A Capacity Analysis Framework for the IEEE 802.11e Contention-Based Infrastructure Basic Service Set

Inanc Inan, *Member, IEEE*, Feyza Keceli, *Member, IEEE*, and Ender Ayanoglu, *Fellow, IEEE*

***Abstract*—We propose a multimedia capacity analysis framework for the Enhanced Distributed Channel Access (EDCA) function of the IEEE 802.11e standard. Our analysis shows that the multimedia capacity of the EDCA function for each Access Category (AC) can accurately be estimated by appropriately weighing the service time predictions of a saturation model over different number of active stations. We propose a simple and generic cycle time model to derive the service time in saturation which we employ in the calculation of an accurate station- and AC-specific queue utilization ratio. Based on the estimated queue utilization ratio, we design a simple model-based admission control scheme. We show that the proposed call admission control algorithm maintains satisfactory user-perceived quality for coexisting voice and video connections in an infrastructure Basic Service Set (BSS) and does not present over- or under-admission problems of previously proposed models in the literature.**

***Index Terms*—Wireless LAN, IEEE 802.11e, Enhanced Distributed Channel Access (EDCA), cycle time, capacity analysis, admission control.**

I. Introduction

The IEEE 802.11 standard [1] defines the Distributed Coordination Function (DCF) which provides best-effort service at the Medium Access Control (MAC) layer of the Wireless Local Area Networks (WLANs). The recently ratified IEEE 802.11e standard [2] specifies the Hybrid Coordination Function (HCF) which enables prioritized and parameterized Quality-of-Service (QoS) services at the MAC layer, on top of DCF. The HCF combines a distributed contention-based channel access mechanism, referred to as Enhanced Distributed Channel Access (EDCA), and a centralized polling-based channel access mechanism, referred to as HCF Controlled Channel Access (HCCA).

The DCF and the EDCA use Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) and slotted Binary Exponential Backoff (BEB) mechanism as the basic access method. The EDCA extends the DCF by defining multiple Access Categories (ACs) with AC-specific Contention Window (CW) sizes, Arbitration Interframe Space (AIFS) values, and Transmit Opportunity (TXOP) limits to support MAC-level QoS [2]. Due to their ease of implementation and satisfactory performance for best-effort data transfer, the distributed contention-based schemes, DCF and EDCA, are widely deployed.

As a direct result of the contention-based nature, DCF and EDCA cannot provide parameterized QoS for real-time applications that require strict QoS guarantees, unless the network load and parameters are tuned such that the network is operating in nonsaturated state [3],[4]. Although the use of an admission control algorithm is recommended in [2] to limit the network load for QoS provisioning, no algorithm is specified. A loose capacity estimation is harmful for admission control, since the quality of ongoing flows will be jeopardized. Conversely, an underestimation of the network capacity results in a fewer number of admitted flows than the network can support.

In this paper, we consider the problem of multimedia capacity estimation and admission control for the EDCA function[1]. Rather than designing a new and complex access model with a large number of states in order to calculate the EDCA capacity (in nonsaturation), we propose a novel, simple, and accurate framework which directly employs the estimations of the comparably simpler saturation analysis[2]. The proposed framework is novel in showing that the saturation figures can effectively be used in model-based network capacity estimation.

In order to assess the performance of EDCA function accu-

Paper approved by R. Fantacci, the Editor for Wireless Networks and Systems of the IEEE Communications Society. Manuscript received March 26, 2007; revised December 31, 2007, July 28, 2008, and March 26, 2009.

I. Inan was with the Department of Electrical Engineering and Computer Science, Center for Pervasive Communications and Computing, Henry Samueli School of Engineering, University of California, Irvine, CA 92697-3975 USA. He is now with Wionics Technologies, Irvine, CA 92618 USA (e-mail: inanc@ieee.org).

F. Keceli was with the Department of Electrical Engineering and Computer Science, Center for Pervasive Communications and Computing, Henry Samueli School of Engineering, University of California, Irvine, CA 92697-3975 USA. She is now with Networks In Motion, Aliso Viejo, CA 92656 USA (e-mail: feyza@ieee.org).

E. Ayanoglu is with the Department of Electrical Engineering and Computer Science, Center for Pervasive Communications and Computing, Henry Samueli School of Engineering, University of California, Irvine, CA 92697-3975 USA (e-mail: ayanoglu@uci.edu).

This work was supported by the Center for Pervasive Communications and Computing, and by National Science Foundation under Grant No. 0434928. Any opinions, findings, and conclusions or recommendations expressed in this material are those of authors and do not necessarily reflect the view of the National Science Foundation.

Portions of the work in this paper have been presented at IEEE Global Communications Conference, November 2007 and at IEEE Global Communications Conference, November 2008.

Digital Object Identifier 10.1109/TCOMM.2009.11.070637

[1] The analysis for DCF is a subset of the proposed generic analysis and is straightforward to derive.

[2] Saturation is the limit reached by the system when each station (or AC) always has a packet to transmit. Conversely, in nonsaturation, the stations (or ACs) experience idle times since they sometimes have no packet to send.

rately for saturation, we propose a simple and generic cycle time model[3]. The proposed cycle time model considers the AIFS and CW differentiation by employing a simple average collision probability analysis and is the first to consider the scenario such that the number of active ACs may vary from station to station. As a direct result, the proposed model also takes the internal collisions into account in the case of a station having more than one active ACs.

According to the proposed multimedia capacity estimation framework, we calculate an approximate station- and AC-specific average service time by weighing the average service time calculated using cycle time model for different number of active stations. Given the average station- and AC-specific traffic load, we translate the average service time into a station- and AC-specific queue utilization ratio[4]. Next, we design a novel centralized EDCA admission control algorithm. The admission decisions of this algorithm are based on the queue utilization ratio. The key point is that the delay guarantee of real-time applications is only possible when the queue utilization ratio of active multimedia flows is smaller than 1 (i.e., when the MAC queue is stable). Comparing the theoretical results with simulations, we show that the proposed call admission control algorithm maintains satisfactory user-perceived quality for coexisting voice and video connections in an infrastructure BSS by limiting the maximum number of admitted flows of each multimedia traffic type. Comparison with extensive simulation results also reveals that the proposed analysis does not result in an overestimation or a significant underestimation of the network capacity. Another feature of the proposed scheme is that it fully complies with the 802.11e standard.

To keep the analysis simple, we assume that no wireless channel errors and coexisting HCCA traffic are present in the design of the saturation cycle time model. In the sequel, we will also discuss the reliability and the possible extensions of the proposed capacity estimation framework when these assumptions are relaxed.

The main contributions of this paper are three-fold; *i)* a simple average cycle time model to evaluate the performance of the EDCA function in saturation for an arbitrary assignment of AC-specific AIFS and CW values and an arbitrary distribution of active ACs at the stations, *ii)* an approximate capacity estimation framework which weighs the saturation service times in order to calculate the nonsaturation service time, and *iii)* a practical model-based admission control algorithm to limit the number of admitted real-time multimedia flows in the 802.11e infrastructure BSS.

II. EDCA Overview

The IEEE 802.11e EDCA is a QoS extension of IEEE 802.11 DCF. The major enhancement to support QoS is that EDCA differentiates packets using different priorities and maps them to specific ACs that are buffered in separate queues at a station. Each AC$_i$ within a station ($0 \leq i \leq i_{max}$, $i_{max} = 3$ in [2]) having its own EDCA parameters contends for the channel independently of the others. Following the convention of [2], the larger the index i is, the higher the priority of the AC is. Levels of services are provided through different assignments of the AC-specific EDCA parameters; AIFS, CW, and TXOP limits.

If there is a packet ready for transmission in the MAC queue of an AC, the EDCA function must sense the channel to be idle for a complete AIFS before it can start the transmission. The AIFS of AC$_i$ is determined by using the MAC Information Base (MIB) parameters as

$$AIFS_i = SIFS + AIFSN_i \times T_{slot}, \qquad (1)$$

where $AIFSN$ is the AC-specific AIFS number, $SIFS$ is the length of the Short Interframe Space and T_{slot} is the duration of a time slot.

If the channel is idle when the first packet arrives at the AC$_i$ queue, the packet can be directly transmitted as soon as the channel is sensed to be idle for AIFS$_i$. Otherwise, a backoff procedure is completed following the completion of AIFS before the transmission of this packet. A uniformly distributed random integer, namely a backoff value, is selected from the range $[0, W_i]$. The backoff counter is decremented at the slot boundary if the previous time slot is idle. Should the channel be sensed busy at any time slot during AIFS or backoff, the backoff procedure is suspended at the current backoff value. The backoff resumes as soon as the channel is sensed to be idle for AIFS again. When the backoff counter reaches zero, the packet is transmitted in the following slot.

The value of W_i depends on the number of retransmissions the current packet experienced. The initial value of W_i is set to $CW_{min,i}$. If the transmitter cannot receive an Acknowledgment (ACK) packet from the receiver in a timeout interval, the transmission is labeled as unsuccessful and the packet is scheduled for retransmission. At each unsuccessful transmission, the value of W_i is doubled until $CW_{max,i}$ is reached. The value of W_i is reset to $CW_{min,i}$ if the transmission is successful, or the packet retransmission limit [1] is reached thus the packet is dropped.

The higher priority ACs are assigned smaller AIFSN. Therefore, the higher priority ACs can either transmit or decrement their backoff counters while lower priority ACs are still waiting in AIFS. This results in higher priority ACs enjoying a relatively faster progress through backoff slots. Moreover, the ACs with higher priority may select backoff values from a comparably smaller CW range. This approach prioritizes the access since a smaller CW value means a smaller backoff delay before the transmission.

Upon gaining access to the medium, each AC$_i$ may carry out multiple frame exchange sequences as long as the total access duration does not go over TXOP$_i$. In a TXOP, the transmissions are separated by SIFS. Multiple frame transmissions in a TXOP can reduce the overhead due to contention. A TXOP limit of zero corresponds to only one frame exchange per access.

An internal (virtual) collision within a station is handled by

[3] The proposed cycle time analysis is based on the fact that a random access system exhibits cyclic behavior. A cycle time is defined as the duration in which an arbitrary tagged user successfully transmits one packet on average [5].

[4] In a typical deployment of an IEEE 802.11e WLAN, i.e., in an infrastructure Basic Service Set (BSS), an Access Point (AP) serves as a gateway between the wired and wireless domains. Since all the measures are station- and AC-specific, the proposed framework considers the potential unbalanced traffic load in the uplink and downlink of the infrastructure 802.11e BSS.

granting the access to the AC with the highest priority. The ACs with lower priority that suffer from a virtual collision act as if an outside collision has occured [2].

III. Related Work

In this section, we provide a brief summary of the studies in the literature that are related to this work.

A. Performance Analysis of EDCA in Saturation

Three major saturation performance models have been proposed for DCF; *i)* assuming constant collision probability for each station, Bianchi [6] developed a simple Discrete-Time Markov Chain (DTMC) and the saturation throughput is obtained by applying regenerative analysis to a generic slot time, *ii)* Cali *et al.* [7] employed renewal theory to analyze a p-persistent variant of DCF with persistence factor p derived from the CW, and *iii)* Tay *et al.* [8] instead used an average value mathematical method to model DCF backoff procedure and to calculate the average number of interruptions that the backoff timer experiences. Having the common assumption of slot homogeneity (for an arbitrary station, constant collision or transmission probability at an arbitrary slot), these models define different renewal cycles all of which lead to accurate saturation performance analysis. These major methods (especially [6]) are modified by several researchers to include an accurate treatment of the QoS features of the EDCA function (AIFS and CW differentiation among ACs) in the saturation analysis [9]–[18].

Our approach in this paper is based on the observation that the transmission behavior in the contention-based 802.11 WLAN follows a pattern of periodic cycles [5]. In this paper, we propose an accurate method of incorporating AIFS and CW differentiation to enable EDCA cycle time analysis. The proposed approach maintains simplicity by employing averaging on the AC- and station-specific collision probability. The proposed cycle time analysis is the generalization of our previously proposed cycle time analysis [19] by considering the possibility of the number of active ACs varying from station to station. The comparison with more complex and detailed theoretical and simulation models reveals that the analytical accuracy is preserved when the proposed cycle time analysis is used.

B. Capacity Analysis and Admission Control in EDCA

The Markov analysis of [6] is also modified by several researchers to include the capacity analysis of the DCF or EDCA function in nonsaturation [20]–[22]. A number of queueing models have also been proposed to analyze delay performance of a station or an AC under the assumption that the traffic is uniformly distributed [3], [4], [23], [24]. Some other queueing models also assumed a MAC queue size of one packet to define a Markovian framework for performance analysis [25], [26].

There are also studies on capacity analysis and admission control considering the infrastructure BSS (where the AP usually has a higher load in the downlink than the stations serving traffic in the uplink). A group of studies mainly concentrated on capacity analysis of only Voice-over-IP (VoIP) traffic for DCF and did not consider traffic differentiation [27]–[32]. Gao *et al.* [33] and Cheng *et al.* [34] calculated VoIP capacity of the WLAN when CW differentiation among uplink and downlink flows are used. A Markov renewal framework for the scenarios where the downlink is always the bottleneck is proposed in [35]. Another group of studies defines parameter adaptation algorithms for QoS enhancement and defines measurement-assisted call admission control algorithms [36]–[42].

Being a very simple extension of the proposed cycle time analysis, our approach in this paper provides an accurate multimedia traffic capacity estimation for the contention-based MAC functions of the 802.11 WLAN. Our design considers the potential unbalanced traffic between the AP and the stations. Under the motivation of previous findings that the optimum operating point of the 802.11 WLAN lies in nonsaturation [4], we define simple tests for centralized admission control of multimedia traffic. Comparison with simulation results for a broad range of traffic types and load shows that the proposed method provides an accurate network capacity estimation and the proposed admission control algorithm prevents both over- and under-admission problems of previously proposed models.

IV. EDCA Cycle Time Analysis

We propose an average cycle time analysis to model the behavior of the EDCA function of any AC at any station in an errorless wireless channel. In this section, we will first define a Traffic Class (TC). Then, we will derive the TC-specific average collision probability. Next, we will calculate the TC-specific average cycle time. Finally, we will relate the average cycle time and the average collision probability to the normalized throughput and service time.

The main assumption for saturation analysis is that each AC always has a frame in service. Note that the performance of EDCA depends on the number of active ACs within the same station as well as the number of active ACs at the other stations due to the fact that the EDCA function acts differently in the case of an internal or an external collision. One of the key differences of our theoretical formulation from the previous work in the literature is as follows. We consider both the possibility of a station running multiple ACs (thus the possibility of internal collisions) and the possibility of the number of active ACs varying from station to station. For example, consider a simple WLAN scenario where an Access Point (AP, labeled STA_0 in the sequel) runs 2 downlink ACs, namely AC$_1$ and AC$_2$. Similarly, assume ν_1 stations $(STA_1, \ldots, STA_{\nu_1},\ \nu_1 > 0)$ only run AC$_1$ and ν_2 other stations $(STA_{\nu_1+1}, \ldots, STA_{\nu_1+\nu_2},\ \nu_2 > 0)$ only have AC$_2$ in the uplink. Although there are 2 distinct ACs active in the system, the downlink AC$_i$ and the uplink AC$_i$ ($i = 1, 2$ for the running example) cannot be expected to have the same performance due to internal and external collision differentiation [2]. In this case, the performance analysis should be carried out individually for 4 different Traffic Classes (TCs) which are uplink AC$_1$, downlink AC$_1$, uplink AC$_2$, and downlink AC$_2$.

We make the following mathematical definitions for the analysis in the sequel.

- Let δ_k $(0 \leq k \leq \nu_{STA})$ be a vector of size 4 which denotes the activity status of ACs at STA_k where ν_{STA} is the total number of stations that have at least one active AC. The value at dimension i of δ_k shows whether AC$_i$ is active or not at STA_k. The entries corresponding to the indices of active (inactive) ACs are labeled with 1 (0). In the example above, $\delta_0 = (0,1,1,0)$, $\delta_1 = (0,1,0,0)$, $\delta_{v_1+1} = (0,0,1,0)$, etc.
- Let ζ be the set of δ_k, i.e., $\zeta = \{\delta_k : 0 \leq k \leq n_{STA}\}$. Above, $\zeta = \{(0,1,0,0),(0,0,1,0),(0,1,1,0)\}$.
- Let ψ_i be the set of δ_k where AC$_i$ is active, i.e., $\psi_i = \{\delta_k : \delta_k(i) = 1, 0 \leq k \leq n\}$. In the example above, $\psi_1 = \{(0,1,0,0),(0,1,1,0)\}$, $\psi_2 = \{(0,0,1,0),(0,1,1,0)\}$, and $\psi_0 = \psi_3 = \{\}$.
- Let $N(S)$ be an operator on a set S which shows the number of elements in the set. Then, the total number of TCs with AC$_i$ active are $N(\psi_i)$ and the total number of TCs is $J = \sum_{i=0}^{i=3} N(\psi_i)$. Note that $N(\zeta) \leq J$ should always hold. In the sequel, each distinct TC is denoted by TC$_j$ $(0 \leq j < J)$. We also define σ_j as the activity status vector of TC$_j$. In the example above, $N(\zeta) = 3$ and $J = 4$. TC$_0$ is the AC$_1$ when only AC$_1$ is active at the station ($\sigma_1 = \psi_1(1)$). TC$_1$ is the AC$_1$ with both AC$_1$ and AC$_2$ are active ($\sigma_2 = \psi_1(2)$). TC$_2$ is the AC$_2$ when only AC$_2$ is active at the station ($\sigma_3 = \psi_2(1)$). TC$_3$ is the AC$_2$ with both AC$_1$ and AC$_2$ are active ($\sigma_4 = \psi_2(2)$).
- Let F be a function with the domain of indices of TCs and the range of indices of ACs. We define this function such as the image of the argument j under function F is the index i of the AC that TC$_j$ uses, i.e., $F(j) = \{i : TC_j \in \psi_i\}$.
- Let G be a function from the domain of the indices of TCs to the range of sets of indices of TCs. We define this function such as the image of the argument j under mapping G is the set of TC indices j' with the same σ, i.e., $G(j) = \{\forall j' : \sigma_j = \sigma_{j'}, 0 \leq j' < J\}$.

A. TC-specific Average Collision Probability

The difference in AIFS of each AC in EDCA creates the so-called *contention zones or periods* as shown in Fig. 1 [11],[12]. In each contention zone, the number of contending TCs may vary. In order to be consistent with the notation of [2], we assume $AIFS_0 \geq AIFS_1 \geq AIFS_2 \geq AIFS_3$. Let $d_j = AIFSN_{F(j)} - AIFSN_3$. Also, let n^{th} backoff slot after the completion of $AIFS_3$ idle interval following a transmission period be in contention zone x. Then, we define $x = \max\Big(F(y) \mid d_y = \max_z(d_z \mid d_z \leq n)\Big)$ which shows contention zone label x is assigned the largest index value within a set of ACs that have the largest AIFSN value which is smaller than or equal to $n + AIFSN_3$.

We define $p_{c_{j,x}}$ $(0 \leq j < J)$ as the conditional probability that TC$_j$ experiences either an external or an internal collision in contention zone x. Note $AIFS_x \geq AIFS_{F(j)}$ should hold for TC$_j$ to transmit in zone x. Following the slot homogeneity assumption of [6], assume that each TC$_j$ transmits with constant probability, τ_j. Also, let the total number TC$_j$ flows be f_j. Then,

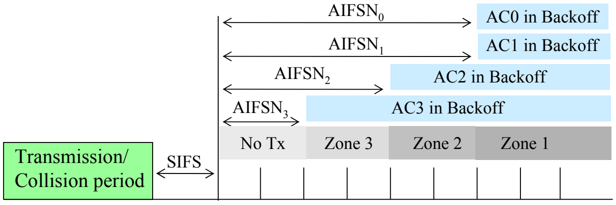


Fig. 1. EDCA backoff after busy medium.

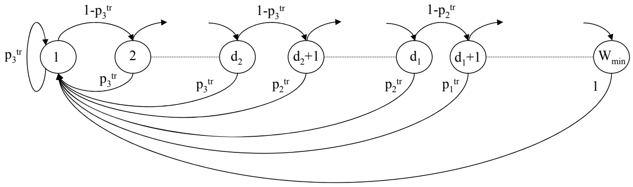


Fig. 2. Transition through backoff slots in different contention zones for the example given in Fig.1.

$$p_{c_{j,x}} = 1 - \frac{\prod\limits_{\forall j' : d_{j'} \leq d_{F^{-1}(x)}} (1 - \tau_{j'})^{f_{j'}}}{\prod\limits_{\forall j' \in G(j) : F(j') \leq F(j)} (1 - \tau_j)}. \tag{2}$$

We use the Markov chain shown in Fig. 2 to find the long term occupancy of the contention zones. Each state represents the n^{th} backoff slot after the completion of the AIFS$_3$ idle interval following a transmission period. The Markov chain model uses the fact that a backoff slot is reached if and only if no transmission occurs in the previous slot. Moreover, the number of states is limited by the maximum idle time between two successive transmissions which is $W_{min} = \min(CW_{F(j),max})$ for a saturated scenario. The probability that at least one transmission occurs in a backoff slot in contention zone x is

$$p_x^{tr} = 1 - \prod_{\forall j' : d_{j'} \leq d_{F^{-1}(x)}} (1 - \tau_{j'})^{f_{j'}}. \tag{3}$$

Note that F^{-1} is not one-to-one. Therefore, we define the image of $F^{-1}(i)$ as a randomly selected TC index j which satisfies $F(j) = i$.

Given the state transition probabilities as in Fig. 2, the long term occupancy of the backoff slots b'_n can be obtained from the steady-state solution of the Markov chain. Then, the TC-specific average collision probability p_{c_j} is found by weighing zone specific collision probabilities $p_{c_{j,x}}$ according to the long term occupancy of contention zones (thus backoff slots)

$$p_{c_j} = \frac{\sum_{n=d_j+1}^{W_{min}} p_{c_{j,x}} b'_n}{\sum_{n=d_j+1}^{W_{min}} b'_n} \tag{4}$$

where x is calculated depending on the value of n as stated previously.

B. TC-Specific Average Cycle Time

Let $E_j[t_{cyc}]$ be average cycle time for a tagged TC$_j$ user. $E_j[t_{cyc}]$ can be calculated as the sum of average duration for *i)* the successful transmissions, $E_j[t_{suc}]$, *ii)* the collisions, $E_j[t_{col}]$, and *iii)* the idle slots, $E_j[t_{idle}]$ in one cycle

$$E_j[t_{cyc}] = E_j[t_{suc}] + E_j[t_{col}] + E_j[t_{idle}]. \tag{5}$$

In order to calculate the average time spent on successful transmissions during a TC$_j$ cycle time, we should find the expected number of total successful transmissions between two successful transmissions of TC$_j$. Let Q_j represent this random variable. Also, let γ_j be the probability that the transmitted packet belongs to an arbitrary user from TC$_j$ given that the transmission is successful. Then,

$$\gamma_j = \frac{\sum_{n=d_j+1}^{W_{min}} b'_n p_{s_{j,n}}/f_j}{\sum_{n=d_j+1}^{W_{min}} (b'_n \sum\limits_{\forall l} p_{s_{l,n}})} \tag{6}$$

where

$$p_{s_{j,n}} = \begin{cases} \dfrac{f_j \tau_j \prod_{j':d_{j'} \le n-1} (1-\tau_{j'})^{f_{j'}}}{\prod_{\forall j' \in G(j):F(j') \le F(j)} (1-\tau_j)}, & \text{if } n \ge d_j + 1 \\ 0, & \text{if } n < d_j + 1. \end{cases} \tag{7}$$

Then, the Probability Mass Function (PMF) of Q_j is

$$\Pr(Q_j = k) = \gamma_j (1-\gamma_j)^k, \quad k \ge 0. \tag{8}$$

We can calculate the expected number of successful transmissions of any TC$_{j'}$ during the cycle time of TC$_j$, $ST_{j',j}$, as

$$ST_{j',j} = f_{j'} E[Q_j] \frac{\gamma_{j'}}{1-\gamma_j}. \tag{9}$$

Inserting $E[Q_j] = (1-\gamma_j)/\gamma_j$ in (9), the intuition that each user from TC$_j$ can transmit successfully once on average during the cycle time of another TC$_j$ user, i.e., $ST_{j,j} = f_j$, is confirmed. Including the own successful packet transmission time of tagged TC$_j$ user in $E_j[t_{suc}]$, we find

$$E_j[t_{suc}] = \sum_{\forall j'} ST_{j',j} T_{s_{j'}} \tag{10}$$

where $T_{s_{j'}}$ is defined as the time required for a successful packet exchange sequence (will be derived in (18)).

To obtain $E_j[t_{col}]$, we need to calculate the average number of users who are involved in a collision, f_{c_n}, at the n^{th} slot after last busy time for given f_j and τ_j, $\forall j$. Let the total number of users transmitting at the n^{th} slot after last busy time be denoted as Y_n. We see that Y_n is the sum of random variables, $Binomial(f_j, \tau_j)$, $\forall j: \ d_j \le n-1$. Employing simple probability theory, we can calculate $f_{c_n} = E[Y_n | Y_n \ge 2]$. After some algebra and simplification,

$$f_{c_n} = \frac{\sum\limits_{j:d_j \le n-1} (f_j \tau_j - p_{s_{j,n}})}{1 - \prod\limits_{j:d_j \le n-1} (1-\tau_j)^{f_j} - \sum\limits_{j:d_j \le n-1} p_{s_{j,n}}}. \tag{11}$$

If we let the average number of users involved in a collision at an arbitrary backoff slot be f_c, then

$$f_c = \sum_{\forall n} b'_n f_{c_n}. \tag{12}$$

We can also calculate the expected number of collisions that an TC$_{j'}$ user experiences during the cycle time of a TC$_j$, $CT_{j',j}$, as

$$CT_{j',j} = \frac{p_{c_{j'}}}{1-p_{c_{j'}}} ST_{j',j}. \tag{13}$$

Then, defining $T_{c_{j'}}$ as the time wasted in a collision period (will be derived in (19)),

$$E_j[t_{col}] = \frac{1}{f_c} \sum_{\forall j'} CT_{j',j} T_{c_{j'}}. \tag{14}$$

Given p_{c_j}, we can calculate the expected number of backoff slots $E_j[t_{bo}]$ that TC$_j$ waits before attempting a transmission. Let $W_{i,k}$ be the CW size of AC$_i$ at backoff stage k [14]. Note that, when the retry limit r is reached, any packet is discarded. Therefore, another $E_j[t_{bo}]$ passes between two transmissions with probability $p_{c_j}^r$ (where $i = F(j)$).

$$E_j[t_{bo}] = \frac{1}{1-p_{c_j}^r} \sum_{k=1}^{r} p_{c_j}^{k-1} (1-p_{c_j}) \frac{W_{i,k}}{2}. \tag{15}$$

Noticing that between two successful transmissions, AC$_j$ also experiences $CT_{j,j}$ collisions,

$$E_j[t_{idle}] = E_j[t_{bo}](CT_{j,j}/f_j + 1) t_{slot}. \tag{16}$$

The transmission probability of a user using TC$_j$ is

$$\tau_j = \frac{1}{E_j[t_{bo}] + 1}. \tag{17}$$

Note that, in [12], it is proven that the mean value analysis for the average transmission probability calculated as in (17) matches the Markov analysis of [6].

The equations (2)-(4), (15), and (17) are a set of nonlinear equations which can be solved numerically for τ_j and p_{c_j}, $\forall j$. Then, the average cycle time for AC$_j$, $\forall j$, can be calculated using (5) where each term in (5) is obtained via (6)-(16).

C. Performance Analysis

Let T_{p_j} be the average payload transmission time for TC$_j$ (T_{p_j} includes the transmission time of MAC and PHY headers), δ be the propagation delay, T_{ack} be the time required for acknowledgment packet (ACK) transmission. Then, for the basic access scheme, we define the time spent in a successful transmission T_{s_j} and a collision T_{c_j} for any TC$_j$ as

$$T_{s_j} = T_{p_j} + \delta + SIFS + T_{ack} + \delta + AIFS_{F(j)} \tag{18}$$
$$T_{c_j} = T_{p_j^*} + ACK\_Timeout + AIFS_{F(j)} \tag{19}$$

where $T_{p_j^*}$ is the average transmission time of the longest packet payload involved in a collision [6]. For simplicity, we assume the packet size to be equal for any TC, then $T_{p_j^*} = T_{p_j}$. Being not explicitly specified in the standards, we set $ACK\_Timeout$, using Extended Inter Frame Space (EIFS) as $EIFS_i - AIFS_i$ ($i = F(j)$). Note that the extensions of (18) and (19) for the RTS/CTS scheme are straightforward [6].

The average cycle time of a TC represents the renewal cycle for each TC. Then, the normalized throughput of TC$_j$ is defined as the successfully transmitted information per renewal

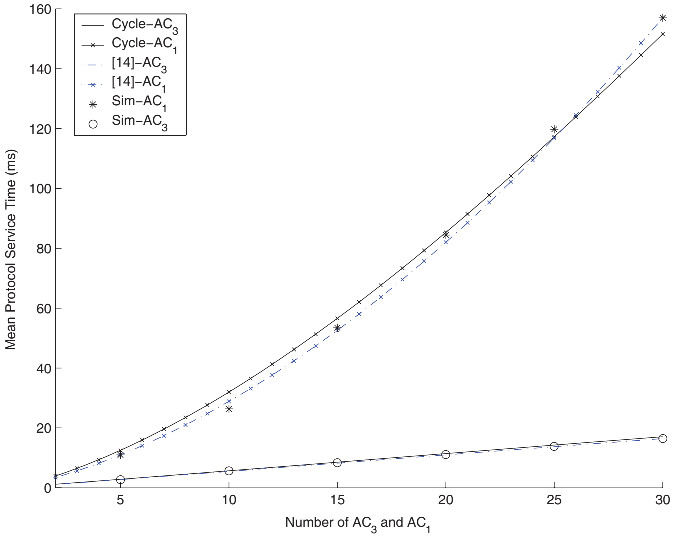


Fig. 3. Analyzed and simulated mean protocol service time of each AC when both N_1 and N_3 are varied from 5 to 30 and equal to each other for the proposed cycle time analysis and the model in [14].

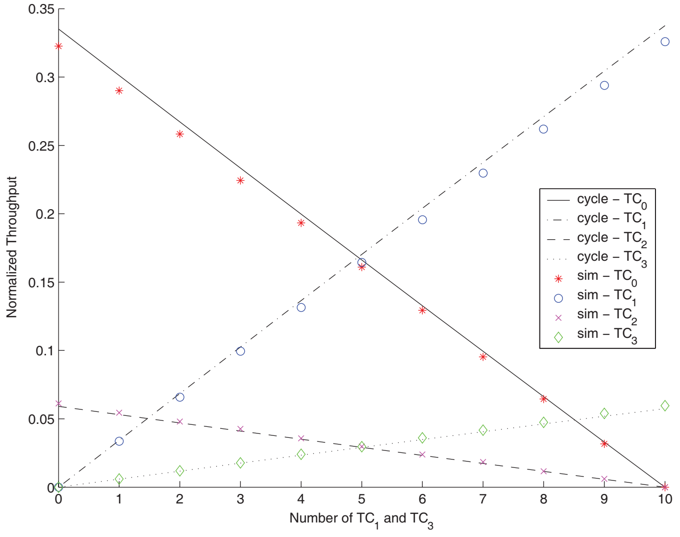


Fig. 4. Analytically calculated and simulated performance of each TC when the number of TC$_1$ and TC$_3$ is varied from 0 to 10 (therefore, TC$_0$ and TC$_2$ vary from 10 to 0).

cycle

$$S_j = \frac{f_j T_{p_j}}{E_j[t_{cyc}]}. \tag{20}$$

The TC-specific cycle time is directly related but not equal to the mean protocol service time. By definition, the cycle time is the duration between successful transmissions. We define the average protocol service time such that it also considers the service time of packets which are dropped due to retry limit. Let $p_{j,drop} = p^r_{c_j}$ be the average packet drop probability. Then, the mean service time $E_j[t_{srv}]$ can be calculated as

$$E_j[t_{srv}] = (1 - p_{j,drop}) E_j[t_{cyc}]. \tag{21}$$

D. Validation

We validate the accuracy of the numerical results by comparing them to the simulation results obtained from ns-2 [43], [44].

In simulations, we consider two ACs, one high priority (AC$_3$) and one low priority (AC$_1$). Unless otherwise stated, each station runs only one AC. For both ACs, the payload size is 1000 bytes. RTS/CTS handshake is turned on. All the stations have 802.11g Physical Layer (PHY) using 54 Mbps and 6 Mbps as the data and basic rate respectively ($T_{slot} = 9\ \mu s$, $SIFS = 10\ \mu s$). We set $AIFSN_1 = 3$, $AIFSN_3 = 2$, $CW_{1,min} = 31$, $CW_{3,min} = 15$, $CW_{1,max} = 255$, $CW_{3,max} = 127$, $r = 7$.

Fig. 3 shows the mean protocol service time of each AC when both ν_1 and ν_3 are varied from 5 to 30 and equal to each other. As the comparison with a more detailed analytical model [14] and the simulation results reveal, the cycle time analysis can predict saturation throughput accurately. Although not included in the figures, a similar discussion holds for the comparison with other detailed and/or complex models of [15]-[17].

In another set of experiments, we test the performance of the system when the stations run multiple ACs so that virtual collisions may occur. The stations run only AC$_1$, only AC$_3$, or both. Similarly to Section IV, we define TC$_0$ as the AC$_3$ when only AC$_3$ is active at the station, TC$_1$ as the AC$_3$ when both AC$_3$ and AC$_1$ are active at the station, TC$_2$ as the AC$_1$ when only AC$_1$ is active at the station, and TC$_3$ as the AC$_1$ when both AC$_3$ and AC$_1$ are active at the station. We keep both ν_1 and ν_3 at 10, and vary the number of TC$_1$ and TC$_3$ from 0 to 10 (therefore, TC$_0$ and TC$_2$ vary from 10 to 0). Fig. 4 shows the normalized throughput of each TC. The predictions of the proposed analytical model follow the simulation results closely. Although not significant for the tested scenario and not apparent in the graphical results, a closer look on the numerical results present the (slightly) higher level of differentiation between AC$_3$ and AC$_1$ which is due to the additional prioritization introduced at the virtual collision procedure.

Due to space limitations, the interested reader is referred to [45] for more analytical and simulation results on the validation of proposed cycle time model.

E. HCCA Coexistence

As previously stated, the HCCA function defines a centrally-controlled polling-based medium access scheme for IEEE 802.11e WLANs. In HCCA, the AP has the highest priority to access the medium among all stations, since it may seize the channel by using a shorter interframe space, namely PIFS, without waiting any backoff time. The AP may either send a poll frame presenting a contention free TXOP to a station, or just start transmitting downlink traffic. Such periods are called Controlled Access Periods (CAPs) where the AP is the owner of the channel and schedule the medium access centrally.

The proposed cycle time approach can also be extended in modeling coexisting EDCA and HCCA traffic. The CAP initiation probability of the AP at an arbitrary slot can be derived from the HCCA traffic parameters. Then, the HCCA traffic can be considered as a separate TC with the transmission probability equal to CAP initiation probability and the successful transmission time equal to the average CAP duration. Considering the HCCA scheme needs more detailed investigation, but by design, the cycle time analysis provides a promising framework for enabling such analysis.

The EDCA runs in the Contention Period (CP). The maximum limit on the CAP duration, namely $dot11CAPLimit$, is defined in [2]. As an alternative rough approximation (which mainly neglects the poll packet collisions at the start of a CAP and assumes that all the CAPs last $dot11CAPLimit$), the TC-specific EDCA cycle time calculated via the proposed model can be normalized considering $dot11CAPLimit$ in order to account for HCCA traffic in EDCA performance analysis.

In this paper, extensions of the cycle time model considering coexisting HCCA traffic, similar to what is described above, are deemed out of scope, and not treated further.

V. Multimedia Capacity Analysis for 802.11e Infrastructure BSS

When working in the saturated case, the contention-based 802.11 MAC suffers from a large collision probability, which leads to low channel utilization and excessively long delay. As shown in [4], the optimal operating point for the 802.11 to work lies in nonsaturation where contention-based 802.11 MAC can achieve maximum throughput and small delay. In [46], it is also shown that a very small increase in system load yields a huge increase (of about two orders of magnitude) of the backoff delay. When the traffic load does not exceed the service rate at saturation, the resulting medium access delay is very small.

In this section, we propose a novel framework where we calculate TC-specific average frame service rate μ via a weighted summation of saturation service rate $E[t_{srv}]$ over varying number of active stations. Defining a TC-specific average queue utilization ratio ρ, we design a simple call admission control algorithm which limits the number of admitted real-time multimedia flows in the 802.11e infrastructure BSS in order to prevent the corresponding TC queues going into saturation. As specified in [2], the admission control is conducted at the AP. Admitted real-time multimedia flows can be served with QoS guarantees, since low transmission delay and packet loss rate can be maintained when the 802.11e WLAN is in nonsaturation [4],[46]. Comparing with simulation results, we show that not only does the proposed admission control algorithm prevent the so-called over-admission or under-admission problems but also efficiently utilizes the network capacity.

A. TC-specific Average Queue Utilization Ratio

Each station runs a QoS reservation procedure with the AP for all of its traffic streams that need parameterized (guaranteed) QoS support. The Station Management Entity (SME) at the AP decides whether the Traffic Stream (TS) is admitted or not regarding the Traffic Specification (TSPEC) in the Add Traffic Stream (ADDTS) request provided by the station. The TSPEC specifies the Traffic Stream Identification Number ($TSID$), the user priority (UP), the mean data rate (R), and the mean packet size (L) of the corresponding TS [2].

Let average frame service rate for TC$_j$ be denoted as μ_j. Also let the average packet arrival rate for TC$_j$ be denoted as λ_j which can easily be calculated employing R of TSs using the same TC at the same station. For simplicity, in the sequel, we assume that TCs at different stations are running TSs with equal TSPEC values (so all traffic parameters remain TC-specific). Though all work in this section can be generalized for varying traffic load and parameters within a TC vary at different stations, we opted not to present this out-of-scope generalization since it would make the model difficult to understand.

We define TC-specific queue utilization ratio ρ as follows

$$\rho_j = \lambda_j/\mu_j, \ \forall j. \tag{22}$$

B. TC-specific Average Frame Service Rate

The TC-specific average queue utilization ratio ρ_j shows the percentage of time on average that TC$_j$ has a frame in service. In other words, ρ_j is the probability that TC$_j$ is active. Our novel approach in calculating μ_j is forming a weighted summation of $E_j[t_{srv}]$ for varying number of active TCs.

Let $P^j_{TC_0,TC_1,...,TC_j,...,TC_{J-1}}(f'_0, f'_1, ..., f'_j, ..., f'_{J-1})$ denote the joint conditional probability that f'_j stations using TC$_j$, $\forall j$, are active given that one TC$_j$ has a frame in service and the total number of TCs is J. Also, let $E_j[t_{srv}(f'_0, f'_1, ..., f'_{J-1})]$ denote the average service time when f'_j stations using TC$_j$, $\forall j$, are active. We use the proposed cycle time model in Section IV to calculate $E_j[t_{srv}(f'_0, f'_1, ..., f'_{J-1})]$[5]. Then, the TC-specific average frame service rate μ_j is calculated as in (23).

Note that the case when $\sum_{j'=0}^{J-1} f'_{j'} = 1$, i.e., there is only one active TC, is not considered by the proposed cycle time model. On the other hand, the cycle time calculation in this case is straightforward. Since no collisions can occur and no other station is active, the successful transmission is performed at AIFS completion. Therefore, $E_j[t_{srv}(f'_0, f'_1, ..., f'_{J-1})] = T_{s_j}$ if $\sum_{j'=0}^{J-1} f'_{j'} = 1$.

We noticed that the distribution of the number of active TCs approximates the sum of independent Binomial distributions with parameters $f_{j'}$ and $\rho_{j'}$, $\forall j'$ as in (24) for the traffic models we used in this study. We confirm the validity of (24) via comparing the analytical estimations with simulation results in Section V-D. On the other hand, we do not argue that the binomial activity distribution holds for any type of traffic model in any scenario. Our observation is that for widely used voice and video traffic models this approximation works well. The proposed framework is generic in the sense that any other activity distribution profile may be used to incorporate other traffic models in other network scenarios.

The fixed-point equations (22)-(24) can numerically be solved for ρ_j and μ_j, $\forall j$.

C. Admission Control Procedure

Upon receiving the ADDTS request, the AP associates the TS with the AC and the TC using the value in the UP field

[5] The proposed capacity estimation framework is generic. Any other accurate saturation analysis method can also be employed for calculating the service time.

$$\frac{1}{\mu_j} = \sum_{0 \le f'_0 \le f_0} \cdots \sum_{1 \le f'_j \le f_j} \cdots \sum_{0 \le f'_{J-1} \le f_{J-1}} E_j[t_{srv}(f'_0, ..., f'_j, ..., f'_{J-1})] \cdot P^j_{TC_0,...,TC_j,...,TC_{J-1}}(f'_0, ..., f'_j, ..., f'_{J-1}) \tag{23}$$

$$P^j_{TC_0,TC_1,...,TC_j,...,TC_{J-1}}(f'_0, f'_1, ..., f'_j, ..., f'_{J-1}) = \binom{f_j - 1}{f'_j - 1} \rho_j^{f'_j - 1} (1 - \rho_j)^{f_j - f'_j} \prod_{\forall j' : j' \neq j} \binom{f_{j'}}{f'_{j'}} \rho_{j'}^{f'_{j'}} (1 - \rho_{j'})^{f_{j'} - f'_{j'}} \tag{24}$$

$$\frac{1}{\mu_j} = \sum_{0 \le f'_1 \le f_1} \cdots \sum_{1 \le f'_j \le f_j} \cdots \sum_{0 \le f'_{J'} \le f_{J'}} E_j[t_{srv}(f'_1, ..., f'_j, ..., f'_{J'-1}, f_{J'}, ..., f_{J-1})] \cdot P^j_{TC_1,...,TC_{J'}}(f'_1, ..., f'_{J'}) \tag{25}$$

and the station MAC address. The traffic stream is admitted if and only if the following tests succeed

$$\rho_j \le \rho_{th}, \ \forall j \tag{26}$$

where $\rho_{th} \le 1$. The tests in (26) ensure that the average traffic arrival rate to all TCs is smaller than the average service rate that can be provided to them. Therefore, the MAC queues of all TCs can be considered to be stable (all TCs remain in nonsaturation on average).

When a real-time flow ends, the source node transmits a Delete Traffic Stream (DELTS) request for the TS [2]. The AP deletes the corresponding entry from the list of admitted flows.

A few remarks on admission control and capacity analysis are as follows.

- The proposed capacity analysis and admission control scheme can easily be extended to the case where some TCs are running best-effort traffic. We actually do a worst-case analysis in Section V-D where the TCs that run best-effort traffic are assumed to be always active. This generalizes (23) as in (25) where J' and $J - J'$ are the number of TCs that run multimedia and best-effort flows respectively. In this case, the admission control tests in (26) are done for TCs that run real-time flows, i.e., $0 \le j \le J'$.
- Although the employed saturation model does not consider wireless channel errors, the admission control scheme can still be effective in an error-prone wireless channel as the admission control decisions are threshold-based. Selecting a comparably smaller $\rho_{th} < 1$ can provide the necessary room for packet retransmissions occuring as a result of wireless channel losses. This may be a simpler approach when compared to a solution that includes the design of a more complex saturation analysis model considering wireless channel errors. The investigation is left as future work.
- The proposed capacity estimation scheme is solely based on mean values and does not consider the worst-case scenario where all the admitted Variable Bit Rate (VBR) multimedia traffic may instantaneously transmit at their peak rate (R_{peak}). Again a wise decision ρ_{th} can limit the channel utilization by multimedia flows thereby leaving room to accommodate bandwidth fluctuations caused by VBR traffic. Alternatively, R_{peak} may be used in the calculation of λ in (22). On the other hand, when R_{peak}/R is very large, this may result in the rejection of many multimedia flows and unnecessarily low channel utilization.
- The TSPECs may also specify a Delay Bound (DB) which denotes the maximum time allowed to transport the frames across the wireless interface including the queueing delay [2]. As also provided in [4, Table I], multimedia services should satisfy QoS requirements in terms of one-way transmission delay, delay variation, and packet loss rate. For example, for voice and video the excellent (acceptable) quality is satisfied if the delay is smaller than 150 ms (400 ms) and the packet loss rate is smaller than $1\% - 3\%$ [4]. Note that packet loss rate includes the dropped packets at the playout buffer of the receiver when the packets are not received within the delay bound. Our capacity analysis does not explicitly consider these metrics in admission control. On the other hand, the proposed call admission control algorithm makes the multimedia TC queues remain stable (TC queues do not go into saturation) by limiting the number of admitted real-time flows. This provides low transmission delays and packet loss ratio due to the low collision probability in nonsaturation [4].
- In the simulations, we observed that the delay experienced by multimedia flows in nonsaturation can go up to 40-50 ms depending on the scenario. In order to guarantee a stochastic delay bound, the admission control tests in (26) should be extended. We may use the method proposed in [34]

  $$\Pr(Q_j > DB_j \cdot \mu_j) \le \epsilon \tag{27}$$

  where Q_j is the queue length of the TC$_j$ and ϵ is the delay violation probability. This test can be extended further for on/off traffic sources and statistical multiplexing at the AP as shown in [34]. On the other hand, in the simulation scenarios we have studied, the addition of this test does not limit the already admitted traffic using (26) since the QoS requirements of the multimedia flows are always satisfied if the system is in nonsaturation state.
- In the simulations, we consider two types of traffic sources; voice and video, where the average packet size of different traffic sources vary. Therefore, T_c is not equal for any TC since $T_{p^*_j} = T_{p_j}$ does not always hold. Due to space limitations, we do not include the calculation of T_{c_j} in this case. We use the method in [6] which has an extensive treatment of the subject.

D. Validation

For the experiments, we use a network topology such that any connection is initiated between a distinct party in the Internet and the WLAN. The traffic is relayed at the AP from (to) the wireless channel to (from) the wired link. The simulations consider three types of traffic sources; voice, video, and background data. The voice traffic models G.711 or G.729 VoIP application as Constant Bit Rate (CBR) traffic (without the use of silence suppression scheme)[6]. The parameters of the VoIP codecs are set as in [32, Table I]. For the video source models, we use the traces of real MPEG-4 video streams [47]. For the particular video source used in the simulations presented in this paper, the average codec bit rate is 174 kbps with an average packet size of 821 bytes. Real-time packets have 40-byte length RTP/UDP/IP header. The background data traffic is modeled by bulk data transfer where every AC using this type of traffic is saturated. Voice flows use AC$_3$, video flows use AC$_2$, and background traffic uses AC$_1$. We set the EDCA parameters as suggested in [2]; $AIFSN_1 = 3$, $AIFSN_2 = 2$, $AIFSN_3 = 2$, $CW_{1,min} = 31$, $CW_{2,min} = 15$, $CW_{3,min} = 7$, $CW_{1,max} = 1023$, $CW_{2,max} = 31$, $CW_{3,max} = 15$, $r = 7$. PHY parameters are set as stated in Section IV-D. The wired link delay is set to 20 ms for all connections.

Besides comparing the performance of the proposed scheme with the simulations, we also study the performance of the following state-of-the-art methodologies in the same scenarios:

- Nonsaturation analysis in the Markov framework: As stated previously, the Markov analysis of [6] is extended to include the capacity analysis of the DCF or EDCA in nonsaturation under the assumption of state independent packet arrival distribution and collision probability. In the comparisons, we use the model in [22] to represent such approaches. Note that the discussion in [22] presents some shortcomings of related studies such as in [20], [21]. We use throughput and average delay estimations obtained from the model to decide on the number of admitted flows.
- Unbalanced uplink/downlink load analysis: Among the works considering the fact that in the downlink the AP has a higher load than the stations, we use the model in [32]. Note that the over-admission problems of the models employing the similar idea [29]–[31] are already shown in [32]. The works in [33], [34] extend [32] when CW differentiation among uplink and downlink flows is employed (to increase VoIP capacity). The investigation of this condition is left as future work.
- Measurement-assisted admission control: The key idea of such approaches is a central unit or each station taking some traffic load/performance measurements on the run and employing the results in the contention parameter adaptation and admission control. To represent such schemes, we use the model in [39]. Note that the specific performance of the model in [39] cannot be a generalization of measurement-assisted methodology performance, as such approaches are heuristic, but the analysis can highlight the advantages and the disadvantages of such methods.

1) Voice Capacity Analysis: In the first set of experiments, we investigate the VoIP capacity of 802.11e WLAN when no other type of traffic coexists. A two-way voice connection is established every ω ms, with the starting time randomly chosen over $[0, \omega]$ ms. We set ω equal to the packet interval duration of the voice codec used. Table I tabulates the maximum number of admitted VoIP connections for different codecs and models. In the simulations, the maximum number of voice connections is obtained in such a way that one more connection results in a packet loss ratio[7] larger than 1%.

As shown in Table I, the analytical results for the proposed model and the simulation results closely follow each other. The model in [22] has over-admission problems (especially more pronounced for lower frame rate VoIP codecs). Such Markov models are known to have sensitive predictions when the number of stations per traffic category[8] does not exceed a threshold [6], [22]. In this specific VoIP scenario, the AP is the only member of one specific traffic category since it has more traffic load than all other stations with equal load. The model in [32] has significant under-admission problems for an arbitrary selection of MAC parameters (especially when the CW settings are small and the underlying PHY is 802.11g)[9]. The measurement-assisted approach in [39] can achieve accurate admission control if the parameters are carefully adjusted. Note that such a heuristic measurement-based algorithm is pretty dependent on the algorithm parameters. Specifically for [39], the performance considerably varies depending on the values of the $SurplusFactor[i]$, $ATL[i]$, the averaging parameter f, etc. In simulations, we have observed that one specific setting is not optimal for any scenario and settings in the presented results might vary from case to case (in Table I, we present the best case results for [39] we obtained by adjusting parameters). The dynamic control of such parameters is a challenging task and out-of-the-scope of this paper.

Fig. 5 shows the packet loss ratio and the average delay of successfully delivered packets[10] for increasing number of active G.711 VoIP connections and codec packet sample interval. These results are obtained via simulation. As the comparison of the results in Table I and Fig. 5 denotes, there is a sudden increase in the downlink packet loss ratio and the average downlink packet delay mainly due to the increasing queueing delay when the queue utilization ratio

[6] The CBR traffic model is used for two reasons; *i)* it provides a worst-case upper bound for the case when the traffic presents on-off traffic characteristics (silence suppression) and *ii)* this enables comparison of voice capacity results with the models proposed in [27]–[33].

[7] A packet drop occurs at the source if the packet cannot be delivered successfully in the maximum limit of retries, r, or there is no available room for the packet in the MAC buffer, and at the sink if the end-to-end delay for the delivered packet exceeds 150 ms [32].

[8] In the context of the Markov framework models, the EDCA function of a traffic category is modeled by a distinct Markov chain.

[9] Our implementation of the model in [32] duplicates the analytical results in [32] which are for a specific DCF MAC parameter set. Although the results are not provided in this paper, the analytical results for the proposed model and our simulation results also confirm the capacity prediction of [32] for the specific DCF scenario.

[10] The presented average delay results are only for the wireless link and exclude the wired link delay. The delay for packets that are not delivered within an end-to-end delay (sum of wireless and wired link delays) of 150 ms are not included in the average delay calculation.

TABLE I
COMPARISON OF THE MAXIMUM NUMBER OF VOIP CONNECTIONS

Sample Period	G.711				G.729			
	Proposed/Simulation	[22]	[32]	[39]	Proposed/Simulation	[22]	[32]	[39]
10 ms	27/27	29	21	26	29/29	30	22	27
20 ms	49/49	52	38	46	56/56	59	43	53
30 ms	70/70	74	53	68	85/85	88	65	82
40 ms	87/87	92	67	84	112/112	117	85	110
50 ms	102/102	108	79	99	139/139	145	106	136
60 ms	115/115	121	89	111	166/166	173	128	162

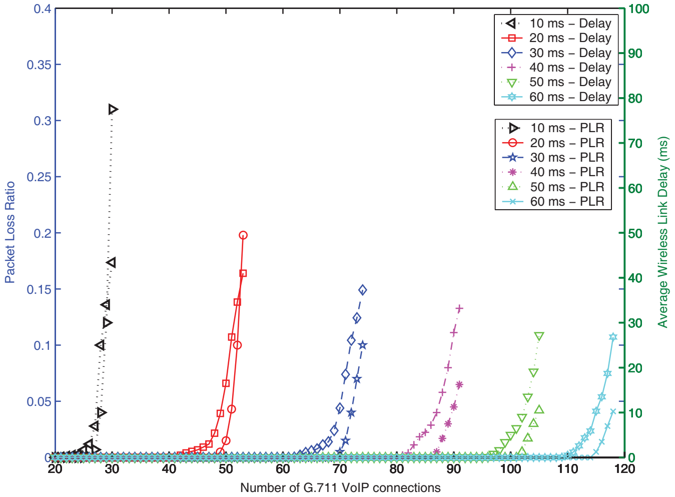


Fig. 5. Packet loss ratio and average delay in the downlink for increasing number G.711 VoIP connections.

exceeds the threshold, ρ_{th}=1[11]. When the load does not exceed the capacity, the packet loss ratio stays smaller than 1% and the average wireless link delay is around 10 ms. In the experiments, the downlink always suffers longer queueing delays and is the main limitation on VoIP capacity. The uplink experiences comparably much smaller packet delays and many fewer packet losses. We do not include uplink results in Fig. 5 in order not to crowd the figure. Although the corresponding results are not presented, a similar discussion holds when VoIP flows employ the G.729 codec.

Fig. 6 shows the Probability Density Function (PDF) of active number of TCs given that the TC at the AP or at the non-AP station (denoted as STA in the figure) is active in a scenario consisting of only VoIP connections (G.711 VoIP codec with 10 ms packet intervals). As previously stated, we analytically calculate $P^{j}_{TC_0,TC_1,...,TC_j,...,TC_{J-1}}(f'_0, f'_1, ..., f'_j, ..., f'_{J-1})$ by assuming that the distribution of the number of active TCs approximates a Binomial distribution with parameters f_j and ρ_j. The comparison in Fig. 6 shows that the PDF of analytical calculation closely follows the PDF obtained through simulation. Although the results are not presented here, a similar discussion holds for other codecs with different packet interval values. The PDF results for simulations are obtained through averaging over several simulation runs with different random number generator seeds and randomized flow start times.

2) Voice Capacity Analysis in the Presence of Background Traffic: In the second set of experiments, we investigate the VoIP capacity when heavy background traffic coexists. Table II shows the number of admitted G.711 VoIP flows for increasing the number of two-way background data connections. The comparison of analytical and simulation results shows that the proposed admission control scheme is highly accurate when a number of TCs (background) are always assumed active while some others (VoIP) are in nonsaturation. As the comparison of Tables I and II presents, the coexistence of background traffic is a big hit on the multimedia capacity of the WLAN. When the number of data connections is 5, the number of admitted flows decreases by around 30%. The decrease ratio goes up to 60% when the number of data connections is increased to 30. Interestingly, the decrease ratio is almost insensitive to packet sampling interval length. Although the results are not presented, a similar discussion holds when VoIP flows employ the G.729 codec.

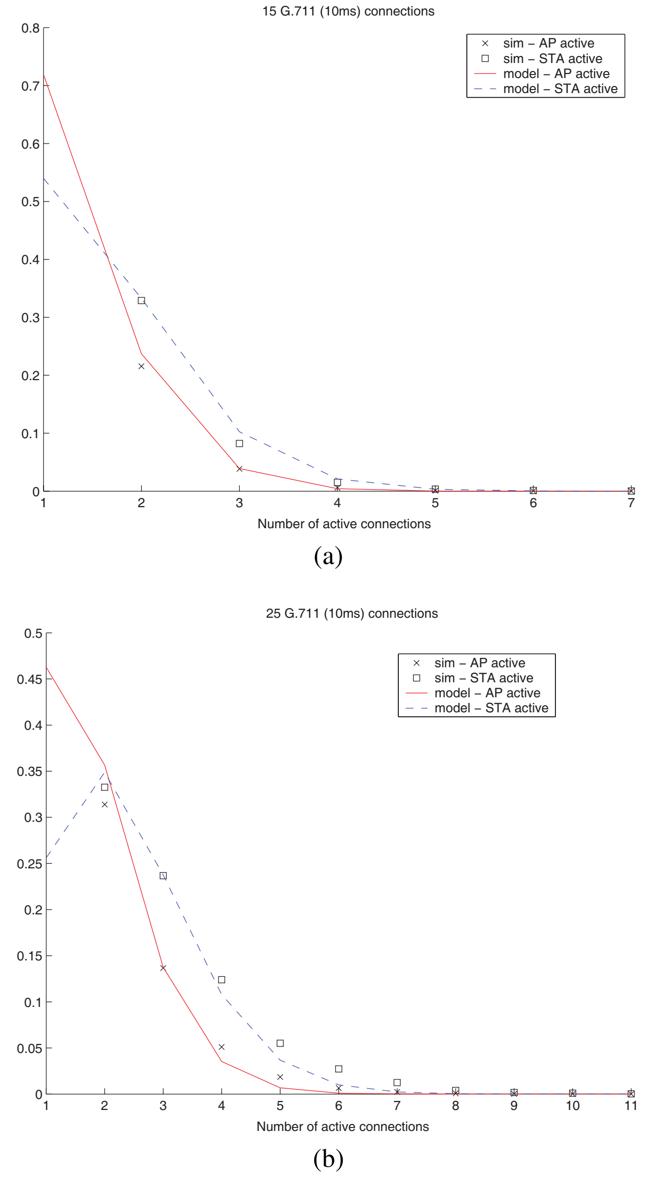
The model in [22] has over-admission problems mainly due to the previously stated reason. We should also note that the number of states/chains in such models may increase with increasing number of traffic categories which may increase the numerical solution complexity significantly. The unbalanced uplink/downlink load analysis of the literature [27]–[32] does not provide such analysis capability for coexisting different types of traffic, therefore the comparison with [32] is not feasible. The measurement assisted approach in [39] employs the original idea of parameter adaptation for best-effort traffic to limit best-effort bandwidth. As parameter adaptation might change the channel capacity and this would lead to an unfair comparison, we do not provide the results of [39] in this specific scenario. On the other hand, the performance dependence of the model in [39] in heuristic parameters on a case by case basis is valid.

3) Voice and Video Capacity Analysis: In the third set of experiments, we investigate the capacity of 802.11e WLAN when both voice and video traffic coexist (using different ACs). Table III shows the number of admitted uplink, downlink, and two-way MPEG-4 flows for increasing the number of VoIP connections. In this scenario, we use the G.711 codec with a 20 ms sample interval. As the results indicate, the analytical and simulation results closely follow each

[11] In the simulations, the MAC buffer size for each node is set to 100 packets. The packet loss ratio and the average delay for successfully delivered packets depend on the buffer size. When the buffer size is smaller, the packet loss ratio is larger and the average delay for successfully delivered packets is smaller. As we have confirmed via simulations (specifically, when the buffer size is 20 packets), the capacity in terms of number of flows stays the same.

TABLE II
COMPARISON OF THE MAXIMUM NUMBER OF G.711 VOIP CONNECTIONS (PROPOSED/SIMULATION/[22]) WHEN HEAVY BACKGROUND TRAFFIC COEXIST.

VoIP Codec	Sample Period	Number of co-existing two-way background data connections					
		5	10	15	20	25	30
G.711	10 ms	19/18/21	16/16/18	14/14/16	12/12/13	11/11/12	10/10/11
	20 ms	35/35/38	29/29/31	26/25/27	23/22/24	21/20/23	19/18/22
	30 ms	49/47/53	41/41/46	36/36/40	32/32/37	29/29/33	27/27/31
	40 ms	62/62/67	52/52/57	45/45/50	40/40/46	37/37/42	34/34/39
	50 ms	73/73/80	61/61/66	53/53/60	47/47/55	43/43/50	40/40/46
	60 ms	83/84/91	69/69/78	60/60/69	54/54/63	49/50/58	45/47/55



(a)

(b)

Fig. 6. The PDF of active number of TCs given that the TC at the AP or at the station (denoted as STA) is active in a scenario consisting of G.711 VoIP connections (10 ms packet intervals). (a) 15 connections. (b) 25 connections. Note that the figures do not present the whole x-axis (activity profile for large number of stations) for better clarity on the comparison of simulation and analysis (especially when the activity probability is not close to zero). The interested reader is referred to [45] for more results.

other. Such a comparison reveals that the proposed capacity prediction and admission control scheme is also effective when different classes of multimedia traffic coexist in the BSS. As the comparison of the number of admitted uplink and downlink flows shows, channel contention overhead is the main limitation on capacity. For the same number of coexisting VoIP connections, the number of admitted downlink flows is larger than the number of admitted uplink flows, as contention overhead is much lower in the downlink scenario. With increasing number of VoIP connections, the difference increases as well. As expected, the two-way video capacity in terms of admitted number of flows is less than the capacity in the uplink only and the downlink only scenarios. The increasing VoIP load does not affect the ratio of downlink to two-way video capacity as significantly as it affects the ratio of the ratio of downlink to uplink video capacity.

Once again, note that the analytical models of [28]–[34] do not provide such analysis capability. The nonsaturation Markov model in [22] has over-admission problems. The accuracy of the measurement-assisted approach in [39] depends on algorithm parameter settings, and the optimal/good settings vary from scenario to scenario for such an approach (we do not change the parameter settings on a case-by-case basis for this scenario).

VI. CONCLUSION

We have designed a practical and simple multimedia capacity prediction and admission control algorithm to limit the number of admitted real-time multimedia flows in the 802.11e infrastructure BSS. Motivated by the previous findings in the literature such that the contention-based 802.11 MAC can achieve high throughput and low delay in nonsaturation, the proposed admission control algorithm is based on simple tests on station- and AC-specific queue utilization ratio estimates. Our novel approach is the calculation of the queue utilization ratio by weighing the average service time predictions of the proposed cycle time saturation model among varying number of active stations. The proposed simple framework is effective in capacity estimation even in the case of coexisting multimedia flows using different ACs with arbitrarily selected MAC parameters. Comparing the theoretical results with simulations, we have shown that the proposed algorithm provides guaranteed QoS for coexisting voice and video connections in an infrastructure BSS (when an uplink/downlink asymmetry exists in terms of traffic load). One of the key insights provided

TABLE III
COMPARISON OF THE MAXIMUM NUMBER OF VIDEO CONNECTIONS (PROPOSED/SIMULATION/[22]/[39]) WHEN VOIP FLOWS COEXIST.

MPEG-4	Number of existing two-way G.711 (20 ms) connections					
	5	10	15	20	25	30
Downlink	109/110/115/109	98/100/102/99	87/88/92/85	76/78/82/75	64/65/69/63	52/54/56/52
Uplink	88/89/91/84	67/69/72/60	57/57/59/50	48/48/51/42	37/37/41/34	28/28/31/22
Two-way	54/55/60/45	46/47/51/42	41/41/45/32	34/34/38/26	28/28/34/23	19/19/25/15

by this study is the accuracy of the proposed approximate capacity estimation framework that uses relatively simpler saturation analysis rather than defining a more complex and hard to implement nonsaturation model.

We have also developed a simple and novel average cycle time model to evaluate the performance of the EDCA function in saturation. The proposed model captures the performance in the case of an arbitrary assignment of AC-specific AIFS and CW values and is the first model to consider an arbitrary distribution of active ACs at the stations. We have shown that the analytical results obtained using the cycle time model closely follow the accurate predictions of the previously proposed more complex analytical models and simulation results.

REFERENCES

[1] IEEE Standard 802.11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, IEEE 802.11 Std., 1999.

[2] IEEE Standard 802.11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Medium Access Control (MAC) Quality of Service (QoS) Enhancements, IEEE 802.11e Std., 2005.

[3] X. Chen, H. Zhai, X. Tian, and Y. Fang, "Supporting QoS in IEEE 802.11e wireless LANs," *IEEE Trans. Wireless Commun.*, pp. 2217-2227, Aug. 2006.

[4] H. Zhai, X. Chen, and Y. Fang, "How well can the IEEE 802.11 wireless LAN support quality of service?" *IEEE Trans. Wireless Commun.*, pp. 3084-3094, Nov. 2005.

[5] K. Medepalli and F. A. Tobagi, "Throughput analysis of IEEE 802.11 wireless LANs using an average cycle time approach," in *Proc. IEEE Globecom '05*, Nov. 2005.

[6] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE J. Sel. Areas Commun.*, pp. 535-547, Mar. 2000.

[7] F. Cali, M. Conti, and E. Gregori, "Dynamic tuning of the IEEE 802.11 protocol to achieve a theoretical throughput limit," *IEEE/ACM Trans. Networking*, pp. 785-799, Dec. 2000.

[8] J. C. Tay and K. C. Chua, "A capacity analysis for the IEEE 802.11 MAC protocol," *Wireless Netw.*, pp. 159-171, July 2001.

[9] Y. Xiao, "Performance analysis of priority schemes for IEEE 802.11 and IEEE 802.11e wireless LANs," *IEEE Trans. Wireless Commun.*, pp. 1506-1515, July 2005.

[10] Z. Kong, D. H. K. Tsang, B. Bensaou, and D. Gao, "Performance analysis of the IEEE 802.11e contention-based channel access," *IEEE J. Sel. Areas Commun.*, pp. 2095-2106, Dec. 2004.

[11] J. W. Robinson and T. S. Randhawa, "Saturation throughput analysis of IEEE 802.11e enhanced distributed coordination function," *IEEE J. Sel. Areas Commun.*, pp. 917-928, June 2004.

[12] J. Hui and M. Devetsikiotis, "A unified model for the performance analysis of IEEE 802.11e EDCA," *IEEE Trans. Commun.*, pp. 1498-1510, Sep. 2005.

[13] H. Zhu and I. Chlamtac, "Performance analysis for IEEE 802.11e EDCF service differentiation," *IEEE Trans. Wireless Commun.*, pp. 1779-1788, July 2005.

[14] I. Inan, F. Keceli, and E. Ayanoglu, "Saturation throughput analysis of the 802.11e enhanced distributed channel access function," in *Proc. IEEE ICC '07*, June 2007.

[15] Z. Tao and S. Panwar, "Throughput and delay analysis for the IEEE 802.11e enhanced distributed channel access," *IEEE Trans. Commun.*, pp. 596-602, Apr. 2006.

[16] J. Zhao, Z. Guo, Q. Zhang, and W. Zhu, "Performance study of MAC for service differentiation in IEEE 802.11," in *Proc. IEEE Globecom '02*, Nov. 2002.

[17] A. Banchs and L. Vollero, "Throughput analysis and optimal configuration of IEEE 802.11e EDCA," *Comp. Netw.*, pp. 1749-1768, Aug. 2006.

[18] Y. Lin and V. W. Wong, "Saturation throughput of IEEE 802.11e EDCA based on mean value analysis," in *Proc. IEEE WCNC '06*, Apr. 2006.

[19] I. Inan, F. Keceli, and E. Ayanoglu, "Performance analysis of the IEEE 802.11e enhanced distributed coordination function using cycle time approach," in *Proc. IEEE Globecom '07*, Nov. 2007.

[20] K. Duffy, D. Malone, and D. J. Leith, "Modeling the 802.11 distributed coordination function in non-saturated conditions," *IEEE Commun. Lett.*, pp. 715-717, Aug. 2005.

[21] P. E. Engelstad and O. N. Osterbo, "Analysis of the total delay of IEEE 802.11e EDCA and 802.11 DCF," in *Proc. IEEE ICC '06*, June 2006.

[22] I. Inan, F. Keceli, and E. Ayanoglu, "Modeling the 802.11e enhanced distributed coordination function," in *Proc. IEEE Globecom '07*, Nov. 2007.

[23] O. Tickoo and B. Sikdar, "A queueing model for finite load IEEE 802.11 random access MAC," in *Proc. IEEE ICC '04*, June 2004.

[24] H. Zhai, Y. Kwon, and Y. Fang, "Performance analysis of IEEE 802.11 MAC protocols in wireless LANs," *Wireless Commun. Mobile Computing*, pp. 917-931, Dec. 2004.

[25] C. H. Foh and M. Zukerman, "A new technique for performance evaluation of random access protocols," in *Proc. European Wireless '02*, Feb. 2002.

[26] J. W. Tantra, C. H. Foh, I. Tinnirello, and G. Bianchi, "Analysis of the IEEE 802.11e EDCA under statistical traffic," in *Proc. IEEE ICC '06*, June 2006.

[27] J. Hui and M. Devetsikiotis, "Metamodeling of Wi-Fi performance," in *Proc. IEEE ICC '06*, June 2006.

[28] K. Medepalli and F. A. Tobagi, "System centric and user centric queueing models for IEEE 802.11 based wireless LANs," in *Proc. IEEE Broadnets '05*, Oct. 2005.

[29] D. P. Hole and F. A. Tobagi, "Capacity of an IEEE 802.11b wireless LAN supporting VoIP," in *Proc. IEEE ICC '04*, June 2004.

[30] S. Garg and M. Kappes, "An experimental study of throughput for UDP and VoIP traffic in IEEE 802.11b networks," in *Proc. IEEE WCNC '03*, Mar. 2003.

[31] ——, "Can I add a VoIP call?" in *Proc. IEEE ICC '03*, May 2003.

[32] L. X. Cai, X. Shen, J. W. Mark, L. Cai, and Y. Xiao, "Voice capacity analysis of WLAN with unbalanced traffic," *IEEE Trans. Veh. Technol.*, pp. 752-761, May 2006.

[33] D. Gao, J. Cai, and C. W. Chen, "Capacity analysis of supporting VoIP in IEEE 802.11e EDCA WLANs," in *Proc. IEEE Globecom '06*, Nov. 2006.

[34] Y. Cheng, L. Cai, X. Ling, W. Song, W. Zhuang, X. Shen, and A. Leon-Garcia, "Improvement of WLAN QoS capability via statistical multiplexing," in *Proc. IEEE Globecom '06*, Nov. 2006.

[35] S. Harsha, A. Kumar, and V. Sharma, "An analytical model for the capacity estimation of combined VoIP and TCP file transfers over EDCA in an IEEE 802.11e WLAN," in *Proc. IEEE IWQoS '06*, June 2006.

[36] L. Zhang and S. Zeadally, "HARMONICA: enhanced QoS support with admission control for IEEE 802.11 contention-based access," in *Proc. IEEE RTAS '04*, May 2004.

[37] D. Pong and T. Moors, "Call admission control for IEEE 802.11 contention access mechanism," in *Proc. IEEE Globecom '03*, Dec. 2003.

[38] J. He, D. Kaleshi, A. Munro, M. Barton, Z. Tang, and Z. Yang, "Management of services differentiation and guarantee in IEEE 802.11e wireless LANs," in *Proc. IEEE VTC '05 - Spring*, May 2005.

[39] Y. Xiao and H. Li, "Voice and video transmissions with global data parameter control for the IEEE 802.11e enhanced distributed channel access," *IEEE Trans. Parallel Distrib. Syst.*, pp. 1041-1053, Nov. 2004.

[40] ——, “Local data control and admission control for QoS support in wireless ad hoc networks," *IEEE Trans. Veh. Technol.*, pp. 1558-1572, Sep. 2004.

[41] Y. Xiao and Y. Pan, “Differentiation, QoS guarantee, and optimization for real-time traffic over one-hop ad hoc networks," *IEEE Trans. Parallel Distrib. Syst.*, pp. 538-549, June 2005.

[42] Y. Xiao, “QoS guarantee and provisioning at the contention-based wireless MAC layer in the IEEE 802.11e wireless LANs," *IEEE Wireless Commun. Mag.*, pp. 14-21, Feb. 2006.

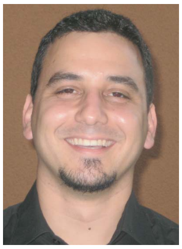
[43] (2006) The Network Simulator, ns-2. [Online]. Available: http://www.isi.edu/nsnam/ns

[44] IEEE 802.11e HCF MAC model for ns-2.28. [Online]. Available: http://newport.eecs.uci.edu/f̃keceli/ns.htm

[45] I. Inan, F. Keceli, and E. Ayanoglu, “Multimedia capacity analysis of the IEEE 802.11e contention-based infrastructure basic service set," ArXiV cs.IT/ 0707.2836, July 2007. [Online]. Available: arxiv.org

[46] P. Serrano, A. Banchs, T. Melia, and L. Vollero, “Performance anomalies of nonoptimally configured wireless LANs," in *Proc. IEEE WCNC ’06*, Apr. 2006.

[47] P. Seeling, M. Reisslein, and B. Kulapala, “Network performance evaluation using frame size and quality traces of single-layer and two-layer video: A tutorial," *IEEE Commun. Surveys Tutorials*, vol. 6, no. 2, pp. 58-78, Third Quarter 2004. [Online]. Available: http://www.eas.asu.edu/trace



Inanc Inan received the B.S. degree from the Middle East Technical University, Ankara, Turkey, in June 2001, the M.S. degree from Bilkent University, Ankara, Turkey, in September 2003, and the Ph.D. degree from University of California, Irvine, in September 2007, all in electrical engineering.

He was with Wireless Networking division of Conexant Systems Inc. from June 2006 to December 2006. Since July 2007, he has been working as a research scientist at Wionics Research - Realtek Group, concentrating mainly on MAC, link, and transport layer protocol research and development for UWB wireless networks. His current research interests include analytical network modeling and simulation, QoS provisioning, fair and efficient resource allocation, protocol and algorithm design for wireless networks.

Feyza Keceli received the B.S. degree from the Middle East Technical University, Ankara, Turkey, in June 2001, the M.S. degree from Bilkent University, Ankara, Turkey, in September 2003, and the Ph.D. degree from University of California, Irvine, in September 2008, all in electrical engineering.

She was with Wireless Networking division of Conexant Systems Inc., from June 2006 to December 2006, where she worked on the design of MAC algorithms for enhancing Bluetooth/WLAN coexistence performance. Since December 2008, she is working at Networks in Motion as a software engineer mainly concentrating on application and system software design and development for localization-based GPS services in mobile wireless networks. Her current research interests include analytical network modeling and simulation, MAC and transport layer protocol design, fair access and QoS provisioning in wireless networks.

Ender Ayanoglu (S’82-M’85-SM’90-F’98) received the B.S. degree from the Middle East Technical University, Ankara, Turkey, in 1980, and the M.S and Ph.D. degrees from Stanford University, Stanford, CA in 1982 and 1986, respectively, all in electrical engineering.

He was with the Communications Systems Research Laboratory of AT&T Bell Laboratories (Bell Labs, Lucent Technologies after 1996) until 1999 and was with Cisco Systems until 2002. Since 2002, he has been a Professor in the Department of Electrical Engineering and Computer Science, the Henry Samueli School of Engineering, University of California, Irvine where he is currently the Director of the Center for Pervasive Communications and Computing and holds the Conexant-Broadcom Endowed Chair.

Dr. Ayanoglu is the recipient of the IEEE Communications Society Stephen O. Rice Prize Paper Award in 1995 and the IEEE Communications Society Best Tutorial Paper Award in 1997. From 1993 to 2008, he was as an Editor of the IEEE Transactions on Communications and from 2004 to 2008 served as its Editor-in-Chief. He served on the Executive Committee of the IEEE Communications Society Communication Theory Committee from 1990 until 2002, and from 1999 to 2002, was its Chair.