Wireless Netw (2009) 15:73–86
DOI 10.1007/s11276-007-0026-9

Statistical multiplexing, admission region, and contention window optimization in multiclass wireless LANs

Yu Cheng · Xinhua Ling · Lin X. Cai · Wei Song ·
Weihua Zhuang · Xuemin Shen · Alberto Leon-Garcia

Published online: 16 February 2007

Abstract This paper presents an analytical model for evaluating the statistical multiplexing effect, admission region, and contention window design in multiclass wireless local area networks (WLANs). We consider distributed medium access control (MAC) which provisions service differentiation by assigning different contention windows to different classes. Mobile nodes belonging to different classes may have heterogeneous traffic arrival processes with different quality of service (QoS) requirements. With bursty input traffic, e.g. on/off sources, our analysis shows that the WLAN admission region under the QoS constraint can be significantly improved, when the statistical multiplexing effect is taken into account. We also analyze the MAC resource sharing between the short-range dependent (SRD) on/off sources and the long-range dependent (LRD) fractional Brownian motion (FBM) traffic, where the impact of the Hurst parameter on the admission region is investigated. Moveover, we demonstrate that the proper selection of contention windows plays an important role in improving the WLAN's QoS capability, while the optimal contention window for each class and the maximum admission region can be jointly solved in our analytical model. The analysis accuracy and the resource utilization improvement from statistical multiplexing gain and contention window optimization are demonstrated by extensive numerical results.

Keywords WLAN performance analysis · Nonsaturated MAC modeling · Statistical multiplexing · Admission region · Contention window optimization · Long-range dependence

1 Introduction

The IEEE 802.11 wireless local area networks (WLANs) have been widely deployed in recent years to provide viable, low-cost wireless Internet access in public *hot spots*, e.g. hotels, airports, office and campus buildings. With the continuous flourish of WLANs, extensive attentions have been incurred both in academia and industry to provision quality of service (QoS) guaranteed multimedia applications over WLANs, for more efficient exploitation of this convenient Internet access infrastructure.

In a multimedia application over the Internet Protocol (IP), the 802.11 wireless link involved in the end-to-end path is prone to become a bottleneck due to the channel

Y. Cheng (✉)
Department of Electrical and Computer Engineering,
Illinois Institute of Technology, Chicago, IL, USA
e-mail: cheng@iit.edu

X. Ling · L. X. Cai · W. Song · W. Zhuang · X. Shen
Department of Electrical and Computer Engineering,
University of Waterloo, Waterloo, ON, Canada
e-mail: x2ling@bbcr.uwaterloo.ca

L. X. Cai
e-mail: lcai@bbcr.uwaterloo.ca

W. Song
e-mail: wsong@bbcr.uwaterloo.ca

W. Zhuang
e-mail: wzhuang@bbcr.uwaterloo.ca

X. Shen
e-mail: xshen@bbcr.uwaterloo.ca

A. Leon-Garcia
Department of Electrical and Computer Engineering,
University of Toronto, Toronto, ON, Canada
e-mail: alberto.leongarcia@utoronto.ca

contention delays and collisions. In order to provide quantitative QoS guarantees, it is necessary to characterize the service capacity, delay, and loss performance of an WLAN in both MAC and IP layers. The IEEE 802.11 medium access control (MAC) protocol [1] contains two access modes, the mandatory distributed coordination function (DCF) mode and the optional point coordination function (PCF) mode. Although PCF is designed with an objective to provision QoS to real-time applications, it is not supported in most current wireless cards, due to its implementation complexity and limited QoS capability [2, 3]. In this paper, we thus focus on the 802.11 DCF.

Most of the existing work on the DCF MAC performance analysis, e.g. [4–7] and the references therein, have focused on deriving the channel throughput or average delay under saturated input traffic. While the saturated modeling is applicable for bulk data transfer applications, it is hardly valid for real-time voice/video applications. In such applications, the traffic arrival process is usually bursty, and modeled as a short-range dependent (SRD) Markovian source or a long-range dependent (LRD) self-similar source [8–10]. For a bursty source in a multiplexing queue, a proper service rate falling between the average arrival rate and the peak rate, which is defined as the *effective bandwidth* [11], needs to be determined to provision certain quantitative QoS guarantees, such as a packet loss probability or a stochastic delay bound. In a WLAN, all the mobile nodes share the wireless link, where the MAC protocol statistically multiplexes the traffic flows from different mobiles over the common communication channel. Thus, the effective bandwidth of a traffic source required for QoS guarantee at the IP layer should be provisioned by the MAC layer under the connection or call admission control (CAC). However, it is very difficult, if not impossible, to establish an efficient analytical tool from the saturated modeling for effective bandwidth provisioning and admission region calculation; currently, the WLAN admission region is mainly determined by measurement studies, computer simulations, or conservative peak-rated based estimation [12–14].

A simple yet accurate analytic model for evaluating the 802.11 DCF in *nonsaturated* case has been presented in [15, 16]. In the nonsaturated model, each node is modeled as a discrete time G/G/1 queue, where the arrival traffic can be from bursty sources and the serving capacity is characterized by MAC analysis. Zhai et al. show in [3] that the maximum DCF MAC capacity and satisfying QoS performance can only be achieved in the nonsaturated case, where the CAC is necessary to maintain the WLAN in a proper operating range for QoS guarantees. In [3], it is also shown that when a WLAN works in the proper operating range, the packet collision probability is quite small and each packet sees an approximately constant service rate; therefore the G/G/1 queue at each node can be well approximated by a G/D/1 queue under admission control. Inspired by the above results, we show in this paper that the single-server queueing analysis at the IP layer can be combined with the nonsaturated DCF analysis at the MAC layer to investigate the statistical multiplexing and admission control of bursty traffic flows over the 802.11 wireless channel. Particularly, such an analytical model enables investigating the impact of traffic self-similarity or long-range dependence on the WLAN capacity. To the best of our knowledge, it is the first time that the LRD traffic over WLAN has been analytically studied.

When heterogeneous applications coexist in a WLAN, the DCF MAC is inefficient in protecting the QoS-critical applications from the QoS-resilient applications. Therefore, the DCF MAC has been enhanced to provision service differentiation, e.g. [17–19] and the references therein. The basic differentiation mechanisms are assigning different service classes different contention window (CW) backoff parameters, different interframe spacing before data transmission (arbitration interframe space, AIFS), or different channel holding time upon the successful channel access (transmission opportunity, TXOP). In this paper, we extend the nonsaturated DCF modeling [15, 16] to analyze a DCF MAC with class differentiation; for simplicity, only contention window based differentiation is considered. We demonstrate that the proposed multiclass DCF analysis is a versatile analytical framework when combined with the IP-layer queueing analysis, where the QoS performance, statistical multiplexing gain, admission region, contention window design, fairness in resource sharing can all be investigated. Particularly, we show through numerical analysis that the proper selection of contention windows plays an important role in improving the MAC capacity for a larger admission region, while the optimal contention window for each class and the maximum admission region can be jointly solved in our analytical model.

In [14], the statistical multiplexing gain in serving bursty traffic is demonstrated by simulation results, where a WLAN with DCF MAC can support up to 76 on/off voice flows, much larger than the capacity of 52 flows by a peak-rate based admission control. However, no analytical approach is available yet in the existing literature to incorporate the statistical multiplexing gain into the WLAN admission region calculation. The objective of this paper is to present such an analytical model. In [20], we have extended the nonsaturated DCF analysis [15, 16] to analyze a WLAN with unbalanced traffic, where downlink flows are aggregated at the access point (AP). In this paper, we use the proposed multiclass DCF analysis to demonstrate that traffic aggregation at the AP can further increase the statistical multiplexing gain over the WLAN, and the contention window differentiation between the AP and the mobile nodes can provision a fair resource sharing between the uplink and downlink aggregate traffic. In [21], we show that the WLAN admission region analysis can significantly facilitate the capacity planning in a cellular/WLAN integrated system.

The remainder of this paper is organized as follows. Section 2 describes the system model, including the network-layer effective bandwidth and the multiclass DCF MAC. Section 3 presents the nonsaturated DCF model with class differentiation. In Section 4, the nonsaturated DCF model is combined with IP-layer queueing analysis to investigate the statistical multiplexing effect, impact of the LRD traffic, admission region, and contention window design in the WLAN. In Section 5, we present extensive numerical results to demonstrate the accuracy of the propose analytical model and the resource utilization improvement results from statistical multiplexing and contention window optimization. Section 6 gives the concluding remarks.

2 System model

2.1 Effective bandwidth and stochastic QoS

We present an analytical model to determine the maximum number of QoS-guaranteed traffic sources that can be supported over the WLAN DCF. Particularly, an effective bandwidth is computed at the network layer according to the packet-level QoS requirement, and the MAC configuration is analytically determined to provision the required serving capacity efficiently. As an illustration, the QoS constraint under consideration is a *stochastic delay bound* d, i.e.

$$P\{D > d\} \leq \epsilon. \tag{1}$$

where D is the queueing delay and ϵ the *delay violation probability*.

2.1.1 On/Off sources

One of the critical multimedia applications is voice over Internet Protocol (VoIP) over WLAN, and the widely adopted traffic model for voice is the on/off model. In an on/off source, the on and off periods are exponentially distributed with average durations of t_{on} and t_{off}, respectively, and the activity factor $p_{on} = t_{on}/(t_{on} + t_{off})$. At the on state, voice traffic is generated at a constant rate of R_p.

Consider that $M(\geq 1)$ on/off sources are multiplexed at a queue with a service capacity of μ, and let Q denote the queue length. It is well-known that a conservative approximation of the overflow probability for on/off sources takes an exponential expression (Eqs. (3-42) and (3-43) in [8]), that is

$$P\{Q > x\} \approx \exp\left[-\frac{M(1-\rho)(\alpha+\beta)}{MR_p - \mu}x\right] \tag{2}$$

where $\rho = p_{on}MR_p/\mu$ is the utilization ratio, and $\alpha = 1/t_{on}$, $\beta = 1/t_{off}$ are the transition rates between the on and off states, respectively. The delay violation probability at d can be equivalently mapped to *buffer overflow probability* at $d\mu$ as

$$P\{Q > d\mu\} \leq \epsilon. \tag{3}$$

From (2), we have

$$P\{Q > d\mu\} \approx \exp\left[-\frac{Md(\mu/p_{on} - MR_p)}{t_{off}(MR_p - \mu)}\right]. \tag{4}$$

Combining (3) and (4), the minimum service rate required to guarantee the QoS, i.e. the effective bandwidth, can then be derived as

$$\mu = \frac{MR_p(t_{off}\log\epsilon - Md)}{t_{off}\log\epsilon - Md/p_{on}}. \tag{5}$$

For VoIP, the delay bound d is set as 150 (400) ms for an excellent (acceptable) voice quality, and ϵ is normally set as 1–3% [28]. From (5), different effective bandwidths will correspond to different levels of QoS; such a QoS-resilience property of VoIP is particularly meaningful for a wireless link, where the channel capacity is usually unstable.

2.1.2 Fractional brownian motion (FBM)

While the SRD Markovian traffic model[1] is reasonably accurate for per-flow or small-scale aggregation traffic modeling, extensive network traffic studies suggest that variable-rate video traffic or large-scale Internet traffic aggregate usually exhibits self-similarity or long-range dependence [10, 22, 23]. Here, we consider the case where the self-similar aggregate traffic can be characterized by an FBM process. The FBM process has a Gaussian marginal distribution, which is justified by the large-scale aggregation over a high speed Internet link, according to the *Central Limit Theorem* [24].

The standard (normalized) FBM process $\{Z(t) : t \geq 0\}$ with Hurst Parameter $H \in [0.5, 1)$ is a centered Gaussian process with stationary increments that possesses the following properties [25]: (a) $Z(0) = 0$, (b) $\mathrm{Var}\{Z(t)\} = t^{2H}$, and (c) $Z(t)$ is sample path continuous. A self-similar input process $\{A(t) : t \geq 0\}$ can then be represented as

$$A(t) = \lambda t + \sigma Z(t) \tag{6}$$

where the mean arrival rate $E\{A(t)\} = \lambda$, and the variance $\mathrm{Var}\{A(t)\} = \sigma^2 t^{2H}$. Note that σ^2 is the variance of traffic in a unit time. When $0.5 < H < 1$, the FBM are both self-similar and long range dependent; when $H = 0.5$, the FBM is self-similar but short range dependent.

[1] The on/off model is the simplest Markovian model.

In a buffer with a stationary Gaussian input, the overflow probability can be accurately estimated by the maximum variance asymptotic (MVA) approach [24, 26]. For the buffer served with capacity μ, define $\kappa = \mu - \lambda$ and $X_t = A(t) - \mu t$. Let m_x be the reciprocal of the maximum of $\sigma^2_{x,t} = \mathrm{Var}\{X_t\}/(x + \kappa t)^2$ for given x, i.e.,

$$m_x = \frac{1}{\max_{t\geq 1} \sigma^2_{x,t}} = \min_{t\geq 1} \frac{(x+\kappa t)^2}{\mathrm{Var}\{X_t\}}. \tag{7}$$

The MVA approximation of the overflow probability is then given by

$$P\{Q > x\} \approx \exp\left(-\frac{m_x}{2}\right). \tag{8}$$

In the self-similar case of an FBM input characterized by λ, σ^2 and H, m_x can be explicitly computed by [26]

$$m_x = \frac{4\kappa^\beta x^{2-\beta}}{S\beta^\beta(2-\beta)^{2-\beta}} \tag{9}$$

where $\beta = 2H$ and $S = \sigma^2$. The explicit expression of m_x leads to the explicit expression of the overflow probability as

$$P\{Q > x\} \approx \exp\left(-\frac{2\kappa^\beta x^{2-\beta}}{S\beta^\beta(2-\beta)^{2-\beta}}\right). \tag{10}$$

Considering $P\{D > d\} = P\{Q > d\mu\}$, we can then solve the effective bandwidth required for QoS guarantee from (10). However, a close-form expression of μ can not be derived; we have to resort to numerical solutions. For convenience, we denote the required effective bandwidth as

$$\mu = f_{FBM}\left(\lambda, \sigma^2, H, d, \epsilon\right). \tag{11}$$

2.2 Multiclass DCF

2.2.1 *IEEE 802.11 DCF*

In the 802.11 DCF mode, a mobile node monitors the channel before starting data transmission. If the channel has been sensed idle for a time interval exceeding a DCF InterFrame Space (DIFS), the node may start transmission; otherwise, the node waits until the channel becomes idle for a DIFS and enters a backoff stage. The time immediately following an idle DIFS is slotted, and a node is allowed to transmit only at the beginning of each *time slot*. At each backoff stage, a random backoff counter is uniformly chosen from $[0, CW - 1]$, where CW is the contention window size in terms of number of slots. The backoff counter decreases by one for each idle time slot and freezes when the channel is sensed busy. When the backoff counter reaches zero, the node starts transmission. After a successful transmission, the contention window is reset to the initial value CW_{min}; the receiver will send back an acknowledge (ACK) frame upon the successful receipt of the data frame after a Short InterFrame Space (SIFS). Upon an ACK timeout, the sender will assume that a collision happened. A collided data frame will be retransmitted according to a new backoff stage, where the contention window size is doubled for each retransmission, up to CW_{max}. A data frame is dropped when the retransmission limit is reached.

The DCF MAC also specifies the optional request-to-send/clear-to-send (RTS/CTS) mechanism to solve the hidden-terminal problem. In the proposed analytical model, we do not consider RTS/CTS for simplicity. However, the model can be readily extended to include RTS/CTS.

2.2.2 *Contention window based differentiation*

With the DCF MAC, all nodes have the same priority to access the channel and on average achieve the same quality of service. Such an undifferentiated access mode is unfavorable when some QoS-critical applications are integrated with the QoS-resilient applications over the WLAN, or when the AP needs to handle aggregated downlink flows that have a much larger traffic load than that in a mobile node [20]. Therefore, we consider an enhanced DCF with class differentiation for more effective QoS provisioning and higher resource utilization.

In the WLAN, mobile nodes belonging to different classes may have heterogeneous bursty traffic arrival processes with different QoS requirements. The multiclass DCF assigns different contention windows to different classes, with the contention windows sizes properly designed to satisfy QoS requirements and achieve efficient resource utilization. Basically, classes with smaller contention windows have a higher priority to access the channel and therefore occupy a larger portion of the serving capacity. Although the CW based differentiation scheme is very simple compared to other service differentiation schemes [17–19], e.g. the 802.11e standard, we will analytically show that such a simple scheme is efficient in provisioning service differentiation. Moreover, the optimal contention window for each class can be numerically solved to maximize the resource utilization under the QoS constraints.

3 Nonsaturated multiclass DCF model

In this section, we extend the nonsaturated analytical model presented in [16, 20] to analyze a DCF with multiple classes, where each class has a unique traffic arrival process and a unique contention window size. The analytical model will be used in Section 4 to evaluate the statistical multiplexing effect, admission region, and contention window optimization.

3.1 Average backoff time

We consider a WLAN supporting S classes with DCF, and N_i nodes are assigned to class i, $i = 1, 2, \ldots, S$. Time is discretized into slots, and each node is modeled as a discrete time G/G/1 queue. Assume that there is no link layer fragmentation, and one IP packet corresponds to one link layer frame. For a class-i node, the average traffic arrival rate is λ_i packets/slot. Define the *packet service time* as the time period from the instant that a packet begins to be serviced by the MAC layer to the instant that it is either successfully transmitted or dropped after several retransmissions. At the steady state, a class-i node achieves an average service rate of μ_i packets/slot and correspondingly a queue utilization ratio of $\rho_i = \frac{\lambda_i}{\mu_i}$. To maintain a stable queue, it is required that $\rho_i < 1$. According to queueing theory, ρ_i is also equal to the probability that the queue is busy, when the buffer size is large enough to guarantee a lossless system [8].

A class-i node is assigned a minimum contention window of $CW_{i,min}$. All the classes have the same retransmission limit of m_r (set as 7 in 802.11 DCF) and the same maximum backoff stage of m_b (set as 5 in 802.11 DCF). Therefore, the maximum contention window of a class-i node is $CW_{i,max} = 2^{m_b} CW_{i,min}$. The contention window for the kth round (re)transmission of a class-i packet is

$$CW_i(k) = \min\left(CW_{i,max}\ ,\ 2^{k-1} CW_{i,min}\right), \quad k = 1, \ldots, m_r + 1 \qquad (12)$$

with the backoff counter randomly chosen over $[0, CW_i(k) - 1]$. Use p_i to denote the collision probability seen by a class-i node, and assume that transmission collisions happen independent of each other. The average backoff time of the node can then be calculated as

$$\overline{W}_i = \sum_{k=1}^{m_r+1} p_i^{k-1} (1 - p_i)^{I\{k<m_r+1\}} \sum_{j=1}^{k} \frac{CW_i(j) - 1}{2} \qquad (13)$$

where the indicator $I\{A\}$ is equal to 1 if A is true, and equal to 0 otherwise. The indicator is used to include the case that a packet is dropped when the retransmission limit is reached.

3.2 Packet collision probability

We now investigate the collision probability of a tagged class-i node. Define q_i the probability that a class-i node transmits a packet in a certain slot. A collision occurs if at least one of the remaining nodes also transmits in the same time slot. Therefore,

$$p_i = 1 - (1 - q_i)^{N_i - 1} \prod_{j=1, j\neq i}^{S} (1 - q_j)^{N_j}. \qquad (14)$$

Conditioning on a busy or non-empty queue, the transmission probability of a class-i node can be approximated as

$$\tau_i = \frac{E[A_i]}{\overline{W}_i + E[A_i]} \qquad (15)$$

where $E[A_i]$ is the average number of transmission attempts the node made during the backoff time. With the collision probability p_i for each transmission attempt, we have

$$E[A_i] = \sum_{k=1}^{m_r+1} k p_i^{k-1} (1 - p_i)^{I\{k<m_r+1\}} = \frac{1 - p_i^{m_r}}{1 - p_i}. \qquad (16)$$

As the node is busy with probability ρ_i and there is no transmission when the node is empty, we can obtain the unconditional transmission probability

$$q_i = \rho_i \cdot \tau_i + (1 - \rho_i) \cdot 0 = \lambda_i \tau_i / \mu_i. \qquad (17)$$

Substituting the probability into (14), we have

$$p_i = 1 - \left(1 - \tau_i \frac{\lambda_i}{\mu_i}\right)^{N_i - 1} \prod_{j=1, j\neq i}^{S} \left(1 - \tau_j \frac{\lambda_j}{\mu_j}\right)^{N_j}, \quad i = 1, \ldots S. \qquad (18)$$

3.3 Average packet service time

To obtain the QoS of each node over the WLAN, we need to solve the average packet service time so that the G/G/1 queue can be analyzed. During the time interval of $1/\mu_i$, the following events may occur:

- a successful transmission by the tagged class-i node
- successful transmissions by the remaining nodes
- collisions due to multiple simultaneous transmissions
- channel idling

We assume that admission control is in place to keep each node in the stable state, i.e. $\rho_i < 1$ ($i = 1, \ldots, S$), and no packet loss happens. Thus, the average number of packets successfully transmitted by a class-j node during $1/\mu_i$ is λ_j/μ_i. Using T_{S_i} to denote the transmission time of a class-i packet (constant packet size considered for simplicity), the total average transmission time during $1/\mu_i$ is $[1 + (N_i - 1)\frac{\lambda_i}{\mu_i}]T_{S_i} + \frac{1}{\mu_i} \sum_{j=1, j\neq i}^{S} N_j \lambda_j T_{S_j}$.

Before a node successfully transmits the packet, the packet may experience collisions. Using T_{C_i} to denote the collision time a class-i node experiences upon each transmission collision, the average collision time till the successful transmission[2] can be calculated as

$$\overline{T}_{C_i} = \sum_{k=1}^{m_r+1} (k-1)T_{C_i} \cdot p_i^{k-1}(1-p_i) \approx \frac{p_i}{1-p_i} T_{C_i}. \qquad (19)$$

Thus, the total average collision time during $1/\mu_i$ is $\frac{1}{2}[(1+(N_i-1)\frac{\lambda_i}{\mu_i})\overline{T}_{C_i} + \frac{1}{\mu_i}\sum_{j=1,j\neq i}^{S} N_j\lambda_j \overline{T}_{C_j}]$. The factor "$\frac{1}{2}$" is used to get rid of the repetitive count of the collision time, considering that most of the collisions occur due to simultaneous transmissions from two nodes.

Based on the above analysis, we can now obtain the average packet service time as

$$\begin{aligned}\frac{1}{\mu_i} &= \left[1+(N_i-1)\frac{\lambda_i}{\mu_i}\right]T_{S_i} + \frac{1}{\mu_i}\sum_{j=1,j\neq i}^{S} N_j\lambda_j T_{S_j} \\ &+\frac{1}{2}\left[\left(1+(N_i-1)\frac{\lambda_i}{\mu_i}\right)\overline{T}_{C_i} + \frac{1}{\mu_i}\sum_{j=1,j\neq i}^{S} N_j\lambda_j \overline{T}_{C_j}\right] \\ &+\overline{W}_i \quad i=1,\ldots,S \end{aligned} \qquad (20)$$

where T_{S_i} and T_{C_i} ($\overline{T}_{C_i}$) can be obtained with the packet length of each class given. When different classes have different T_{S_i}, the calculation of T_{C_i} ($\overline{T}_{C_i}$) is given in the appendix.

For all the classes, given the arrival rates $\vec{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_S]$, the minimum contention windows $\overrightarrow{CW} = [CW_{1,min}, CW_{2,min}, \ldots, CW_{S,min}]$, the number of nodes $\vec{N} = [N_1, N_2, \ldots, N_S]$, the equation sets of (18) and (20) can be solved numerically to obtain $\vec{p} = [p_1, p_2, \ldots, p_S]$ and $\vec{\mu} = [\mu_1, \mu_2, \ldots, \mu_S]$. Note that τ_i in the equation sets is a function of p_i by combining (13), (15), and (16). With the service time solved, the QoS of each node can then be obtained by analyzing the G/G/1 queue [15, 16].

4 Statistical multiplexing, admission region, and contention window optimization

In the previous section, the analytical model is developed from the perspective of QoS analysis, which is not convenient for the admission region design. Exhaustive search is one approach to find the admission region, by comparing the QoS achieved under a certain $\vec{N}$ with the QoS specifications. However, such an approach is only applicable to a WLAN with two classes [20], which will incur unacceptable computing overhead when several (>2) classes are supported by the WLAN.

As mentioned in Section 1, the maximum MAC throughput and the satisfying QoS can only be achieved in a nonsaturated operating point [3, 14], where the collision probability is small and the packet service time is close to a constant rate for fixed-size packets (i.e. the service time variance due to the collision and backoff is negligible). Such observations indicate that each node in a WLAN can be well modeled as a G/D/1 queue, with CAC applied to maintain the WLAN at a nonsaturated operating point. When the traffic arrival process is known, the single-server queueing analysis can be used to determine the appropriate service rate $\vec{\mu} = [\mu_1, \mu_2, \ldots, \mu_S]$ that satisfies the QoS requirement of each class. In the case with available $\vec{\mu} = [\mu_1, \mu_2, \ldots, \mu_S]$, the equation sets of (18) and (20) can then be used to solve $\vec{p} = [p_1, p_2, \ldots, p_S]$ and the admission region $\vec{N} = [N_1, N_2, \ldots, N_S]$. Particularly, we consider voice traffic modeled as on/off sources and video traffic modeled as FBM sources; the associated single server queueing analysis is presented in Section 2.1.

4.1 Contention window and admission region

With the service rates $\vec{\mu} = [\mu_1, \mu_2, \ldots, \mu_S]$ predetermined by (5) or (11) under the QoS constraints, the minimum contention windows $\overrightarrow{CW} = [CW_{1,min}, CW_{2,min}, \ldots, CW_{S,min}]$ and the admission region $\vec{N} = [N_1, N_2, \ldots, N_S]$ can be jointly solved, if the resource sharing in the WLAN is further constrained by a certain *fairness policy*.

Using V_i to denote the normalized throughput of a class-i node, we have

$$V_i = (1-p_i^{m_r+1})\rho_i\tau_i \approx \rho_i\tau_i. \qquad (21)$$

A possible fairness policy can be the requirement that the throughputs of different classes follow a proportional relationship as

$$\frac{N_i V_i}{N_1 V_1} = r_i, \quad \text{for } i = 2, \ldots, S. \qquad (22)$$

Without loss of generality, we assume that class-1 requires a high priority to access the channel and is assigned a small contention window as $CW_{1,min} \in [1, 32]$. By changing $CW_{1,min}$ from 1 to 32 and solving the equation sets of (18), (20), and (22) repeatedly, we can find an optimal $CW^*_{1,min}$, which maximizes the total resource utilization as $\sum_{i=1}^{S} N_i^* V_i^*$. Correspondingly, we obtain the optimal contention windows for all the other classes $[CW^*_{2,min}, CW^*_{3,min}, \ldots, CW^*_{S,min}]$ and the maximum admission region $[N_1^*, N_2^*, \ldots, N_S^*]$.

[2] The collision time associated with the dropped packets is ignored. When the admission control is applied to maintain the collision probability at a small value, the packet dropping probability at the MAC layer is close to zero.

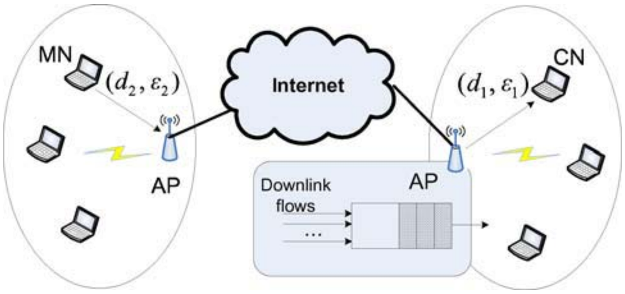


Fig. 1 End-to-end communications with downlink flows multiplexed at AP

4.2 Multiplexing at AP

The joint contention window and admission region solution can be applied to analyze the statistical multiplexing at the AP. As an illustration, we consider a WLAN supporting N mobile nodes with an AP and each node has a two-way voice conversation through the AP with a correspondence node (CN) outside of the WLAN. The traffic in both directions are modeled as on/off flows with the same parameters. All the downlink flows are multiplexed at the AP. The AP is assigned of class-1 with a minimum contention window of $CW_{1,min}$, and all the mobile nodes are assigned of class-2 with a minimum contention window of $CW_{2,min} = rCW_{1,min}$. The factor r is the CW differentiation ratio between the two classes.

Assume that the CNs are in another two-class WLAN, as shown in Fig. 1. The AP and the mobile node specify a stochastic delay bound of d_1 (for downlink in the destination WLAN) and d_2 (for uplink in the source WLAN) respectively to guarantee an end-to-end delay bound of $d = d_1 + d_2$, assuming a negligible delay experienced in the wireline network. Denote the QoS requirements as $P\{D_1 > d_1\} \le \epsilon_1$ and $P\{D_2 > d_2\} \le \epsilon_2$. It is not difficult to see that

$$\begin{aligned} P\{D_1 \le d_1\}P\{D_2 \le d_2\} &\le P\{D_1 + D_2 \le d_1 + d_2\} \\ &= P\{D \le d_1 + d_2\}. \end{aligned} \tag{23}$$

After some manipulation of (23), we can have

$$\begin{aligned} P\{D > d_1 + d_2\} &\le P\{D_1 > d_1\} + P\{D_2 > d_2\} \\ &\quad - P\{D_1 > d_1\}P\{D_2 > d_2\} \\ &\approx P\{D_1 > d_1\} + P\{D_2 > d_2\} \\ &\le \epsilon_1 + \epsilon_2. \end{aligned} \tag{24}$$

Therefore, the delay violation probabilities for the AP (ϵ_1) and the mobile node (ϵ_2) can also be assigned as $\epsilon = \epsilon_1 + \epsilon_2$.

To satisfy the QoS requirement, the minimum packet service rate of the two classes should satisfy

$$\mu_1 = \frac{NR_p(t_{off} \log \epsilon_1 - Nd_1)}{t_{off} \log \epsilon_1 - Nd_1/p_{on}} \tag{25}$$

$$\mu_2 = \frac{R_p(t_{off} \log \epsilon_2 - d_2)}{t_{off} \log \epsilon_2 - d_2/p_{on}}. \tag{26}$$

In the two-class WLAN, the equation sets of (18) and (20) are simplified as

$$p_1 = 1 - \left(1 - \tau_2 \frac{p_{on} R_p}{\mu_2}\right)^N \tag{27}$$

$$p_2 = 1 - \left(1 - \tau_1 \frac{p_{on} N R_p}{\mu_1}\right)\left(1 - \tau_2 \frac{p_{on} R_p}{\mu_2}\right)^{(N-1)} \tag{28}$$

$$\frac{1}{\mu_1} = T_{S_1} + N \frac{p_{on} R_p}{\mu_1} T_{S_2} + \frac{1}{2}\left[\overline{T}_{C_1} + N \frac{p_{on} R_p}{\mu_1} \overline{T}_{C_2}\right] + \overline{W}_1 \tag{29}$$

$$\begin{aligned} \frac{1}{\mu_2} &= \left[1 + (N-1)\frac{p_{on} R_p}{\mu_2}\right] T_{S_2} + \frac{p_{on} N R_p}{\mu_2} T_{S_1} \\ &\quad + \frac{1}{2}\left[\left(1 + (N-1)\frac{p_{on} R_p}{\mu_2}\right)\overline{T}_{C_2} + \frac{p_{on} N R_p}{\mu_2} \overline{T}_{C_1}\right] \\ &\quad + \overline{W}_2 \end{aligned} \tag{30}$$

To investigate the impact of the contention window on the admission region, we consider $CW_{1,min}$ pre-configured, and solve (p_1, p_2, r, N) from the equation set (25)–(30).

4.3 Resource sharing between voice and video

The equation sets of (18) and (20) in fact represents a general analytical framework which enables the analysis of resource sharing between voice and video traffic over a WLAN. Consider a practical scenario, where the mobile nodes in a WLAN also download video traffic in addition to the two-way voice communications shown in Fig. 1. All the video traffic is multiplexed into a video downlink buffer, indexed as class-3, which coexists with the class-1 voice downlink buffer at the AP. There is no up-link video traffic. Assume that the aggregate video traffic is modeled as an FBM process with parameters $(\lambda_v, \sigma_v^2, H_v)$. We investigate the impact of the self-similar video traffic on the voice admission region N (i.e., the number of two-way voice conversations).

Specifically, in the voice/video integrated scenario, there are three classes of traffic: class-1 for downlink aggregated voice traffic, class-2 for uplink per-flow voice traffic, and class-3 for downlink aggregated video traffic. The traffic arrival rates are $\lambda_1 = Np_{on}R_p$, $\lambda_2 = p_{on}R_p$, and $\lambda_3 = \lambda_v$. The number of channel competing queues for each class are $N_1 = 1$, $N_2 = N$, and $N_3 = 1$. Assume that the delay in the wireline network is negligible, and the uplink delay in the WLAN under CAC is also very small [3, 14]; thus, the end-to-end delays for voice and video are mainly due to the queueing delays at the down-link buffers. According to a QoS constraint of stochastic delay bound, the effective bandwidth requirements of μ_1 and μ_3 can be determined according to (5) and (11), respectively. Applying (18) and (20) to the three-class context, we can have six equations regarding p_1, p_2, p_3, $1/\mu_1$, $1/\mu_2$, $1/\mu_3$, respectively. To achieve balanced uplink and downlink throughput for the two-way voice traffic, we can have one extra equation as

$$\mu_1\left(\frac{1}{\mu_1}-\overline{W}_1\right)=\mu_2\left(\frac{1}{\mu_2}-\overline{W}_2\right) \tag{31}$$

which indicates that an uplink voice frame observes the same channel busyness ratio (CBRO) [3] as that observed by an downlink voice frame at the balance point. Now with the seven equations obtained above, we can then solve (p_1, p_2, p_3, μ_2, N, $CW_{2,min}$, $CW_{3,min}$), if $CW_{1,min}$ is also given in addition to μ_1, μ_3, N_1, and N_3.

It is noteworthy that according to Section 4.1, we may solve (p_1, p_2, p_3, N, $CW_{2,min}$, $CW_{3,min}$) from the six equations obtained from (18) and (20), with μ_1, μ_2, μ_3, N_1, N_3, and $CW_{1,min}$ given. We find such an approach sometimes results in negative solutions, due to the complex correlations among the multiple parameters. Therefore, we analyze a relaxed case, where μ_2 is determined by the MAC model. Equation (31) in fact implies a downlink/uplink fairness policy that the downlink throughput at the AP should be approximately N times of the uplink throughput at a mobile node.

5 Numerical results

In this section, we use numerical results to demonstrate that the admission region can be significantly improved by exploiting the statistical multiplexing and properly configuring the contention windows of different classes. The impact of LRD traffic on the WLAN capacity is also investigated. We use Maple [27] to obtain all the numerical results.

We consider a WLAN with the same parameters as those used in [14], which are listed in Table 1. In addition, ACK frames are transmitted at a basic rate of 1 Mbps, and DATA frames at the channel rate of 11 Mbps. The on/off voice traffic is considered with $t_{on} = t_{off} = 300$ ms. During the on periods, a one-way voice flow is generated at a rate of 32 Kbps with a fixed packet size of 160 bytes. Correspondingly, $R_p = 25$ packets/s, or 5×10^{-4} packets/slot. The packet transmission time and the packet collision time are constant as

$$T_S = T_{DATA} + SIFS + T_{ACK} + DIFS = 707.27us$$

$$T_C = T_{DATA} + ACK_{timeout} + DIFS = T_S = 707.27us.$$

Table 1 IEEE 802.11 DCF parameters

Bit rate for DATA frames	11 Mbps
Bit rate for ACK frames	1 Mbps
Bit rate for PLCP & Preamble	1 Mbps
Slot time	20 μs
SIFS	10 μs
DIFS	50 μs
PLCP & preamble	24 bytes
MAC header	28 bytes
IP header	20 bytes
DATA frame	PLCP & preamble + MAC header + IP header + payload
ACK fame	PLCP & preamble + 14 bytes

5.1 Accuracy of the multiplexing analysis

In the first example, we analyze an 802.11 DCF network supporting one-way voice flows without AP. The analytical results are then compared with the simulation results presented in [14] to demonstrate the accuracy of the analysis.

Without the AP, only class-2 mobile nodes exist. We set $CW_{2,min} = 32$, $m_b = 5$, and $m_r = 7$. Each node generates an on/off voice flow. The QoS requirement at each node is set as a stochastic delay bound of $d_2 = 150$ ms with $\epsilon_2 = 0.01$ for excellent voice quality. The modeling equations of the single-class network can be simplified as

$$\mu_2 = \frac{R_p(t_{off}\log\epsilon_2 - d_2)}{t_{off}\log\epsilon_2 - d_2/p_{on}} = 22.77\text{packets/s}$$

$$p_2 = 1-\left(1-\tau_2\frac{p_{on}R_p}{\mu_2}\right)^{(N-1)} \tag{32}$$

$$\frac{1}{\mu_2} = \left[1+(N-1)\frac{p_{on}R_p}{\mu_2}\right]T_S + \frac{1}{2}\left[1+(N-1)\frac{p_{on}R_p}{\mu_2}\right]\overline{T}_C + \overline{W}_2 \tag{33}$$

where $\overline{T}_C = \frac{p_2}{1-p_2}T_C$ and τ_2 is a function of p_2 according to (13), (15), and (16). From (32) and (33), we can obtain $p = 0.5048$ and $N = 70.43$ ($\lfloor N \rfloor = 70$). Correspondingly, the average backoff time $\overline{W}_2 = 111.87$ slots.

In [14], a peak-rate based admission control is proposed. Based on the observation that WLAN achieves the maximum throughput with satisfactory delay performance when the channel busyness ratio reaches 0.92, the admission region based on peak rate allocation can be calculated as

Table 2 Admission regions under different burstness parameter and stochastic delay bound, without AP

		$d_2 = 150$	$d_2 = 300$	$d_2 = 400$
$p_{on} = 0.5$	μ_2 (pkts/s)	22.77	21.11	20.42
	N	70.43	69.74	69.36
	u_b	0.9510	0.9518	0.9523
$p_{on} = 0.4$	μ_2 (pkts/s)	21.80	19.72	18.70
	N	87.71	86.47	85.80
	u_b	0.9511	0.9523	0.9529
$p_{on} = 0.3$	μ_2 (pkts/s)	20.35	17.65	16.41
	N	115.50	113.09	111.80
	u_b	0.9516	0.9536	0.9544

$\frac{0.92}{25\times707.27\times10^{-6}} \approx 52$. However, the simulation results show that the WLAN can in fact support up to 76 on/off voice flows with a satisfactory delay performance. Our analytical admission region of 70 is quite close to the simulation results, which demonstrates that the proposed analytical model can effectively exploit the statistical multiplexing of on/off flows over the DCF MAC channel. Furthermore, the channel busyness ratio obtained from our analysis is $u_b = (\frac{1}{\mu_2} - \overline{W}_2) * \mu_2 \approx 0.95$, which is also very close to the optimal operating point of 0.92 found by simulations in [3, 14]. The conservativeness of the analytical admission region is mainly due to two reasons: (1) the expression (4) used to determine the packet service rate is a conservative estimation of the buffer overflow probability [8]; (2) the discrete G/G/1 queue at each node is approximated by a G/D/1 queue.

5.2 Impact of the burstness and delay bound

In this example, we further investigate the impact of the burstness and the stochastic delay bound on the admission region. The admission regions are calculated from the equation set (32) and (33) under different p_{on} and d_2, while all the other configurations are the same as those used in the previous example. The results are listed in Table 2, from which we can have the following observations.

(1) With a given delay bound, the burstier the traffic source (i.e., the smaller the p_{on}), the lager the admission region. The reason is that a larger statistical multiplexing gain can be achieved over burstier traffic sources.

(2) With a given burstness parameter (i.e. the activity factor) p_{on}, when the delay bound is increased, the calculated admission region does not increase, but decreases slightly. Such an observation is in contradiction with the fact that a looser QoS constraint can allow more traffic flows in service. The analytical deviation is due to that in the nonsaturated DCF model, the G/G/1 queue interacts with the MAC model only through first-order statistics, i.e. the average traffic arrival rate and average packet service rate. In this example, when the traffic arrival rate is given and the average service rate is predetermined according to the delay bound, the queue is then decoupled from the MAC model as described by (32) and (33). In other words, the on/off queueing effect is shielded from the MAC model. In (33), when the required service rate μ_2 decreases with the relaxed delay bound, the MAC channel will see the input queue with an increased packet transmission probability $\rho = \frac{p_{on}Rp}{\mu_2}$, which leads to a higher collision probability; therefore the obtained admission region decreases slightly. This analytical deviation can be avoided when the G/G/1 queue and the MAC model are well coupled, which will be demonstrated in the next example.

(3) Under all the configurations, the calculated channel busyness ratio is steadily around 0.95, which is in consistent with the observation in [3] that the optimal operating point is insensitive to the number of active nodes.

5.3 Multiplexing at AP and contention window optimization

In the third example, we illustrate that the statistical multiplexing gain can be further improved by multiplexing at the AP, in a networking scenario as shown in Fig. 1. For N mobile nodes having two-way voice conversations, there are $2N$ on/off flows. For the QoS guarantee, we first consider a configuration that peak rate allocation ($\mu_2 = R_p$) is used in each mobile node for $d_2 = 0$, and the AP guarantees a stochastic delay bound of $d_1 = 150$ ms with $\epsilon_1 = 0.01$. Such a network can be analyzed by solving the equation set of (25), (27)–(30).

To investigate the impact of the contention window on the admission region, we change the value of $CW_{1,min}$ from 1 to 86 and solve (p_1, p_2, r, N) for each value, correspondingly. The obtained "$2N$ vs. $CW_{1,min}$" and "r vs. $CW_{1,min}$" curves are plotted in Fig. 2, with $p_{on} = 0.5$ and 0.3, respectively. From the figure, we have the following observations.

(1) An optimal $CW_{1,min}$ exists to maximize the admission region, and the optimal CW differentiation ratio also varies with $CW_{1,min}$. In Fig. 2(a), under $t_{off} = 300$ ms, $p_{on} = 0.5$, the maximum admission region $2N \approx 89.41$ ($\lfloor 2N \rfloor = 89$) is achieved at $CW^*_{1,min} = 12$, where $r \approx 24$; the region is increased by $\frac{89-70}{70} \approx 27.14\%$, compared to the case without the AP, due to the extra statistical multiplexing gain at the AP. Note that the contention window differentiation is implicitly required by the fairness rule of balanced downlink/uplink throughput.

(2) Comparing Fig. 2(a) to Fig. 2(b), we can see that when the input sources are burstier with a smaller $p_{on} = 0.3$, the maximum admission region increases considerably due to a lower average source rate and efficient statistical multiplexing. In Fig. 2(b), the maximum admission

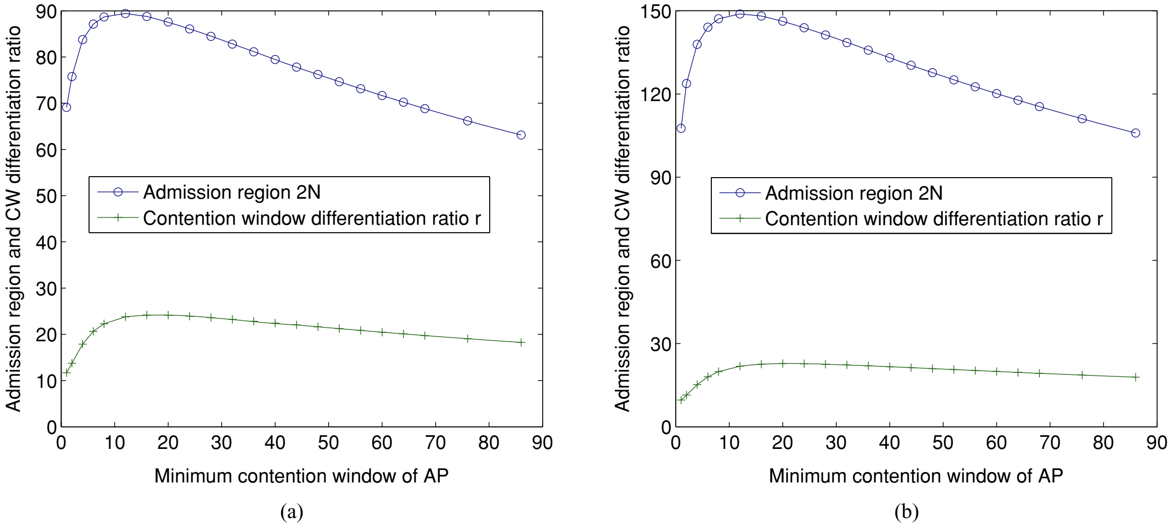


Fig. 2 The impact of the contention window on the admission region when downlink flows are aggregated at the AP, with $d_1 = 150$ ms, $d_2 = 0$, $t_{off} = 300$ ms. (a) $p_{on} = 0.5$, (b) $p_{on} = 0.3$

region $2N \approx 148.86$ is also achieved at $CW^*_{1,min} = 12$, with an increase of $\frac{148-115}{115} \approx 28.70\%$ compared to the case without the AP (the results in Table 2).

(3) We have the same $CW^*_{1,min}$ in Fig. 2(a) and 2(b), showing that the $CW^*_{1,min}$ is insensitive to p_{on}.

(4) We also calculate the curves as those given in Fig. 2, with t_{off} changing to 600 ms for p_{on} set as 0.5 and 0.3, respectively. We have the same observations as those in the case of $t_{off} = 300$ ms. When we compare the corresponding curves obtained with $t_{off} = 300$ ms and 600 ms, respectively, we can see that the variation of t_{off} affects the size of $CW^*_{1,min}$, which changes to 8 when t_{off} increases to 600 ms. In addition, the maximum admission region under $t_{off} = 600$ is slightly smaller than the corresponding one under $t_{off} = 300$. The reason is that with a given p_{on}, a larger t_{off} implies a larger t_{on}. As $R_p > \mu_1$, a mobile node performs like a saturated node at the "on" state; a longer "on" period is prone to cause more collisions and thus a smaller admission region.

QoS-Relaxation Effect – With downlink multiplexing at the AP, we also investigate the impact of QoS requirement on the admission region. In Fig. 3, the admission regions under two different stochastic delay bounds are compared. With a looser delay constraint of 400 ms, the maximum admission region achieved is larger than that with a delay constraint of 150 ms, where the $CW^*_{1,min}$ for these two cases are 12 and 8, respectively. The QoS-relaxation effect is more obvious when $CW_{1,min}$ is larger than the optimal size.

In Section 5.2, we pointed out that the QoS-relaxation effect is not properly captured, when the G/G/1 queue is decoupled from the MAC model. However, with the AP multiplexing, we can see that the equation set for the G/G/1 queue and the MAC model, i.e. (25), (27)–(30), are well coupled in solving the admission region. To further demonstrate the QoS-relaxation effect, we recalculate the admission regions under the burstness and delay configurations given in Table 2, with $CW_{1,min} = 10$ and $r = 20$. The results are listed in Table 3, where the admission region increases with the relaxed delay constraint. The effective bandwidth of a downlink flow, i.e. μ_1/N, is also presented in the table, which is very close to the average packet arrival rate, reflecting the efficient resource utilization by statistical multiplexing [11]. We also give the channel busyness ratios seen by a mobile node and by the AP, i.e. u_b^{MN} and u_b^{AP}. In all the configurations, we have $u_b^{MN} \approx u_b^{AP}$, which implies that the contention window differentiation guarantees the fair resource sharing between the uplink and downlink traffic. Moreover,

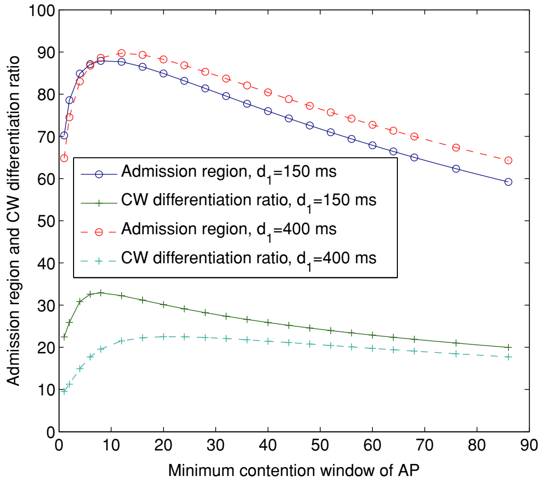


Fig. 3 The impact of the QoS requirement on the admission region when downlink flows are aggregated at the AP, with $d_2 = 0$, $t_{off} = 600$ ms, $p_{on} = 0.5$

Table 3 Admission regions under different burstness parameter and stochastic delay bound, with downlink multiplexing at AP

		$d_2 = 150$	$d_2 = 300$	$d_2 = 400$
$p_{on} = 0.5$	$\frac{\mu_1}{N}$ (pkts/s)	13.68	13.11	12.96
	$2N$	88.32	90.16	90.65
	u_b^{MN}	0.9015	0.8994	0.8988
	u_b^{AP}	0.9166	0.9182	0.9186
$p_{on} = 0.4$	$\frac{\mu_1}{N}$ (pkts/s)	10.94	10.47	10.36
	$2N$	110.43	112.76	113.36
	u_b^{MN}	0.9014	0.8992	0.8986
	u_b^{AP}	0.9166	0.9183	0.9187
$p_{on} = 0.3$	$\frac{\mu_1}{N}$ (pkts/s)	8.13	7.82	7.74
	$2N$	147.88	150.71	151.43
	u_b^{MN}	0.9008	0.8988	0.8983
	u_b^{AP}	0.9170	0.9185	0.9189

Table 4 Admission regions under different QoS configurations

	ϵ_1				
	0.0005	0.0025	0.0050	0.0075	0.0095
$d_1 = 50$ ms	82.62	84.27	85.04	85.51	85.79
$d_1 = 100$ ms	86.82	87.80	88.21	88.45	88.58

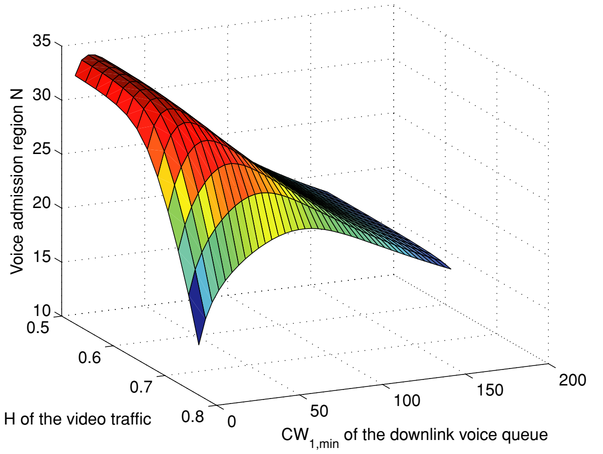


Fig. 4 The impact of the LRD video traffic on the voice admission region

we again find that the channel busyness ratios are all close to the optimal operating point of 0.92 given in [3].

Delay over Two Hops – The statistical multiplexing effect at the AP can also be illustrated by investigating the admission regions under different delay configurations over the uplink and downlink. In Table 4, we list the admission regions obtained at $CW_{1,min} = 12$, under different combinations of $(d_1, d_2, \epsilon_1, \epsilon_2)$ while maintaining $d_1 + d_2 = 150$ ms and $\epsilon_1 + \epsilon_2 = 0.01$. From the table we can see that, the looser the QoS constraint on the AP, the larger the admission region. The reason is that the looser QoS constraint allows better exploitation of the statistical multiplexing effect at the AP; the peak rate allocation at mobiles nodes (i.e. maximum delay allowed at the AP) can therefore lead to the maximum admission region.

5.4 Impact of the self-similar traffic

In this example, we give a numerical analysis of the scenario described in Section 4.3, where a downlink video queue, indexed as class-3, shares the resources with two-way voice conversations. We consider on/off voice flows with $t_{off} = 300$ ms and $p_{on} = 0.5$. The downlink video traffic is modeled as a FBM process, and we assume that it consumes the bandwidth approximately as much as 20 voice flows. Particularly, $\lambda_v = 20 p_{on} R_p$, $\sigma_v^2 = 0.8\lambda_v$, and the value of H is changed to reflect the degree of long-range dependence. The stochastic delay bounds for both video and voice traffic (at the downlink queues) are set as 150 ms with violation probability bound of 0.01. Given a value of H, we can solve the admission region N of voice traffic according to the analytical model presented in Section 4.3. Specifically, we solve the parameters $(p_1, p_2, p_3, \mu_2, N, CW_{2,min}, CW_{3,min})$ from (18), (20), and (31).

With H changing from 0.5 to 0.74 at a step size of 0.02 and $CW_{1,min}$ changing from 8 to 160 at a step size of 4, the three-dimensional plot "N vs. H and $CW_{1,min}$" is shown in Fig. 4. From the figure, we have the following observations. (1) For a given value of H (i.e., a given level of LRD in the video traffic), an optimal $CW_{1,min}^*$ exists in the three-class scenario which leads to the maximum voice admission region. The optimal contention window configuration for other two classes can also be obtained correspondingly. (2) For a given $CW_{1,min}$, the impact of the H value on the voice admission region is also clearly demonstrated in the figure. When the value of H increases, the voice admission region N decreases considerably. The reason is that a larger Hurst parameter represents a stronger long-range dependence within the video traffic, which requires a larger service capacity to maintain the QoS level of the video traffic [25, 26]; therefore, the leftover capacity for voice traffic reduces. For example, when $H = 0.5$, the FBM process is a self-similar Gaussian process but with short-range dependence; we find that the optimal voice admission region is $\lfloor 2N^* \rfloor = 66$, which is less by 22 voice flows than the admission region without video traffic (compared to the result given in Table 3). The result is intuitively reasonable, as the video traffic under our consideration is approximately equivalent to an aggregate of 20 voice flows and an SRD Gaussian process is a reasonable model for the traffic aggregate according to the Central Limit Theorem. However, when H increases to 0.74, the admission region $\lfloor 2N^* \rfloor$ reduces to 44. (3) The value of H also impacts the configuration of the optimal contention windows. For a better illustration, the "N vs. $CW_{1,min}$ curves under different H values are plotted in Fig. 5. From the figure, it can be seen

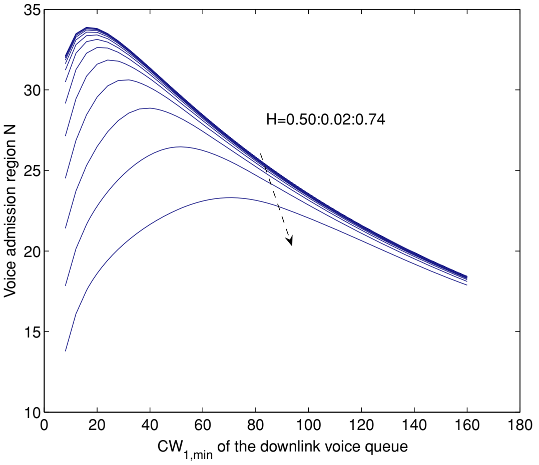


Fig. 5 The impact of the LRD video traffic on the contention window configuration

that $CW^*_{1,min}$ increases along with the increment of H in the particular example considered here.

6 Conclusions

In this paper, we present an analytical framework to study the QoS provisioning and resource allocation in a multiclass DCF WLAN. In particular, the G/G/1 based nonsaturated DCF model [15, 16] is extended to include the class differentiation. Mobile nodes belonging to different classes may have heterogeneous traffic arrival processes with different QoS requirements, and each class is assigned a unique contention window by the DCF for services differentiation. Assuming admission control in place, the single server queueing analysis for both SRD on/off sources and LRD FBM traffic is integrated with the multiclass nonsaturated DCF model, to analytically investigate the admission region, statistical multiplexing gain, impact of self-similarity or long-range dependence, and optimal configuration of contention windows. The accuracy of the proposed analysis and the statistical multiplexing gain are validated by comparing our analytical results with the simulation results presented in [14]. We also present numerical results to demonstrate that the statistical multiplexing can be further improved by multiplexing downlink flows at the access point, while the fair resource sharing between the downlink and uplink traffic is achieved by an optimal contention window design. Moreover, the admission region in a video/voice integrated scenario is also numerically investigated, where the video traffic is modeled as an FBM process. For further work, we are working to achieve the generality of the proposed analytical framework for analyzing any CSMA/CA based MAC protocol.

Appendix: Calculation of the transmission collision time

When all the nodes in a WLAN have the same packet load and therefore the same packet transmission time, the duration of a transmission collision is the same as a packet transmission time. However, if the nodes associated with different classes have different packet loads and therefore different packet transmission times, the duration of a transmission collision will depend on the nodes involved in the collision.

With a random selection of the backoff counter, it is not difficult to prove that most of the collisions happen between two nodes. Consider a tagged class-i node in transmission, and use $p_{s|i}$ to denote the conditional probability that a class-s node is transmitted in the same time slot with the tagged class-i node, leading to a collision. We have

$$p_{i|i} = \binom{N_i - 1}{1} q_i \left(1 - q_i\right)^{N_i - 2} \cdot \prod_{j=1, j\neq i}^{S} (1 - q_j)^{N_j} \quad (34)$$

$$p_{s|i} = \binom{N_s}{1} q_s (1 - q_s)^{N_s - 1} \cdot (1 - q_i)^{N_i - 1} \cdot \prod_{j=1, j\neq i,s}^{S} (1 - q_j)^{N_j} \quad \text{for } s \neq i. \quad (35)$$

The duration of a transmission collision that a class-i node experiences can be calculated as

$$T_{C_i} \approx \sum_{j=1}^{S} \frac{p_{j|i}}{P_{norm,i}} \max(T_{S_j}, T_{S_i}), \quad i = 1, \ldots, S \quad (36)$$

where

$$P_{norm,i} = \sum_{j=1}^{s} p_{j|i} = P\{\text{tagged node collides with another}\}. \quad (37)$$

From (37) and (19), we can obtain $\overline{T}_{C_i}$.

References

1. *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications,* IEEE standards 802.11 (January 1997).
2. J.Y. Yeh and C. Chen, Support of multimedia services with the IEEE 802.11 MAC protocol, in: *Proc. IEEE ICC* (2002) pp. 600–604.
3. H. Zhai, X. Chen and Y. Fang, How well can the IEEE 802.11 wireless LAN support quality of service? IEEE Trans. Wireless Commun. 4(6) (November 2005) 3084–3094.
4. G. Bianchi, Performance analysis of the IEEE 802.11 distributed coordination function, IEEE J. Sel. Areas Commun. 18(3) (March 2000) 535–547.
5. F. Cali, M. Conti and E. Gregori, Dynamic tuning of the IEEE 802.11 protocol to achieve a theoretical throughput limit, IEEE/ACM Trans. Networking 8(6) (December 2000) 785–799.

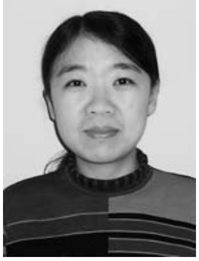
6. Y. TAy and K. Chua, A capacity analysis for the IEEE 802.11 MAC protocol, Wireless Networks 7(2) (March 2001) 159–171.
7. Z. Hadzi-Velkov and B. Spasenovski, Saturation throughput-delay analysis of IEEE 802.11 DCF in fading channel, in: *Proc. IEEE ICC* (2003) pp. 121–126.
8. M. Schwartz, *Broadband Integrated Networks* (New Jersey, Prentice Hall, 1996).
9. U.K. Sarkar, S. Ramakrishnan and D. Sarkar, Modeling full-length video using Markov-modulated Gamma-based framework, IEEE/ACM Trans. Networking 11 (August 2003) 638–649.
10. K. Park and W. Willinger, Self-similar network traffic: an overview, in: *Self-Similar Network Traffic and Performance Evaluation,* eds. K. Park and W. Willinger, John Wiley & Sons, Inc. (2000) pp. 1–38.
11. F.P. Kelly, Notes on effective bandwidth, in: *Stochastic Networks: Theory and Applications*, eds. F.P. Kelly, S. Zachary and I. Ziedins, Oxford, U.K., Oxford Univ. Press (1996) pp. 141–168.
12. S. Garg and M. Kappes, An experimental study of throughput for UDP and VoIP traffic in IEEE 802.11b networks, in: *Proc. IEEE WCNC* (2003) pp. 1748–1753.
13. D.P. Hole and F.A. Tobagi, Capacity of an IEEE 802.11b wireless LAN supporting VoIP, in: *Proc. IEEE ICC* (2004) pp. 196–201.
14. H. Zhai, J. Wang and Y. Fang, Providing statistical QoS guarantee for voice over IP in the IEEE 802.11 wireless LANs, IEEE Wireless Commun. 13(1) (February 2006) 36–43.
15. O. Tickoo and B. Sikdar, Queueing analysis and delay mitigation in IEEE 802.11 random access MAC based wireless networks, in: *Proc. IEEE INFOCOM* (2004) pp. 1404–1413.
16. O. Tickoo and B. Sikdar, A queueing model for finite load IEEE 802.11 random access MAC, in: *Proc. IEEE ICC* (2004) pp. 175–179.
17. I. Ada and C. Castelluccia Differentiation mechanisms for IEEE 802.11, in: *Proc. IEEE INFOCOM* (2001) pp. 209–218.
18. Y. Xiao, H. Li and S. Choi, Protection and guarantee for voice and video traffic in IEEE 802.11e wireless LANs, in: *Proc. IEEE INFOCOM* (2004) pp. 2152–2162.
19. G. Bianchi, I. Tinnirello and L. Scalia, Understanding 802.11e contention-based prioritization mechanisms and their coexistence with legacy 802.11 stations, IEEE Network 19(4) (July/August 2005) 28–34.
20. L.X. Cai, X. shen, J.W. Mark, L. Cai and Y. Xiao, Voice capacity analysis of WLAN with unbalanced traffic, IEEE Trans. Trans. Veh. Technol. 55(3) (May 2006) 752–761.
21. W. Song, W. Zhuang and Y. Cheng, Load balancing for cellular/WLAN integrated networks, IEEE Network, to appear.
22. J. Beran, R. Sherman, M. Taqqu and W. Willinger, Long-range dependence in variable-bit-rate video traffic, IEEE Trans. Commun., 43 (February/March/April 1995) 1566–1579.
23. M.M. Krunz and A.M. Makowski, Modeling video traffic using M/G/∞ input processes: A compromise between Markovian and LRD models, IEEE J. Select. Areas Commun. 16 (June 1998) 733–748.
24. J. Choe and N. Shroff, A central-limit-theorem-based approach for analyzing queue behavior in high-speed networks, IEEE/ACM Trans. Networking 6 (October 1998) 659–671.
25. I. Norros, On the use of fractal brownian motion in the theory of connectionless networks, IEEE J. Select. Areas Commun. 13 (August 1995) 953–962.
26. J. Choe and N. Shroff, Queueing analysis of high-speed multiplexers inculding long-range dependent arrival processes, in: *Proc. IEEE INFOCOM'99* 2 (March 1999) 617–624.
27. Andre Heck, *Introduction to Maple* (3rd ed.). (Springer-Verlag, New York, 2003).
28. International Telecommunicaion Union, One-way transmission time (May 2003).

Yu Cheng received the B.E. and M.E. degrees in Electrical Engineering from Tsinghua University, Beijing, China, in 1995 and 1998, respectively, and the Ph.D. degree in Electrical and Computer Engineering from the University of Waterloo, Waterloo, Ontario, Canada, in 2003. From September 2004 to July 2006, he was a postdoctoral research fellow in the Department of Electrical and Computer Engineering, University of Toronto, Ontario, Canada. Since August 2006, he has been with the Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, Illinois, USA, as an Assistant Professor. His research interests include service and application oriented networking, autonomic network management, Internet performance analysis, resource allocation, wireless networks, and wireless/wireline interworking. He received a Postdoctoral Fellowship Award from the Natural Sciences and Engineering Research Council of Canada (NSERC) in 2004.

Xinhua Ling received the B. Eng. degree in Radio Engineering from Southeast University, Nanjing, China in 1993 and the M. Eng. degree in Electrical Engineering from the National University of Singapore, Singapore in 2001. He is currently pursuing the Ph.D. degree in the Department of Electrical and Computer Engineering at the University of Waterloo, Ontario, Canada. From 1993 to 1998, he was an R&D Engineer in Beijing Institute of Radio Measurement, China. From February 2001 to September 2002, he was with the Centre for Wireless Communications (currently Institute for Infocom Research), Singapore, as a Senior R&D Engineer, developing the protocol stack for UE in the UMTS system. His general research interests are in the areas of cellular, WLAN, WPAN, mesh and ad hoc networks and their internetworking, focusing on protocol design and performance analysis.

Lin X. Cai received the B.Sc. degree in computer science from Nanjing University of Science and Technology, Nanjing, China, in 1996 and the MASc. degree in electrical and computer engineering from the University of Waterloo, Waterloo, Canada, in 2005. She is currently working toward the Ph.D. degree in the same field at the University of Waterloo.

Her current research interests include network performance analysis and protocol design for multimedia applications over wireless networks.



Wei Song received the B.S. degree in electrical engineering from Hebei University, China, in 1998 and the M.S. degree in computer science from Beijing University of Posts and Telecommunications, China, in 2001. She is currently working toward the Ph.D. degree at the Department of Electrical and Computer Engineering, University of Waterloo, Canada. Her current research interests include resource allocation and quality-of-service (QoS) provisioning for the integrated cellular networks and wireless local area networks (WLANs).

Weihua Zhuang received the Ph.D. degree in electrical engineering from the University of New Brunswick, Canada. Since October 1993, she has been with the Department of Electrical and Computer Engineering, University of Waterloo, Canada, where she is a Professor. Dr. Zhuang is a co-author of the textbook *Wireless Communications and Networking* (Prentice Hall, 2003). Her current research interests include multimedia wireless communications, wireless networks, and radio positioning. She received the Outstanding Performance Award in 2005 and 2006 from the University of Waterloo and the Premier's Research Excellence Award in 2001 from the Ontario Government for demonstrated excellence of scientific and academic contributions. She is the Editor-in-Chief of *IEEE Transactions on Vehicular Technology* and an Editor of *IEEE Transactions on Wireless Communications*.

Xuemin (Sherman) Shen received the B.Sc.(1982) degree from Dalian Maritime University (China) and the M.Sc. (1987) and Ph.D. degrees (1990) from Rutgers University, New Jersey (USA), all in electrical engineering. He is a Professor and the Associate Chair for Graduate Studies, Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research focuses on mobility and resource management in interconnected wireless/wired networks, UWB wireless communications systems, wireless security, and ad hoc and sensor networks. He is a co-author of three books, and has published more than 300 papers and book chapters in wireless communications and networks, control and filtering. Dr. Shen serves as the Technical Program Committee Chair for IEEE Globecom'07, General Co-Chair for Chinacom'07 and QShine'06, the Founding Chair for IEEE Communications Society Technical Committee on P2P Communications and Networking. He also serves as a Founding Area Editor for *IEEE Transactions on Wireless Communications*; Associate Editor for *IEEE Transactions on Vehicular Technology*; *KICS/IEEE Journal of Communications and Networks*; *Computer Networks* (Elsevier); *ACM/Wireless Networks*; and *Wireless Communications and Mobile Computing* (John Wiley), etc. He has also served as Guest Editor for *IEEE JSAC*, *IEEE Wireless Communications*, and *IEEE Communications Magazine*. Dr. Shen received the Excellent Graduate Supervision Award in 2006, and the Outstanding Performance Award in 2004 from the University of Waterloo, the Premier's Research Excellence Award in 2003 from the Province of Ontario, Canada, and the Distinguished Performance Award in 2002 from the Faculty of Engineering, University of Waterloo. Dr. Shen is a registered Professional Engineer of Ontario, Canada.

Alberto Leon-Garcia received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Southern California, in 1973, 1974, and 1976 respectively. He is a Full Professor in the Department of Electrical and Computer Engineering, University of Toronto, ON, Canada, and he currently holds the Nortel Institute Chair in Network Architecture and Services. In 1999 he became an IEEE fellow for "*For contributions to multiplexing and switching of integrated services traffic*".

Dr. Leon-Garcia was Editor for Voice/Data Networks for the *IEEE Transactions on Communications* from 1983 to 1988 and Editor for the *IEEE Information Theory Newsletter* from 1982 to 1984. He was Guest Editor of the September 1986 Special Issue on Performance Evaluation of Communications Networks of the *IEEE Selected Areas on Communications*. He is also author of the textbooks *Probability and Random Processes for Electrical Engineering* (Reading, MA: Addison-Wesley), and *Communication Networks: Fundamental Concepts and Key Architectures* (McGraw-Hill), co-authored with Dr. Indra Widjaja.