Queueing Analysis and Delay Mitigation in IEEE 802.11 Random Access MAC based Wireless Networks[0]

Omesh Tickoo and Biplab Sikdar
Department of Electrical, Computer and Systems Engineering,
Rensselaer Polytechnic Institute, Troy, NY, USA.

***Abstract*— In this paper, we present an analytic model for evaluating the queueing delays at nodes in an IEEE 802.11 MAC based wireless network. 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. We model each node as a discrete time $G/G/1$ queue and derive the service time distribution while 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.**

I. Introduction

The IEEE 802.11 MAC [10] 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 propose and 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 [4], [15]. Work has also been conducted on improving the 802.11 MAC by using channel adaptive backoff schemes as reported in [3], [19] while [16] 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 [6], [7], [17], [18] while [14] 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 [5]. Delay analysis for the PCF mode of operation has been proposed in [6], [17] but no such analysis been reported for the DCF case. This paper addresses this void 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 which 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 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

[0]This work supported in part by NSF under contract number 0313095 and Intel Corporation.

0-7803-8355-9/04/$20.00 ©2004 IEEE.

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. Extensive 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 IV presents the extension of the model to the proposals for collision free bursts and IEEE 802.11e. Finally, Section V 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 [10] 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 source 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 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

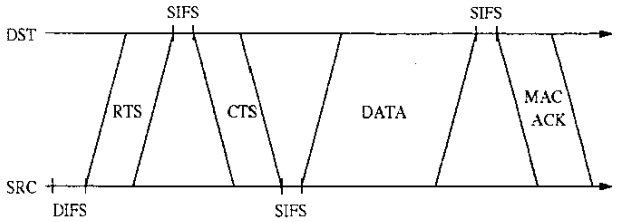


Fig. 1. Basic operation of the CSMA/CA protocol.

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 Figure 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 lengths of each packet is assumed to be arbitrary and the channel transmission rate is C bits/sec.

A. Modeling the Backoff Mechanism

In order to model the MAC layer queueing delays and losses, we first analyze the back-off mechanism associated with the exponential back-off mechanism of 802.11 MAC protocol's Collision Avoidance mechanism. In Figure 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 (i.e. an RTS transmission) results in a collision by p. The lower and upper bounds on the contention window associated with backoffs are denoted by CW_{min} and CW_{max} and we use the notation $m = \log_2(CW_{max}/CW_{min})$. 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

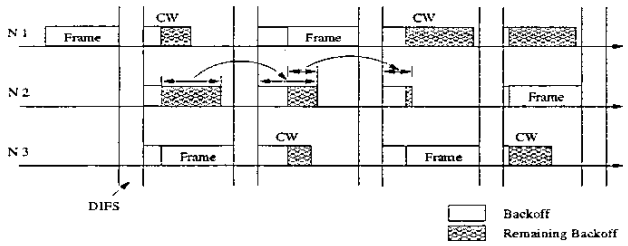


Fig. 2. The backoff mechanism of 802.11 MAC. The frame transmission time includes the RTS/CTS exchange and the MAC layer ACK. CW = Contention Window.

initiating transmission by BC. This back-off counter is calculated from

$$BC = \text{int}\left(\text{rnd}() \cdot CW(k)\right) \tag{1}$$

where the function `rnd()` returns a pseudo-random number uniformly distributed in $[0, 1]$ and $CW(k)$ 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 $1 \leq BC \leq CW(k)$ 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 CW_{min} [10]. For each unsuccessful attempt, the value of CW is doubled until it reaches the upper limit of CW_{max} specified by the protocol. Then, at the end of k unsuccessful attempts, $CW(k)$ is given by

$$CW(k) = \min\left(CW_{max}, 2^{k-1}CW_{min}\right) \tag{2}$$

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 $CW = W$ is given by

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

where $k \leq m$. Note that the second case ($W = CW_{max}$) includes all cases where the number of collisions is greater than m. The probability that back-off counter $BC = i$, $1 \leq i \leq CW_{max}$, 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 [15], [16] 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 extend the model to obtain an approximate expression for collision probabilities in the general 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 in the saturated case is given by

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

Now consider a network with N nodes operating in discrete time where the packet arrival rate at each node is given by λ packets per slot while the packet service rate of the network is denoted by μ packets per slot. A packet is backlogged on arrival if at the instant of arrival, the system is non-empty. We approximate the probability that the system is empty when an arbitrary arrival occurs by

$$\pi_0 = 1 - \frac{N\lambda}{\mu} \tag{6}$$

which is exact only for the $M/M/1$ case. Then, for any arbitrary packet, with probability π_0, the backoff window is 0 and with probability $1-\pi_0$, it is backlogged. Then, the average backoff window size for general (non-saturated) arrival rates is given by

$$\overline{W} = \frac{N\lambda}{\mu}\frac{1-p-p(2p)^m}{1-2p}\frac{CW_{min}}{2} \tag{7}$$

Note that while while an arrival to an idle node at an instant where some other nodes have non empty queues but are in backoff, will not be backlogged. In our analysis we neglect the occurrence of such cases. However, as we verify later in the section for simulation results, this approximation still leads to reasonably accurate results. Now, following the arguments of [15], [16] and considering the fact that only those nodes with a nonempty queue (the probability of which is again approximated by $1 - N\lambda/\mu$ since each node can get at most $1/N$ of the server's capacity) can actually collide with packets from other nodes, the packet collision probability can be obtained by solving

$$p = 1 - \left(1 - \frac{N\lambda}{\mu}\frac{(1-2p)}{1-p-p(2p)^m}\frac{2}{CW_{min}}\right)^{N-1} \tag{8}$$

In Section III-C we compare the results of this rather approximate analysis with the simulation results where we find a reasonably close match for most cases.

B. The Queueing Model

To obtain the delays and losses experienced by 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 20μsec, 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 directy 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 $P[BO = 0] = \pi_o(1-p)$. Now with probability $1-\pi_0$ 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 $1, \cdots, CW_{min}$. In case of a successful transmission after a single collision (with probability $p(1-p)$), the pmf of the number of backoff slots is obtained through $U_{1,CW_{min}} * U_{1,2CW_{min}}$ 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 $U_{1,CW_{min}}, U_{1,2CW_{min}}, \cdots, U_{1,2^mCW_{min}}(i)$ while the remaining terms are $U_{1,2^mCW_{min}}(i)$ since the backoff window is constrained by $CW_{max} = 2^mCW_{min}$. Then, the probability the tagged node experiences i backoff slots, $i > 0$, is given by

$$\begin{aligned} P[BO = i] = (1-\pi_0)\Big[&(1-p)U_{1,CW_{min}}(i) + p(1-p)\\ &\Big[U_{1,CW_{min}} * U_{1,2CW_{min}}(i)\Big] + \cdots + p^m(1-p)\\ &\Big[U_{1,CW_{min}} * U_{1,2CW_{min}} * \cdots * U_{1,2^mCW_{min}}(i)\Big]\\ &+p^{m+1}(1-p)\Big[U_{1,CW_{min}} * \cdots * U_{1,2^mCW_{min}} *\\ &U_{1,2^mCW_{min}}(i)\Big] + \cdots\Big] \end{aligned} \quad (9)$$

with the corresponding pgf $BO(z)$. 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.

Now, since the average window size is $\overline{W}$ (Eqn. (5)), the probability that a node attempts a transmission in an arbitrary slot is given by $(1-\pi_0)/\overline{W}$. Then, the probability that a given slot is active, q, is given by

$$q = 1 - \left(1 - \frac{1-\pi_0}{\overline{W}}\right)^N \quad (10)$$

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

$$P[j \text{ slots active}|BO = i] = \binom{i}{j} q^j(1-q)^{i-j} \quad (11)$$

for $j = 0, \cdots, i$. Unconditioning on i, we have

$$P[j \text{ slots active}] = \sum_{i=j}^{\infty} \binom{i}{j} q^j(1-q)^{i-j} P[BO = i] \quad (12)$$

Also, the probability that a slot results in a collision given that it is active, q_c, is given by

$$\begin{aligned} q_c &= P[\text{collision}|\text{slot active}]\\ &= \frac{1 - \left(1-\frac{1}{\overline{W}}\right)^N - \frac{N}{\overline{W}}\left(1-\frac{1}{\overline{W}}\right)^{N-1}}{1 - \left(1-\frac{1}{\overline{W}}\right)^N} \end{aligned} \quad (13)$$

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

$$P[k \text{ collisions}|j \text{ active slots}] = \binom{j}{k} q_c^k(1-q_c)^{j-k} \quad (14)$$

Now, each collision is of duration $T_{COLL} = DIFS + \tau_{RTS}$ where τ_{RTS} is the time required to transmit a RTS packet. Thus each collision between two transmissions from the tagged node adds T_{COLL} 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 $T_{COLL} = DIFS + \tau_{pkt}$ where τ_{pkt} is the packet transmission time. Also, each successful transmission by other nodes between the two successful transmissions of the tagged noded adds a time proportional to the packet length of the transmitted packet to the service time at the tagged node. In

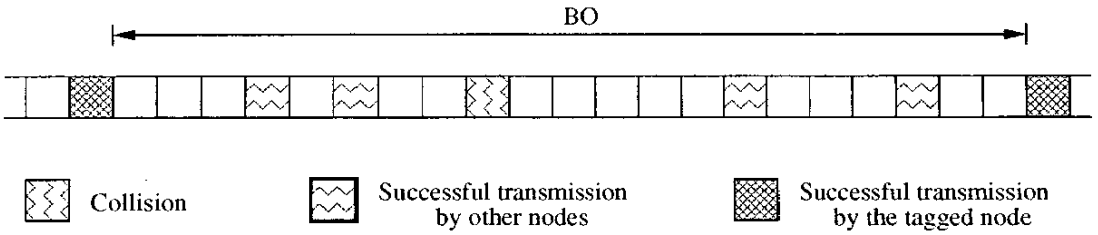


Fig. 3. Interleaving of transmissions and collisions contributing to the service time.

our analysis we allow for general packet length distributions and the probability that a packet transmission takes n slots (which is dependent on the packet length and the channel rate) is denoted by $l(n)$ with the corresponding pgf $L(z)$. Then, the contribution of j successful transmissions to the service time of the tagged node is given by

$$P[\sum^{j} \text{pkt time} = i] = l * l * \cdots * l(i) = l^{(j)}(i) \quad (15)$$

where $l^{(j)}()$ represents the $j-$fold convolution of $l(n)$. The contribution of the successful transmissions of the other competing stations and the collisions, X, to service time of the tagged node is then given by

$$P[X = n] = \begin{cases} \binom{j}{k} q^k (1-q)^{j-k}. & n = kT_{COLL} + i \\ l^{(j-k)}(i) P[\text{SA} = k] & \\ 0 & \text{otherwise} \end{cases} \quad (16)$$

where $P[\text{SA} = k]$ represents the probability that there are k active slots and is given by Equation (12). The above expression evaluates the probability of the event where there are k slots active between two transmissions from the tagged node, j of which result in collisions contributing kT_{COLL} slots to the service time while the $k - j$ successful transmissions contribute i slots. 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 n. The pgf of the final service time, $B(z)$, which comprises of the backoff slots (BO), the delay due to other stations transmitting (X) and the length of the packet to be served (l) is then given by

$$B(z) = BO(z)X(z)L(z) \quad (17)$$

Using standard discrete time queueing theory [2], 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

$$U(z) = [1 - A'(1)B'(1)] \frac{(z-1)B(A(z))}{z - B(A(z))} \quad (18)$$

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

$$V_{int}(z) = \frac{[1 - A'(1)B'(1)](z-1)B(z)[1 - A(B(z))]}{A'(1)[1 - B(z)][z - A(B(z))]} \quad (19)$$

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		

TABLE I
SIMULATION SETTINGS

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 $\overline{F}$. This adds a fractional component to the system time of $V_{frac} = 1 - \overline{F}$. The total system time is then given by $V = V_{int} + V_{frac}$ whose mean can be expressed as

$$\overline{V} = 1 - \overline{F} + B'(1) + \frac{[A'(1)]^2 B''(1) + A''(1)B'(1)}{2[1 - A'(1)B'(1)]} \quad (20)$$

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

$$\overline{Q} = A'(1)\overline{V} \quad (21)$$

Eqn. (20) 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.

C. Simulation Results

To validate our analytic model, we conducted extensive simulations using the simulator *ns-2* [8] 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 1500×500 meters and the nodes were randomly distributed over this region. The routing protocol used for the simulations was Dynamic Source Routing (DSR) [11] and we also verified our results for routing using Destination Sequenced Distance Vector (DSDV) [12]. The interface queues at each mode used a Droptail policy and the interface queue length was set at 50 packets. All sources and receivers have an omni-directional antenna of height 1.5m 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 1.

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

$$a(n) = \begin{cases} 1-p & n=0 \\ p & n=1 \end{cases} \quad (22)$$

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

In Figure 4 we compare the simulation results for the collision probabilities as obtained from the simulations and the approximate expression in Equation (8). We see that while for the 10 node case we have a good match with the simulation results, for the 20 node case we have some deviation. However, the saturation values of the collision probabilities when the load on the network approaches 1 match closely with the simulation. Note that a cause of the error is the fact that in the characterization of π_0, we take the nominal packet transmission time as the service time for simplicity. However, as the analysis shows, the service time is always greater the nominal packet transmission time due to the delays associated with channel access. In order to get more accurate estimates of the collision probabilities, an iterative technique similar to the one in [13] can be used. Under this iterative strategy, we start with the nominal packet transmission time as the service time and then compute the actual service as given by Equation (17). This service time can then be used to recalculate the collision probability which is then used again to find the new service time. This process continues till the values of the service time and the collision probabilities converge.

Figure 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 results. As expected, the system saturates more quickly for the 20 node cases 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.

IV. Extension to 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 [9], 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 current proposal 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.

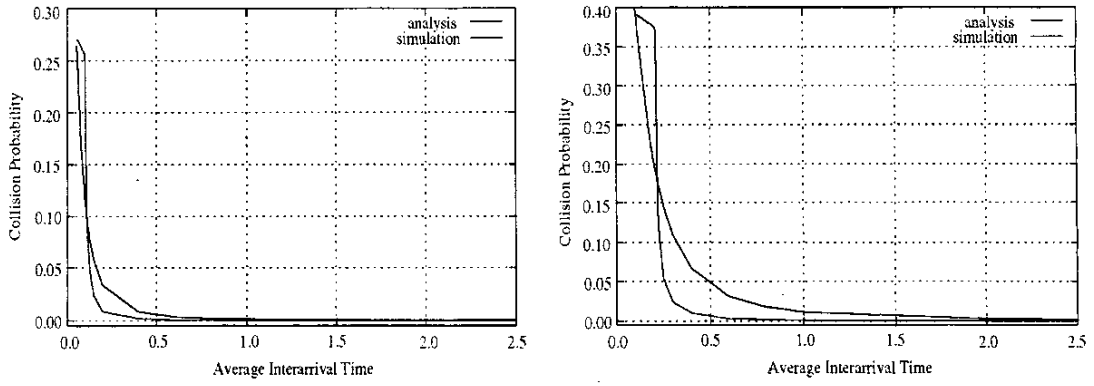
To obtain the delay and buffer occupancy characteristics, we argue that the queue at each node in this case can be modeled by a discrete time $G/G/1$ queue with server interruptions. To justify the model, note that at the MAC layer with collision free bursts, once the channel is successfully accessed 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 model, the length of each slot corresponds to the time to transmit a packet. Note that in the previous section, the length of each slot was 20μs which was the duration of a backoff slot. We now term a 20μs 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 $l(n)$ mini-slots, the 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 n slots, $C(n)$, 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 the pmf of the size of an arbitrary burst.

Recall that the probability that there are n arrivals in an arbitrary slot is given by $a(n)$. 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 $a(n)$, though no more than M packets can be scheduled in a burst, irrespective of $a(n)$. 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

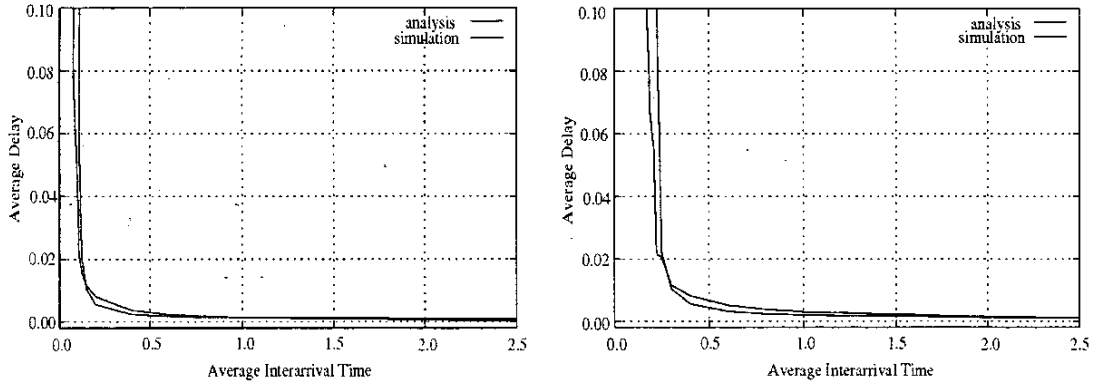
$$P[\alpha = i] = \frac{a(i)}{1-a(0)}, \qquad i = 1, 2, \cdots \quad (23)$$

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



(a) 10 nodes (b) 20 nodes

Fig. 4. Comparison of the collision probabilities.



(a) 10 nodes (b) 20 nodes

Fig. 5. Comparison of the average packet delays.

load conditions, the size of an arbitrary burst or the available time, C, is given by

$$P[\beta = i] = \begin{cases} \sum_{j=0}^{\infty} \frac{1}{j+1}\alpha(i+jM) & i = 1, \cdots, M-1 \\ \alpha(M) + \sum_{j=1}^{\infty}\sum_{k=0}^{M-1} \frac{j}{j+1}\alpha(k+jM) & i = M \end{cases} \tag{24}$$

Now for high load conditions,

$$P[\beta' = i] = \begin{cases} 1 & i = M \\ 0 & \text{otherwise} \end{cases} \tag{25}$$

The batch size distribution, which is equivalent to the available time distribution, can then be approximated as

$$P[C = i] = (1-\rho)P[\beta = i] + \rho\,\delta(M) \qquad i = 1, \cdots, M \tag{26}$$

where $\rho = E[A]/E[B]$ is the load on the system and $\delta(\cdot)$ is the delta function. Note that the above is an approximation which is accurate at low and high loads. 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.

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 the previous section, the probability that there are j active mini-slots between two successive transmissions of the tagged node, with k of them resulting in collisions are again given by Equations (12) and (13). The average backoff window size $\overline{W}$ and the collision probability are again obtained using Equations (7) and (8) respectively. Similarly, the probability that there are j active slots between two successive transmissions of the tagged node with k of them resulting in collisions are again given by Equations (12) and (13). 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

$$P[\text{BL} = i \mid C = k] = l * l * \cdots * l(i) = l^{(k)}(i) \tag{27}$$

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

$$P[\mathrm{BL}=i]=\sum_{k=1}^{M}P[C=k]l^{(k)}(i) \quad (28)$$

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

$$\mathrm{BL}^{(j)}(i)=\mathrm{BL}*\mathrm{BL}*\cdots*\mathrm{BL}(i) \quad (29)$$

Following the arguments of the previous section, the contribution of the successful transmissions of the other competing stations and the collisions, X, to service time of the tagged node is then given by

$$P[X=n]=\begin{cases}\binom{j}{k}q^k(1-q)^{j-k}. & n=kT_{COLL}+i\\ \mathrm{BL}^{(j-k)}(i)P[\mathrm{SA}=k] & \\ 0 & \text{otherwise}\end{cases} \quad (30)$$

where $P[\mathrm{SA}=k]$ again represents the probability that there are k active slots and is given by Equation (12). As in the previous section, the above expression needs to be evaluated for all possible values of i, j and k which result in a given value of n. The pgf of the final interrupt time in terms of mini-slots, $B(z)$, which comprises of the backoff slots (BO) and the delay due to other stations transmitting (X) is then given by

$$B(z)=BO(z)X(z) \quad (31)$$

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

$$D(i)=\sum_{j=(2i-1)E[l]/2}^{(2i+1)E[l]/2}b(j), \qquad i=0,1,\cdots \quad (32)$$

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 expressions of [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

$$\sigma=\frac{E[C]}{E[C]+E[D]} \quad (33)$$

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 at the end of an arbitrary available slot, at the end of an arbitrary interrupted slot and just after any slot respectively. Then

$$U(z)=\sigma U_C(z)+(1-\sigma)U_D(z) \quad (34)$$

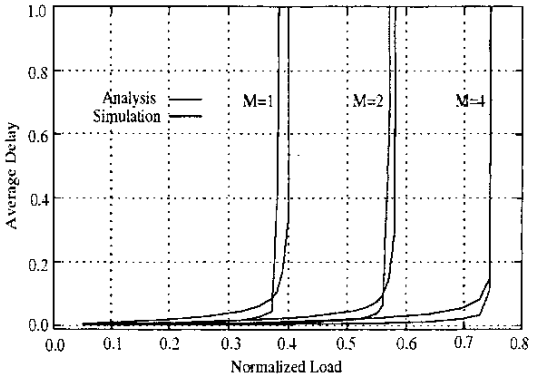


Fig. 6. Comparison of the average packet delays for different burst sizes.

and using results from [2], it can be shown that

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

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 Appendix 1. The average queue length is then given by

$$\overline{Q}=\overline{U}+(1-\overline{F})A'(1) \quad (35)$$

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

$$\overline{V}=(1-\overline{F})+\frac{\overline{U}}{A'(1)} \quad (36)$$

The optimal value of M for a given input load can be obtained by differentiating Eqn. (36) 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 Figure 6 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.

V. Conclusions

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 some recent 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 I: 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 [2]. 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

$$C(z) = C_1(z) + C_2(z) \tag{37}$$

where $C_1(z)$ is a polynomial

$$C_1(z) = \sum_{i=1}^{I} m_i z^i \tag{38}$$

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

$$C_2(z) = \frac{\sum_{j=1}^{J} n_j z^j}{\prod_{k=1}^{K} (1 - \nu_k z)^{w_k}} \tag{39}$$

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

$$\Phi(z) = \sum_{i=1}^{I} m_i [A(z)]^i z^{I-i} \tag{40}$$

$$\Psi(z) = \sum_{j=1}^{J} n_j [A(z)]^j z^{J-j} \tag{41}$$

$$\Pi(z) = \sum_{k=1}^{K} [z - \nu_k A(z)]^{w_k} \tag{42}$$

$$X^*(z) = \sum_{i=1}^{I} x^*(i) [A(z)]^i z^{I-i} \tag{43}$$

$$X^{**}(z) = \sum_{j=1}^{J} x^{**}(j) [A(z)]^j z^{J-j} \tag{44}$$

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

$$Y(A(z)/z) = \frac{\Pi(z) X^*(z) + z^I X^{**}(z)}{z^I \Pi(z)} \tag{45}$$

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

$$D_0(z) = \frac{(z-1)\Big[\Pi(z) X^*(z) + z^I X^{**}(z)\Big]}{z^I \Pi(z) - D(A(z))\Big[\Pi(z)\Phi(z) + z^I \Psi(z)\Big]} \tag{46}$$

and the procedure for doing so is outlined below. When the condition for stability is satisfied (i.e. $A'(1) < \sigma$), the denominator of Eqn. (46) 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 and I. Tinnirello, "Kalman filter estimation of the number of competing terminals in an IEEE 802.11 network," *Proceedings of IEEE INFOCOM*, San Francisco, CA, April 2003.

[2] H. Bruneel and B. Kim, *Discrete-Time Models for Communication Systems Including ATM*, Kluwer Academic Publishers, Boston, 1993.

[3] F. Cali, M. Conti and E. Gregori, "IEEE 802.11 protocol: design and performance evaluation of an adaptive backoff mechanism," *IEEE journal on selected areas of Communications*, vol. 18, no. 9, pp. 1774-1786, September 2000.

[4] H. Chhaya and S. Gupta, "Performance modeling of asynchronous data transfer methods of IEEE 802.11 MAC protocols," *Wireless Networks*, vol. 3, pp. 217-234, 1997.

[5] S. Choi, J. del Prado, S. Nandgopalan and S. Mangold, "IEEE 802.11e contention-based channel access (EDCF) performance evaluation," *Proceedings of IEEE ICC*, Anchorage, Ak, May 2003.

[6] C. Coutras, S. Gupta and N. Shroff, "Scheduling of real-time traffic in IEEE 802.11 wireless LANs," *Wireless Networks*, vol. 6, no. 6, pp. 457-466, November 2000.

[7] B. Crow, I. Widjaja, J. Kim and P. Sakai, "Investigation of the IEEE 802.11 Medium Access Control (MAC) sublayer functions," *Proceedings of IEEE INFOCOM*, pp. 126-133, Kobe, Japan, March 1997.

[8] K. Fall and K. Varadhan, editors, "ns notes and documentation," The VINT Project, UC BERKELY, LBL, USC/ISI, and Xerox PARC, November 1997.
[9] D. Heyman and T. Lakshman, "Source models for VBR broadcast-video traffic," *IEEE/ACM Transactions on Networking*, vol. 4, no. 1, pp. 40-48, February 1996.
[10] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, IEEE standards 802.11, January 1997.
[11] D. B. Johnson and D. A. Maltz, "Dynamic source routing in ad hoc wireless networks," *Mobile Computing*, edited by T. Imielinsky and H. Korth, Chapter 5, pp. 153-181, Kluwer Academic Publishers, 1996.
[12] C. E. Perkins and P. Bhagwat, "Highly dynamic Destination Sequenced Distance-Vector routing (DSDV) for mobile computers," *Proceedings of ACM SIGCOMM*, pp. 234-244, August 1994.
[13] B. Sikdar and D. Manjunath, "Queueing analysis of scheduling policies in copy networks of space based multicast packet switches," *IEEE/ACM Transactions on Networking*, vol. 8, no. 3, pp. 396-406, June 2000.
[14] J. Sobrinho and A. Krishnakumar, "Real-time traffic over the IEEE 802.11 medium access control layer," *Bell Labs Technical Journal*, vol. 1, no. 2, pp. 172-187, Autumn 1996.
[15] Y. Tay and K. Chua, "A capacity analysis for the IEEE 802.11 MAC protocol," *Wireless Networks*, vol. 7, no. 2, pp. 159-171, March, 2001.
[16] O. Tickoo and B. Sikdar, "On the impact of IEEE 802.11 MAC on traffic characteristics," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 2, pp. 189-203, February 2003.
[17] M. Veeraraghavan, N. Cocker and T. Moors, "Support of voice services in IEEE 802.11 wireless LAN," *Proceedings of IEEE INFOCOM*, pp. 488-497, Anchorage, Alaska, April 2001.
[18] M. Visser and M. El Zarki, "Voice and data transmission over an 802.11 wireless network," *Proceedings of IEEE PIMRC*, pp. 648-652, Toronto, Canada, September 1995.
[19] J. Weinmiller, H. Woesner, J.-P. Ebert and A. Wolisz, "Modified backoff algorithms for DFWMAC's distributed coordination function," *Proceedings of the 2nd ITG Fachtagung Mobile Kommunikation*, pp. 363-370, Neu-Ulm, Germany, September 1995.