Modeling Queueing and Channel Access Delay in Unsaturated IEEE 802.11 Random Access MAC Based Wireless Networks

Omesh Tickoo and Biplab Sikdar, *Member, IEEE*

***Abstract*—In this paper, we present an analytic model for evaluating the queueing delays and channel access times at nodes in wireless networks using the IEEE 802.11 Distributed Coordination Function (DCF) as the MAC protocol. The model can account for arbitrary arrival patterns, packet size distributions and number of nodes. Our model gives closed form expressions for obtaining the delay and queue length characteristics and models each node as a discrete time $G/G/1$ queue. The service time distribution for the queues is derived by accounting for a number of factors including the channel access delay due to the shared medium, impact of packet collisions, the resulting backoffs as well as the packet size distribution. The model is also extended for ongoing proposals under consideration for 802.11e wherein a number of packets may be transmitted in a burst once the channel is accessed. Our analytical results are verified through extensive simulations. The results of our model can also be used for providing probabilistic quality of service guarantees and determining the number of nodes that can be accommodated while satisfying a given delay constraint.**

***Index Terms*—Delay modeling, IEEE 802.11, queueing analysis.**

I. Introduction

THE IEEE 802.11 MAC [11] has become ubiquitous and gained widespread popularity as a layer-2 protocol for wireless local area networks. While efforts have been made to support the transmission of real time traffic in such networks, they primarily use centralized scheduling and polling techniques based on the point coordination function (PCF). For ad hoc scenarios, a more reasonable model of operation is that of random access and the distributed coordination function (DCF) where it is substantially more difficult to provide delay guarantees, and the performance of the MAC protocol can easily become the bottleneck due to factors like channel contention delays and collisions. In order to provide such guarantees, it is necessary to be able to characterize the delays and other performance metrics in these networks. In this paper we focus on developing a generic analytic model for the delay and queue length characteristics in IEEE 802.11 MAC based networks in the random access mode. Based on the insights gained from this analytic framework, we then evaluate the performance of techniques to better support delay sensitive (real time) traffic.

Existing work on the performance of the 802.11 MAC has focused primarily on its throughput and capacity under saturated conditions using Markovian, mean value and fixed point analysis methods [5], [17], [12]. Work has also been conducted on improving the 802.11 MAC by using channel adaptive backoff schemes as reported in [4], [21] while [18] investigates the impact of such schemes on the traffic characteristics. The effectiveness of polling based mechanisms using the Point Coordination Function to support voice services in the 802.11 based LANs has been studied in [7], [8], [19], and [20], while [16] considers scenarios without access points. A simulation based comparison of the delays in 802.11b and 802.11e in the DCF mode is presented in [6]. A theoretical lower limit on the delay under saturated conditions for the DCF mode has been evaluated in [22] while the channel access time under saturated conditions is evaluated in [23]. Delay analysis for the PCF mode of operation has been proposed in [7], [19], [15] but no such analysis been reported for the DCF case. This paper addresses this void in the existing literature and presents analytic models for the queue characteristics in wireless network operating in the random access mode and analyzes their ability to support real time traffic.

We propose a detailed analytic model based on a discrete time $G/G/1$ queue which allows for the evaluation of the networks under consideration for general traffic arrival patterns and arbitrary number of users. Our analysis gives expressions for the probability generating function for the queue lengths and the delays. Thus, probabilistic service guarantees in terms of both the delays and packet loss probabilities can be evaluated and used for purposes like call admission control and providing statistical delay bounds. The results of the queueing model can also be used to evaluate the number of connections that can be supported for a given delay or loss constraint. The key to the model is the characterization of the service time distribution which needs to account for the channel access time resulting from the random access mechanism. Our model accounts for the collision avoidance and exponential backoff mechanism of 802.11, the delays in the channel access due to other nodes transmitting and the delays caused by collisions. The results obtained from this model have been verified through extensive simulations.

This paper also evaluates the effectiveness of some techniques to reduce the delays in the network red that arise due the channel access time in multiple-access protocols. In particular, we evaluate the proposal of IEEE 802.11e where a node on successfully accessing the channel, is allowed to send

Manuscript received January 1, 2006; revised January 1, 2007; first published February 25, 2008; last published August 15, 2008 (projected); approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor R. Srikant. This work was supported in part by the National Science Foundation under Grant 0313095 and by Intel Corporation. This paper was presented in part at the IEEE INFOCOM 2004, Hong Kong.

O. Tickoo was with the Department of Electrical, Computer and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180 USA. He is now with Intel Corporation, Hillsboro, OR 97124 USA (e-mail: omesh.tickoo@intel.com).

B. Sikdar is with the Department of Electrical, Computer and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180 USA (e-mail: sikdab@rpi.edu).

Digital Object Identifier 10.1109/TNET.2007.904010

M consecutive packets instead of one, thereby reducing the delay arising from the channel access by a factor of $M-1$. We extend our queueing model to account for this variation of the MAC protocol and derive expressions for obtaining the delay characteristics in IEEE 802.11 networks with "collision free bursts". The collision free bursts also smoothen the fine time scale burstiness of the traffic thereby further aiding in the reduction of the delays and losses. Simulations have been used to verify the effectiveness of this mechanism and are presented in the paper.

The rest of the paper is organized as follows. In Section II we present a brief overview of the IEEE 802.11 MAC protocol. In Section III we present the detailed queueing model and present the simulation results to verify the model. Section V presents the extension of the model to the proposals for collision free bursts in IEEE 802.11e. Finally, Section VI presents a discussion of the results and concluding remarks.

II. Overview of the IEEE 802.11 MAC

The IEEE 802.11 MAC layer is responsible for a structured channel access scheme and is implemented using a Distributed Coordination Function based on the Carrier Sense Medium Access with Collision Avoidance (CSMA/CA) protocol. An alternative to the DCF is also provided in the form of a Point Coordination Function which is similar to a polling system for determining the user having the right to transmit. We only describe the relevant details of the DCF access method and refer the reader to [11] for other details on the IEEE 802.11 standard.

The CSMA/CA based MAC protocol of IEEE 802.11 is designed to reduce the collisions due to multiple sources transmitting simultaneously on a shared channel. In a network employing the CSMA/CA MAC protocol, each node with a packet to transmit first senses the channel to ascertain whether it is in use. If the channel is sensed to be idle for an interval greater than the Distributed Inter-Frame Space (DIFS), the node proceeds with its transmission. If the channel is sensed as busy, the node defers transmission till the end of the ongoing transmission. The node then initializes its *backoff timer* with a randomly selected *backoff interval* and decrements this timer every time it senses the channel to be idle. The timer has the granularity of a *backoff slot* (which we denote by δ) and is stopped in case the channel becomes busy and the decrementing process is restarted when the channel becomes idle for a DIFS again. The node is allowed to transmit when the backoff timer reaches zero. Since the backoff interval is chosen randomly, the probability that two or more stations will choose the same backoff value is very low. The details of the exact implementation of the backoff mechanism are described in Section III-A. Along with the Collision Avoidance, 802.11 uses a positive acknowledgment (ACK) scheme. All the packets received by a node implementing 802.11 MAC must be acknowledged by the receiving MAC. After receiving a packet the receiver waits for a brief period, called the Short Inter-Frame Space (SIFS), before it transmits the ACK.

There is another particular feature of wireless local area networks (LANs), known as the "hidden node" problem, that the 802.11 MAC specification addresses. Two stations that are not within hearing distance of each other can lead to collisions at a third node which receives the transmission from both sources. To take care of this problem, 802.11 MAC uses a reservation

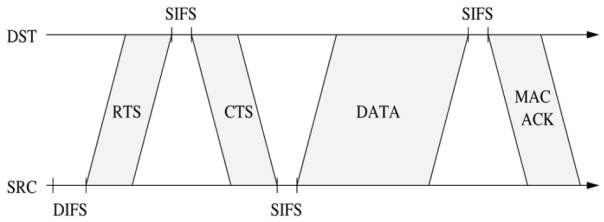

Fig. 1. Basic operation of the CSMA/CA protocol.

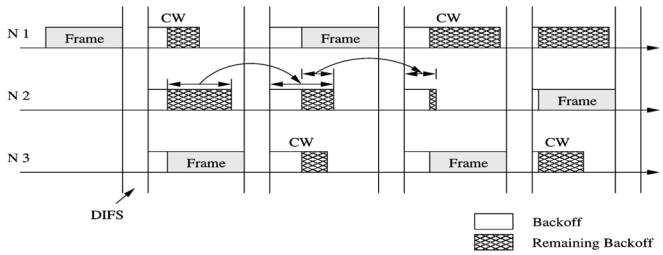

Fig. 2. Backoff mechanism of 802.11 MAC. The frame transmission time includes the RTS/CTS exchange and the MAC layer ACK. CW = Contention Window.

based scheme. A station with a packet to transmit sends an Ready To Send (RTS) packet to the receiver and the receiver responds with a Clear To Send (CTS) packet if it is willing to accept the packet and is currently not busy. This RTS/CTS exchange, which also contains timing information about the length of the ensuing transaction, is detected by all the nodes within hearing distance of either the sender or receiver or both and they defer their transmissions till the current transmission is complete.

The basic operation of the CSMA/CA based MAC protocol of IEEE 802.11 is shown in Fig. 1 and it shows the exchange of various packets involved in each successful transmission and the spacing between these packets.

III. Queueing Model for the 802.11 DCF

In this section we introduce a discrete time $G/G/1$ queue for modeling nodes in a random access network based on the 802.11 MAC. We assume a network with N nodes using the DCF of IEEE 802.11 to schedule their transmissions. We assume the use of RTS and CTS messages for channel reservation. The analysis can be easily extended for the cases where such messages are absent. The packet arrival process and the length of each packet is assumed to be arbitrary and the channel transmission rate is C bits/sec. Finally, the paper does not consider the hidden node problem.

A. Modeling the Backoff Mechanism

In order to model the MAC layer queueing delays and losses, we first analyze the exponential back-off scheme of 802.11 MAC protocol's Collision Avoidance mechanism. In Fig. 2 we show the details of this backoff mechanisms. With multiple nodes contending for the channel, once the channel is sensed idle for a DIFS, each node with a packet to transmit decrements its backoff timer. The node whose timer expires first begins transmission and the remaining nodes stop their timers and defer their transmission. Once the current node finishes transmission, the process repeats again and the remaining nodes start decrementing their timer from where they left off.

In the following analysis we denote the probability that an arbitrary packet transmission, or an RTS transmission if RTS-CTS exchange is used, results in a collision by p (since hidden nodes are not considered in this paper, there are no CTS collisions). The lower and upper bounds on the contention window associated with backoffs are denoted by $\mathrm{CW}_{\min}$ and $\mathrm{CW}_{\max}$ and we use the notation $m = \log_2(\mathrm{CW}_{\max}/\mathrm{CW}_{\min})$. Once a node goes into collision avoidance or the exponential back-off phase, we denote the number of slots that it waits beyond a DIFS period before initiating transmission by BC. This back-off counter is calculated from

$$\mathrm{BC} = \texttt{int}(\texttt{rnd}() \cdot \texttt{CW(k)}) \tag{1}$$

where the function `rnd()` returns a pseudo-random number uniformly distributed in $[0, 1]$ and $\mathrm{CW}(k)$ represents the contention window after k unsuccessful transmission attempts. Note that in case the `int()` operation is done using a `ceil()` function, the effective range for BC becomes $1 \leq \mathrm{BC} \leq \mathrm{CW}(k)$ since the probability of `rnd()` $= 0$ is 0 assuming a continuous distribution. For the rest of this paper we assume that a ceil() function is used to do the `int()` operation.

The first attempt at transmitting a given packet is performed assuming a CW value equal to the minimum possible value of $\mathrm{CW}_{\min}$ [11]. For each unsuccessful attempt, the value of CW is doubled until it reaches the upper limit of $\mathrm{CW}_{\max}$ specified by the protocol. Then, at the end of k unsuccessful attempts, $\mathrm{CW}(k)$ is given by

$$\mathrm{CW}(k) = \min(\mathrm{CW}_{\max}, 2^{k-1}\mathrm{CW}_{\min}). \tag{2}$$

Also, let the probability that a transmission attempt is unsuccessful, i.e., the probability of a collision be denoted by p. Then, the probability that $\mathrm{CW} = W$ is given by

$$\Pr\{\mathrm{CW} = W\} = \begin{cases} p^{k-1}(1-p) & \text{for } W = 2^{k-1}\mathrm{CW}_{\min} \\ p^m & \text{for } W = \mathrm{CW}_{\max} \end{cases} \tag{3}$$

where $k \leq m$. Note that the second case ($W = \mathrm{CW}_{\max}$) includes all cases where the number of collisions is greater than m. The probability that back-off counter $\mathrm{BC} = i, 1 \leq i \leq \mathrm{CW}_{\max}$, is then given by

$$\Pr\{BC = i\} = \begin{cases} \left[\sum_{k=0}^{m-1} \frac{p^k(1-p)}{2^k CW_{\min}} + \frac{p^m}{CW_{\max}}\right] & 1 \leq i \leq CW_{\min} \\ \left[\sum_{k=j}^{m-1} \frac{p^k(1-p)}{2^k CW_{\min}} + \frac{p^m}{CW_{\max}}\right] & 2^{j-1}CW_{\min} + 1 \leq i \leq 2^j CW_{\min} \\ \frac{p^m}{CW_{\max}} & 2^{m-1}CW_{\min} + 1 \leq i \leq CW_{\max}. \end{cases} \tag{4}$$

In [17], [18] the collision probability p was derived for the saturated network case where each node always has a packet to send and each incoming packet is immediately backlogged. In this paper, we develop a model to obtain an expression for the collision probabilities in the general, unsaturated case. In the saturated case where each packet is backlogged immediately, each packet starts out with a window of $CW_{\min}$. With probability $1 - p$ the transmission is successful and the average backoff window of such a packet is $CW_{\min}/2$. With probability $p(1-p)$ the first transmission fails and the packet is successfully transmitted in the second attempt (using a backoff window of $2CW_{\min}$) which adds $CW_{\min}$ to the average backoff window seen by the packet. Continuing along these lines for cases with larger number of losses, the average backoff window seen by packets at the nodes when the node experiences a collision rate of p is given by

$$\begin{aligned} \bar{W} &= (1-p)\frac{CW_{\min}}{2} + p(1-p)\frac{2CW_{\min}}{2} + \cdots \\ &\quad + p^m(1-p)\frac{2^m CW_{\min}}{2} + p^{m+1}\frac{2^m CW_{\min}}{2} \\ &= \frac{1-p-p(2p)^m}{1-2p}\frac{CW_{\min}}{2}. \end{aligned} \tag{5}$$

The equation above may also be used for relating the collision rate to the average window size for non-saturated cases with a rather small error creeping in due to the fact that in non-saturated nodes, it is not necessary that all packets experience backoff at least once. We also note that in the IEEE 802.11 standards [11] Sections 9.2.5.2 and 9.2.5.5 specify conditions where the backoff process should be invoked even for the first attempt at transmitting a packet. Also, all nodes must perform a backoff after every transmission with the more fragments bit set to 0, even if there are no packets currently queued up. These increase the likelihood that an arbitrary packet arrival experiences some backoff slots before it is transmitted. Finally, at low loads where some errors may be introduced by (5), the packet transmission time rather than the backoff time dominates the packet delay, making the impact of such errors quite small. Consequently, we use (5) to characterize the average window size for a given collision probability. Note however, that the collision probability is a function of the load at each node and we proceed to evaluating it.

Now consider a network with N nodes operating in discrete time where the packet arrival rate at each node is λ packets per slot, the packet service rate of the network is denoted by μ packets per slot and the queue utilization at a node is denoted by ρ. To evaluate the collision probabilities when the nodes are unsaturated, we consider a tagged node which transmits in a given slot. Now, a collision occurs if one or more of the remaining $N-1$ nodes also transmit in this slot. Then, letting $P[SE]$ denote the probability that a node does not transmit in a slot, we have

$$p = 1 - P[\mathrm{SE}]^{N-1} \tag{6}$$

where we have used the widely used decoupling approximation [1], [17] which assumes that the event that a node does not transmit in a slot is independent of similar decisions by the other nodes. Now, using QE to represent "queue empty" and QNE to denote "queue not empty" for ease of notation, $P[SE]$ is given by

$$\begin{aligned} P[SE] &= P[SE \,|\, \mathrm{QE}]P[\mathrm{QE}] + P[SE \,|\, \mathrm{QNE}]P[\mathrm{QNE}] \\ &= 1 \cdot (1-\rho) + \rho P[SE \,|\, \mathrm{QNE}]. \end{aligned}$$

since if a queue is empty, it does not transmit with probability 1 and the probability that a queue is empty is given by $1-\rho$. Note that a queue is non-empty in a slot either if it is backlogged or if a new arrival occurs in that slot while the queue was empty.

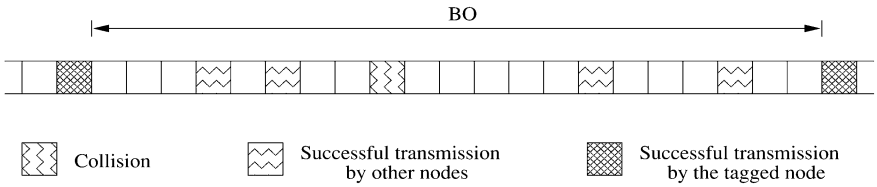

Fig. 3. Interleaving of transmissions and collisions contributing to the service time.

Now, considering the fact that we are interested in stable queues and backoff slots are two orders of magnitude smaller than typical data packet lengths, the probability of the latter case is quite small. Also, a backlogged queue will not transmit in a slot with probability $(\bar{W}-1)/\bar{W}$. Then, $P[SE \mid \text{QNE}]$ can be approximated by $(\bar{W}-1)/\bar{W}$. Consequently,

$$P[SE] = (1-\rho) + \rho\frac{\bar{W}-1}{\bar{W}} = 1 - \frac{\rho}{\bar{W}} \tag{7}$$

and combining (5), (6) and (7) the loss rate p is given by

$$p = 1 - \left(1 - \rho\frac{(1-2p)}{1-p-p(2p)^m}\frac{2}{\text{CW}_{\min}}\right)^{N-1}. \tag{8}$$

To determine ρ, we now characterize the average time to serve a packet. For each packet, the node spends $\bar{W}$ slots in backoff. Also, with the long term fairness of exponential backoff, in the case where all nodes have the same traffic arrival rates, on an average $\rho(N-1)$ transmissions from other nodes occur between two transmissions from the tagged node. This contributes $\rho(N-1)T_S$ slots to the service time where T_S is the average length of a packet in units of backoff slots. Now, with each packet transmission resulting in a collision with probability p, the average number of collisions per successful transmission is given by $p/(1-p)$. The contribution due to the collisions of packets of other nodes is thus given by $\rho(N-1)T_C p/2(1-p)$ where T_C is the time of a collision in units of slots and the factor of 2 in the denominator represents the first degree approximation that only two nodes are involved in a collision. Finally, adding the time to transmit the packet of the tagged node (T_S), its backoff time $(\bar{W})$, and any collision that it may have, we get,

$$\frac{1}{\mu} = \rho(N-1)\left[T_S + T_C\frac{p}{2(1-p)}\right] + \bar{W} + T_S + T_C\frac{p}{2(1-p)} \tag{9}$$

Then, using the fact that $\rho = \lambda/\mu$ for a stable system and substituting $1/\mu = \rho/\lambda$ and $\bar{W}$ from (5) in the equation above, we have

$$\rho = \frac{\lambda\left[T_S + T_C\frac{p}{2(1-p)}\right] + \lambda\frac{(1-p-p(2p)^m)}{1-2p}\frac{CW_{\min}}{2}}{1-\lambda(N-1)\left[T_S + T_C\frac{p}{2(1-p)}\right]}. \tag{10}$$

We can now substitute ρ in (8) to obtain p by solving

$$p = 1 - \left(1 - \frac{\lambda + \lambda\left[T_S + T_C\frac{p}{2(1-p)}\right]\frac{1-2p}{(1-p-p(2p)^m)}\frac{2}{CW_{\min}}}{1-\lambda(N-1)\left[T_S + T_C\frac{p}{2(1-p)}\right]}\right)^{N-1}. \tag{11}$$

B. Queueing Model

To obtain the delays and losses experienced by a packet at each node, we model the system as a discrete time $G/G/1$ queue. The unit of time or the slot length corresponds to the length δ of a backoff slot. Note that in real networks the packet arrival process may be a continuous time process and we account for the fact that the arrival may occur anywhere in the slot. Also, since δ is of the order of 20 μsec, the error introduced by the discretization is quite small. We denote by $a(n)$ the probability that n messages arrive in a given slot at a given node with the corresponding probability generating function (pgf) $A(z)$. Also, $b(n)$ denotes the the probability that the service time of a packet takes n slots with the corresponding pgf $B(z)$. Now, $b(n)$ depends on the number of nodes contending for the channel as well as the packet length distribution and we now characterize its distribution.

We define the service time of a packet to be the time from the instant the packet reaches the head of the queue in the node to the instant it successfully departs from the queue. Thus it has two components: (1) the time till the node successfully accesses and reserves the channel for use and (2) the time required to transmit the packet. While the second part is essentially characterized by the packet length distribution, the first part needs a more detailed analysis. To characterize the time required to successfully access the channel, we refer to Fig. 3. Between any two successful transmissions by a tagged node, other nodes may successfully transmit a number of packets or may be involved in a number of collision, each of which add to the channel access time of the tagged node. Note that transmission attempts by the tagged node which result in collisions are also included in this access time characterization.

We first characterize the number of backoff slots that the tagged node has to wait between two successful transmissions. When a packet comes in and finds that the system is empty, it directly proceeds with a transmission and if successful, depart without experiencing any backoff slots. Thus, the probability that the number of backoff slots, BO, is zero is approximated by $P[\text{BO}=0] = (1-\rho)(1-p)$. Now with probability ρ the packet goes into backoff at least once. Now, note that if the tagged node successfully transmits the packet in its first attempt (with probability $1-p$) the number of backoff slots is uniformly distributed between $1,\ldots,CW_{\min}$. In case of a successful transmission after a single collision (with probability $p(1-p)$), the probability mass function (pmf) of the number of backoff slots is obtained through $U_{1,\text{CW}_{\min}} * U_{1,2\text{CW}_{\min}}$ and so on, where $U_{a,b}$ denotes a uniform distribution between a and b and $*$ represents the convolution operation. For a sequence of k, $k > m$, successive collisions for the same packet, we have k convolutions the first m of which are $U_{1,\text{CW}_{\min}}, U_{1,2\text{CW}_{\min}}, \ldots, U_{1,2^m\text{CW}_{\min}}(i)$ while the remaining terms are $U_{1,2^m\text{CW}_{\min}}(i)$ since the backoff

window is constrained by $CW_{\max} = 2^m \mathrm{CW}_{\min}$. Then, the probability the tagged node experiences i backoff slots, $i > 0$, is given by

$$\begin{aligned} P[\mathrm{BO} = i] = \rho[&(1-p)U_{1,\mathrm{CW}_{\min}}(i) + p(1-p) \\ &\times [U_{1,\mathrm{CW}_{\min}} * U_{1,2\mathrm{CW}_{\min}}(i)] + \cdots + p^m(1-p) \\ &\times [U_{1,\mathrm{CW}_{\min}} * U_{1,2\mathrm{CW}_{\min}} * \cdots * U_{1,2^m\mathrm{CW}_{\min}}(i)] \\ &+ p^{m+1}(1-p)[U_{1,\mathrm{CW}_{\min}} * \cdots * U_{1,2^m\mathrm{CW}_{\min}} * \\ &\times U_{1,2^m\mathrm{CW}_{\min}}(i)] + \cdots] \end{aligned} \tag{12}$$

with the corresponding pgf $\mathrm{BO}(z)$. Note that the maximum number of retransmission attempts allowed for each packet is governed by the long retry count (SLRC) (short retry count (SSRC) for transmissions without the RTS-CTS exchange) which forms the limit on the summation above. However, its effect may be neglected since the term $p^k(1-p)$ becomes negligibly small as k increases.

To evaluate the service time seen by a packet waiting at the tagged node, we now characterize how many of the backoff slots experienced by it were followed by collisions or successful transmissions by other nodes. We term such slots as active slots. Now, since the average window size is $\bar{W}$ ((5)), and a queue is active with probability ρ, the probability that a node attempts a transmission in an arbitrary slot is given by $\rho/\bar{W}$. Note that since we are looking at a backoff slot between two successful transmissions from the tagged node, if the tagged node transmits in any of the backoff slots the slot must be accompanied by a collision. Denote by "TX n" and "TX other" the event that the tagged node transmits in an arbitrary slot and the event that at least one of the remaining $N-1$ nodes transmits in an arbitrary slot, respectively. Then, the probability that a given slot is active (i.e., contains a transmission attempt by at least one of the N nodes and in case the tagged node transmits it experiences a collision), q, is given by

$$\begin{aligned} q &= P[\mathrm{TX\ other}](1 - P[\mathrm{TX}n]) \\ &\quad + P[\mathrm{TX\ other}]P[\mathrm{TX}n] \\ &= P[\mathrm{TX\ other}] \\ &= 1 - \left(1 - \frac{\rho}{\bar{W}}\right)^{N-1}. \end{aligned} \tag{13}$$

Then, given that the tagged node experiences i backoff slots before it successfully transmits a packet, the pmf of the number of active slots within the backoff slots is given by

$$P[j\mathrm{slots\ active} \,|\, \mathrm{BO} = i] = \binom{i}{j} q^j (1-q)^{i-j} \tag{14}$$

for $j = 0, \ldots, i$. We next obtain the probability that a slot results in a collision given that it is active, q_c. A collision can occur in an active slot in one of two ways: 1) the tagged node transmits and at least one of the other nodes also transmits in the slot or 2) the tagged node does not transmit in the slot but two or more of the other nodes do. Now we know that if an active slot contains a transmission by the tagged node, it results in a collision i.e., at least one additional node also transmits in the slot. Then q_c is given by

$$\begin{aligned} q_c = P[\mathrm{collision} \,|\, \mathrm{slot\ active}] &= \frac{P[\mathrm{collision,\ active}]}{P[\mathrm{slot\ active}]} \\ &= \frac{\left(1-\frac{\rho}{\bar{W}}\right)\left[1 - \left(1-\frac{\rho}{\bar{W}}\right)^{N-1} - \frac{(N-1)\rho}{\bar{W}}\left(1-\frac{\rho}{\bar{W}}\right)^{N-2}\right]}{1 - \left(1-\frac{\rho}{\bar{W}}\right)^{N-1}} \\ &\quad + \frac{\frac{\rho}{\bar{W}}\left[1 - \left(1-\frac{\rho}{\bar{W}}\right)^{N-1}\right]}{1 - \left(1-\frac{\rho}{\bar{W}}\right)^{N-1}} \\ &= \frac{1 - \left(1-\frac{\rho}{\bar{W}}\right)^{N-1} - \frac{(N-1)\rho}{\bar{W}}\left(1-\frac{\rho}{\bar{W}}\right)^{N-1}}{1 - \left(1-\frac{\rho}{\bar{W}}\right)^{N-1}}. \end{aligned} \tag{15}$$

Thus the probability that out of j active slots k result in collisions is given by

$$P[k\mathrm{collisions} \,|\, j\ \mathrm{active\ slots}] = \binom{j}{k} q_c^k (1-q_c)^{j-k}. \tag{16}$$

Now, each collision is of duration $T_C = \mathrm{DIFS} + \tau_{\mathrm{RTS}}$ where τ_{RTS} is the time required to transmit an RTS packet. Thus each collision between two transmissions from the tagged node adds T_C slots to the service time at the tagged node. Note that in situations where RTS-CTS packets are not used to reserve the channel, the duration of a collision is given by $T_C = \mathrm{DIFS} + \tau_{\mathrm{pkt}}$ where τ_{pkt} is the packet transmission time. Also, each successful transmission by other nodes between the two successful transmissions of the tagged node adds a time proportional to the packet length of the transmitted packet to the service time at the tagged node. In our analysis we allow for general packet length distributions and the probability that a packet transmission takes ν slots (which is dependent on the packet length and the channel rate and includes the duration of the RTS, CTS and ACK exchange) is denoted by $l(\nu)$ with the corresponding pgf $L(z)$. Then, the contribution of j successful transmissions to the service time of the tagged node is given by

$$P\left[\sum^{j} \mathrm{pkt\ time} = u\right] = l * l * \cdots * l(u) = l^{(j)}(u) \tag{17}$$

where $l^{(j)}()$ represents the $j-$fold convolution of $l(\nu)$. Note that in the expression above, we have assumed that the all nodes have the same packet size distribution. Analysis for the more general case of arbitrary packet size distributions at different nodes is presented in Section IV.

Consider a scenario where the tagged packet experiences i backoff slots of which j are active and among these j active slots, k slots have collisions. If the $j-k$ successful packet transmission by the other nodes contribute u slots, then the pmf of the conditional channel access time for the successful transmission of the tagged packet, Y, is given by

$$P[Y = s \,|\, i, j, k] = \begin{cases} l^{(j-k)}(u) & s = i + kT_C + u \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

where $l^{(j-k)}()$ represents the $j-k$-fold convolution of the packet size distributions of the $j-k$ successful transmissions.

Now, the joint probability distribution of i backoff slots, j active slots and k collisions, $P[i,j,k]$ can be evaluated using

$$P[i,j,k] = P[j,k \mid i]P[i]. \tag{19}$$

The pmf of the backoff slots, i, is $P[i] = P[\mathrm{BO} = i]$ and is given in (12). Also, by combining (14) and (16) which characterize the number of active slots given i and the number of collisions in these active slots, respectively, we have

$$P[j,k \mid i] = \binom{i}{j} q^j (1-q)^{i-j} \binom{j}{k} q_c^k (1-q_c)^{j-k}. \tag{20}$$

The probability mass function of the service time Y can now be obtained by unconditioning (18) on i, j and k using the expressions for $P[j,k \mid i]$ and $P[i]$:

$$\begin{aligned} P[Y=s] &= \sum_{i,j,k} P[Y=s \mid i,j,k]P[i,j,k]I(s) \\ &= \sum_{i}^{\infty}\sum_{j}^{i}\sum_{k}^{j} l^{(j-k)}(u)P[j,k \mid i]P[\mathrm{BO}=i]I(s) \\ &= \sum_{i}^{\infty}\sum_{j}^{i}\sum_{k}^{j} \left[l^{(j-k)}(u) \binom{i}{j} q^j (1-q)^{i-j} \right. \\ &\quad \left. \times \binom{j}{k} q_c^k (1-q_c)^{j-k} P[\mathrm{BO}=i]I(s) \right] \end{aligned} \tag{21}$$

where $I(s)$ is an indicator function which equals 1 when $s = u + i + kT_C$ and 0 otherwise. Note that the above expression needs to be evaluated for all possible values of i, j and k which result in a given value of s. As described in the derivation for $P[\mathrm{BO} = i]$ in (12), the number of possible backoff slots i in the expression above extends to infinity because we have not considered the fact that a packet may be dropped by the MAC layer after a certain number of unsuccessful retransmissions. However, as argued before, the error caused by this is quite negligible since the probability of a packet experiencing an extremely large number of collisions and thus backoff slots is extremely small (i.e., $p^k(1-p) \to 0$ as k increases). The pgf of the final service time, $B(z)$, is then obtained by the convolution of the channel access time $(Y(z))$ and the length of the packet to be served (l) and is given by

$$B(z) = Y(z)L(z). \tag{22}$$

Using standard discrete time queueing theory [3, Ch. 1, (1.21), p. 14], the pgf of the system occupancy of the $G/G/1$ queue at random slot boundaries (beginning of a slot), $U(z)$, is given by

$$U(z) = [1 - A'(1)B'(1)]\frac{(z-1)B(A(z))}{z - B(A(z))} \tag{23}$$

and the pgf of the integer part of the system time (where system time is defined as the total time spent in the system from the arrival instant to the service completion time) can be shown to be [3, Ch. 1, (1.59), p. 31]

$$V_{\mathrm{int}}(z) = \frac{[1 - A'(1)B'(1)](z-1)B(z)[1 - A(B(z))]}{A'(1)[1 - B(z)][z - A(B(z))]} \tag{24}$$

Allowing arrivals to occur at any point in the slot, we denote the distance of the arrival point from the start of the slot by F with mean $\bar{F}$. This adds a fractional component to the system time of $V_{\mathrm{frac}} = 1 - F$. The total system time is then given by $V = V_{\mathrm{int}} + V_{\mathrm{frac}}$ whose mean can be expressed as [3, Ch. 1, (1.63), p. 31]

$$\bar{V} = 1 - \bar{F} + B'(1) + \frac{[A'(1)]^2 B''(1) + A''(1)B'(1)}{2A'(1)[1 - A'(1)B'(1)]}. \tag{25}$$

The average queue size at each node can then be obtained using Little's law and is given by

$$\bar{Q} = A'(1)\bar{V}. \tag{26}$$

Equation (25) can now be solved to obtain the number of nodes that can be supported for arbitrary arrival traffic patterns while providing a specified delay guarantee.

Note that the second moment of the time a packet spends in the system is given by

$$\begin{aligned} \overline{V^2} &= E[(V_{\mathrm{int}} + V_{\mathrm{frac}})^2] \\ &= E[V_{\mathrm{int}}^2] + E[V_{\mathrm{frac}}^2] + 2E[V_{\mathrm{int}}]E[V_{\mathrm{frac}}] \end{aligned} \tag{27}$$

To obtain the terms in the equation above, we note that

$$E[V_{\mathrm{frac}}] = 1 - \bar{F} \tag{28}$$

$$E\left[V_{\mathrm{frac}}^2\right] = 1 + \overline{F^2} - 2\bar{F} \tag{29}$$

and after differentiating $V_{\mathrm{int}}(z)$ given by (24) once and twice, respectively, and taking the limit as $z \to 1$, we obtain

$$E[V_{\mathrm{int}}] = B'(1) + \frac{[A'(1)]^2 B''(1) + A''(1)B'(1)}{2A'(1)[1 - A'(1)B'(1)]} \tag{30}$$

$$\begin{aligned} E\left[V_{\mathrm{int}}^2\right] &= \frac{A''(1)B''(1)\left[1 + A'(1)B'(1)\right]}{2A'(1)[1 - A'(1)B'(1)]^2} \\ &\quad + \frac{[A''(1)]^2[B'(1)]^3 + [A'(1)]^3[B''(1)]^2}{2A'(1)[1 - A'(1)B'(1)]^2} \\ &\quad + \frac{A'''(1)[B'(1)]^2 + [A'(1)]^2 B'''(1)}{3A'(1)[1 - A'(1)B'(1)]} \\ &\quad + \frac{A''(1)[B'(1)]^2 + A'(1)B''(1)}{A'(1)[1 - A'(1)B'(1)]} \\ &\quad + B'(1) + \frac{[A'(1)]^2 B''(1) + A''(1)B'(1)}{2A'(1)[1 - A'(1)B'(1)]} \end{aligned} \tag{31}$$

Higher order moments of the total time that a packet spends in the system can similarly be obtained by differentiating $V_{\mathrm{int}}(z)$ an appropriate number of times and taking the limit as $z \to 1$.

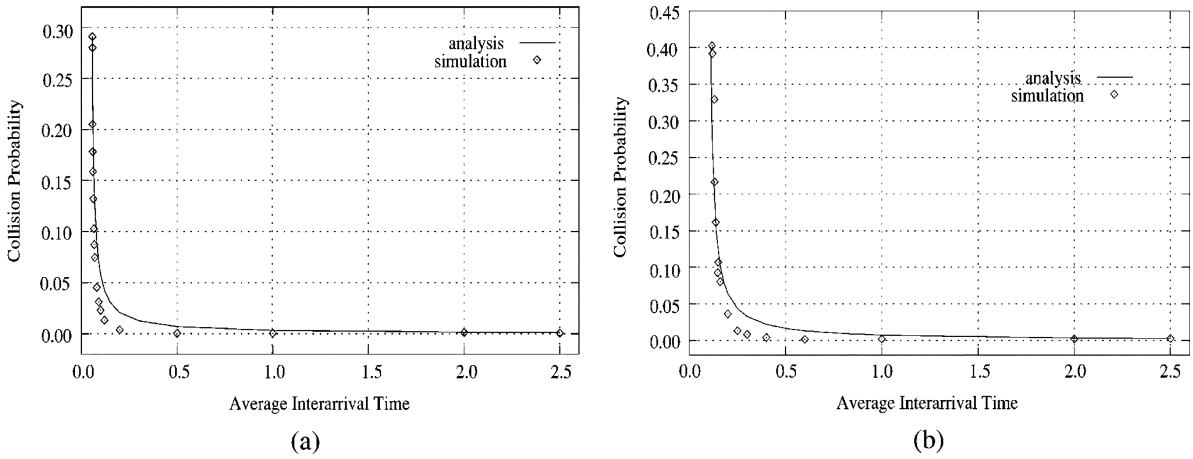

Fig. 4. Comparison of the collision probabilities. (a) 10 nodes. (b) 20 nodes.

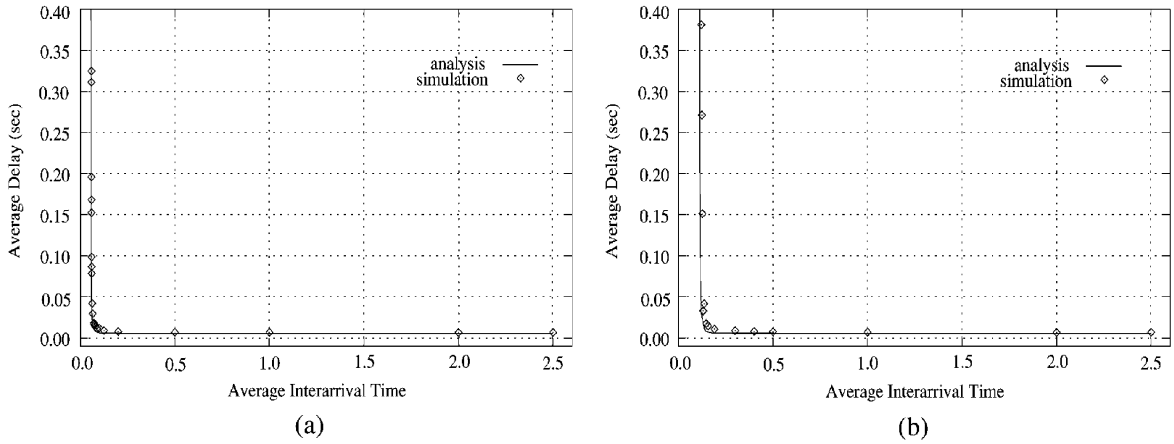

Fig. 5. Comparison of the average packet delays. (a) 10 nodes. (b) 20 nodes.

C. Simulation Results

To validate our analytic model, we conducted extensive simulations using the ns-2 simulator [9] for different network topologies, number of nodes as well as the load on the network. In this section, we report on our simulation results for the case of 10 and 20 nodes and omit the others since they are similar. The simulations for the results reported in this section were carried out for a rectangular region of 670×670 meters and the nodes were randomly distributed over this region. The routing protocol used for the simulations was Ad-hoc On-demand Distance Vector routing (AODV) [14]. We also verified our results for routing using Destination Sequenced Distance Vector (DSDV) [13]. The interface queues at each node used a Droptail policy and the interface queue length was set at 5000 packets. All sources and receivers have an omni-directional antenna of height 1.5 m with transmitter and receiver gains of 1 each. The simulations were run for a simulated time of 1800 seconds. All other parameter settings for the physical and MAC layers for these simulations are given in Table I.

Each node was the source for one flow as well as the sink for another flow. Thus the 10 node case corresponds to 10 flows while the 20 node case had 20 active flows. The arrival process at each node, $(a(n))$, was assumed to follow the distribution

$$a(n) = \begin{cases} 1-\gamma & n = 0 \\ \gamma & n = 1 \end{cases} \tag{32}$$

TABLE I
SIMULATION SETTINGS

Physical Layer		802.11 MAC	
Propagation	2 ray gnd	RTS size	44 bytes
Channel	Wireless	CTS size	38 bytes
Rx Threshold	3.652e-10	DIFS	50 μsec
Bandwidth	2 Mbps	SIFS	10 μsec
Frequency	914 MHz	Slot size	20 μsec
Loss Factor	1.0		

resulting in an average inter-arrival time of $1/\gamma$. The sources used UDP as the transport protocol and the packet sizes were assumed to be 1000 bytes.

In Fig. 4 we compare the results for the collision probabilities as obtained from the simulations with those obtained from our analysis (the expression in (11)). The results are plotted for both the 10 node as well as the 20 node case. In both cases, we see the close match between the analytic and simulation results with a small deviation in the knee region. We also note that when the nodes become saturated, the expression for the collision probability p reduces to the expressions in [17] and [1] (which only consider the saturated case) and in these scenarios, our results are consistent.

Fig. 5 compares the simulation and analytic results for the average delays for the 10 and 20 node cases. For both scenarios, we see the close match between the analytic and the simulation

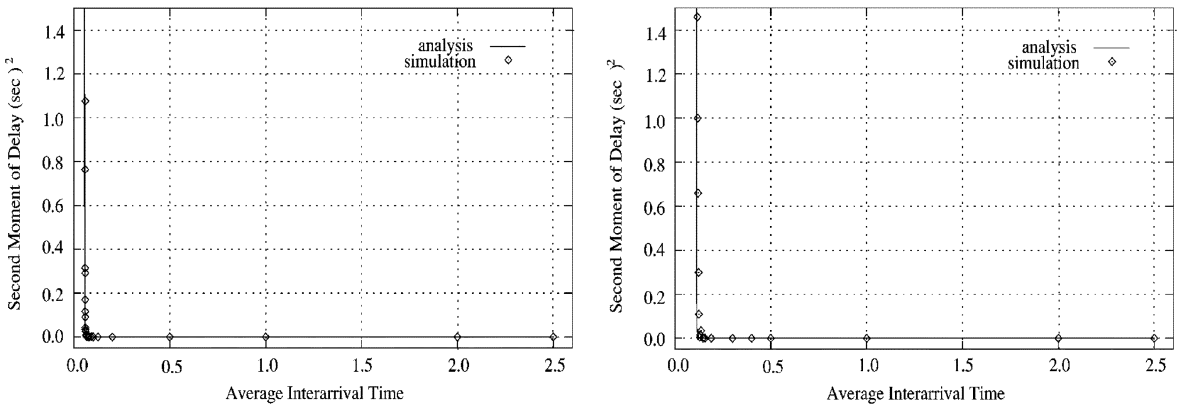

Fig. 6. Comparison of the second moment of packet delays. (a) 10 nodes. (b) 20 nodes.

results. As expected, the system saturates more quickly for the 20 node case, at approximately half the load of the 10 node case. Similar results were also obtained for other topologies and network sizes, validating the analytic model for the delay in an 802.11 based network.

Finally, Fig. 6 compares the simulation and analytic results for the second moment of the delays experienced by the packets. For both the 10 and 20 node cases, we note the close match between the simulation and analytic results.

IV. Extension to Heterogeneous Traffic

In the previous section, we assumed that the traffic arrival rates and the distribution of the packet sizes were the same at all nodes. In this section, we now extend this analysis to consider heterogeneous traffic conditions at the nodes.

We again assume that there are N nodes in the network and we use the same notation as used in the previous section but add a subscript to make it node specific. We denote the packet arrival rate at each node and its utilization by λ_n and $\rho_n, 1 \le n \le N$, respectively. We also denote by $p_n, 1 \le n \le N$ the probability that an arbitrary transmissions attempt of node n experiences a collision. Then, the average backoff window of node n is given by

$$\bar{W}_n = \frac{1 - p_n - p_n(2p_n)^m}{1 - 2p_n} \frac{\mathrm{CW}_{\min}}{2}. \tag{33}$$

Following the derivation of Section III, the probability that node n does not transmit in an arbitrary slot is given by

$$P_n[SE] = 1 - \frac{\rho_n}{\bar{W}_n} \tag{34}$$

and thus the collision rate experienced by packets from node n is given by

$$p_n = 1 - \prod_{\substack{i=1 \\ i \neq n}}^{N} \left(1 - \rho_i \frac{(1-2p_i)}{1 - p_i - p_i(2p_i)^m} \frac{2}{\mathrm{CW}_{\min}}\right). \tag{35}$$

Similarly, the average time to serve a packet from node n is given by

$$\frac{1}{\mu_n} = \sum_{\substack{i=1 \\ i \neq n}}^{N} \rho_i \left[T_{S_i} + T_{C_i} \frac{p_i}{2(1-p_i)}\right] + \bar{W}_n + T_{S_n} + T_{C_n} \frac{p_n}{2(1-p_n)} \tag{36}$$

and substituting $1/\mu_n = \rho_n/\lambda_n$ in the equation above, we have

$$\rho_n = \sum_{\substack{i=1 \\ i \neq n}}^{N} \lambda_n \rho_i \left[T_{S_i} + T_{C_i} \frac{p_i}{2(1-p_i)}\right] + \lambda_n \left[\bar{W}_n + T_{S_n} + T_{C_N} \frac{p_n}{2(1-p_n)}\right] \tag{37}$$

Equations (35) and (37) then give us a set of $2N$ equation in terms of $2N$ unknowns (p_n and $\rho_n, 1 \le n \le N$) which can be solved numerically.

For the rest of the analysis in this section, consider an arbitrary node $n (1 \le n \le N)$. Following the analysis of Section III, the probability that the tagged node experiences i backoff slots is given by

$$\begin{aligned} P[\mathrm{BO}_n = i] &= \rho_n \left[(1-p_n) U_{1,\mathrm{CW}_{\min}}(i) + p_n(1-p_n) \right. \\ &\quad \times [U_{1,\mathrm{CW}_{\min}} * U_{1,2\mathrm{CW}_{\min}}(i)] + \cdots + p_n^m (1-p_n) \\ &\quad \times [U_{1,\mathrm{CW}_{\min}} * U_{1,2\mathrm{CW}_{\min}} * \cdots * U_{1,2^m \mathrm{CW}_{\min}}(i)] \\ &\quad + p_n^{m+1}(1-p_n)[U_{1,\mathrm{CW}_{\min}} * \cdots * U_{1,2^m \mathrm{CW}_{\min}} \\ &\quad \left. * U_{1,2^m \mathrm{CW}_{\min}}(i)] + \cdots \right] \end{aligned} \tag{38}$$

with the corresponding pgf $\mathrm{BO}_n(z)$. Following the analysis of Section III-B, the probability that any of the backoff slots experienced by the tagged node is active (i.e., contains a transmission attempt by at least one of the N nodes and in case the tagged node transmits it experiences a collision), q_n, is given by

$$q_n = 1 - \prod_{\substack{i=1 \\ i \neq n}}^{N} \left(1 - \frac{\rho_i}{\bar{W}_i}\right). \tag{39}$$

Then, given that the tagged node experiences i backoff slots before it successfully transmits a packet, the pmf of the number of active slots within the backoff slots is given by

$$P[j \text{ slots active} \,|\, \mathrm{BO}_n = i] = \binom{i}{j} q_n^j (1-q_n)^{i-j} \quad (40)$$

for $j = 0, \ldots, i$. The probability that a slot results in a collision given that it is active, q_{n_c}, is then given by

$$q_{n_c} = \frac{1 - \prod_{\substack{i=1 \\ i\neq n}}^{N} \left(1 - \frac{\rho_i}{\bar{W}_i}\right) - \sum_{\substack{i=1 \\ i\neq n}}^{N} \frac{\rho_i}{\bar{W}_i} \prod_{\substack{j=1 \\ j\neq i,n}}^{N} \left(1 - \frac{\rho_j}{\bar{W}_j}\right)}{1 - \prod_{\substack{i=1 \\ i\neq n}}^{N} \left(1 - \frac{\rho_i}{\bar{W}_i}\right)} \quad (41)$$

and the probability that out of j active slots k result in collisions is given by

$$P[k \text{ collisions} \,|\, j\text{active slots}] = \binom{j}{k} q_{n_c}^k (1-q_{n_c})^{j-k} \quad (42)$$

Now each collision adds $T_C = \mathrm{DIFS} + \tau_{\mathrm{RTS}}$ slots to the tagged node's service time. As the next step, we now need to evaluate the impact of successful transmissions of other nodes on the service time of the tagged node. We assume that the packet size distribution at node $n, 1 \le n \le N$ follows the pmf $l_n(\nu)$ (i.e., $l_n(\nu)$ denotes the probability that a packet transmission by node n requires ν slots) with corresponding pgf $L_n(z)$. A successful transmission during the backoff slots between two successful transmissions of the tagged node does not involve any transmissions from the tagged node itself. Then, the probability that an arbitrary backoff slot between two successful transmissions of the tagged node contains a successful transmission is given by

$$P[\mathrm{succ}] = \left(1 - \frac{\rho_n}{\bar{W}_n}\right) \sum_{\substack{i=1 \\ i\neq n}}^{N} \frac{\rho_i}{\bar{W}_i} \prod_{\substack{j=1 \\ j\neq i,n}}^{N} \left(1 - \frac{\rho_j}{\bar{W}_j}\right). \quad (43)$$

Then, given that a slot contains a successful transmission, the probability that it belongs to node $i, i \neq n$, is given by

$$R_n[i] = P[i \,|\, \mathrm{succ}] = \frac{P[i, \mathrm{succ}]}{P[\mathrm{succ}]} = \frac{\frac{\rho_i}{\bar{W}_i} \prod_{\substack{j=1 \\ j\neq i,n}}^{N} \left(1 - \frac{\rho_j}{\bar{W}_j}\right)}{\sum_{\substack{i=1 \\ i\neq n}}^{N} \frac{\rho_i}{\bar{W}_i} \prod_{\substack{j=1 \\ j\neq i,n}}^{N} \left(1 - \frac{\rho_j}{\bar{W}_j}\right)} \quad (44)$$

Then the probability that a successful transmission by the other nodes adds ν slots to the service time of the tagged node is given by

$$\tilde{l}_n(\nu) = \sum_{\substack{i=1 \\ i\neq n}}^{N} R_n[i] l_i(\nu) \quad (45)$$

and the contribution of an arbitrary number of successful transmissions, say j, is given by

$$P\left[\sum^{j} \text{pkt time} = u\right] = \tilde{l}_n * \tilde{l}_n * \cdots * \tilde{l}_n(u) = \tilde{l}_n^{(j)}(u) \quad (46)$$

where $\tilde{l}_n^{(j)}()$ represents the $j-$fold convolution of $\tilde{l}_n(\nu)$.

As in Section III-B, we again consider a scenario where the tagged packet experiences i backoff slots of which j are active and among these j active slots, k slots have collisions. If the $j-k$ successful packet transmission by the other nodes contribute u slots, then the pmf of the conditional channel access time for the successful transmission of the tagged packet, Y, is given by

$$P[Y_n = s \,|\, i, j, k] = \begin{cases} \tilde{l}_n^{(j-k)}(u) & s = i + kT_C + u \\ 0 & \text{otherwise} \end{cases}. \quad (47)$$

Now, the joint probability distribution of i backoff slots, j active slots and k collisions, $P_n[i, j, k]$ can be evaluated using

$$P_n[i, j, k] = P_n[j, k \,|\, i] P_n[i]. \quad (48)$$

The pmf of the backoff slots i is $P_n[i] = P[\mathrm{BO}_n = i]$ and is given in (38). Also, by combining (40) and (42) we have

$$P_n[j, k \,|\, i] = \binom{i}{j} q_n^j (1-q_n)^{i-j} \binom{j}{k} q_{n_c}^k (1-q_{n_c})^{j-k} \quad (49)$$

The probability mass function of the service time Y can now be obtained by unconditioning (47) on i, j and k and is given by

$$\begin{aligned} P[Y_n = s] &= \sum_{i,j,k} P[Y_n = s \,|\, i, j, k] P[i, j, k] I(s) \\ &= \sum_{i}^{\infty} \sum_{j}^{i} \sum_{k}^{j} \left[\tilde{l}_n^{(j-k)}(u) \binom{i}{j} q_n^j (1-q_n)^{i-j} \right. \\ &\quad \left. \times \binom{j}{k} q_{n_c}^k (1-q_{n_c})^{j-k} P[\mathrm{BO}_n = i] I(s) \right] \end{aligned} \quad (50)$$

where $I(s)$ is an indicator function which equals 1 when $s = u + i + kT_C$ and 0 otherwise. Again, the expression above needs to be evaluated for all possible values of i, j and k which result in a given value of s. The pgf of the final service time, $B_n(z)$, is then obtained by the convolution of the channel access time $(Y_n(z))$ and the length of the packet to be served (l_n) and is given by

$$B_n(z) = Y_n(z) L_n(z). \quad (51)$$

Finally, since arrivals may occur at any point in the slot, we denote the distance of the arrival point from the start of the slot at the tagged node by F_n with mean $\bar{F}_n$. The mean of the total system time is then given by

$$\bar{V}_n = 1 - \bar{F}_n + B_n'(1) + \frac{[A_n'(1)]^2 B_n''(1) + A_n''(1) B_n'(1)}{2A_n'(1)[1 - A_n'(1) B_n'(1)]} \quad (52)$$

and the average queue size at the tagged node is given by

$$\bar{Q}_n = A_n'(1) \bar{V}_n \quad (53)$$

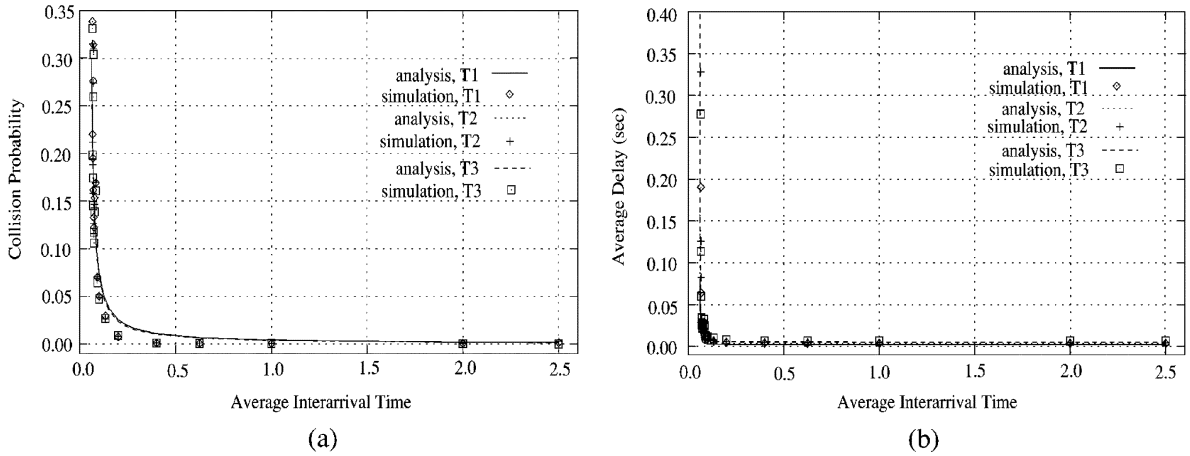

Fig. 7. Collision probabilities and packet delays with heterogeneous loads. T1, T2 and T3 denote nodes with 256, 512 and 1000 byte packets, respectively. (a) Collision probabilities. (b) Queueing delay.

where $A_n(z)$ denotes the pgf of the packet arrival process at the tagged node.

A. Simulation Results

We now validate the analysis presented in this section using simulations carried out using the ns-2 simulator. We use the same simulation settings as specified in Section III-C except for the number of nodes and the traffic pattern at each node. In the simulation results reported here, we have 12 nodes in the network with each node being the source of one flow and the the sink of another, resulting in 12 flows. Four of these flows had packets of 256 bytes, another four had packets of 512 bytes and the remaining four flows had packets of 1000 bytes. In addition, the packet arrival rate at the nodes with 512 and 1000 byte packets was 1.25 and 1.5 times, respectively, of the packet arrival rate at the nodes with 256 byte packets.

Fig. 7 shows the collision probabilities and the average packet delays of the three types of flows for various average packet inter-arrival times. The x-axis of the figure marks the average inter-arrival time at the nodes with 256 byte packets and the corresponding inter-arrival times at other nodes can be obtained by multiplying these values by 1.25 and 1.5, respectively. We first note that the analytic results match quite well with the simulation results. Also, there is not much of a difference between the collision rates experienced by flows of different types though nodes with lower arrival rates experience slightly higher collision rates. This is because nodes with higher loads often find that the lower load nodes are not competing with them for transmission slots, resulting in lower collision rates. Also, the delays at the nodes with larger packet sizes and arrival rates increases much faster than that of nodes with smaller packets and lower arrival rates. Since this is a bit difficult to see in Fig. 7(b), in Fig. 8 we have shown the delays for only the nodes with 256 and 1000 byte packet and zoomed in on the arrival rates to better illustrate this difference.

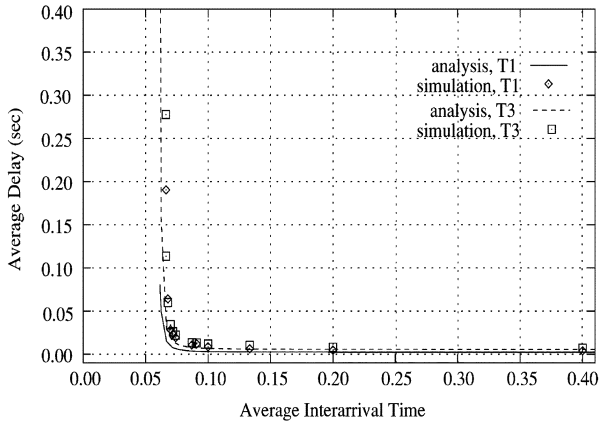

Fig. 8. Average packet delays for the nodes with 256 and 1000 byte packet for a subset of the arrival rates.

V. Extension to IEEE 802.11e and Collision Free Bursts

The major contributor to the delay in 802.11 based networks is the delay introduced by the channel contention. Intuitively, this delay can be reduced if instead of transmitting just one packet, the node is allowed to transmit a burst of packets once it successfully accesses and reserves the channel. This reduces the per packet channel contention delay by a factor of $M-1$ where M is the burst size. Considering the fact that multimedia traffic like VBR video is typically bursty [10], this scheme will be particularly well suited for real time traffic.

IEEE 802.11e provides an Enhanced DCF (EDCF) mode which provides differentiated channel access to frames of different priorities. In addition, there is a provision which allows a station to transmit multiple MAC frames consecutively after a single channel access as long as the whole transmission time does not exceed the transmission opportunity (TXOP) limit. In this section, we extend our model to account for such scenarios and consider the case where a station may transmit M consecutive packets for each successful channel access.

To obtain the delay and buffer occupancy characteristics, we argue that the queue at each node in this case can be modeled by a discrete time $G/G/1$ queue with server interruptions. To justify the model, note that at the MAC layer with collision free bursts, once the channel is successfully accessed and reserved, a maximum of M packets can be served contiguously signifying the time when the server is "available". However, once this set of packets has been transmitted, the server is "interrupted" for a duration equal to the time till the next successful channel access and reservation by the node. In this new server interruption model, the length of each slot corresponds to the time required to transmit a packet. Note that in the previous section, the length

of each slot was 20 μs which was the duration of a backoff slot. We now term a 20 μs slot a "mini-slot" to distinguish it from the "service time slots" used in the analysis of this section. Since we allow for variable packet lengths with pmf $l(\nu)$ mini-slots, the expected length of each slot for the interrupted server model is given by $20E[l]\mu$s. Note that with this model for the slot length, only the first moment of the delays resulting from our model is valid.

We now develop the expressions for the available and interrupted states. We denote the available and interrupted states by C and D, respectively. The probability that the available state lasts η slots, $C(\eta)$, corresponds to the number of packets scheduled in each burst. The number of packets that can be scheduled in one burst is bounded above by M and we now derive an approximate pmf of the size of an arbitrary burst.

Recall that the probability that there are i arrivals in an arbitrary slot is given by $a(i)$. The characterization of size of a scheduled burst is based on the following observations. When the load is low, the queue sizes are likely to be very small and the size of the burst scheduled would be dependent primarily on $a(i)$, though no more than M packets can be scheduled in a burst, irrespective of $a(i)$. However, for high load cases, a queue would very likely have M packets queued up once it gets access to the channel and thus the burst size would usually be M. Now consider an arbitrary slot with an arrival. Conditioned on the fact that there is an arrival, the number of packets in the burst, α, is given by

$$P[\alpha = i] = \frac{a(i)}{1 - a(0)}, \quad i = 1, 2, \ldots. \tag{54}$$

For $\alpha \leq M$, all the packets are scheduled in a single burst. However, for $\alpha > M$, we need $\lceil \alpha/M \rceil$ bursts with the first $\lceil \alpha/M \rceil - 1$ bursts being of size M and the last one of size $\alpha - M\lceil \alpha/M \rceil - M$ packets. Note that under high load conditions, the last burst would also most likely be of size M since additional packets are likely to have queued up during the transmission of the first $\lceil \alpha/M \rceil - 1$ bursts. To obtain the size of an arbitrary burst, we then need to quantify the burst sizes resulting from each possible value of α. Then, for low load conditions, the size of an arbitrary burst or the available time, C, is approximated by (55), shown at the bottom of the page. Note that (55) is not exact since it assumes that at low loads an arrival always sees an empty queue which may be justified by the fact that at low loads, the probability that an arbitrary arrival finds the queue non-empty is quite low. Also, note that at low loads (or equivalently at low arrival rates) the likelihood of back to back batch arrivals in successive slots is quite low which further justifies the approximation in the equation above. Now for high load conditions where an arbitrary arrival is quite likely to see a non-empty queue and M packets are likely to accumulate between two successive transmission from a node, we have

$$P[\beta' = i] = \begin{cases} 1 & i = M \\ 0 & \text{otherwise} \end{cases} \tag{56}$$

Equations (55) and (56) are exact characterizations of the burst size as $\rho \to 0$ and $\rho \to 1$. Combining the two equations into one which is exact at these two extreme values, we approximate the batch size distribution, which is equivalent to the available time distribution, at arbitrary loads by

$$P[C = i] = (1 - \rho)P[\beta = i] + \rho\delta(M), \quad i = 1, \ldots, M \tag{57}$$

where $\rho = E[A]/E[B]$ is the load on the system and $\delta(\cdot)$ is the delta function. As noted earlier, the expression above is an approximation which is accurate at low and high loads and is used here to maintain analytical tractability. As our simulation results show, because of this approximation, we marginally overestimate the delay at moderate loads. However, the magnitude of the errors are well within acceptable limits justifying the use of this approximation.

With this characterization of the size of a burst we can now model the interrupted time distribution. The interrupted time corresponds to the time spent between two successful transmissions from the tagged node and comprises of the time spent in backoff and the contributions from the successful transmissions of other nodes and collisions resulting from its own as well as other node's transmissions. As in Section III, the probability that there are j active mini-slots in i backoff slots between two successive transmissions of the tagged node, with k of them resulting in collisions are again given by (14) and (16). The average backoff window size $\bar{W}$ and the collision probability are again obtained using (5) and (11), respectively. Now, the length the transmissions resulting from each of these active slots depends on the size of the scheduled burst and the packet size distribution. With the pmf of the packet length (in mini-slots) denoted by $l(n)$ and given that there are k packets scheduled in the burst, the pmf of the burst length (BL) (in mini-slots) is given by

$$P[\mathrm{BL} = \nu \,|\, C = k] = l * l * \cdots * l(\nu) = l^{(k)}(\nu). \tag{58}$$

Unconditioning on the number of packets in the burst, we have

$$P[\mathrm{BL} = \nu] = \sum_{k=1}^{M} P[C = k] l^{(k)}(\nu). \tag{59}$$

We now consider the case when there are j successful transmissions from other nodes between the two successive transmissions of the tagged node. The pmf of the total contribution from the bursts of each of these transmissions is then given by

$$\mathrm{BL}^{(j)}(u) = \mathrm{BL} * \mathrm{BL} * \cdots * \mathrm{BL}(u). \tag{60}$$

Now consider a scenario where the tagged node experiences i backoff slots between two of its successful transmissions with j and k denoting the number of active slots and slots with collision in these i backoff slots. The pmf of j conditioned on i and that of k conditioned on j are given in (14) and (16), respectively. Following the arguments of the previous sections, the pmf of the delay introduced in the service time of a packet from the

$$P[\beta = i] = \begin{cases} \sum_{j=0}^{\infty} \frac{1}{j+1} \alpha(i + jM) & i = 1, \ldots, M-1 \\ \alpha(M) + \sum_{j=1}^{\infty} \sum_{k=0}^{M-1} \frac{j}{j+1} \alpha(k + jM) & i = M \end{cases} \tag{55}$$

tagged node by the collisions and successful transmissions of other nodes, X, in this case is given by

$$P[X = s \mid i,j,k] = \begin{cases} \mathrm{BL}^{(j-k)}(u) & s = i + kT_C + u \\ 0 & \text{otherwise} \end{cases} \tag{61}$$

As in Section III-B, the joint probability distribution of i, j and k, $P[i,j,k]$ can be evaluated using

$$P[i,j,k] = P[j,k \mid i]P[i] \tag{62}$$

where $P[i] = P[\mathrm{BO} = i]$ and is given in (12) and $P[j,k \mid i]$ is given in (20). The pmf of X can now be obtained by unconditioning (61) on i, j and k using the expressions for $P[j,k \mid i]$ and $P[i]$:

$$\begin{aligned} P[X = s] &= \sum_{i,j,k} P[X = s \mid i,j,k]P[i,j,k]I(s) \\ &= \sum_{i}^{\infty}\sum_{j}^{i}\sum_{k}^{j}\left[\mathrm{BL}^{(j-k)}(u)\binom{i}{j}q^j(1-q)^{i-j}\right. \\ &\quad \left.\times \binom{j}{k}q_c^k(1-q_c)^{j-k}P[\mathrm{BO} = i]I(s)\right] \end{aligned} \tag{63}$$

where $I(s)$ is an indicator function which equals 1 when $s = u + i + kT_C$ and 0 otherwise. As in the previous sections, the above expression needs to be evaluated for all possible values of i, j and k which result in a given value of s. Note that the delay characterized by X, which comprises of the backoff slots (BO) and the delay due to other stations transmitting is also the interrupt time experienced by the queue at the tagged node. The pgf of the interrupt time in terms of mini-slots, $B(z)$, is then

$$B(z) = X(z). \tag{64}$$

Aggregating the distribution for $b(n)$ in blocks of $E[l]$, we can then obtain the interrupted time distribution in terms of the average service time slots. Then the pmf of the interrupted time is given by

$$D(i) = \sum_{j=(2i-1)E[l]/2}^{(2i+1)E[l]/2} b(j), \quad i = 0,1,\ldots \tag{65}$$

where $b(j) = 0$ for $j < 0$. Note that loss of resolution resulting from the aggregation in the above expression introduces some errors in the final calculation, the magnitude of which increases as the packet sizes increase.

Using the analysis for infinite buffered discrete time queues with general arrivals, general service time distributions and general server interruptions presented in [3, Sec. 3.2], we can now derive the queue length characteristics at each node. Denoting by σ the fraction of time for which the channel is available, we have

$$\sigma = \frac{E[C]}{E[C] + E[D]} \tag{66}$$

and the condition for the stability of the queue is given by $A'(1) < \sigma$. Let $U_C(z), U_D(z)$ and $U(z)$ denote the pgf of the equilibrium buffer occupancy as observed just after the end of an arbitrary available (i.e., service) slot, just after the end of an arbitrary interrupted slot and just after any slot, respectively. Then

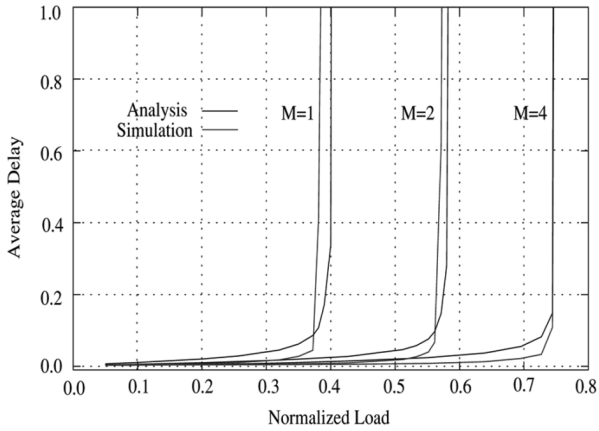

Fig. 9. Comparison of the average packet delays for different burst sizes.

$$U(z) = \sigma U_C(z) + (1-\sigma)U_D(z) \tag{67}$$

and using [3, (3.94), p. 107], it can be shown that

$$\begin{aligned} U(z) &= \frac{(z-1)^2A(z)[1-D(A(z))]Y(A(z)/z)}{(E[C]+E[D])(A(z)-1)(A(z)-z)W(z)} \\ &\quad + \frac{(z-1)(A(z)-A^2(z))[1-C(A(z)/z)D(A(z))Y(1)]}{(E[C]+E[D])(A(z)-1)(A(z)-z)W(z)} \end{aligned}$$

where $W(z) = 1 - C(A(z)/z)D(A(z)), Y(1) = [1 - A'(1)/\sigma]E[C]$ and the methodology for obtaining $Y(A(z)/z)$ is outlined in the Appendix. Now, since $U(z)$ denotes the pgf of the buffer occupancy just after any slot, the expected number of packets in the queue after an arbitrary slot is given by $\bar{U} = U'(1)$. Also, with F denoting the time of an arrival relative to the start of the slot in which it arrives, $1 - \bar{F}$ denotes the fraction of a slot which includes the new arrivals in the slot. Then, noting that the average number of arrivals in a slot is given by $A'(1)$, the average queue length at any arbitrary instant of time is given by

$$\bar{Q} = \bar{U} + (1-\bar{F})A'(1) \tag{68}$$

and using Little's law, the average system time is given by

$$\bar{V} = (1-\bar{F}) + \frac{\bar{U}}{A'(1)}. \tag{69}$$

The optimal value of M for a given input load can be obtained by differentiating (69) with respect to M and equating it to zero. The same expression can also be used to evaluate the number of connections that can be supported subject to a delay guarantee.

A. Simulation Results

To verify the analytic model of the previous subsection, we now compare the analytic results with those obtained using the ns-2 simulator. In Fig. 9, we show the results for a 10 node topology for burst sizes of $M = 1$, $M = 2$ and $M = 4$. The arrival stream at each node was a batch arrival process with the with fixed batches of size 4. The probability of a batch arriving at any slot was modeled by a Bernoulli process. In the figure, we plot the average delays as a function of the normalized load. We see the good match between the simulation and the analysis results. The slight difference in the analytic and simulation delays

for the moderate load cases is due to the approximation in the burst size characterization. However, we note that the difference is well within acceptable limits, justifying the use of the approximation for the sake of reducing computational complexity.

VI. Conclusion

The performance of the MAC protocol is critical in order for a network to support delay sensitive and real time applications and can easily form the performance bottleneck due to factors like channel contention delays and collisions. In this paper we present an analytic model to evaluate the performance of the IEEE 802.11 MAC in terms of its delays and queue lengths and evaluate its capability to support delay sensitive traffic. The performance evaluation is done by developing a queueing model for each node in the network which accounts for the intricacies of the MAC protocol and its behavior as a function of the number of users in the network. The developed model can be used for a number of purposes like admission control and determining the number of connections that can be supported for a given delay or loss constraint.

Each node is modeled as a discrete time $G/G/1$ queue and we allow for arbitrary number of nodes, arrival patterns and packet size distributions. We present a detailed analysis for the service time distribution which accounts for factors like the channel access delay due to the shared medium, impact of packet collisions and the resulting backoffs as well as the packet size distribution. Our analytic results have been verified using extensive simulations.

A key observation from the queueing model is that the primary contributor to the delay is the channel access and reservation time associated with each packet transmission. We also extend our model to proposals in IEEE 802.11e to reduce these delays which allow a node to schedule a burst of packets once they gain channel access. Each node in now modeled as a discrete time $G/G/1$ queue with interruptions. The analytic results were again verified using simulations.

Appendix
Evaluating Y(Z)

This Appendix outlines a methodology to obtain the function $Y(A(z)/z)$ in terms of $C(z)$ under the assumption that $C(z)$ is a rational function of z, and is taken from [3]. Since any rational function of Z can be expressed as a ratio of two polynomials and $C(z)$ vanishes at $z = 0$ (since the length of an available time is at least 1), C(z) can be written as

$$C(z) = C_1(z) + C_2(z) \tag{70}$$

where $C_1(z)$ is a polynomial

$$C_1(z) = \sum_{i=1}^{I} m_i z^i \tag{71}$$

and $C_2(z)$ is the ratio of two polynomials where the degree of the numerator is not higher than that of the denominator:

$$C_2(z) = \frac{\sum_{j=1}^{J} n_j z^j}{\prod_{k=1}^{K} (1 - \nu_k z)^{w_k}} \tag{72}$$

where $1/\nu_k$ are the zeros of the denominator and w_k are the corresponding multiplicities. Now define the functions

$$\Phi(z) = \sum_{i=1}^{I} m_i [A(z)]^i z^{I-i} \tag{73}$$

$$\Psi(z) = \sum_{j=1}^{J} n_j [A(z)]^j z^{J-j} \tag{74}$$

$$\Pi(z) = \sum_{k=1}^{K} [z - \nu_k A(z)]^{w_k} \tag{75}$$

$$X^*(z) = \sum_{i=1}^{I} x^*(i) [A(z)]^i z^{I-i} \tag{76}$$

$$X^{**}(z) = \sum_{j=1}^{J} x^{**}(j) [A(z)]^j z^{J-j} \tag{77}$$

where $x^*(i)$ and $x^{**}(j)$ are unknown constants to be determined. Then, $Y(A(z)/z)$ is given by

$$Y(A(z)/z) = \frac{\Pi(z) X^*(z) + z^I X^{**}(z)}{z^I \Pi(z)}. \tag{78}$$

The unknown quantities $x^*(i)$ and $x^{**}(j)$ can be determined using the following equation

$$D_0(z) = \frac{(z-1)[\Pi(z) X^*(z) + z^I X^{**}(z)]}{z^I \Pi(z) - D(A(z))[\Pi(z)\Phi(z) + z^I \Psi(z)]} \tag{79}$$

and the procedure for doing so is outlined below. When the condition for stability is satisfied (i.e., $A'(1) < \sigma$), the denominator of (79) has exactly $I+J$ zeros inside the unit disk of the complex plane, one of which equals unity. It can also be shown that the $I+J$ zeros of the denominator are the zeros of the numerator as well. This condition provides us with $I+J-1$ linear equations in the unknowns $x^*(i)$ and $x^{**}(j)$ (no equation is obtained for the zero $z = 1$), which, together with the normalizing equation $D_0(1) = 1$, can be used to determine the unknown parameters and thus $Y(A(z)/z)$.

References

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

[2] G. Bianchi and I. Tinnirello, "Kalman filter estimation of the number of competing terminals in an IEEE 802.11 network," in *Proc. IEEE INFOCOM*, San Francisco, CA, Apr. 2003.

[3] H. Bruneel and B. Kim, *Discrete-Time Models for Communication Systems Including ATM*. Boston, MA: Kluwer, 1993.

[4] F. Cali, M. Conti, and E. Gregori, "IEEE 802.11 protocol: Design and performance evaluation of an adaptive backoff mechanism," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 9, pp. 1774–1786, Sep. 2000.
[5] H. Chhaya and S. Gupta, "Performance modeling of asynchronous data transfer methods of IEEE 802.11 MAC protocols," *Wireless Netw.*, vol. 3, pp. 217–234, 1997.
[6] S. Choi, J. del Prado, S. Nandgopalan, and S. Mangold, "IEEE 802.11e contention-based channel access (EDCF) performance evaluation," in *Proc. IEEE ICC*, Anchorage, AK, May 2003.
[7] C. Coutras, S. Gupta, and N. Shroff, "Scheduling of real-time traffic in IEEE 802.11 wireless LANs," *Wireless Netw.*, vol. 6, no. 6, pp. 457–466, Nov. 2000.
[8] B. Crow, I. Widjaja, J. Kim, and P. Sakai, "Investigation of the IEEE 802.11 medium access control (MAC) sublayer functions," in *Proc. IEEE INFOCOM*, Kobe, Japan, Mar. 1997, pp. 126–133.
[9] , K. Fall and K. Varadhan, Eds., *NS Notes and Documentation, The VINT Project, UC BERKELY, LBL, USC/ISI, and Xerox PARC*. Berkeley: Univ. California Berkeley, Nov. 1997.
[10] D. Heyman and T. Lakshman, "Source models for VBR broadcast-video traffic," *IEEE/ACM Trans. Netw.*, vol. 4, no. 1, pp. 40–48, Feb. 1996.
[11] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, IEEE standards 802.11, Jan. 1997.
[12] A. Kumar, E. Altman, D. Miorandi, and M. Goyal, "New insights from a fixed point analysis of single cell IEEE 802.11 WLANs," *IEEE/ACM Trans. Netw.*, vol. 15, no. 4, pp. 588–601, Aug. 2007.
[13] C. Perkins and P. Bhagwat, "Highly dynamic destination sequenced distance-vector routing (DSDV) for mobile computers," in *Proc. ACM SIGCOMM*, Aug. 1994, pp. 234–244.
[14] C. Perkins and E. Royer, "Ad-hoc on-demand distance vector routing," in *Proc. IEEE Workshop Mobile Comput. Syst. Appl.*, New Orleans, LA, Feb. 1999, pp. 90–100.
[15] B. Sikdar, "An analytic model for the delay in IEEE 802.11 PCF MAC based wireless networks," *IEEE Trans. Wireless Commun.*, vol. 6, no. 4, pp. 1542–1560, Apr. 2007.
[16] J. Sobrinho and A. Krishnakumar, "Real-time traffic over the IEEE 802.11 medium access control layer," *Bell Labs Tech. J.*, vol. 1, no. 2, pp. 172–187, Autumn 1996.
[17] Y. Tay and K. Chua, "A capacity analysis for the IEEE 802.11 MAC protocol," *Wireless Netw.*, vol. 7, no. 2, pp. 159–171, Mar. 2001.
[18] O. Tickoo and B. Sikdar, "On the impact of IEEE 802.11 MAC on traffic characteristics," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 2, pp. 189–203, Feb. 2003.
[19] M. Veeraraghavan, N. Cocker, and T. Moors, "Support of voice services in IEEE 802.11 wireless LAN," in *Proc. IEEE INFOCOM*, Anchorage, AK, Apr. 2001, pp. 488–497.
[20] M. Visser and M. El Zarki, "Voice and data transmission over an 802.11 wireless network," in *Proc. IEEE PIMRC*, Toronto, Canada, Sep. 1995, pp. 648–652.
[21] J. Weinmiller, H. Woesner, J.-P. Ebert, and A. Wolisz, "Modified backoff algorithms for DFWMAC's distributed coordination function," in *Proc. 2nd ITG Fachtagung Mobile Kommunikation*, Neu-Ulm, Germany, Sep. 1995, pp. 363–370.
[22] Y. Xiao and J. Rosdahl, "Throughput and delay limits of IEEE 802.11," *IEEE Commun. Lett.*, vol. 8, no. 8, pp. 355–357, Aug. 2002.
[23] A. Zanella and F. De Pellegrini, "Statistical characterization of the service time in saturated IEEE 802.11 networks," *IEEE Commun. Lett.*, vol. 9, no. 3, pp. 225–227, Mar. 2005.

Omesh Tickoo received the B.E. degree in electronics and communication engineering from Karnataka Regional Engineering College, Surathkal, India, and the M.S. and Ph.D. degrees in computer and systems engineering from Rensselaer Polytechnic Institute, Troy, NY.

He is currently a Research Scientist for Intel Corporation, Hillsboro, OR. His research interests include next generation hardware architectures, network traffic modeling and multimedia streaming over wireless networks.

Biplab Sikdar (S'98–M'02) received the B.Tech. degree in electronics and communication engineering from North Eastern Hill University, Shillong, India, the M.Tech. degree in electrical engineering from Indian Institute of Technology, Kanpur, India, and the Ph.D. degree in electrical engineering from Rensselaer Polytechnic Institute (RSI), Troy, NY, in 1996, 1998 and 2001, respectively.

He is currently an Associate Professor in the Department of Electrical, Computer and Systems Engineering at RSI. His research interests include wireless MAC protocols, network routing and transport protocols, network security and queueing theory.

Dr. Sikdar is a member of Eta Kappa Nu and Tau Beta Pi.