Stochastic Delay Guarantees and Statistical Call Admission Control for IEEE 802.11 Single-Hop Ad Hoc Networks

Atef Abdrabou and Weihua Zhuang, *Fellow, IEEE*

***Abstract*—This paper presents a new approach to provide stochastic delay guarantees via fully distributed model-based call admission control for IEEE 802.11 single-hop ad hoc networks. We propose a novel stochastic link-layer channel model to characterize the variations of the channel service process in a non-saturated case using a Markov-modulated Poisson process (MMPP) model. We use the model to calculate the effective capacity of the IEEE 802.11 channel. The channel effective capacity concept is the dual of the effective bandwidth theory. Our approach offers a tool for distributed statistical resource allocation in ad hoc networks, which combines both efficient resource utilization and quality-of-service (QoS) provisioning to a certain probabilistic limit. Simulation results demonstrate that the MMPP link-layer model and the calculated effective capacity can be used effectively in allocating resources with stochastic delay guarantees.**

***Index Terms*—IEEE 802.11 MAC, ad hoc network, queuing model, resource allocation, call admission control, delay.**

I. Introduction

THE IEEE 802.11 medium access control (MAC) protocol has been widely studied and deployed. It is a very strong candidate for ad hoc networks mainly because of its distributed nature and low implementation cost. Supporting multimedia applications in IEEE 802.11 based ad hoc networks is a challenging task. The increasing demand for bandwidth from multimedia applications, the quality-of-service (QoS) constraints (such as delay bound), and the distributed control of ad hoc networks represent the main challenges. In this paper, we focus on the delay bound as a QoS constraint. As the IEEE 802.11 distributed coordination function (DCF) allocates the channel bandwidth equally among the nodes in an ad hoc network, resource allocation such as call admission control (CAC) is vital for QoS provisioning. An effective CAC scheme for ad hoc networks should work in a distributed manner and should use the wireless bandwidth efficiently (i.e. with a minimal amount of information exchanges). Indeed, as fully statistical (model based) CAC is efficient in bandwidth usage (involves only computations with minimal signaling exchanges) and does not need assistance from a central controller, it is very suitable for ad hoc networks.

Manuscript received May 28, 2007; revised October 10, 2007; accepted October 23, 2007. The associate editor coordinating the review of this paper and approving it for publication was X. Zhang.

The authors are with the Centre for Wireless Communications (CWC), Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: {alotfy, wzhuang}@bbcr.uwaterloo.ca).

Digital Object Identifier 10.1109/T-WC.2008.070564

Basically, the service time distribution of the IEEE 802.11 DCF channel (server) follows a very complex general distribution that can be evaluated only numerically, specially in a non-saturated case [1] [2]. This implies that a G/G/1 queuing analysis is required in order to provide statistical CAC for different multimedia traffic types [1]. The G/G/1 analysis is difficult when the arrival and service time distributions are complicated. Moreover, only first order statistics such as average waiting time can be used in CAC if it is based on a standard queuing analysis by using the Little's theorem. The resource allocation (e.g. CAC) based on first order statistics can guarantee only that the average end-to-end delay in a packet transmission does not exceed a delay bound. However, this may not be efficient for real-time multimedia applications specially if the actual packet service time is not close to its average value as in the case of IEEE 802.11 DCF [1]. On the other hand, CAC decisions that depend on stochastic bounds, such as $\Pr(D > D_{max}) \leq \epsilon$ (where D represents the total packet delay, D_{max} is the delay bound, and ϵ is the QoS violation probability upper bound), is more effective, but unfortunately cannot be realized by the standard queuing analysis.

To the best of our knowledge, the majority of the related works are either measurement-based admission control (e.g., [3]) or model-assisted measurement-based admission control (e.g., [4]- [6]). In [5], a CAC strategy is proposed based on the saturated throughput estimate. However, it is difficult to provide QoS guarantees in the saturated case since the node queues would be unstable. In [6], a centralized CAC algorithm has been proposed based on the effective bandwidth concept to guarantee a certain buffer loss rate. The CAC decision is based on a comparison between the effective bandwidth and the difference between the saturated throughput and the unsaturated throughput in average values without considering the randomness of the service time. Also, the saturation throughput is not necessarily the maximum throughput that the network can reach [7]. Our objective in this paper is to achieve stochastic delay guarantees by using a fully distributed model-based CAC algorithm. In order to realize our objective, we propose a link-layer stochastic channel model for IEEE 802.11 networks. We aim at characterizing statistically the IEEE 802.11 DCF channel capacity (service) variations at different traffic loads [1]. The model offers a tool for statistical CAC in

[1]We use *capacity* and *service* interchangeably throughout this paper, referring to the random process that describes the number of packets successfully transmitted over a time interval t.

order to provide stochastic performance bounds without the need of a queuing analysis. Our link-layer channel model is based on the effective capacity link model presented in [8] for a wireless channel with capacity varying randomly with time. It is different from a physical-layer channel model that is used to predict the characteristics of the physical layer, although both channel models have a similar objective. The effective capacity for a channel is the dual of the effective bandwidth theory, which has been developed for wired networks [8], [9]. The effective bandwidth theory addresses the problem of finding the capacity to bound the queue for a random source traffic process served by a fixed capacity channel. However, by considering a random time-varying channel, the problem of bounding the queue can be addressed in a similar way by finding the effective capacity of the channel. It has been shown in [10] that the effective capacity for a channel model can also be extended to work with statistical traffic sources. We propose to use source traffic and channel modeling in making the CAC decisions without consuming the limited processing power of the ad hoc network nodes or the bandwidth of the channel in frequent measurements or traffic monitoring. The effective bandwidth approach has been used before to solve the classical resource allocation problem of finding the number of multiplexed traffic sources sharing a first-in first-out (FIFO) buffer with fixed server capacity under a probabilistic QoS constraint [9]. In fact, our approach tackles the CAC problem in a single-hop IEEE 802.11 ad hoc network in a way similar to the classical one by introducing the effective capacity of the IEEE 802.11 channel. The IEEE 802.11 DCF as a server resembles a FIFO statistical multiplexer that multiplexes the traffic from different traffic sources but on a distributed fashion.

This paper presents two main contributions. First, we propose an MMPP link-layer channel model for the IEEE 802.11 DCF. The MMPP model has been used extensively in characterizing the arrival process of statistically multiplexed multimedia traffic sources [9]. However, we use the MMPP model here in a novel way to characterize the service process (not the arrival process) of the IEEE 802.11 DCF shared channel and to derive its effective capacity. To the best of our knowledge, there is no related work in the literature that addresses the effective capacity calculation for IEEE 802.11 DCF either in an ad hoc mode or in an infrastructure-based mode (WLANs). Moreover, our resource allocation technique (by using the effective bandwidth and the effective capacity) offers a step ahead of the other proposed schemes in the literature, as our scheme provides stochastic delay guarantees instead of average delay guarantees. We show that the derived effective capacity is sufficiently accurate in computing the number of nodes that can be admitted under the QoS constraint in terms of the delay bound. Second, we introduce a simple distributed model-based CAC algorithm for an IEEE 802.11 single-hop ad hoc network.

The rest of the paper is organized as follows. Section II provides the necessary background for the effective bandwidth theory and the effective capacity for a channel. It also illustrates the basic equations used throughout the paper in order to calculate both the effective bandwidth and the effective capacity. The system model is introduced in Section III. Section IV consists of four parts, where we first illustrate the behavior of IEEE 802.11 DCF under different traffic loads, present our proposed MMPP link-layer channel model, then show the applicability of the MMPP link-layer model to the case of heterogeneous traffic sources and provide the distributed CAC algorithm. Section V presents the simulation results to validate the proposed link-layer channel model and to demonstrate the performance of the proposed CAC algorithm. Section VI concludes this research.

II. Preliminaries

A. The Effective Bandwidth of a Traffic Source

The effective bandwidth approach is to show that the queue length and the corresponding delay at a node can be bounded exponentially for different stochastic traffic types if an amount of bandwidth equal to the effective bandwidth of the source is provided by the channel [9].

Consider a queue of infinite buffer size served by a channel of constant service rate c. Let D denote the total delay (queuing delay + service time) that a source packet experiences. By using the large deviation theory [8] [11], it can be shown that the probability ϵ that D exceeds a delay bound of D_{max} is given by

$$\epsilon = \Pr\{\mathrm{D} \geq \mathrm{D_{max}}\} \approx e^{-\theta_b D_{\max}} \tag{1}$$

where the exponent θ_b is the solution of

$$\theta_b = c\eta_b^{-1}(c). \tag{2}$$

In (2), $\eta_b^{-1}(.)$ is the inverse function of $\eta_b(.)$ which is the effective bandwidth of the traffic source, given by

$$\eta_{\mathrm{b}}(\mathrm{x}) = \lim_{\mathrm{t}\to\infty} \frac{1}{t}\frac{1}{x}\log E[e^{xA(t)}], \forall x > 0 \tag{3}$$

where $A(t)$ represents the arrival process of the source, i.e. the number of packet arrivals in the interval [0, t]. Thus, the source which has a delay bound of D_{max} will experience a delay-bound violation probability of at most ϵ if the constant channel capacity c is at least equal to its effective bandwidth [8].

B. The Effective Capacity of a Channel

The effective capacity model is the dual of the effective bandwidth theory when the channel capacity is time varying. Let $S(t)$ denote the service process of the channel (the amount of data that the channel can carry) in bits over the time interval [0, t]. The effective capacity function can be defined as [8]

$$\eta_c(x) = -\lim_{t\to\infty} \frac{1}{t}\frac{1}{x}\log E[e^{-xS(t)}], \forall x > 0. \tag{4}$$

Similar to the effective bandwidth theory, it can be shown that the probability of the delay D exceeding a certain delay bound D_{max} satisfies [10]

$$Pr\ \{D \geq D_{\max}\} \approx e^{-\theta_c D_{\max}} \tag{5}$$

where the exponent θ_c is the solution of

$$\theta_c = u\eta_c^{-1}(u). \tag{6}$$

Therefore, a source should limit its data rate to a maximum of u in order to ensure that its delay bound (D_{max}) is violated with a probability of at most ϵ.

It has been shown in [10] that, if both the traffic source rate and the channel capacity are time varying, both the effective bandwidth of the source and the effective capacity of the channel should be equal in order to satisfy the stochastic delay bound. Then for a large enough D_{max}, the total delay satisfies

$$\frac{1}{D_{\max}} \log \Pr(D > D_{\max}) = -\theta_s \tag{7}$$

where θ_s is given by

$$\theta_s = r\eta_c(r) \tag{8}$$

and r is the unique solution of the equation

$$\eta_c(r) = \eta_b(r). \tag{9}$$

III. System Model

We consider an IEEE 802.11 DCF single-hop ad hoc network, with a single and error-free physical channel. All the nodes can hear each other, so there are no hidden or exposed terminals. The network nodes are either active nodes (traffic sources) or just receivers. In what follows, unless ambiguity occurs, the term node refers to an active node. Consider the network in a non-saturated condition [7]. All the traffic sources are iid exponential on-off traffic sources (i.e. the on and off times are independent exponential random variables). It has been shown in [9] that the on-off sources can be used successfully to model different multimedia traffic types. Each active node i has a traffic source with average on time $1/\alpha_i$, average off time $1/\beta_i$, a constant data rate R_i during an on time period and the QoS requirement captured by $D_{i_{max}}$ and ϵ.

Following the carrier-sense multiple access with collision avoidance (CSMA/CA) protocol as described in [12], the contention window (CW) size initially is set to CW_{min}. After an unsuccessful transmission, the CW size is doubled up to a maximum value

$$CW_{\max} = 2^{m_b} \times CW_{\min} \tag{10}$$

where m_b is the number of backoff stages and has a typical value such as 5 [12]. The random access employs the four-way RTS-CTS-DATA-ACK handshaking technique as described in [7], [12]. Under the assumption of a fixed packet size, the packet transmission time T_s in time slots is given by [7]

$$T_s = T_{RTS} + T_{CTS} + 3\ SIFS + T_{ACK} + T + DIFS \tag{11}$$

and the packet collision time in time slots is given by [7]

$$T_c = T_{RTS} + DIFS \tag{12}$$

where T_{RTS}, T_{CTS} and T_{ACK} represent the transmission times in time slots for the RTS, CTS and ACK packets, respectively; T is the data packet transmission time in time slots, which is constant for a fixed packet size; DIFS is the distributed inter-frame spacing and the SIFS is the short inter-frame spacing as defined in [12].

A. Service Time Statistics

In this subsection we address the first and the second order statistics of the service time distribution of the IEEE 802.11 DCF. These statistics help us in specifying the network operation region and in formulating our proposed MMPP model, as described in Subsections IV-A and IV-B.

The service time distribution of the IEEE 802.11 is complicated since, between two successful packet transmissions of any node, three different random variables (in the case of a fixed packet size) are involved; namely, the time W spent in the idle backoff time slots, the time T_{cl} wasted in collisions happened either to other nodes or to the node under consideration, and the time T_{st} consumed in the successful transmissions of the other nodes. The analysis of the unsaturated case is harder than the saturated counterpart since every node may or may not have backlogged packets in its queue based on the value of the node queue utilization factor ρ (the probability of non-empty queue). In the unsaturated case, the system (from the point of view of any node that wants to transmit a packets) can be viewed as having different states. Each state has a number of nodes with backlogged packets. The system spends a random time in each state before transferring to another state. In fact, the first and second order statistics of the packet service time for any node can be obtained in a unified way if we conditioned all the associate random variables on the number of the nodes having backlogged packets, n, in the system [13]. If the exact stationary distribution of the states is known, both the average service time and the variance can be calculated. The actual state distribution is computationally complex even for much simpler types of CSMA-based MAC protocols [14]. We use the average service rate conditioned on n in our proposed model.

Let p_n be the collision probability that a packet of the node under consideration will see, given n other nodes having backlogged packets ($n+1$ nodes compete for transmission). We have

$$p_n = 1 - (1 - \frac{1}{\overline{W}_n})^n \tag{13}$$

where $\overline{W}_n$ is the conditional average backoff time (idle time slots) given n nodes having backlogged packets, represented by [15]

$$\begin{aligned}\overline{W_n} = E[E[W_n|Bo=k]] &= \sum_{k=0}^{m_b} p_n^k(1-p_n)\frac{2^k CW_{\min}}{2} \\ &\quad + p_n^{m_b+1}\frac{2^{m_b} CW_{\min}}{2} \\ &\approx \frac{1-p_n-p_n(2p_n)^{m_b}}{1-2p_n}\left(\frac{CW_{\min}}{2}\right)^2\end{aligned} \tag{14}$$

with Bo being the backoff stage.

The variance of W_n can then be calculated using the following equation

$$Var[W_n] = Var[E[W_n|Bo=k]] + E[Var[W_n|Bo=k]].$$

The first term on the right hand side can be derived as

$$Var\left(E[W_n|Bo=k]\right) \approx \frac{1-p_n-p_n\left(4p_n\right)^{m_b}}{1-4p_n}\frac{CW_{\min}^2}{4} - \overline{W_n}^2$$

while the second term approximately equals to

$$E\left(Var[W_n|Bo=k]\right) \approx \frac{1-p_n-p_n\left(4p_n\right)^{m_b}}{1-4p_n}\frac{CW_{\min}^2}{12}$$

and this finally leads to

$$Var[W_n] \approx \frac{1-p_n-p_n(4p_n)^{m_b}}{1-4p_n}\left(\frac{CW_{\min}^2}{3}\right) - \overline{W_n}^2. \quad (15)$$

The time spent in successful transmissions for the n nodes (having backlogged packets) between two successful transmissions of the node under consideration follows a geometric distribution [13] with parameter κ

$$\Pr\{T_{st} = sT_s|n\} = \kappa(1-\kappa)^s, \quad s = 0,1,2,... \quad (16)$$

where κ equals to $1/(n+1)$ as the IEEE 802.11 is shown to be fair both on short and long term basis [16]. Therefore, the conditional average of T_{st} is given by

$$E[T_{st}|n] = \frac{1-\kappa}{\kappa}T_s = nT_s \quad (17)$$

and the conditional variance is

$$Var[T_{st}|n] = n(n+1)T_s^2. \quad (18)$$

The conditional average and variance of T_{cl} can be obtained following the same way as in [13]

$$E[T_{cl}|n] = \frac{n+1}{2}\frac{p_n}{1-p_n}T_c \quad (19)$$

$$Var[T_{cl}|n] = \left(\frac{n+1}{2}\frac{p_n}{1-p_n} + \left(\frac{n+1}{2}\frac{p_n}{1-p_n}\right)^2\right)T_c^2. \quad (20)$$

The total conditional average of the service time T_t equals to the sum of the above conditional averages plus the packet transmission time (T_s) of the node under consideration, given by

$$E[T_t|n] = (n+1)T_s + \frac{n+1}{2}\frac{p_n}{1-p_n}T_c + \overline{W_n} \quad (21)$$

and hence the conditional service rate is

$$\mu_n = \frac{1}{(n+1)T_s + \frac{n+1}{2}\frac{p_n}{1-p_n}T_c + \overline{W_n}}. \quad (22)$$

As the calculation of the service rate needs the stationary distribution of the states, another approach based on the first order statistics followed by [17] leads to the following average service rate

$$\mu = \frac{1}{\rho(N-1)\left[T_s + \frac{T_c}{2}\frac{p}{1-p}\right] + \overline{W} + T_s + \frac{T_c}{2}\frac{p}{1-p}}. \quad (23)$$

In (23), p is the unconditional collision probability, given by

$$p = 1 - (1-\frac{\rho}{\overline{W}})^{N-1} \quad (24)$$

where N is the number of active nodes, and $\overline{W}$ is the average backoff window given by

$$\overline{W} \approx \frac{1-p-p(2p)^{m_b}}{1-2p}\frac{CW_{\min}}{2}. \quad (25)$$

However, the service time variance can not be obtained using the same approach but only by numerical techniques [1] [2].

IV. The MMPP Link-Layer Model and the CAC Algorithm

A. IEEE 802.11 Behavior Under Different Traffic Loads

We study in this subsection the IEEE 802.11 DCF operation region (in terms of traffic load) over which our model can work with sufficient accuracy. In fact, the traffic load directly affects the packet collision probability p, which controls the service time distribution of the IEEE 802.11 DCF [1] [2]. We can identify three different regions of operation for the IEEE 802.11 DCF. The first region is characterized by a low traffic load where the IEEE 802.11 packet service time becomes almost deterministic as has been shown by computer simulations in [2]. In this region, the collision probability is small, i.e., very few collisions occur. Therefore, the collision time and the backoff time (the contention window size most likely at CW_{min}) can be neglected as compared with the packet transmission time T_s. In Appendix we show that at a low traffic load (low ρ), the ratio of the standard deviation of the service time $std(T_t)$ to the average service time $E[T_t]$ is approximately given by

$$q = \frac{std[T_t]}{E[T_t]} \approx \frac{\sqrt{(N-1)\rho\left((N-1)\rho+3\right)}}{(N-1)\rho+1}.$$

The service time distribution becomes more accurately deterministic as the value of q becomes smaller than one. This requires that ρ(N-1) be sufficiently smaller than 1. The collision probability p at low ρ based on (24) can be approximated to

$$p = 1 - \left(1-\frac{\rho}{\overline{W}}\right)^{N-1} \approx \frac{(N-1)\rho}{\overline{W}}.$$

Since ρ(N-1) should be smaller than one, this implies that

$$p\overline{W} < 1$$

where $\overline{W}$ can be approximated (by neglecting the higher orders of p) using (25) to

$$\overline{W} \approx \frac{1-p}{1-2p}\frac{CW_{\min}}{2}.$$

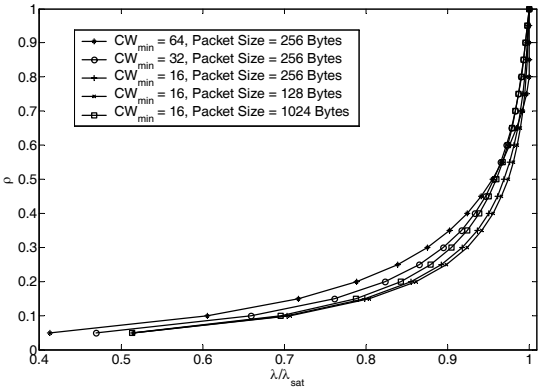
This leads to (by neglecting the second order of p)

$$p \le \frac{2}{4+CW_{\min}}. \quad (26)$$

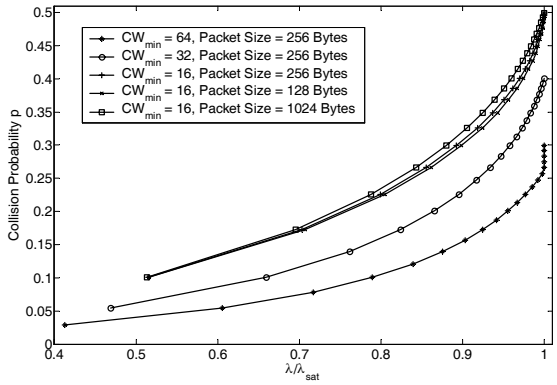
By solving (24) at the upper bound of (26) to calculate ρ, we use the following equation to obtain the value of the traffic load λ_l corresponding to the upper bound of the first region

$$\lambda_l \approx \frac{\rho}{T_s(\rho\,(N-1)+1)}. \quad (27)$$

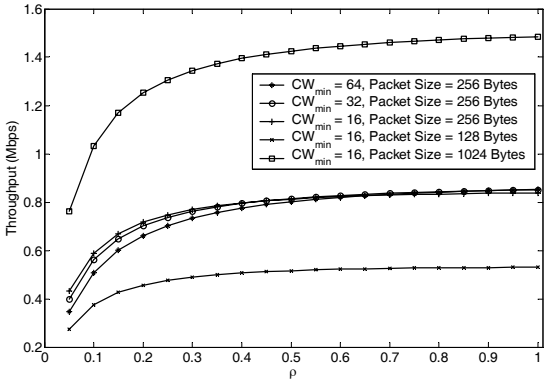
The above equation is derived from (23) by neglecting the ratio of both $\overline{W}$ and T_c with respect to T_s.



(a) Utilization factor variations with λ/λ_{sat}.



(b) Collision probability variations with λ/λ_{sat}.



(c) Throughput variations with ρ.

Fig. 1. IEEE 802.11 behavior under different traffic loads for different contention window and packet sizes.

We study the behavior of IEEE 802.11 in the second and the third regions by solving (23) and (24) simultaneously according to the parameters given in Table I in Section V and by using the fact that $\rho = \lambda/\mu$. Figures 1(a) and 1(b) show the relation between the normalized average traffic load λ/λ_{sat} (where λ_{sat} is the saturation traffic load) and ρ, and p respectively for 20 nodes and different minimum contention window and packet sizes. Figure 1(c) shows the network throughput versus the utilization factor ρ for the same number of nodes, but different minimum contention window

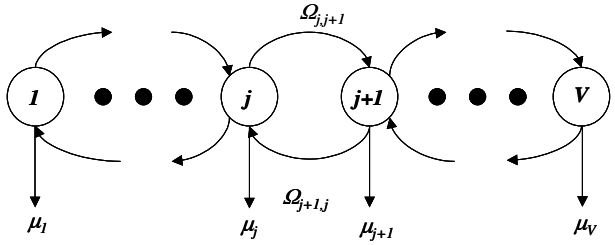


Fig. 2. The MMPP link-layer model.

and packet sizes, respectively. As we can see from Figure 1(a), the utilization factor ρ is very sensitive to the traffic load when it approaches saturation ($\lambda > 0.8\ \lambda_{sat}$). It increases up to its maximum value at (ρ = 1) with a very large slope irrespective of the contention window or the packet size used. The collision probability is also sensitive to the traffic load and increases more rapidly when $\lambda > 0.8\ \lambda_{sat}$ regardless of the used packet size or contention window size as can be seen in Figure 1(b). We define the second region of operation as $\lambda_l < \lambda \leq 0.8\ \lambda_{sat}$ and the third region as $\lambda > 0.8\ \lambda_{sat}$. Since we are concerned with the packet delay, driving the network to work in the third region (beyond $0.8\ \lambda_{sat}$) may lead to a large delay if the average traffic load fluctuates toward saturation. Moreover, from Figure 1(c), it can be seen that if the network is allowed to work in the third region, only a small amount of the network throughput (less than 10% of the saturation throughput) would be gained. Therefore, the proposed MMPP link-layer channel model characterizes only the second region of operation. Also, the proposed CAC algorithm restricts the node admission to keep the network in the second region.

B. MMPP Link-Layer Model for IEEE 802.11

We use an MMPP model to approximate the channel service process $S(t)$ when a certain average traffic load λ is applied to each active node. We assume that all the traffic sources have the same traffic parameters (average on time, average off time, and the data rate during the on time). We relax this assumption later in Subsection IV-C. The process $S(t)$ is modeled from the perspective of the node under study by a Markov chain that has V states. While in state i, the process behaves as a Poisson process with a state dependent parameter μ_i as shown in Figure 2. Each state in the Markov chain represents the number of active nodes that have backlogged packets as seen by the node under study whenever it wants to transmit a packet. Note that an active node (i.e. a traffic source) may or may not have backlogged packets at a given instant. Consider that there are n nodes with backlogged packets when the process is in state j. We approximate the service process $S(t)$ at state j by a Poisson process with a rate μ_j, where μ_j is given by (22). The Poisson approximation is based on our previous work in [18] where it is shown that the IEEE 802.11 DCF has a kind of memoryless behavior when all the competing nodes have backlogged packets.

The state transitions are limited to the adjacent states, because only one node can send a packet at a time and that the traffic sources are random and not synchronized. To find the rates of transitions we approximate the busy period

of the queue of any active node by an exponential random variable. If state $j+1$ and state j represents $m-1$ and m nodes having backlogged packets respectively, then the rate of transition $\Omega_{j,j+1}$ from state j to state j+1 can be calculated as the reciprocal of the average busy period of the queue [19] multiplied by m as

$$\Omega_{j,j+1} = m\left(\frac{\mu_j(\alpha+\beta)}{R} - \beta\right). \tag{28}$$

The rate of transition $\Omega_{j+1,j}$ from state j+1 to j simply equals to $(N-m)\,\beta$ since both the on time and the off time of the traffic sources follow an exponential distribution. The model captures the states when the node under consideration is competing with two active nodes or more. We ignore the states when the node under consideration is sending alone or competing just with one node (i.e. just one node has backlogged packets) as these states will not last for a significant time for the traffic loads in the considered region of operation. These leads to V equal to $N-2$. We found by the computer simulations that the state corresponds to two nodes with backlogged packets becomes insignificant, when the value of the traffic load is high enough (closer to 0.8 λ_{sat} than to λ_l). The model accuracy is affected by the number of nodes in the network since the assumption of constant and independent collision probability of [7] becomes more reasonable as the number of nodes increases.

From the MMPP model for $S(t)$, the effective capacity of the IEEE 802.11 DCF can be derived (using the results in [11] and [8]) as

$$\eta_c(x) = \frac{sp\left(Q + (e^{-x}-1)\,\Phi\right)}{x} \tag{29}$$

where Q is the transition rate matrix, Φ = diag(μ_1, μ_2, …, μ_V), and $sp(\mathrm{A})$ is the spectral radius of matrix A.

C. The MMPP Model with Heterogeneous On-Off Sources

The MMPP link-layer model can be applied to the case of heterogeneous on-off sources (sources with different traffic parameters) if we use homogeneous sources with equivalent statistics to represent them approximately. We match the average, the variance, and the autocovariance of the heterogeneous sources with the homogeneous ones in order to obtain the traffic parameters of them in a way similar to that in [9]. The autocovariance function of an on-off source is given by [9]

$$C(\tau) = R^2 u(1-u)e^{-(\beta+\alpha)\tau}$$

where u is the probability that the source is in the on state and given by

$$u = \frac{\beta}{\beta+\alpha} \tag{30}$$

and $1/\alpha$ is the average on time, $1/\beta$ is the average off time, and R is the constant data rate during the on time period. In order to compute the parameters (α, β, and R) of the equivalent homogeneous sources we solve the following equations

$$MuR = \sum_{l=1}^{L} M_l R_l u_l \tag{31}$$

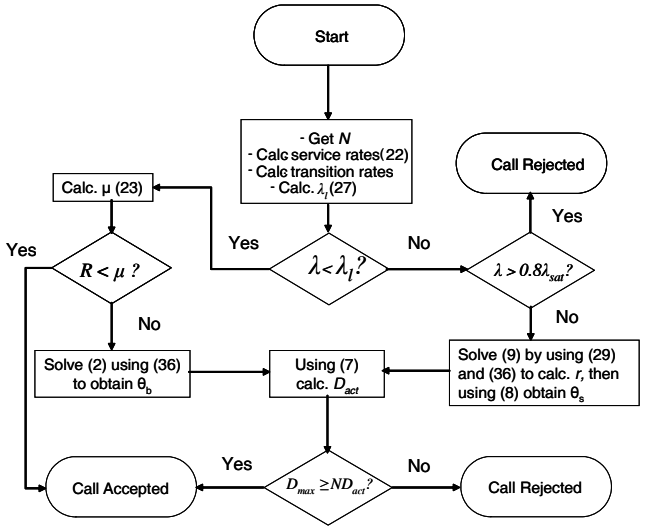


Fig. 3. The distributed model-based CAC algorithm.

$$MR^2u(1-u) = \sum_{l=1}^{L} M_l R_l^2 u_l(1-u_l) \tag{32}$$

$$MR^2u(1-u)e^{-\left(\frac{\beta}{u}\right)} = \sum_{l=1}^{L} M_l R_l^2 u_l(1-u_l)e^{-\left(\frac{\beta_l}{u_l}\right)} \tag{33}$$

$$M = \sum_{l=1}^{L} M_l \tag{34}$$

where L is the number of source groups with the same traffic parameters and M_l is the number of sources per group and M is the number of equivalent sources. By (30)-(34), we can obtain all the parameters for the equivalent homogeneous sources. We use those parameters to compute the effective capacity as described in Subsection IV-B.

D. The Distributed Model-based CAC Algorithm

Our distributed model-based CAC algorithm is based on the MMPP link-layer channel model. We assume that: (i) The traffic source model parameters are known at each active node; (ii) No active nodes leave the network during the execution of the algorithm. The following are the steps of the algorithm:

Step 1: A new node that wants to join the network exchanges information with the network and knows the number of active nodes in the network and also the traffic source parameters, following a procedure such as those given in [20] [21]. If the sources have different parameters, the node calculates equivalent homogeneous traffic source parameters using (30)-(34). The new node then calculates the service rates and the transition rates of the Markov chain using (22) and (28).

Step 2: The new node calculates its average traffic rate (λ) using the following equation

$$\lambda = Ru. \tag{35}$$

If $\lambda < \lambda_l$, the node goes to Step 3; otherwise the node jumps to Step 4.

Step 3: The node calculates the service rate μ using (23)-(25). If $R < \mu$, the node can be admitted to the network;

TABLE I
IEEE 802.11 SYSTEM PARAMETERS [12]

System Parameter	Value
Packet payload	256 Bytes
PHY header	128 bits
ACK	112 + PHY header
RTS	160 + PHY header
CTS	112 + PHY header
Slot Time	50 μs
SIFS	28 μs
DIFS	128 μs
Basic Rate	1 Mbps
Data Rate	2 Mbps
CW_{min}	16
Backoff stages (m_b)	5

otherwise, the node solves (2) after replacing c with μ to get the value of θ_b. The effective bandwidth for an on-off source is given by [11]

$$\eta_b(x) = \left(\frac{R}{2} - \frac{\beta+\alpha}{2x}\right) + \sqrt{\left[\frac{R}{2} - \frac{\beta+\alpha}{2x}\right]^2 + \frac{\beta R}{x}}. \quad (36)$$

The node then proceeds to Step 5.

Step 4: The node compares the value of λ and the value of λ_{sat} after its admission. If $\lambda > 0.8\lambda_{sat}$, the node does not admit itself in order to prevent the network from being driven to the region of operation beyond $0.8\lambda_{sat}$ (the third region as in Subsection IV-A). If $\lambda \leq 0.8\lambda_{sat}$, the node solves (9) by using (36) and (29) in order to calculate r. By applying the value of r in (8), the node obtains θ_s.

Step 5: Let D_{act} denote the delay that results in a violation probability less than or equal to ϵ from the perspective of the node under study (if it uses the channel all the time to send its packets). By replacing D_{max} with D_{act} in (7) and using the values of θ or θ_s obtained by Step 3 or Step 4 respectively, the delay bound D_{act} can be calculated. If more than one service class is available, D_{max} represents the strictest delay bound among the different service classes. Since all the other nodes equally share the same channel with the node under study, if $D_{max} \geq ND_{act}$ the node can admit itself into the network, otherwise it cannot.

Figure 3 illustrates the fully distributed CAC procedure. Every node that wants to join the network can do the calculations to know if it can admit itself to the network or not with a minimal amount of information. This implies more efficient usage of the scarce bandwidth of the wireless channel. Also, the algorithm does not depend on any measurements or traffic monitoring, which is very essential for battery-powered ad hoc network nodes.

V. MODEL VALIDATION AND SIMULATION RESULTS

We verify the MMPP model and the effective capacity approach using the ns-2 simulator [22]. The simulation model simulates nodes moving in an unobstructed plane following the *random waypoint* model [23] with a maximum speed of 1 m/s. In the simulation, a node chooses its speed and its destination randomly and then moves to the destination. The simulation is done for a network having a variable number of mobile nodes over an area of 250x250 m^2. The node radios have a transmission range of 250 m and a carrier-sense range

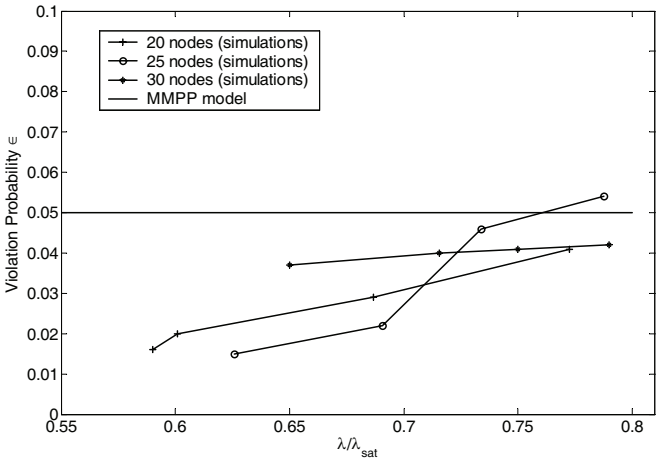


Fig. 4. Violation probability variations with λ/λ_{sat} .

of 550 m. Only half of the nodes are active traffic sources, the other half are only receivers. The network represents a single-hop ad hoc network, where every sender sends data packets to one unique receiver.

A. Model Validation

In order to validate our approach, we simulate on-off exponential traffic sources with the same (α, β) parameters as those used in the MMPP model. We calculate the delay bound for a violation probability ϵ of 0.05 for different number of nodes and different traffic loads by using the procedure described in Subsection IV-D. This delay bound is then used as an input to the ns2 simulator in order to measure the actual violation probability at different traffic loads. Table II shows the calculated delay bounds (using the IEEE 802.11 parameters given in Table I) in seconds for different traffic loads. Here, we validate the model only for $\lambda_l < \lambda \leq 0.8\ \lambda_{sat}$ (which is the operating region characterized by the model). The results in Table II indicate that, when the traffic load in the network increases, the delay bound required to satisfy ϵ increases as less network resources become available for each active node with increasing traffic load. Figure 4 shows the measured violation probability compared with the 5% value obtianed by calculating the effective capacity using the MMPP model for 20, 25 and 30 nodes respectively. The figure shows that using the MMPP model to calculate the effective capacity is generally conservative. As the traffic increases towards 0.8 λ_{sat}, the model becomes more accurate. When the traffic load increases to more than $0.8\lambda_{sat}$, the model becomes slightly optimistic since in this region the queue utilization is very sensitive to the variation of the traffic load.

B. Average-Delay-based CAC and the Proposed Model-based CAC

Figures 5 and 6 compare the CAC based on average delay guarantees and the CAC based on the effective capacity approach and the MMPP link-layer model which provide stochastic delay guarantees for the same delay bound D_{max}. Figure 5 shows the relation between the number of admitted nodes based on the average delay, the number of admitted

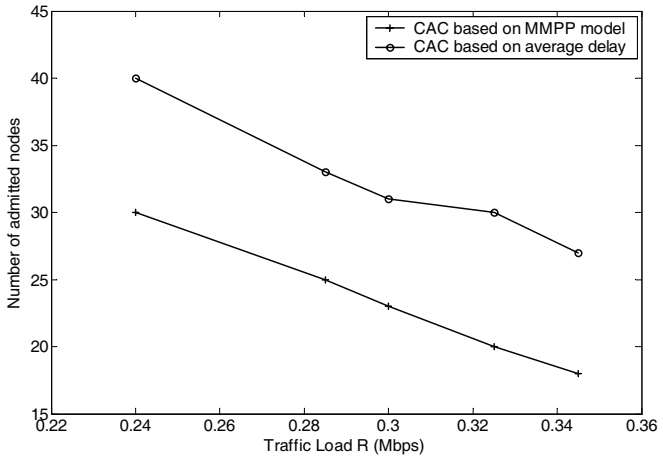


Fig. 5. Number of admitted nodes at different traffic loads for MMPP model and average delay based CAC.

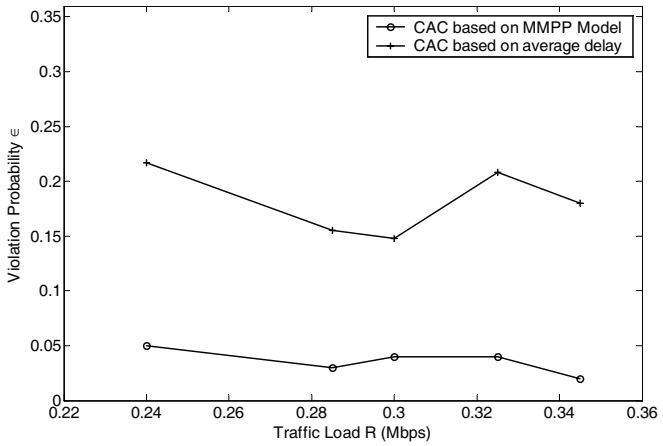


Fig. 6. Violation probability at different traffic loads for MMPP model and average delay based CAC.

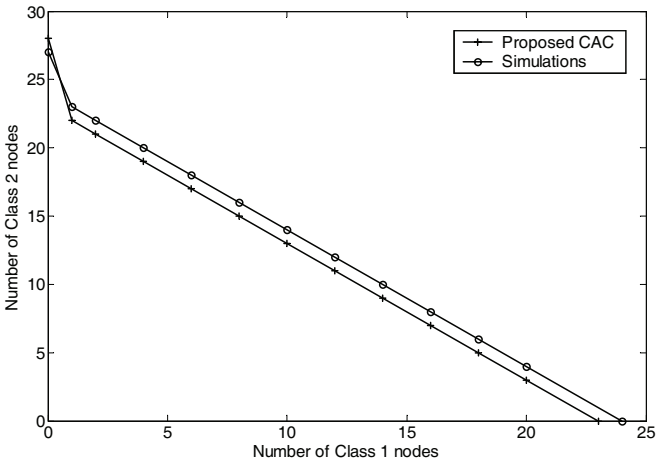


Fig. 7. Admission region for homogeneous sources with two service classes.

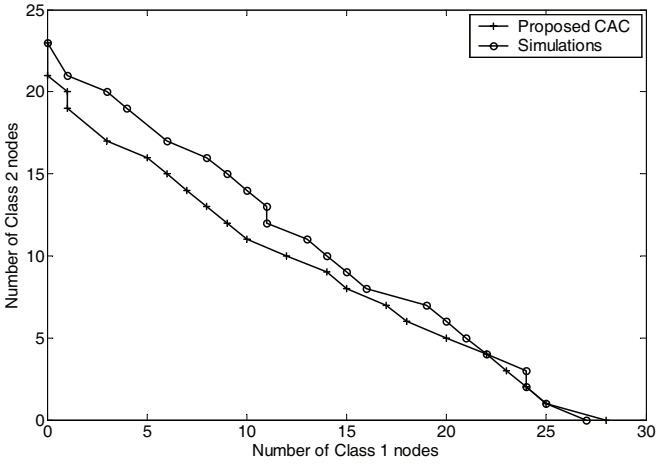


Fig. 8. Admission region for heterogeneous sources with two service classes.

nodes based on our proposed approach and the traffic load is represented by the peak rate of the traffic source R. The figure shows that we can admit more nodes based on the average delay criterion. However, this comes with the expense of having a much higher violation probability ϵ as shown in Figure 6. In fact, this result is aligned with that in [9] which illustrates how the effective bandwidth can be used to provide stochastic QoS guarantees for a number of traffic sources sharing the buffer of an FIFO statistical multiplexer served by a fixed capacity server. It has been shown in [9] that the CAC based on the average source rate results in a larger number of traffic sources that can be admitted into the multiplexer buffer to achieve certain stochastic QoS guarantee. The number of sources that can be admitted decreases if the CAC is based on the effective bandwidth concept [9] and decreases even more if the CAC is based on the peak rate of the sources where a strict deterministic QoS guarantee is provided (i.e. transmission of every packet should satisfy the delay bound). The similarity between the results shown in Figures 5 and 6 and those given in [9] illustrates that the effective capacity approach using the MMPP model is effective. The IEEE 802.11 DCF operates in a way similar to a statistical multiplexer in the sense that the shared channel multiplexes statistically the traffic from different sources but on a distributed manner.

Figures 5 and 6 also show that the CAC based on the first order statistic (average total delay) is not effective for real time applications, as it provides no control on the violation probability. Actually, the IEEE 802.11 server capacity variations should be taken into consideration as its service time distribution does not have a negligible variance. The MMPP model captures the service variations and makes the CAC decision based on the stochastic delay bound requirement. It does not require the variance of the service time distribution of the non-saturated IEEE 802.11 DCF, which is quite complicated to obtain as we indicate in Subsection III-A, but is essential for any conventional queuing analysis.

C. The Admission Region

Figures 7 and 8 show samples of the admission region of two service classes for two different cases. The first case as shown in Figure 7 represents traffic sources of the same parameters ($\alpha_1 = \alpha_2 = 2.5\ s^{-1}$, $\beta_1 = \beta_2 = 0.2\ s^{-1}$, $R_1 = R_2 = 325$ Kbps) but with different delay requirements ($D_{1_{max}} = 1.5$ s and $D_{2_{max}} = 2.4$ s) for both service classes. In the sample of the second case shown in Figure 8, the parameters of the traffic sources are $\alpha_1 = \alpha_2 = 2.5\ s^{-1}$, $\beta_1 = \beta_2 = 0.2\ s^{-1}$, $R_1 = 325$ Kbps, and $R_2 = 380$ Kbps.

TABLE II
VARIATION OF CALCULATED DELAY BOUND WITH NORMALIZED TRAFFIC LOAD (λ/λ_{sat})

λ/λ_{sat} (20 nodes)	Delay Bound (s)	λ/λ_{sat} (25 nodes)	Delay Bound (s)	λ/λ_{sat} (30 nodes)	Delay Bound (s)
0.56	1.45	0.62	1.34	0.68	1.25
0.60	1.67	0.69	1.60	0.72	1.37
0.69	2.10	0.73	2.46	0.75	2.17
0.77	3.17	0.79	2.70	0.78	2.28

The delay requirement for class 1 is $D_{1_{max}}$ = 2.4 s and for class 2 is $D_{2_{max}}$ = 2.7 s. The CAC in Figure 8 is done by finding the equivalent homogeneous sources for both service classes and then applying the CAC procedure as described in Subsection IV-D. As we can see from Figure 7, when the traffic sources have the same parameters, the IEEE 802.11 server deals with all of them in a similar way and hence the CAC procedure admits only the number of sources that satisfy the service class with the strictest delay criterion (class 1). When no class 1 nodes are available, the IEEE 802.11 can serve more of class 2 nodes. This is a typical behavior when homogeneous traffic sources are multiplexed in an FIFO buffer [24]. Figure 7 also shows that our proposed CAC approach is in a good match with the simulation results. Figure 8 shows a comparison between the admission regions obtained by our proposed CAC approach and by computer simulations. The proposed CAC admits the number of equivalent sources that satisfy the strictest delay bound among the two classes. The figures shows that the proposed CAC algorithm based on equivalent source parameters is also in a good agreement with the simulation results. The figure is also similar to the FIFO admission region shown in [9].

VI. CONCLUSION

In this paper, we propose a new approach to achieve stochastic delay guarantees to IEEE 802.11 single hop ad hoc networks. Our approach tackles the CAC problem in IEEE 802.11 DCF in a way that resembles the classical one of finding the number of traffic sources that can be admitted in an FIFO statistical multiplexer. We present an MMPP link-layer channel model for IEEE 802.11 DCF. The model aims at characterizing the random service process variations in order to provide an effective capacity for the IEEE 802.11 DCF channel. The effective capacity model is the dual of the effective bandwidth theory. It can be used to allocate network resources in order to provide stochastic QoS guarantees for multimedia traffic sources served by a channel of time varying capacity. We also illustrate that the IEEE 802.11 behaves differently according to the traffic load in the network. Based on this illustration and by using the effective capacity model, we propose a distributed statistical CAC algorithm for IEEE 802.11 single-hop ad hoc networks. We validate the model and the algorithm by computer simulations. It is shown that the our model can be used effectively in allocating network resources and providing a stochastic guarantee for the delay bound.

ACKNOWLEDGEMENTS

This work was supported by the Natural Science and Engineering Research Council (NSERC) of Canada.

APPENDIX

In order to simplify the derivation of the service time standard deviation for a low traffic load, we assume that T_{st}, W and T_{cl} are independent random variables. The assumption is reasonable since, as the traffic load is low, the backoff window size will be minimum most of the time and hence will not have a significant effect on T_{st}. This implies that the variance of T_t conditioned on the number of nodes having backlogged packets is given by

$$Var[T_t|n] = Var[T_{st}|n] + Var[T_{cl}|n] + Var[W_n]. \quad (37)$$

Actually, T_{cl} is very small and can be ignored compared to T_{st} and the same holds for W as p is very small. Therefore, the conditional expectation of the service time in (21) is approximated by

$$E[T_t|n] \approx E[T_s t|n] = (n+1)T_s$$

and the conditional variance in (37) is approximated by

$$Var[T_t|n] \approx Var[T_{st}|n] = n(n+1)T_s^2.$$

The variance of the service time can be obtained by using

$$Var[T_t] = Var[E[T_t|n]] + E[Var[T_t|n]]. \quad (38)$$

We approximate the stationary state distribution by a binomial distribution of parameters $N-1$ and ρ in order to roughly estimate the variance. This leads to

$$E[Var[T_t|n]] \approx [(N-1)\rho((N-1)\rho - \rho + 1) + (N-1)\rho]\, T_s^2$$

and

$$Var[E[T_t|n]] \approx \left[(N-1)\rho - (N-1)\rho^2\right] T_s^2.$$

By ignoring the second order of ρ and by using (38), we obtain

$$\frac{std[T_t]}{E[T_t]} \approx \frac{\sqrt{(N-1)\rho((N-1)\rho+3)}}{(N-1)\rho+1}.$$

REFERENCES

[1] O. Tickoo and B. Sikdar, "A queueing model for finite load IEEE 802.11 random access MAC," in *Proc. ICC 2004*, vol. 1, June 2004, pp. 175–179.
[2] H. Zhai, Y. Kwon, and Y. Fang, "Performance analysis of IEEE 802.11 MAC protocols in wireless LANs," *Wireless Commun. & Mob. Comput.*, vol. 4, pp. 917–931, 2004.
[3] Y. Xiao and H. Li, "Local data control and admission control for QoS support in wireless ad hoc networks," *IEEE Trans. Veh. Technol.*, vol. 53, pp. 1558–1572, Sept. 2004.
[4] S. Valaee and B. Li, "Distributed call admission control for ad hoc networks," in *Proc. IEEE VTC'02*, Sept. 2002, pp. 1244–1248.
[5] D. Pong and T. Moors, "Call admission control for IEEE 802.11 contention access mechanism," in *Proc. IEEE Globecom'03*, Dec. 2003, pp. 3514–3518.
[6] L. Lin, H. Fu, and W. Jia, "An efficient admission control for IEEE 802.11 networks based on throughput analysis of unsaturated traffic," in *Proc. IEEE Globecom'05*, Dec. 2005, pp. 3017–3021.

[7] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE J. Select. Areas Commun.*, vol. 18, no. 3, pp. 535–547, Mar. 2000.

[8] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630-643, July 2003.

[9] M. Schwartz, *Broadband Integrated Networks*. Prentice Hall, 1998.

[10] D. Wu and R. Negi, "Effective capacity-based quality of service measures for wireless networks," in *Proc. ACM Mob. Nets. and App. (MONET)*, vol. 11, Feb. 2006, pp. 91–99.

[11] G. Kesedis, J. Walrand, and C. S. Chang, "Effective bandwidth for multiclass Markov fluids and other ATM sources," *IEEE/ACM Trans. Networking*, vol. 1, pp. 424–428, Aug. 1993.

[12] IEEE Standard for Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications, ISO/IE 8802-11: 1999(E), Aug. 1999.

[13] K. Medepalli and F. Tobagi,"System centric and user centric queueing models for IEEE 802.11 based wireless LANs," in *Proc. IEEE Broadband Networks*, vol. 1, Oct. 2005, pp. 612–621.

[14] F. Tobagi and L. Kleinrock, "Packet switching in radio channels–part IV: stability considerations and dynamic control in carrier sense multiple access," *IEEE Trans. Commun.*, vol. 25, no. 10, pp. 1103–1119, Oct. 1977.

[15] Y. Tay and K. Chua, "A capacity analysis for the IEEE 802.11 MAC protocol," *Wireless Networks*, vol. 7, Kluwer Academic Publisher, 2001, pp. 159–171.

[16] G. Berger-Sabbatel, A. Duda, M. Heusse, and F. Rousseau, "Short-term fairness of 802.11 networks with several hosts," in *Proc. 6th IFIP/IEEE Intl. Conf. Mob. Wirel. Communi. Net.*, Oct. 2004, pp. 263–274.

[17] L. X. Cai, X. Shen, J. Mark, L. Cai, and Y. Xiao, "Voice capacity analysis of WLAN with unbalanced traffic," *IEEE Trans. Veh. Technol.*, vol. 55, pp. 752–761, May 2006.

[18] A. Abdrabou and W. Zhuang, "Service time approximation IEEE 802.11 ad hoc networks," in *Proc. IEEE Infocom'07*, Anchorage, Alaska, May 2007.

[19] Y. Liu and W. Gong, "On fluid queueing systems with strict priority," *IEEE Trans. Automat. Contr.*, vol. 48, pp. 2079–2088, Dec. 2003.

[20] W. Chen, N. Jain, and S. Singh, "ANMP: ad hoc network management protocol," *IEEE J. Select. Areas Commun.*, vol. 17, no. 8, pp. 1506–1531, Aug. 1999.

[21] C.-C. Shen, C. Jaikaeo, C. Srisathapornphat, and H. Zhuochuan, "The Guerrilla management architecture for ad hoc networks," in *Proc. IEEE MILCOM'02*, vol. 1, Oct. 2002, pp. 467–472.

[22] The VINT Project. The UCB/LBNL/VINT Network Simulator-ns (version 2). http://mash.cs.berkeley.edu/ns.

[23] J. Broch, D. Maltz, D. Jonthon, Y. Hu, and J. Jetcheva, "A performance comparison of multi-hop wireless ad-hoc network routing protocols," in *Proc. ACM/IEEE Mobicom'98*, pp. 85–97.

[24] A. Berger and W. Whitt, "Extending the effective bandwidth concept to networks with priority classes," *IEEE Commun. Mag.*, vol. 36, no. 8, pp. 78–83, Aug. 1998.

Atef Abdrabou received the B.Sc. degree in 1997 and the M.Sc. degree in 2003, both in electrical engineering, from Ain Shams University, Cairo, Egypt. He is currently working toward his Ph.D. degree at the Department of Electrical and Computer Engineering, University of Waterloo, Ontario, Canada. His current research interests include quality-of-service provisioning and resource allocation for multihop ad hoc wireless networks.

Weihua Zhuang (M'93-SM'01) received the B.Sc. and M.Sc. degrees from Dalian Maritime University, China, and the Ph.D. degree from the University of New Brunswick, Canada, all in electrical engineering. In October 1993, she joined the Department of Electrical and Computer Engineering, University of Waterloo, Canada, where she is a Professor. She is a co-author of the textbook Wireless Communications and Networking (Prentice Hall, 2003), a co-recipient of a Best Paper Award from IEEE ICC 2007, a Best Student Paper Award from IEEE WCNC 2007, and the Best Paper Award from 2007 Int. Conf. Heterogeneous Networking for Quality, Reliability, Security and Robustness (QShine'07). Her current research interests include wireless communications and networks, and radio positioning.

Dr. Zhuang received the Outstanding Performance Award in 2005 and 2006 from the University of Waterloo for outstanding achievements in teaching, research, and service, and the Premier's Research Excellence Award (PREA) in 2001 from the Ontario Government for demonstrated excellence of scientific and academic contributions. She is the Editor-in-Chief of IEEE Transactions on Vehicular Technology, and an Editor of IEEE Transactions on Wireless Communications, EURASIP JOURNAL ON WIRELESS COMMUNICATIONS AND NETWORKING, and INTERNATIONAL JOURNAL OF SENSOR NETWORKS.