Analysis of the Distribution of the Backoff Delay in 802.11 DCF: A Step Towards End-to-end Delay Guarantees in WLANs*

Albert Banchs

Universidad Carlos III de Madrid
Departamento de Ingeniería Telemática
banchs@it.uc3m.es

Abstract. In this paper we present an analytical method to study the distribution of the backoff delay in an 802.11 DCF WLAN under saturation conditions. We show that, with our method, the probability that the delay is below a given threshold can be computed accurately and efficiently. We also discuss how our analysis can be used to perform admission control on the number of accepted stations in the WLAN in order to provide delay assurances to real-time applications.

1 Introduction

As 802.11 WLANs see their capacity increased (from the traditional 2 Mbps channel capacity to 11 Mbps in 802.11b and 54 Mbps in 802.11a), these networks become better suited for the transport of real-time traffic. Since the performance of real-time applications is largely dependent on delay, there arises the need for an analysis of the delay in this type of networks.

To the date, the analysis of the delay in 802.11 WLAN has received some attention. The analyses of [1–3] are limited to the average delay, which is insufficient to assess the performance of real-time applications, as these applications require not only a low average delay but a low delay for all (or most of) their packets. The analyses of [4, 5] overcome this limitation by introducing probability generating functions (pgf's), which allow the computation of the probability distribution function (pdf) of the delay. However, computing pdf values with this method is very costly computationally and hence the approaches of [4, 5] are of little practical use to perform e.g. admission control functionality. This paper presents an original method to compute the delay distribution of 802.11 DCF that, in contrast to the previous analyses, is both accurate and efficient.

The analysis of the delay in this paper focuses on the backoff component of the delay under saturation conditions, hereafter referred to with *saturation delay*. By backoff delay we understand the time elapsed since a packet starts its backoff process until it is successfully transmitted[1]. This is one of the main components of the end-to-end delay. With saturation conditions we mean that all the stations in the WLAN always have

* This work has been performed within the IST FP6 Integrated Project DAIDALOS.

[1] In case the packet is discarded, we consider its backoff delay equal to ∞.

packets to transmit. Note that assuming saturation conditions corresponds to the worst case and thus provides us with an upper bound on the backoff delay.

The rest of the paper is structured as follows. In Section 2 we present a brief overview of the 802.11 DCF protocol. In Section 3 we propose a method to analyze the distribution of the saturation delay. In Sections 4 we evaluate the performance (namely, accuracy and computational efficiency) of the method proposed. The results obtained show that, with our method, the probability that the delay falls below a certain value can be computed accurately and efficiently. In Section 5 we discuss how our algorithm to compute the saturation delay distribution can be used to perform admission control in a WLAN with real-time traffic in order to provide this traffic type with end-to-end delay guarantees. Finally, in Section 6 we present our concluding remarks.

2 802.11 DCF

The DCF access method of the IEEE 802.11 standard [6] is based on the CSMA/CA protocol. A station with a new packet to transmit senses the channel and, if it remains free for a DIFS time, it transmits. If the channel is sensed busy, the station waits until the channel becomes idle for a DIFS time, after which it starts a backoff process. Specifically, it generates a random backoff time before transmitting.

The backoff time is chosen from a uniform distribution in the range $(0, CW - 1)$, where the CW value is called Contention Window, and depends on the number of transmissions failed for the packet. At the first transmission attempt, CW is set equal to a value CW_{min}, and it is doubled after each unsuccessful transmission, up to a maximum value CW_{max}.

The backoff time is decremented once every time interval T_e for which the channel is detected empty, "frozen" when a transmission is detected on the channel, and reactivated when the channel is sensed empty again for a DIFS time (if the transmission is detected as successful) or an EIFS time (if it is detected as unsuccessful). The station transmits when the backoff time reaches zero.

If the packet is correctly received, the receiving station sends an ACK frame after a SIFS time. If the ACK frame is not received within an ACK Timeout time, a collision is assumed to have occurred and the packet transmission is rescheduled according to the given backoff rules. If the number of retransmissions reaches a predefined Retry Limit, the packet is discarded. Upon completing the transmission (either with a success or with a discard), the transmitting station resets the CW to its initial value and starts a new backoff process; before this ends, a new packet cannot be transmitted.

The use of the Request to Send (RTS) / Clear to Send (CTS) mechanism is optional in 802.11. When this option is applied, upon the backoff counter reaching zero, the transmitting station sends an RTS frame to the receiving station, which responds with a CTS frame. The packet is then sent when the transmitting station receives the CTS.

3 Saturation Delay Analysis

In this section we propose an analytical model to compute the distribution of the saturation delay. We first analyze the simplified case in which all packets have the same

fixed length and the RTS/CTS mechanism is not used, and then propose two extensions of the basic analysis to account for these cases.

3.1 Basic Analysis

Let us consider a WLAN with N stations operating under saturation conditions and sending packets of a fixed packet length l. Our objective is to compute the probability that, under these conditions, a packet transmission of a tagged station experiences a saturation delay smaller than a given value D. We denote this probability by $P(d < D)$.

Fig. 1 illustrates the different components of the saturation delay. Applying the theorem of the total probability, $P(d < D)$ can be decomposed as follows

$$P(d < D) = \sum_{i=0}^{R} P(d < D/i \text{ col}) P(i \text{ col}) \tag{1}$$

where $P(i \text{ col})$ represents the probability that a packet suffers i collisions before being successfully transmitted and R is the Retry Limit.

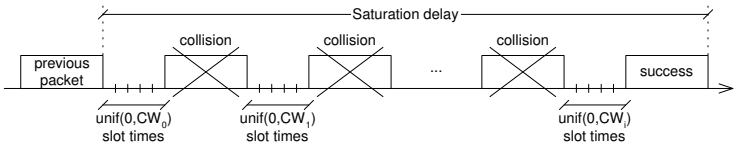

Fig. 1. Saturation delay.

Let us define a slot time as the time interval between two consecutive backoff time decrements of the tagged station. Note that, according to this definition, a slot time may be either empty or contain the transmission of one or more stations. Applying to the previous equation the theorem of the total probability on the total number of slot times the tagged station counts down before transmitting successfully, we have

$$P(d < D) = \sum_{i=0}^{R} \sum_{j=0}^{W_i} P(d < D/i \text{ col}, j \text{ slots}) P(j \text{ slots}/i \text{ col}) P(i \text{ col}) \tag{2}$$

where $W_i = \sum_{k=0}^{i} CW_k - 1$, with $CW_k = min(2^k CW_{min}, CW_{max})$, and $P(j \text{ slots}/i \text{ col})$ is the probability that the sum of the $i + 1$ backoff times of the packet equals j,

$$P(j \text{ slots}/i \text{ col}) = P\left(\sum_{k=0}^{i} unif(0, CW_k - 1) = j \right) \tag{3}$$

where $unif(0, C)$ represents a discrete random variable uniformly distributed on $\{0, 1, \ldots, C\}$.

As the probability mass function (pmf) of a sum of discrete random variables is equal to the convolution of the individual pmf's, we can compute $P(j \text{ slots}/i \text{ col})$ as follows

$$P(j \text{ slots}/i \text{ col}) = (f_0 * f_1 * \ldots * f_i)_j \tag{4}$$

being f_k the pmf of $unif(0, CW_k - 1)$. We compute the above convolution with Fast Fourier Transforms (FFT's), as FFT provides a very efficient means of computing convolutions.

Let τ be the probability that a station transmits in a slot time in a WLAN with N stations under saturation conditions. Following the analysis of [7], we compute τ by solving the non-linear equation resulting from the following two equations:

$$p = 1 - (1 - \tau)^{N-1} \tag{5}$$

and

$$\tau = \frac{2(1-2p)(1-p^{R+1})}{W(1-(2p)^{m+1})(1-p) + (1-2p)[(1-p^{R+1}) + W2^m p^{m+1}(1-p^{R-m})]} \tag{6}$$

where R is the Retry Limit, $W = CW_{min} + 1$, m is such that $CW_{max} = 2^m CW_{min}$ and p is the probability that a transmission attempt collides.

The first approximation upon which we base our analysis is the same as [8]: we assume that a station other than the tagged station transmits at each slot time with a constant and independent probability τ. With this assumption, the probability that the tagged station suffers i collisions before transmitting successfully can be computed according to

$$P(i \text{ col}) = P_c^i P_s = \left(1 - (1-\tau)^{N-1}\right)^i (1-\tau)^{N-1} \tag{7}$$

where P_s corresponds to the probability that a transmission of the tagged station is successful (i.e. none of the other $N-1$ stations transmits) and P_c to the probability that it collides (i.e. some other station transmits).

Our second approximation[2] is to assume that the saturation delay given i collisions and j slot times is a gaussian random variable, which we denote by d_{ij}. Note that, assuming independence between different slot times (which is given by the first approximation) and a number of slot times large enough (which is the typical case), the Central Limit Theorem assures that this approximation is accurate.

With the above approximation, it is enough to know the average and the typical deviation of d_{ij} (which we denote by m_{ij} and σ_{ij}, respectively) to compute $P(d < D/i \text{ col}, j \text{ slots})$,

$$P(d < D/i \text{ col}, j \text{ slots}) = \begin{cases} 0.5 + 0.5\, erf\left(\frac{D-m_{ij}}{\sqrt{2}\sigma_{ij}}\right), & \frac{D-m_{ij}}{\sigma_{ij}} \geq 0 \\ 0.5\, erfc\left(-\frac{D-m_{ij}}{\sqrt{2}\sigma_{ij}}\right), & \frac{D-m_{ij}}{\sigma_{ij}} < 0 \end{cases} \tag{8}$$

[2] This approximation is the key difference between our model and the analyses of [4, 5]; while, with our approximation, we only need to compute the average and typical deviation values of d_{ij}, which can be done efficiently, [4, 5] compute all the possible values of d_{ij} and their probability, which, as d_{ij} can take a very large number of different values, is very costly computationally.

Given the assumption of independence between different slot times, m_{ij} can be computed as the sum of the average duration all slot times in d_{ij},

$$m_{ij} = j\, m_n + i\, T_c + T_s \tag{9}$$

where m_n is the average duration of a slot time in which the tagged station does not transmit, T_c is the duration of a slot time that contains a collisions and T_s is the duration of a slot time that contains a successful transmission.

The duration of a slot time that contains a successful transmission is equal to [9]

$$T_s = T_{PLCP} + \frac{H + l}{C} + SIFS + \frac{ACK}{C} + DIFS \tag{10}$$

where T_{PLCP} is the PLCP (Physical Layer Convergence Protocol) preamble and header transmission time, H is the MAC overhead (header and FCS), ACK is the length of an ACK frame and C is the channel bit rate.

Similarly, the duration of a slot time that contains a collision is equal to

$$T_c = T_{PLCP} + \frac{H + l}{C} + EIFS \tag{11}$$

The average duration of a slot time in which the tagged station does not transmit, m_n, is computed as

$$m_n = P_{s,n}\, T_s + P_{c,n}\, T_c + P_{e,n}\, T_e \tag{12}$$

where $P_{s,n}$ represents the probability that a slot time in which the tagged station does not transmit contains a successful transmission, $P_{c,n}$ the probability that it contains a collision and $P_{e,n}$ the probability that it is empty.

$P_{s,n}$, $P_{e,n}$ and $P_{c,n}$ can be computed from τ and N as

$$P_{s,n} = (N-1)\tau(1-\tau)^{N-2}, \qquad P_{e,n} = (1-\tau)^{N-1} \tag{13}$$

and

$$P_{c,n} = 1 - P_{s,n} - P_{e,n} \tag{14}$$

With the assumption of independence between different slot times, the typical deviation σ_{ij} can be computed from

$$\sigma_{ij}^2 = j\, \sigma_n^2 \tag{15}$$

with

$$\sigma_n^2 = P_{s,n}\, T_s^2 + P_{c,n}\, T_c^2 + P_{e,n}\, T_e^2 - m_n^2 \tag{16}$$

which closes the analysis.

3.2 RTS/CTS

In case the RTS/CTS option is used, successful packets are preceded by a RTS/CTS exchange, while collisions occur with RTS frames instead of data packets. Accordingly, the durations of the slot times containing a successful transmission and a collision have to be computed as in [9] for the RTS/CTS case. With this only modification, the analysis of the previous clause can be used to compute the saturation delay distribution for the RTS/CTS case.

3.3 Non fixed packet lengths

Next, we extend our basic model to the case when packet lengths are not fixed but follow a certain distribution. Specifically, we consider that a packet length takes a value l of the set L with probability P_l, being L the set of all possible packet lengths. For simplicity, we assume that all stations transmit the same packet length distribution; however, the analysis would be very similar in the case when this condition does not hold.

In order to account for non fixed packet lengths, we have to modify the expressions to obtain the m_{ij} and σ_{ij} values. m_{ij} is computed as

$$m_{ij} = j\, m_n + i\, m_c + m_s \tag{17}$$

where m_n is the average duration of a slot time in which the tagged station does not transmit, m_c is the average duration of a slot time in which the tagged station collides and m_s is the average duration of a slot time in which the tagged station transmits a packet successfully.

The average duration of a slot time in which the tagged station does not transmit, m_n, is computed as

$$m_n = \sum_{l \in L} P_{s,l,n}\, T_{s,l} + \sum_{l \in L} P_{c,l,n}\, T_{c,l} + P_{e,n}\, T_e \tag{18}$$

where $P_{s,l,n}$ represents the probability that a slot time in which the tagged station does not transmit contains a successful transmission of a packet of length l, $P_{c,l,n}$ the probability that it contains a collision with the longest packet involved of length l and $T_{s,l}$ and $T_{c,l}$ are the slot time durations in each case.

$P_{s,l,n}$ and $P_{c,l,n}$ are computed as

$$P_{s,l,n} = (N-1)\tau(1-\tau)^{N-2} P_l \quad \text{and} \quad P_{c,l,n} = (1 - P_{s,l,n} - P_{e,n})\, P_{c,l} \tag{19}$$

where $P_{c,l}$ is the probability that the longest packet involved in a collision is of length l. Neglecting the collisions of more than two stations,

$$P_{c,l} = 2\, P_l \sum_{k \in L_l} P_k - P_l^2 \tag{20}$$

where L_l is the set of all the packet lengths smaller than or equal to l.

The duration of a slot time that contains a successful transmission of a packet of length l, $T_{s,l}$, and the duration of a slot time that contains a collision of two packets, the longest of length l, $T_{c,l}$, can be computed following Eqs. (10) and (11).

Finally, the typical deviation σ_{ij} for the non fixed packet length case can be computed from

$$\sigma_{ij}^2 = j\, \sigma_n^2 + i\, \sigma_c^2 + \sigma_s^2 \tag{21}$$

with

$$\sigma_n^2 = \sum_{l \in L} P_{s,l,n}\, T_{s,l}^2 + \sum_{l \in L} P_{c,l,n}\, T_{c,l}^2 + P_{e,n}\, T_e^2 - m_n^2, \tag{22}$$

$$\sigma_c^2 = \sum_{l \in L} P_{c,l}\, T_{c,l}^{*2} - m_c^2 \quad \text{and} \quad \sigma_s^2 = \sum_{l \in L} P_l\, T_{s,l}^2 - m_s^2 \tag{23}$$

4 Performance Evaluation

Next, we evaluate the accuracy and computational efficiency of the model proposed. The values of the system parameters used to obtain the results, both for the analytical model and the simulation runs, have been taken from the 802.11b physical layer. The packet length has been taken equal to 1000 bytes for the fixed packet length case, and derived from the measurements of Internet traffic presented in [10] for the non-fixed packet length case. Simulations are performed with an event-driven simulator developed by us, that closely follows the 802.11 DCF protocol details for each independently transmitting station.

Figs. 2, 3 and 4 illustrate the cumulative distribution function (cdf) of the saturation delay –i.e. $P(d < D)$ as a function of D– for our basic model, RTS/CTS extension and non-fixed packet lengths extension, respectively. Analytical results are represented with lines and simulations with points. Simulation results are given with a 95% confidence interval below 0.1%. Results show that our analysis is very accurate; in all cases, and for all values of D and N, simulations coincide almost exactly with analytical results. In addition, results corroborate the intuition that delays are smaller for the RTS/CTS and non-fixed packet lengths cases (the latter due to smaller packets being transmitted).

In order to evaluate the computational efficiency of our method, we measured the times required to compute the cdf values given in Figs. 2, 3 and 4. Measurements have been taken in a Pentium 4 PC with 2.66 GHz of CPU speed and 192 MB of RAM, running under the Linux operating system. We obtained that, for all models (basic, RTS/CTS and non fixed packet lengths) and different values of N (2, 10, 30 and 100), the time required to compute the 20 cdf values given in each of the graphs, ranged from 0.37 to 0.45 seconds. These results show that, with the model proposed, the times required to compute the $P(d < D)$ values keep very low (in all cases below 0.5 seconds for 20 points) and, moreover, are practically constant (almost independent of the model and N). We believe that these results, even though taken in a single platform and running not necessarily optimized code, do proof the low computational cost of our algorithm. Note that the times measured (in the order of 0.5 seconds) are fully acceptable to take an admission control decision; moreover, as (following the discussion of the next section)

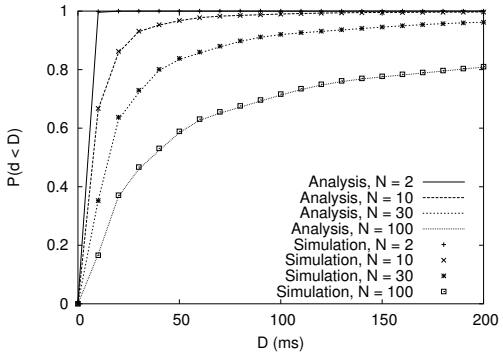

Fig. 2. Saturation delay cdf: Basic Model.

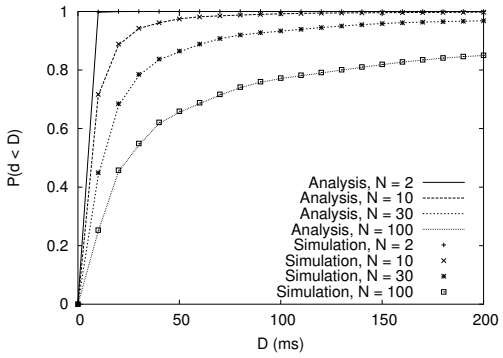

Fig. 3. Saturation delay cdf: RTS/CTS.

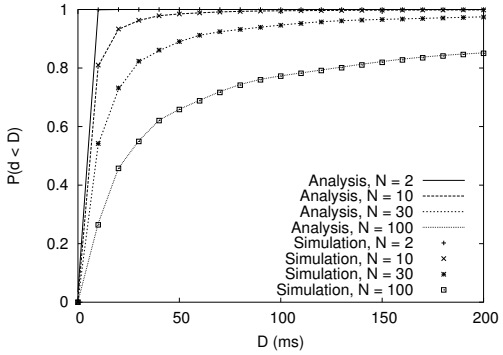

Fig. 4. Saturation delay cdf: Non fixed packet lengths.

in some situations one $P(d < D)$ value may be enough for admission control, the time involved in taking an admission control decision may even be much smaller.

5 Discussion on End-to-end Delay Guarantees in WLANs

The method we have proposed in this paper allows computing the distribution of the backoff delay under saturation conditions. The backoff delay is one of the main components of the end-to-end delay, but not necessarily the only one. Real-time applications require end-to-end delay (i.e. the sum of all the delay components) to be below a certain threshold (at least for most of the packets), or otherwise their performance is unsatisfactory. In this section we discuss how our method can be used to derive the worst-case distribution of the end-to-end delay, and thus allow providing end-to-end delay guarantees by means of admission control.

The fact that our model assumes saturation conditions represents the worst possible case for a tagged station, as this station will experience the largest delays when all the other stations have always packets to transmit. Therefore, this is the case that should be considered if our goal is to provide end-to-end delay guarantees by limiting the number

of stations in the WLAN by performing admission control. Many of the previous delay analyses of DCF (namely, [1–3]) also assume saturation conditions.

If we consider an end-to-end communication between two WLAN stations, or a WLAN station and the Access Point, then the end-to-end delay consists of two main components: the backoff and the queuing delays. The first is the time elapsed since a packet starts its backoff process until it is successfully transmitted, while the second is the time elapsed since the generation of a packet until it reaches the first position of the transmission buffer. The backoff component of the delay is accurately characterized in the present paper. An open issue is the computation the queuing delay.

The problem of computing the queuing delay in the above case can be seen as analyzing a classical G/G/1 queue, in which the arrivals follow the process given by the packet arrivals at the station, and the queue service time follows the distribution of the backoff delay (which has been characterized in this paper). This problem can be dealt with classical queuing theory [11] – this is the approach taken by [4, 5].

The 802.11 standard allows that a station, once it gets access to the channel, sends not only one but multiple packets separated by SIFS times. This option is appropriate e.g. for voice sources, because of the stringent delay requirements of their packets, and also because the short length of voice packets would make the protocol overhead very high otherwise. For a tagged station using this option, and sending all the packets waiting for transmission in its buffer every time it gets access to the channel, the end-to-end delay consists of the backoff delay only, and therefore the model presented in this paper can be used to characterize the end-to-end delay.

6 Summary and Final Remarks

As the capacity of WLANs and their use by real-time applications increases, there arises the need for better understanding and predicting the delay behavior in this type of networks. In this paper we have proposed a method to compute accurately and efficiently the distribution of the backoff delay in 802.11 DCF under saturation conditions. The method proposed is a first step towards an admission control algorithm that, by limiting the number of stations in the WLAN, ensures end-to-end delays low enough for real-time applications.

The backoff delay experienced by a station can be interpreted as the service time seen by its internal queue. Then, classical queuing theory can be used to derive the queuing delay, given the characterization of the backoff delay obtained in this paper. If a station sends all its waiting packets when it accesses the channel, the backoff delay derived here is the only component of the end-to-end delay.

Our model to analyze the backoff delay of a tagged station assumes that all other stations always have packets to transmit. As this corresponds to the worst case for the delay of the tagged station, the results obtained represent an upper bound and are therefore appropriate for providing the tagged station with delay guarantees. However, our analysis could also be reused for non-saturation conditions, if the τ probabilities under non-saturation conditions were given (a rough approximation to compute them is proposed in [4]).

In the literature, there have been many protocol proposals for WLAN that, unlike DCF, have been designed specifically to satisfy the delay requirements of real-time applications (see e.g. [12–14]). The PCF scheme of 802.11 [6] was also designed with a similar intention. However, none of these (including PCF) is widely deployed today, which leaves DCF as the only option to provide real-time traffic communication in today's WLANs.

The IEEE 802.11 WG is currently undergoing a standardization activity to extend the 802.11 protocol with QoS support, leading to the upcoming 802.11e standard. The EDCA access mechanism of 802.11e is an extension of the DCF protocol. We believe that our analysis here provides a basis that can be extended to analyze the delay of 802.11e EDCA.

References

1. E. Ziouva and T. Antonkopoulos, "CSMA/CA Performance under high traffic conditions: throughput and delay analysis," *Computer Communications*, vol. 25, no. 1, pp. 313–321, January 2002.
2. P. Chatzimisios, A.C. Boucouvalas, and V. Vitsas, "Packet delay analysis of IEEE 802.11 MAC protocol," *IEE Electronics Letters*, vol. 39, no. 18, pp. 1358–1359, September 2003.
3. B. Li and R. Battiti, "Performance Analysis of An Enhanced IEEE 802.11 Distributed Coordination Function Supporting Service Differentiation," in *Proceedings of QofIS'03*, Stockholm, Sweden, October 2003.
4. O. Tickoo and B. Sikdar, "Queueing Analysis and Delay Mitigation in IEEE 802.11 Random Access MAC based Wireless Networks," in *Proceedings of IEEE INFOCOM'04*, Hong Kong, China, March 2004.
5. H. Zhai and Y. Fang, "Performance of Wireless LANs Based on IEEE 802.11 MAC Protocols," in *Proceedings of IEEE PIMRC'03*, 2003.
6. IEEE 802.11, *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications*, Standard, IEEE, August 1999.
7. H. Wu, Y. Peng, K. Long, S. Cheng, and J. Ma, "Performance of Reliable Transport Protocol over IEEE 802.11 Wireless LAN: Analysis and Enhancement," in *Proceedings of IEEE INFOCOM'02*, New York City, New York, June 2002.
8. F. Cali, M. Conti, and E. Gregori, "Dynamic Tuning of the IEEE 802.11 Protocol to Achieve a Theoretical Throughput Limit," *IEEE/ACM Transactions on Networking*, vol. 8, no. 6, pp. 785–799, December 2000.
9. G. Bianchi, "Performance Analysis of the IEEE 802.11 Distributed Coordination Function," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 3, pp. 535–547, March 2000.
10. K. Claffy, G. Miller, and K. Thompson, "The nature of the beast: Recent traffic measurements from an internet backbone," in *Proceedings of INET'98*, Geneve, Switzerland, July 1998.
11. H. Bruneel and B. Kim, *Discrete-Time Models for Communication Systems Including ATM*, Kluwer Academic Publishers, 1993.
12. V. Kanodia, C. Li, B. Sadeghi, A. Sabharwal, and E. Knightly, "Distributed Multi-Hop with Delay and Throughput Constraints," in *Proceedings of MOBICOM'01*, Rome, Italy, July 2001.
13. J. L. Sobrinho and A.S. Krishnakumar, "Real-Time Traffic over the IEEE 802.11 Medium Access Control Layer," *Bell Labs Technical Journal*, 1996.
14. S. Chevrel et al., "Analysis and optimisation of the HIPERLAN Channel Access Contention Scheme," *Wireless Personal Communications*, vol. 4, pp. 27–39, 1997.