity monitoring.

On the other hand, T_{busy} , T_{fail} , and T_{succ} can be formulated by considering the frame exchanges and MAC layer timing parameters involved in a successful or collided transmission cycle. For example, T_{succ} can be expressed by the duration of DATA and ACK frame for pure 802.11a/b/g traffic, or by the duration of CTS, DATA, and ACK frame when the tagged node operates at hybrid 802.11g environment and has CTS-to-self protection turned on:

$$T_{succ}^{pure} = \left(T_p + T_{PHY} + \frac{8 \times L_{DATA}}{R_{DATA}} \right) + \delta + SIFS + \left(T_p + T_{PHY} + \frac{8 \times L_{ACK}}{R_{ACK}} \right) + \delta + DIFS \quad (1a)$$

¹ Previous work has shown that this assumption has very meager effects on model accuracy [12].

$$\begin{aligned}
T_{succ}^{1lg_hybrid} &= \left(T_p^{CCK} + T_{PHY}^{CCK} + \frac{8 \times L_{CTS}}{R_{CTS}} \right) + \delta + SIFS \\
&+ \left(T_p^{OFDM} + T_{PHY}^{OFDM} + \frac{8 \times L_{DATA}}{R_{DATA}} \right) + \delta + SIFS \quad (1b) \\
&+ \left(T_p^{OFDM} + T_{PHY}^{OFDM} + \frac{8 \times L_{ACK}}{R_{ACK}} \right) + \delta + DIFS.
\end{aligned}$$

where L_{CTS} , L_{DATA} , and L_{ACK} is the size (in bytes) of CTS, DATA and ACK frame, respectively. R_{CTS} , R_{DATA} , and R_{ACK} is the data rate (in bps) of CTS, DATA, and ACK, respectively. SIFS is the mandatory Short IFS inserted between frames. σ is the propagation delay.

In the case when the attempt of packet transmission fails, considering the ACK timeout effect, T_{fail} is expressed with the longest data frame involved in the collision. In other words, the L_{DATA} in Equation 1a and 1b is the size (in bytes) of the longest data frame involved in the collision.

The duration of busy slot, T_{busy} , can be expressed either by T_{succ} when the busy slot is occupied by successful transmission of the background traffic, or by T_{fail} when the busy slot is occupied by packet collisions. In the case when the wireless nodes operate at finite data rates, T_{busy} can be collected by the long-term statistics of channel activity monitoring.

Finally, note that, with infinitely high data rate, the time duration to carry the payload of CTS, DATA, and ACK frames become infinitesimal. As a result, depending on the network operates in pure 802.11a/b/g, or in hybrid 802.11b/g environment, T_{succ} , T_{fail} , and T_{busy} can be expressed by

$$\begin{aligned}
T_{succ}^{pure} &= T_{fail}^{pure} = T_{busy}^{pure} \\
&= 2T_p + 2T_{PHY} + 2\delta + SIFS + DIFS \quad (2a)
\end{aligned}$$

$$\begin{aligned}
T_{succ}^{1lg_hybrid} &= T_{fail}^{1lg_hybrid} = T_{busy}^{1lg_hybrid} \\
&= T_p^{1lb} + T_{PHY}^{1lb} + 2T_p^{1lg_pure} + 2T_{PHY}^{1lg_pure} + 3\delta + 2SIFS + DIFS. \quad (2b)
\end{aligned}$$

where T_p and T_{PHY} is the preamble and PHY layer overhead. SIFS is the mandatory Short IFS inserted between frames. δ is the propagation delay.

C. Throughput and Delay Model of Legacy DCF

To derive the throughput and delay of IEEE 802.11 MAC, we need to construct the detailed service time distribution, which will then be applied to standard queuing theory model. In this subsection, we derive the detailed service time distribution by carefully examining the variable amount of time spent on busy and silent slots and the corresponding occurring probabilities throughout the backoff stages.

We first define the occurring probability $F_{k,n-k}^j$ that, in any single backoff stage j with backoff timer selected from 0 to W_j , there are exactly k busy time slots and $(n-k)$ idle slots is,

$$F_{k,n-k}^j = \frac{1}{W_j} C_k^{i+k} P_{busy}^k (1 - P_{busy})^{n-k}, \quad 0 \leq k \leq n \leq W_j. \quad (3)$$

Moreover, we know that such combination of number of busy and idle slots can be a cumulative effect from successive backoff stages. Therefore, we then define an intermediate term

$$\begin{aligned}
&S_{k,n-k}^j \text{ for probability of backoff counter being frozen } k \text{ times} \\
&\text{and idle } (n-k) \text{ times that up to back off stage } j, \\
&S_{k,n-k}^0 = F_{k,n-k}^0, \quad 0 \leq k \leq n \leq W_0, \quad \text{for } j = 0. \\
&S_{k,n-k}^j = P_{fail} \sum_{m=0}^{k-1} \left(\sum_{i=0}^{n-k} S_{m,i}^{j-1} \times F_{k-j-m,n-k-i}^j \right) \\
&\text{, for } 1 \leq j \leq m, \quad j \leq k \leq n \leq \sum_{i=0}^j W_i - 1.
\end{aligned} \quad (4)$$

For stage 0, this intermediate term equals Equation 3. For stage greater than zero (i.e. $j=1,2,\dots,m$), this intermediate term includes all possible cases, from combination of previous stage(s) and the current stage, which result in k busy slots and $(n-k)$ idle slots. Finally, for all m backoff stages, the overall probability of backoff counter being frozen exactly k times and idle $(n-k)$ times is

$$\begin{aligned}
S_{k,n-k} &= (1 - P_{fail}) \sum_{i=0}^{m-2} S_{k,n-k}^i + S_{k,n-k}^{m-1}, \\
&0 \leq k \leq n \leq \sum_{i=0}^m W_i - 1.
\end{aligned} \quad (5)$$

Since Equation 5 covers all possible combinations of busy and idle slot time, we have the probability distribution function (pdf) of MAC layer service time, B_t , in the sequences of time points at the multiples of the busy medium time (T_{busy}) and slot time (T_{slot}) plus the successful transmission time (T_{succ}).

$$\begin{aligned}
\text{Prob}[B_t] &= S_{k,n-k}, \quad \text{for } t = k * T_{busy} + (n-k) * T_{slot} + T_{succ} \\
&0 \leq k \leq n \leq \sum_{i=0}^m (W_i + 1) - 1.
\end{aligned} \quad (6)$$

Furthermore, the probability generating function (pgf) of B_t can be expressed with a reasonable system clock unit, e.g. in μs or in system slot time,

$$\begin{aligned}
B(z) &= \sum_{k=0}^{\infty} \text{Prob}[B_t] \times z^t \\
&= S_{0,0} z^{T_{succ}} + S_{1,0} z^{1 * T_{busy} + T_{succ}} + S_{1,1} z^{1 * T_{busy} + 1 * T_{slot} + T_{succ}} + S_{2,0} z^{2 * T_{busy} + T_{succ}} \dots
\end{aligned} \quad (7)$$

We know that the average MAC layer service time can be obtained by $B'(1)$, the first derivative of $B(z)$ at $z=1$. Hence, with the LDATA bytes long payload in IEEE 802.11 DATA frame, maximum achievable throughput can be expressed by

$$\text{Throughput} = \frac{8 * L_{DATA}}{B'(1) + T_{succ}}. \quad (8)$$

On the other hand, to derive the total packet delay, we can apply standard discrete time queuing theory [13]-[15] with the statistical characteristics of the arrival and service process. Here we assume the first and second moment of the arrival distribution are known and can be expressed in closed form. $A'(1)$ and $A''(1)$ represent the first and second derivative of the pgf of arrival distribution, $A(z)$, at $z=1$, respectively.

According to [13], if the arrival process is a general independent (GI) arrival process, i.e., the numbers of packets entering the system during the consecutive time units are

assumed to be independent and identically distributed (i.i.d.), the mean system time, i.e. packet delay in our case, of GI/G/1 queue system can be expressed as

$$\overline{Delay}_{GI/G/1} = 1 - \bar{X} + B'(1) + \frac{[A'(1)]^2 B''(1) + A''(1) B'(1)}{2[1 - A'(1) B'(1)]}. \quad (9)$$

where $B'(1)$ and $B''(1)$ are the first and second derivative of the pgf of MAC layer service time, i.e. $P_{T^{serv}}(z)$, at $z=1$. \bar{X} is the mean distance of the arrival point from the start of the unit time slot. When unit time is small, \bar{X} is negligible.

However, Equation 9 is not applicable for applications with deterministic arrival process, e.g. VoIP. Therefore, we refer to [14] for models of discrete-time D/G/1 queues. The average delay of such system is

$$\overline{Delay}_{D/G/1} = B'(1) - \frac{A'(1)(A'(1)-1) - B''(1)}{2(A'(1) - B'(1))} + \sum_{r=1}^{N-1} \frac{1}{1 - z_r}. \quad (10)$$

where N is the inter-arrival time, in system time unit, of the deterministic arrival process. z_r are the roots of solving $zN - B(z) = 0$ on or inside the unit circle.

Finally, for packet arrivals that are neither General Independent process nor Deterministic process, [15] provides an upper bound of the system waiting time,

$$\bar{W} \leq \frac{A''(1) + B''(1)}{2[A'(1) - B'(1)]}. \quad (11)$$

And hence the upper bound of total system delay is

$$\overline{Delay}_{G/G/1} \leq B'(1) + \bar{W}. \quad (12)$$

D. Throughput and Delay Model of EDCA

The QoS-enabled IEEE 802.11e EDCA mechanism provides prioritized medium access by assigning different AIFS and backoff window parameters (CW_{min} and CW_{max}) to different traffic categories. In particular, AIFS provides advanced opportunity to high priority traffic to access the medium by shorten the amount of time a station defers access to the channel following a busy time slot. However, AIFS changes the way we construct the occurring probability of busy and idle slot combinations. Therefore, we need to re-define Equation 3.

In order to perform theoretical throughput and delay analysis, we again assume the ‘‘best-case scenario’’. That is, we assume that only the tagged node utilize the short AIFS traffic category, i.e. AC_VO or AC_VI with $AIFSn=2$, while all other competing traffic utilize the AC_BE traffic category with $AIFSn=3$. In this way, only the tagged node has higher advanced priority to access the medium and thus is considered as ‘‘best-case’’ scenario.

Under this setting, what happens in the last backoff time slot of the tagged node decides two different scenarios. 1) When last backoff slot ($cw=1$) was an idle slot, the transmission is subject to collision. 2) When last backoff slot was a busy slot, the tagged node un-freezes the backoff timer one time slot before all other traffic. As a result, the backoff timer of tagged node expires before all other traffic un-freeze the timer, and thus the transmission is guaranteed to be successful without collision.

Here we first consider the case last backoff slot was an idle slot. We define the occurring probability $FCC_{k,n-k}^j$ that, in any single backoff stage j , there are exactly k busy time slots and $(n-k)$ idle slots is,

$$FCC_{k,n-k}^j = \frac{1}{W_j} C_k^{n-k-1} P_{busy}^k (1 - P_{busy})^{n-2k-1}, \quad (13)$$

$$0 \leq k \leq (W_j + 1) / 2 - 1, \quad 2k + 1 \leq n \leq W_j.$$

Note that this formulation differs from Equation 3 in the occurring probability of idle slots. The very first time slot and the time slots after busy slot always happen before all other traffic with probability 1. The slots other than these special slots and busy slots in Equation 3 are all classified as idle slot with occurring probability $(1 - P_{busy})$.

We then consider the case last backoff slot was a busy slot. We define the occurring probability $FNC_{k,n-k}^j$ that, in any single backoff stage j , there are exactly k busy time slots and $(n-k)$ idle slots is,

$$FNC_{k,n-k}^j = \frac{1}{W_j} C_{k-1}^{n-k-1} P_{busy}^k (1 - P_{busy})^{n-2k}, \quad (14)$$

$$1 \leq k \leq (W_j + 1) / 2 - 1, \quad 2k \leq n \leq W_j.$$

Similarly, in this formulation, only $(n-2k)$ idle slots happen with probability $(1 - P_{busy})$. All other idle slots happen with probability 1.

Subsequently, the intermediate terms defining all possible cases, from combination of previous stage(s) and the current stage, which result in k busy slots and $(n-k)$ idle slots follow similar derivations as Equations 4 to 7.

$$SCC_{k,n-k}^0 = FCC_{k,n-k}^0, \quad 0 \leq k \leq (W_j + 1) / 2 - 1, \quad 2k + 1 \leq n \leq W_j.$$

$$SCC_{k,n-k}^j = P_{fail} \sum_{m=0}^{k-1} \left(\sum_{i=0}^{n-k} SCC_{m,i}^{j-1} \times FCC_{k-j-m,n-k-i}^j \right)$$

$$, \text{ for } 1 \leq j \leq m, \quad j \leq k \leq \sum_{i=0}^j [(W_j + 1) / 2 - 1] - 1,$$

$$2k + 1 \leq n \leq \sum_{i=0}^j W_j - 1 \quad (15)$$

$$SNC_{k,n-k}^0 = FNC_{k,n-k}^0, \quad 1 \leq k \leq (W_j + 1) / 2 - 1, \quad 2k \leq n \leq W_j.$$

$$SNC_{k,n-k}^j = P_{fail} \sum_{m=0}^{k-1} \left(\sum_{i=0}^{n-k} SCC_{m,i}^{j-1} \times FNC_{k-j-m,n-k-i}^j \right)$$

$$, \text{ for } 1 \leq j \leq m, \quad j \leq k \leq \sum_{i=0}^j [(W_j + 1) / 2 - 1] - 1,$$

$$2k \leq n \leq \sum_{i=0}^j W_j - 1 \quad (16)$$

Finally, for all m backoff stages, the overall probability of backoff counter being frozen exactly k times and idle $(n-k)$ times is

$$S_{k,n-k} = \sum_{i=0}^{m-1} SNC_{k,n-k}^i + (1-p) \sum_{i=0}^{m-2} SCC_{k,n-k}^i + SCC_{kn-k}^{m-1}, \quad (19)$$

$$0 \leq k \leq \sum_{i=0}^j [(W_j + 1)/2 - 1] - 1.$$

As a result, we can plug in the obtained probability distribution function (pdf) of MAC layer service time, B_t , and corresponding probability generating function to Equations 8-12 to get the theoretical throughput and delay limit for IEEE 802.11e EDCA.

IV. RESULT

In this section, we use the queuing system based packet throughput and delay model to quantify and explore the theoretical throughput and delay limits of different IEEE 802.11 MAC specifications.

A. P_{busy} and T_{busy}

From the analysis model in Section III, we can see that MAC layer packet service time, and subsequently throughput and packet delay, is a direct function of two parameters: P_{busy} , slot busy probability, and T_{busy} , average slot busy interval. Therefore, it is important before we proceed to present the numerical evaluations of the proposed model, we first quantify and understand the implications of these two parameters.

P_{busy} is an indicator for how busy the network is and it is usually a function of number of nodes in the network, their traffic patterns and corresponding traffic load. It is obvious from Section III that the busier the network is, the more often a packet will wait on busy slots, and consequently the MAC layer service time is longer. Unfortunately, no existing model can be used to quantify P_{busy} with arbitrary number of nodes and traffic loads, such that we might have difficulties in relating the amount of P_{busy} with real-world scenarios and quantifiable metrics such as number of nodes or traffic loads².

Nevertheless, in real-world traffic scenarios, P_{busy} can always be obtained by gathering the long-term statistics from channel activity monitoring. We thus adapt the usage model scenarios suggested by IEEE 802.11 Task Group N (TGn) [17] to further illustrate the relationship between P_{busy} and real-world non-saturated traffic. As summarized in Table II, we use different combinations of high-bandwidth multimedia (video and audio) and data networking applications to emulate futuristic high-performance wireless network scenarios, such as digital home, digital office, and public hotspots. We then obtain P_{busy} of each scenario through simulations. Later, we use these scenarios to evaluate the accuracy of the proposed packet delay and throughput model under particular P_{busy} .

On the other hand, T_{busy} is an indicator for how long a busy slot takes. The longer T_{busy} is, the longer it takes to wait on busy slots, and consequently the MAC layer service time is longer. Recall from Equation 1 in Section III, it is obvious

² If we assume all nodes transmit in saturation mode, then the models in [16] can accurately quantify and related P_{busy} with number of nodes in the network.

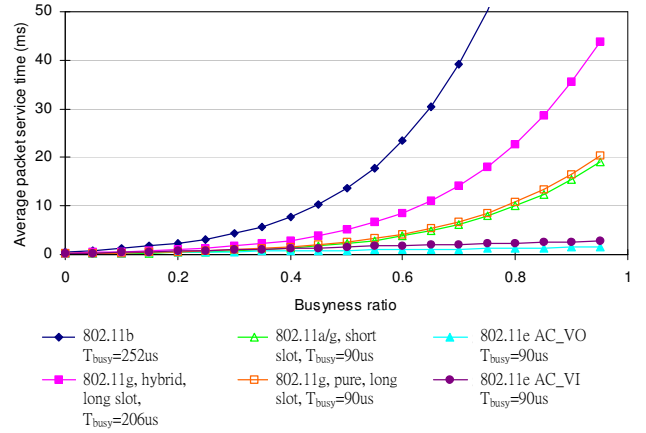


Figure 3. Average MAC layer service time of different 802.11 specifications

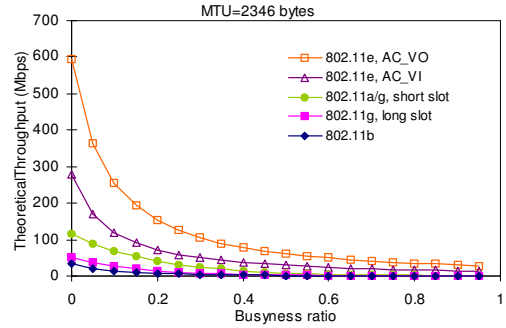


Figure 4. Theoretical throughput limit of different 802.11 specifications

that, when we consider infinite data rate, T_{busy} is determined by the PHY and MAC overhead specified in different versions of 802.11 standard. Later we will see how does different length of T_{busy} , potentially varies about an order of magnitude for different versions of 802.11 standard, impact on throughput and delay limits of IEEE 802.11 MAC protocols.

B. MAC Layer Packet Service Time

To study the theoretical packet throughput and delay limit of IEEE 802.11 MAC, we first examine the MAC layer packet service time. Note that, even with infinitely high data rate, this is the minimum required time that the packets need to wait during MAC backoff due to finite protocol overhead of background traffic in busy slots.

Figure 3 plots the average MAC layer service time of different versions of 802.11 standard in the presence of non-saturated background traffic. We can see that packet service time of all 802.11 specifications increases with P_{busy} . In a network with busyness ratio P_{busy} as low as 0.5-0.6, the MAC layer packet service time can be in the order of tens of milliseconds. As we will see in later sections, this significant amount of medium access time limits the achievable packet throughput and delay of IEEE 802.11-based protocols.

Besides, we can refer the MAC parameters and overhead of different IEEE 802.11 specifications in Table I and see such parameters do affect the packet service time significantly. For example, because the minimum contention window (CW_{min}) of 802.11g is only half of that of 802.11b, it takes roughly half of

the time for a packet from 802.11g station to be served, compared to a packet from 802.11b station. Due to shorter T_{busy} , packet service time of pure 802.11a/g network is further decreased, compared to 802.11g station in hybrid 802.11 b/g networks. Moreover, through the medium access advantage of 802.11e AIFS differentiation, the packet service time in 802.11e networks reduces by an order of magnitude, compare to legacy 802.11 protocols. On the other hand, by comparing pure 802.11g in long slot and short slot setting, we find it interesting that shorter slot time does not help in reducing the packet service time too much. It is because shorter slot time only saves the time spent in idle slots by couple of microseconds, which are relatively insignificant compared to the duration of busy slots in hundreds of microseconds.

From the observations we made above, we find that, through the progression of recent PHY and MAC amendments on 802.11 standard, the minimum medium access time has been reduced significantly. In the following sections, we will examine such effects on packet throughput and delay.

C. Theoretical Throughput Limit

In this section, we examine the theoretical maximum throughput of different 802.11 specifications with infinite data rate. From previous section we know that, even with infinitely high data rate, there is a minimum medium access time that the packets need to wait during MAC backoff. As a result, the maximum amount of data delivered in a given period of time is bounded by this minimum medium access time and thus a theoretical throughput limit of 802.11 MAC exists.

Figure 4 plots the theoretical throughput limits with arbitrary background traffic at different level of busyness ratio. When there is no background traffic, i.e. $P_{\text{busy}} = 0$, our result is consistent with the theoretical throughput limit presented in [10]. As P_{busy} increases, we can see the maximum throughput decreases exponentially. In a network with P_{busy} about 0.5, the throughput limit has decreased for more than an order of magnitude. Moreover, as the latest IEEE 802.11n proposal aims to provide 100Mbps effective throughput at the MAC layer, our result indicates that theoretically, even with infinite data rate, such goal can only be achieved at the condition that the busyness ratio of the network is less than 0.3 (or 0.2) with 802.11e AC_VO (or AC_VI) being employed. Such QoS performance boundary is not identified before.

D. Theoretical Delay Limit

From the results we present above, the existence of finite MAC service time, in the case of infinitely high data rate, bounds the minimum packet delay that can be achieved by IEEE 802.11-based MAC. Moreover, recall from Section III, depending on the type of packet arrival process, additional backlog waiting time will be added to the total packet delay. In this section, by using the queuing system based MAC layer packet delay model, we present the results of theoretical delay limit of different 802.11 MAC specifications in the presence of arbitrary background traffic.

Figure 5 shows the theoretical packet delay limit of different 802.11 specifications with infinite operation data rate. Here we use the deterministic arrival process of a G. 711 VoIP application with 10ms inter-arrival time as a case study. We can see the delay limits increase exponentially as P_{busy} increases. It is because the MAC layer service time increases

TABLE II TGN USAGE MODELS IN HIGH PERFORMANCE NETWORKS

Application	Offered Load (Mbps)	Traffic Type	Number of applications in		
			TGn-1: DIGITAL HOME	TGn-2: DIGITAL OFFICE	TGn-3: PUBLIC HOTSPOT
VoIP	0.096	UDP	3	30	15
Video Conferencing /Video Phone	0.5	TCP	1	10	0
A/V Streaming	2-4	UDP	1	0	10
STDV	4	UDP	1	1	1
HDTV	19.2-24	UDP	2	0	0
Internet File Transfer	N/A	TCP	1	0	10
Local File Transfer	N/A	TCP	0	10	2
P_{busy} (with unlimited data rate)			0.159	0.217	0.47

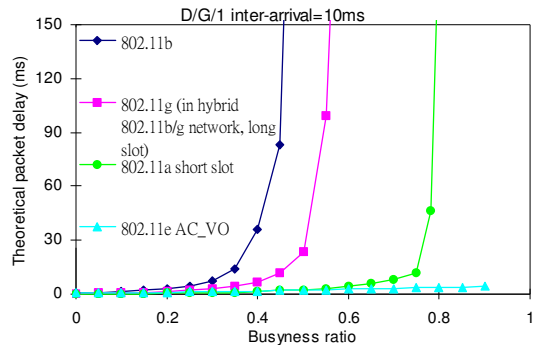


Figure 5. Theoretical delay limit of different 802.11 specifications

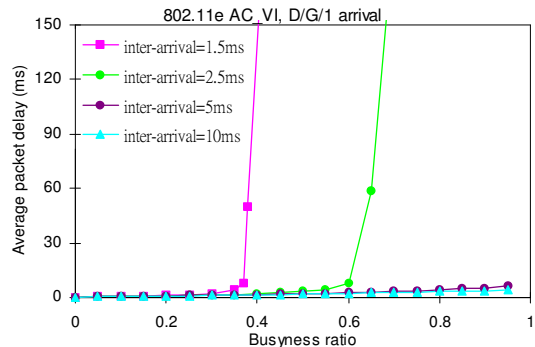


Figure 6. Theoretical delay limit of 802.11e AC_VI with different arrival process

as the network get busier, and thus increases the backlog waiting time significantly. An observation that worth special attention in this figure is that there is a point where the packet delay becomes unbounded. Recall from the queuing model in Section III, this is because the fact that whenever the queue service time approaches or even exceeds the packet arrival time (i.e. 10ms in this case), the queue becomes unable to handle packets in a timely manner, and eventually packets become indefinitely backlogged. The existence of such turning point demonstrates an important performance limitation of

TABLE III COMPARISON OF PACKET DELAY FROM MODEL AND SIMULATIONS

		P_{Busy}	Delay (Model)	Delay (Simulation)	Error (%)
802.11b	TGn-1	0.159	2.000ms	1.945ms	2.75%
	TGn-2	0.217	3.239ms	3.120ms	3.67%
	TGn-3	0.47	∞	656.669ms	N/A
802.11g	TGn-1	0.159	0.996ms	0.957ms	3.92%
	TGn-2	0.217	1.471ms	1.450ms	1.43%
	TGn-3	0.47	15.020ms	13.926ms	7.28%
802.11e, AC_VI	TGn-1	0.159	0.533ms	0.502ms	5.82%
	TGn-2	0.217	0.609ms	0.664ms	9.03%
	TGn-3	0.47	0.949ms	0.912ms	3.90%
802.11e, AC_VI	TGn-1	0.159	0.899ms	0.810ms	9.90%
	TGn-2	0.217	1.060ms	1.075ms	1.42%
	TGn-3	0.47	1.798ms	1.773ms	1.39%

802.11-based MAC that is not discovered in previous literature: with the presence of background traffic, the packet delay of 802.11-based MAC can be boundlessly high even with infinite operating data rate. Our model identifies a network condition boundary, such as $P_{\text{busy}} > 0.45$ for 10ms-frame G. 711 VoIP in 802.11b networks, that any delay-sensitive application may never be able to meet the delay requirement beyond such boundary.

Furthermore, from previous section we know the MAC layer service time reduces as the MAC parameters, such as CW_{min} , protocol overhead, and AIFS, improve from 802.11b, to 802.11g, to 802.11a, and then to 802.11e. Therefore, we see the turning point of boundless delay moves toward higher busyness ratio in the same order. On the other hand, we also explore the effects of arrival process on packet delay limits, particularly for the turning point of boundless packet delay. As the packet delay limit of 802.11e appears to be finite in all level of busyness ratio in Figure 5, we plot Figure 6 for the packet delay limit of prioritized 802.11e MAC with decreasing arrival time. We can see a turning point of unbounded delay eventually emerges as the arrival time decreases less than 2.5ms. It is because when arrival time decreases, it poses stricter delay constraints to the queue system, consequently results in packet delay limit curve moving toward less busy environments. Our result indicates the performance bottleneck also exists for the latest QoS-enabled IEEE 802.11e EDCA.

E. Model Validation with TGn Usage Scenarios

We further validate the accuracy of our model with ns-2 simulations of non-saturated TGn scenarios listed in Table II. With the traffic characteristics specified in Table II, each TGn scenario corresponds with a particular P_{busy} value. We collect simulation results of packet delay for G.711 10ms inter-arrival VoIP traffic and compare to the delay derived from the analytical model under the same P_{busy} value. Table III shows a close match between our model and simulations (error < 10%).

V. DISCUSSION

A. Effects of Competing Traffic Packet Data Rates and Payload Sizes

In real-world IEEE 802.11-based wireless network deployments, the operating data rate is not only finite, it is changing dynamically. A wireless node usually degrade the

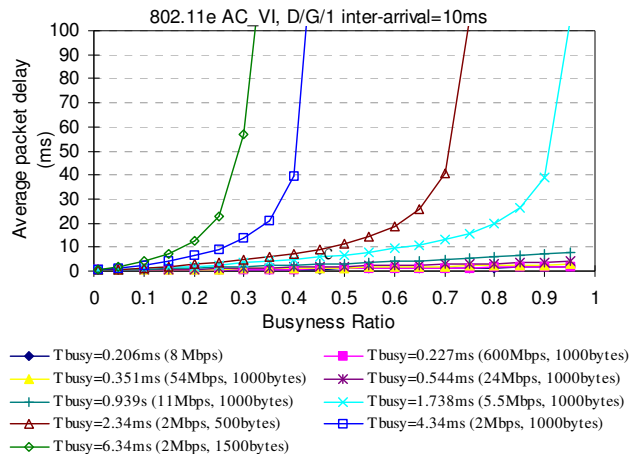


Figure 7. Average packet delay limit with competing traffic operates at different data bit rates and different payload sizes

data bit rate (to incorporate a more resilient modulation scheme) due to increased distance or obstructions such as walls between the access point and the wireless node, or due to repeated unsuccessful frame transmissions. Besides, the payload sizes of the competing traffic also vary from time to time. It is apparent from the equations in section III, that it takes more time for a wireless station to wait on busy slots when other wireless nodes operate at slower data rate or longer payload size. As a result, increased busy slot time increases the service time and delay and results in significant performance degradation.

This problem becomes more important in hybrid IEEE 802.11 networks. As specified in the latest IEEE 802.11n proposal[8], the fastest data rate is expected as high as 600Mbps. Meanwhile the network is backward compatible to legacy IEEE 802.11a/b/g devices, which support data rate as low as 1Mbps. In a network that the operating data rates vary in such a wide range, the resulting packet service time, and consequently network throughput and delay, may also fluctuate greatly. Therefore, it is essential to investigate the performance under the scenarios that nodes operate at different data bit rates and different payload sizes.

Figure 7 plots the average packet delay of prioritized 802.11e AC_VI access category with competing traffic operates at different data bit rates and different payload sizes. We vary the data rate from infinite to 2Mbps, in which cases the average busy slot length increases an order of magnitude. As a result, we can see delay performance varies significantly. For the cases the operating data rate is higher than 11Mbps (up to infinite data rate), the average packet delay maintains below 10ms for all medium busyness ratio. When the operating data rate drops to 5.5Mbps, the turning point for boundless delay emerges. In the case when all other nodes operate at 2Mbps, the average packet delay might have reached to an unacceptable level (> 100ms) with busyness ratio as little as 0.4. On the other hand, when we fix the data rate at 2Mbps and increase the payload sizes of background traffic, we can see the delay turns to be infinitely high at even less busyness ratio ($P_{\text{busy}} \sim 0.3$).

We should keep in mind that data bit rates and payload sizes of competing traffic may change from time to time and are not controlled by any other node in the network. In other words, even the considered node operates at the highest possible data rate and uses the highest priority access category, the performance of the considered node can be constrained by the operating characteristics of other nodes in the network. Such performance limitation is not properly identified and quantified in previous literature.

B. Performance Improvements on Frame Bursting and Block Acknowledgement

From the observations we made in previous subsections, we know that the major contributor to the delay in 802.11 based networks is the delay introduced during backoff stages. It is especially inefficient in terms of channel utilization that such backoff medium contention has to be repeated for every arriving packet. On the other hand, the amendments of IEEE 802.11 standard specify two frame aggregation schemes to improve channel utilizations by aggregating multiple transmissions in one medium contention, namely Frame-Bursting (FB) and Block Acknowledgment (BTA).

Frame-Bursting inserts a burst (say, m number) of DATA frames and corresponding ACK frames back-to-back without initiating another round of random backoff. In addition, Frame-Bursting does not require any explicit signaling between the source and receiver nodes, and hence can be implemented in any IEEE 802.11 based networks. On the other hand, Block ACK mitigates the inefficiency of protocol overhead by placing a burst of DATA frames separated by a SIFS period without being acknowledged. At the end of the burst, the sender initiates an explicit Block ACK Request (BAR) to enquire the number of frames successfully received by the receiver. The receiver then responds with a Block ACK (BA) frame. The number of frames in a BTA burst (say n) is broadcasted by the access point or pre-negotiated between the sender and the receiver. It is obvious that, with the same number of data frames in a burst ($m=n$), BTA transmits fewer frames and thus saves more overhead compared to FB.

Assuming the considered node always has data packets to send, we can express the theoretical throughput of FB and BTA aggregation schemes by slight modifications in Equation 8

$$T_{put_{FB}} = \frac{8 * L_{DATA} * m}{B'(1) + (T_p + T_{PHY} + SIFS) * 2m}. \quad (14)$$

$$T_{put_{BTA}} = \frac{8 * L_{DATA} * n}{B'(1) + (T_p + T_{PHY} + SIFS) * (n + 2)}. \quad (15)$$

Figure 8 shows the theoretical throughput of Frame-bursting and Block ACK schemes with different burst sizes in the presences of non-saturated competing traffic. In particular, we can see that, with the same burst size $m=n=16$, the performance improvements from BTA is greater than that from FB in low busyness environments. The difference in throughput improvements between BTA and FB, however, becomes insignificant as the wireless medium becomes busier. It is because the waiting time in backoff contention becomes the dominant factor that limits the theoretical throughput under such network conditions.

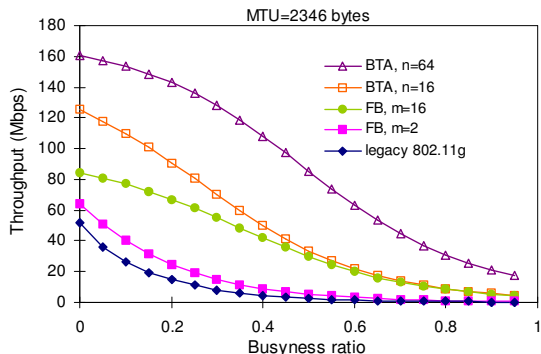


Figure 8. Theoretical throughput of Frame-bursting and Block ACK schemes with different burst sizes

VI. CONCLUSION

In this paper, we investigate the theoretical limits of IEEE 802.11 MAC throughput and delay performance in the presence of non-saturated background traffic. A queuing system based analytical model is proposed to evaluate the throughput and delay bounded by PHY and MAC overhead and backoff waiting time even the operating data rate is infinitely high. We present a detailed analysis for theoretical throughput and delay limits in different IEEE 802.11 specifications. We identify a performance bottleneck beyond which the packet delay becomes infinitely high. Such bottleneck exists for all IEEE 802.11 contention-based DCF and EDCA MAC protocol, although the exact turning point depends on the packet arrival pattern in consideration.

We also show that such theoretical limits are functions of the MAC layer parameters the nodes operate on and the busyness of the wireless medium caused by competing traffic. We study the effects of factors like backoff contention windows, protocol overhead, system slot time, and interframe space time on throughput and delay limits. One of the key observations is that the advanced medium access opportunity enabled by AIFSn in high priority EDCA voice and video access categories is the primary contributor which significantly improves the QoS in terms of maximum achievable throughput and minimum achievable delay. The effects of Frame-bursting and Block ACK frame aggregation schemes on packet performance are also discussed. We plan to extend the analytic model to discuss the packet delay performance of Frame-bursting and Block ACK schemes with arbitrary arrival processes.

REFERENCES

- [1] IEEE Standard for Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, Sep. 1999
- [2] IEEE Standard for Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Higher Data Rate Extension in the 2.4 GHz Band (802.11b), Sep. 1999.
- [3] IEEE Standard for Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, Amendment 4: Further Higher Data Rate Extension in the 2.4 GHz Band (802.11g), June. 2003
- [4] IEEE Standard for Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, Amendment 1: high-speed physical layer in the 5 GHz band (802.11a), Sep. 1999.
- [5] V.K. Jones, R. DeVegt, and J. Terry, "Interest for HDR extension to 802.11a," IEEE 802.11-02-081r0, Jan. 2002.

- [6] M. Tzannes, T. Cooklev, and D. Lee, "Extended Data Rate 802.11a," IEEE 802.11-02-232r0, Mar. 2002.
- [7] S. Hori, Y. Inoue, T. Sakata, and M. Morikura, "System capacity and cell radius comparison with several high data rate WLANs," IEEE 802.11-02-159r1, Mar. 2002.
- [8] HT MAC Specification, Interoperability MAC Specification v1.24, Enhanced Wireless Consortium publication, Jan. 2006.
- [9] IEEE Standard for Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, Amendment 7: Medium Access Control Quality of Service (QoS) Enhancements (802.11e), Jan. 2005.
- [10] Y. Xiao and J. Rosdahl, "Throughput and delay limits of IEEE 802.11," *IEEE Communications Letters*, vol. 6, Aug. 2002, pp. 355 – 357.
- [11] Y. Xiao and J. Rosdahl, "Performance analysis and enhancement for the current and future IEEE 802.11 MAC protocols," *ACM Mobile Computing and Communications Review*, vol. 7, April 2003, pp. 6-19.
- [12] G. Bianchi and I. Tinnirello, "Kalman Filter Estimation of the Number of Competing Terminals in an IEEE 802.11 network," INFOCOM 2003
- [13] H. Bruneel and B. G. Kim, "Discrete-time models for communication systems including ATM," Kluwer Academic Publisher, 1993.
- [14] LD Servi, "D/G/I queue with vacation," *Oper. Res.* (1986).
- [15] L. Kleinrock, *Queueing Systems, Volume II: Computer Applications*, Wiley Interscience, New York, 1976.
- [16] G. Bianchi, "Performance Analysis of the IEEE 802.11 Distributed Coordination Function", *JSAC*, March 2000
- [17] A. P. Stephens et al., 802.11 TGN Functional Requirements, IEEE 802.11-03/802r23, May 2004.
- [18] S. Wang and A. Helmy, "Performance Limits and Analysis of Contention-based IEEE 802.11 MAC", USC Technical Report, unpublished.