A Cross-Layer Approach for WLAN Voice Capacity Planning

Yu Cheng, *Member, IEEE*, Xinhua Ling, *Student Member, IEEE*, Wei Song, Lin X. Cai, Weihua Zhuang, *Senior Member, IEEE*, Xuemin Shen, *Senior Member, IEEE*

***Abstract*— This paper presents an analytical approach to determining the maximum number of on/off voice flows that can be supported over a wireless local area network (WLAN), under a quality of service (QoS) constraint. We consider multiclass distributed coordination function (DCF) based medium access control (MAC) that can provision service differentiation via contention window (CW) differentiation. Each on/off voice flow specifies a stochastic delay bound at the network layer as the QoS requirement. The downlink voice flows are multiplexed at the access point (AP) to alleviate the MAC congestion, where the AP is assigned a smaller CW compared to that of the mobile nodes to guarantee the aggregate downlink throughput. There are six-fold contributions in this paper: 1) a nonsaturated multiclass DCF model is developed; 2) a cross-layer framework is proposed, which integrates the network-layer queueing analysis with the multiclass DCF MAC modeling; 3) the channel busyness ratio control is included in the framework to guarantee the analysis accuracy; 4) the framework is exploited for statistical multiplexing gain analysis, network capacity planning, contention window optimization, and voice traffic rate design; 5) a head-of-line outage dropping (HOD) scheme is integrated with the AP traffic multiplexing to further improve the MAC channel utilization; 6) performance of the proposed cross-layer analysis and the associated applications are validated by extensive computer simulations.**

***Index Terms*— Cross-layer analysis, WLAN voice capacity, contention window optimization, head-of-line outage dropping, nonsaturated MAC modeling**

I. Introduction

In RECENT years, extensive efforts have been made both in academia and industry to provision voice over Internet Protocol (VoIP) over IEEE 802.11 wireless local area networks (WLANs) [1]–[4]. In order to support the toll-quality voice or other multimedia applications with quality of service (QoS) requirements, the WLAN is required to provision some quantitative performance guarantees, e.g. packet loss rate, delay bound, or delay jitter. For Internet QoS provisioning, a bottleneck link is usually modeled as a queueing system at the network layer for QoS analysis. The traffic arrival process in voice/video applications is usually bursty with variable data rate [5], [6]; a proper service rate falling between the average arrival rate and the peak rate, which is defined as the *effective bandwidth* [7], needs to be determined to satisfy the quantitative QoS requirements. Given the link capacity, which is constant in wireline networks, the *network capacity* in terms of the number of effective-bandwidth guaranteed (i.e., QoS guaranteed) traffic flows can then be obtained.

With the WLAN Internet access, the wireless link involved in the end-to-end path is prone to become a bottleneck due to the limited spectrum, channel contention delays, and possible collisions. The QoS analysis and flow (or call) level capacity planning in the WLAN in fact incurs a *cross-layer* design problem, where the effective bandwidth required by the network layer for QoS guarantee needs to be provisioned by the medium access control (MAC) layer under the call admission control (CAC). To the best of our knowledge, there was no such a cross-layer analytical tool available before our work presented in this paper, by which we can determine the network capacity and the proper configuration of the MAC, according to network-layer QoS requirements. The IEEE 802.11 MAC protocol contains two access modes, the mandatory distributed coordination function (DCF) mode and the optional point coordination function (PCF) mode. In this paper, we consider the 802.11 DCF.

The DCF mode is contention based and distributed and, hence, renders effective resource allocation and quantitative QoS control very difficult. Most of the previous works on the DCF MAC performance analysis, e.g. [8]–[10] and the references therein, focus on deriving the channel throughput or average delay from a Markov chain 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 that are normally associated with a bursty arrival process. It is very difficult, if not impossible, to derive an efficient analytical tool from the saturated modeling for effective bandwidth provisioning and call-level network capacity analysis.

In [11], [12], a simple yet accurate analytic model is developed for evaluating the 802.11 DCF in nonsaturated case, where each node is modeled as a discrete time G/G/1 queue[1]. While the nonsaturated model is used in [11], [12] to derive the MAC service time distribution for input traffic with a general arrival distribution, no hint is provided there on how to apply the model to analytically obtain a capacity region under certain

Manuscript received May 15, 2006; revised December 3, 2006.

Y. Cheng is with the Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, IL 60616, USA (email: cheng@iit.edu).

X. Ling, W. Song, L. X. Cai, W. Zhuang, and X. Shen are with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (email: {x2ling, wsong, lcai, wzhuang, xshen}@uwaterloo.ca).

Digital Object Identifier 10.1109/JSAC.2007.070505.

[1] According to the standard Kendall's notation, for G/G/1 and G/D/1 queueing systems, "*G*" stands for a "general" distribution, "*D*" stands for a "deterministic" service time, and "1" represents the number of service channel.

QoS requirements. In [2], it is shown that the maximum DCF MAC capacity and satisfying QoS performance can only be achieved in the nonsaturated case. It is also shown that, when a WLAN works in the proper operation range under CAC, 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. Inspired by the above results, we present in this paper that the queueing analysis at the network layer can be combined with the nonsaturated DCF MAC modeling to form a *cross-layer analytical framework* to investigate the statistical multiplexing gain, QoS guarantee, and call-level network capacity over the WLAN. While the proposed cross-layer framework is generic in traffic models and applications, we focus on the voice over WLAN in this paper.

It should be emphasized that maintaining the WLAN within the proper unsaturated operation range by CAC is essential for generating accurate analytical results from the proposed cross-layer framework; otherwise, the heavy MAC collision will lead to large service time fluctuation and invalidate the G/D/1 modeling. The simulation studies in [2] show that, at the optimal operation point, the *channel busyness ratio* (CBRO) is stable around 0.9 (without request/clear to send, RTS/CTS) or 0.95 (with RTS/CTS) independent of the packet size and number of mobile nodes. Such an observation is exploited in [1] for a measurement-based call admission and rate control in voice/data integrated WLANs. In this paper, we integrate the CBRO control within the cross-layer framework, where the admission region (or call-level network capacity) taking account of the statistical multiplexing among on/off voice flows can be accurately and analytically calculated.

The packet-level buffering and QoS adaption are commonly used to improve the sustainability of a multimedia IP application session under various network load conditions. However, it has been manifested [1]–[3], [11], [12] that the QoS performance over the DCF MAC shows a "good-or-bad" sharp-turning behavior around the operation point, in which case the QoS adaption is ineffective. We demonstrate in this paper that downlink traffic multiplexing at the AP, in two-way conversations, can improve the channel utilization and facilitate the downlink QoS adaption by adjusting the resource allocation to the AP. Moreover, the aggregate downlink rate at the AP is differentiated with the per-flow uplink rate at a mobile node by a multiclass DCF MAC that provisions service differentiation via contention window (CW) differentiation. We extend the nonsaturated DCF modeling of [11], [12] to include the class differentiation. Given the packet-level QoS requirements, both the network capacity and the MAC layer contention windows for the AP and the mobile nodes can be *jointly* determined from our cross-layer framework. In addition, we investigate a head-of-line outage dropping (HOD) scheme, where a head-of-line packet being served by the MAC layer will be dropped when it exceeds the delay bound. We show that the QoS of the WLAN applying the HOD scheme degrades gradually instead of the sharp-turning behavior as the traffic load increases, which can considerably facilitate the QoS adaption and the measurement-based admission control.

In Section II, we give more review on related work. Section III describes the system model. Section IV presents the nonsaturated DCF model with class differentiation. In Section V, the nonsaturated DCF model is combined with the CBRO control and network layer queueing analysis to form the cross-layer analytical framework. Applications of the framework to statistical multiplexing gain analysis, network capacity planning, voice codec rate design are also presented. In Section VI, extensive numerical analysis and computer simulation results are presented to demonstrate the analysis accuracy and the efficiency of the HOD scheme, statistical multiplexing gain, and contention window optimization. Section VII gives the concluding remarks.

II. Related Work

A comprehensive discussion of the cross-layer optimization issues for efficient wireless multimedia transmissions is given in [13]. Various possible interactions from the physical layer (PHY) up to the application layer (APP) in the wireless protocol stack are addressed in [14]–[16]. However, little work on NET/MAC cross-layer design for provisioning QoS guaranteed realtime applications, as demonstrated in this paper, had been reported in the literature.

The DCF MAC has been enhanced to provision service differentiation, e.g. in [17], [18] and the references therein. The standardized differentiation mechanisms, as defined in the enhanced distributed channel access (EDCA) of IEEE 802.11e, include differentiating CW backoff parameters, interframe spacing before data transmission (arbitration interframe space, AIFS), and channel holding times upon the successful channel access (transmission opportunity, TXOP). In this paper, we extend the nonsaturated DCF modeling of [11], [12] to analyze a DCF MAC with CW differentiation. While the CW differentiation is only a subset of the differentiation schemes provided by EDCA, it should not be difficult to include AIFS differentiation into the analytical framework. Integrating TXOP into the nonsaturated DCF model is still an open issue. The CW differentiated DCF provides us simple yet effective service differentiation MAC, which facilitates our effort in revealing insight into the NET/MAC cross-layer design.

The voice capacity of WLANs has been investigated by measurement studies or analytical estimation under simplified assumptions of channel collision [19]–[21]. However, all the studies deal with constant-rate voice flows and do not consider the possible statistical multiplexing gain achievable by exploiting the on/off effect in voice. While the capacity region of on/off voice flows has been studied by computer simulations in [1], the capacity region and associated statistical multiplexing gain are analytically determined under a QoS specification in this paper. It is shown in [3] that the unbalanced traffic distribution due to downlink aggregation can easily make AP the QoS bottleneck under the standard 802.11 DCF. We show in this paper that the CW differentiation can provision a fair resource sharing between the uplink and downlink traffic. A multiplexing-multicast (M-M) scheme is proposed in [4] to improve the downlink performance by aggregating packets from multiple voice flows into a big multicast packet. While the study in [4] focus on suppressing the packet header overhead by multicasting, we emphasize the statistical multiplexing gain by traffic aggregation.

In WLANs, the excessive packet delay due to heavy collision tends to result in head-of-line (HOL) blocking problem

[22], where the subsequent packets in the queue have to wait before the HOL packet is served or dropped. The HOD scheme proposed in this paper to address the HOL blocking issue is closely related to the active queue management scheme [23], [24] that has been well studied in wireline networks; however, there is no much work quantitatively demonstrating the effect of HOD over a WLAN. Particularly, this paper is the first work (as far as we know) that reveals the effect of HOD in alleviating the "good-or-bad" sharp-turning DCF behavior.

III. System Model

In this paper, we develop an analytical framework for WLAN voice capacity planning. A typical application of the network capacity planning is to determine the call-level network capacity or the *admission region*, which guarantees the packet-level QoS for each admitted VoIP flow. The analytical framework can also help to determine the MAC protocol configuration parameters, i.e. the contention window, to tune the WLAN into the optimal operation point for maximum admission region. Another typical scenario in network planning is the *rate planning*. In telephone networks, the call-level capacity is usually planned by the Erlang-B formula to guarantee a target call blocking probability. In order to fit the predetermined call-level capacity and the packet-level QoS requirements, the voice traffic source rate (i.e. the codec rate) needs to be determined properly.

A. IEEE 802.11 DCF MAC

The basic access mode of the IEEE 802.11 MAC layer protocol is Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) protocol based distributed coordination function. In the DCF mode, each time the channel is sensed idle for a time interval exceeding a DCF InterFrame Space (DIFS), the node starts a new or continues an existing 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 time slots in the current stage. The backoff counter decreases by one for each idle time slot and stops when the channel is sensed busy. When the backoff counter reaches zero, the node starts transmission at the beginning of the next time slot. After a successful transmission, the contention window is reset to 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). If the sender does not receive the ACK within a certain time, i.e. ACK timeout, it assumes a collision and then arranges a retransmission according to a new backoff stage with doubled contention window size 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 extended to include the RTS/CTS mechanism.

B. Multiclass DCF

With the DCF MAC, all nodes have the same priority to access the channel and achieve the same quality of service on average. Such a homogeneous access mode is unfavorable when different nodes have different service requirements. Particularly in our case, the AP needs to handle aggregated downlink flows, corresponding to a much larger traffic load than that from a mobile node; it is preferable that the MAC allocates the AP a larger serving capacity to guarantee the downlink throughput. Therefore, we consider an enhanced DCF with class differentiation for more effective QoS provisioning and higher resource utilization.

In the multiclass WLAN, mobile nodes belonging to different classes may have different 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 with 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. We will show that the optimal contention window for each class can be analytically determined to maximize the resource utilization under the QoS constraints.

IV. Nonsaturated Multiclass DCF Model

A. Average Backoff Time

We consider a WLAN supporting S classes with DCF, where class i has N_i nodes, $i = 1, 2, \cdots, S$. Time is discretized into slots, and each node is modeled as a discrete time G/G/1 queue. Assume 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 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 [5].

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 (equal to 7 in 802.11 DCF) and the same maximum backoff stage of m_b (equal to 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

$$CW_i(k) = \min(CW_{i,max}, 2^{k-1} CW_{i,min}), k = 1, \cdots, m_r + 1 \tag{1}$$

with the backoff counter randomly chosen over $[0, CW_i(k) - 1]$. Let p_i denote the packet collision probability seen by a class-i node, and assume that the collision probabilities associated with different nodes are independent of each other. The average backoff time of the node in terms of time slots is then given by

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

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.

B. Packet Collision Probability

We now derive the collision probability of a tagged class-i node. Let q_i denote 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,

$$p_i = 1 - (1-q_i)^{N_i-1} \prod_{j=1, j\neq i}^{S} (1-q_j)^{N_j}. \quad (3)$$

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

$$\tau_i = \frac{E[A_i]}{\overline{W}_i + E[A_i]} \quad (4)$$

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

$$E[A_i] = \sum_{k=1}^{m_r+1} k p_i^{k-1}(1-p_i)^{I\{k<m_r+1\}} = \frac{1-p_i^{m_r}}{1-p_i}. \quad (5)$$

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

$$q_i = \rho_i \cdot \tau_i = \lambda_i \tau_i / \mu_i. \quad (6)$$

Substituting (6) into (3), we have

$$p_i = 1 - (1-\tau_i \frac{\lambda_i}{\mu_i})^{N_i-1} \prod_{j=1, j\neq i}^{S} (1-\tau_j \frac{\lambda_j}{\mu_j})^{N_j}, \quad i = 1, \cdots S. \quad (7)$$

C. 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 an admission control scheme is implemented to keep each node in the stable state, i.e., $\rho_i < 1$ $(i = 1, \cdots, S)$, 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. Letting T_{S_i} denote the transmission time of a class-i packet (assuming a constant packet size 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 a packet, the packet may experience collisions. Letting T_{C_i} denote the collision time that a class-i node experiences upon each transmission collision, the average collision time till the successful transmission[2] can be calculated as

$$\begin{aligned}\overline{T}_{C_i} &= \sum_{k=1}^{m_r+1} (k-1) T_{C_i} \cdot p_i^{k-1}(1-p_i) \\ &= \frac{p_i[1-(m_r+1)p_i^{m_r} + m_r p_i^{m_r+1}]}{1-p_i} T_{C_i} \approx \frac{p_i}{1-p_i} T_{C_i}.\end{aligned} \quad (8)$$

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 obtain the average packet service time for class $i = 1, \cdots, S$ as

$$\begin{aligned}\frac{1}{\mu_i} = &\left[1 + (N_i - 1)\frac{\lambda_i}{\mu_i}\right] T_{S_i} + \frac{1}{\mu_i} \sum_{j=1,j\neq i}^{S} N_j \lambda_j T_{S_j} + \\ &\frac{1}{2}\left[\left(1 + (N_i - 1)\frac{\lambda_i}{\mu_i}\right)\overline{T}_{C_i} + \frac{1}{\mu_i}\sum_{j=1,j\neq i}^{S} N_j \lambda_j \overline{T}_{C_j}\right] + \overline{W}_i\end{aligned} \quad (9)$$

where T_{S_i} and T_{C_i} $(\overline{T}_{C_i})$ can be obtained from the packet length of each class given.

For all the classes, given the arrival rates $\vec{\lambda} = [\lambda_1, \lambda_2, \cdots, \lambda_S]$, the minimum contention windows $\overrightarrow{CW} = [CW_{1,min}, CW_{2,min}, \cdots, CW_{S,min}]$, the numbers of nodes $\vec{N} = [N_1, N_2, \cdots, N_S]$, (7) and (9) can be solved numerically to obtain $\vec{p} = [p_1, p_2, \cdots, p_S]$ and $\vec{\mu} = [\mu_1, \mu_2, \cdots, \mu_S]$. Note that τ_i in the equations is a function of p_i by combining (2), (4), and (5). With $\vec{\mu}$ solved, the QoS of each node can then be obtained by analyzing the G/G/1 queue [11], [12].

V. Cross-Layer Analytical Framework

In the previous section, the DCF MAC model is developed from the perspective of QoS analysis, which will be integrated with the network-layer queueing analysis in this section to form a cross-layer analytical framework for network capacity planning. Specifically, each node in the WLAN under consideration is modeled as a G/D/1 queue, with CAC being applied to maintain the WLAN at a nonsaturated operation point. For the given traffic arrival process, the single-server queueing analysis techniques is used to determine the appropriate service rates $\vec{\mu} = [\mu_1, \mu_2, \cdots, \mu_S]$ that satisfy the QoS requirements of all the classes. With $\vec{\mu} = [\mu_1, \mu_2, \cdots, \mu_S]$, (7) and (9) are then used to compute $\vec{p} = [p_1, p_2, \cdots, p_S]$ and the admission region $\vec{N} = [N_1, N_2, \cdots, N_S]$. In the following, we focus on voice capacity analysis to illustrate the applications of the cross-layer framework. Before going into details of the cross-layer framework, we first give a more thorough examination of the optimal operation point.

[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 negligible.

A. Optimal Operation Point of a WLAN

We demonstrate the constraint effect of the optimal operation point (OOP) through simulating an 802.11b WLAN (without RTS/CTS) supporting on/off voice flows. Each node in the WLAN generates a single on/off flow; the on/off parameters and voice source packet rate follow those used in [1]. The constant voice packet size and therefore the constant T_S and T_C are considered. More details of our simulation system are given in Section VI. In simulations, the channel busyness ratio and channel utilization (CU) are estimated by

$$\text{CBRO} = (V_S T_S + V_C T_C)/T_{sim} \tag{10}$$

$$\text{CU} = V_S T_S / T_{sim} \tag{11}$$

where V_S and V_C denote the numbers of successful and collided transmissions, respectively, in a simulation run, and T_{sim} denotes the system time duration of the simulation run for a given flow number.

From the simulations, we find that the CU starts to drop when the CBRO exceeds 0.9 under 76 voice flows, which is consistent with the simulation results presented in [1]. The reason for the decreasing CU beyond the optimal operation point (identified by the CBRO of 0.9) is that the traffic load is too high to be effectively handled by the DCF MAC. The curves of the packet service time are shown in Fig. 1[3]. In the unsaturated region below the OOP, the average packet service time is small (implying a short contention delay) and remains steady (reflected by a very small standard deviation). The packet sojourn time (i.e., departure time - arrival time) is almost equal to the packet service time (i.e., departure time - MAC service start time), implying a queueing delay close to zero. However, in the operation region exceeding the OOP, the MAC channel becomes saturated due to a large number of collisions, which results in sharply increased packet service time and sojourn time. It is noteworthy that the "good-or-bad" sharp turning behavior at the MAC layer makes the network-layer queueing system ineffective. Such a contradiction can be solved by downlink traffic multiplexing at the AP.

B. Traffic Multiplexing at AP

As we know, each VoIP conversation incurs two VoIP flows, uplink and downlink, respectively. We consider the case that each mobile node communicates through the AP with a correspondence node (CN) outside the WLAN. With the standard DCF MAC, in order to achieve symmetric data rate in uplink and downlink directions, each downlink voice flow needs to occupy a separate queue at the AP for channel contention. In practice, all the downlink flows are multiplexed into a single downlink queue for simplicity, although traffic multiplexing tends to make AP the performance bottleneck over the standard DCF MAC [3]. Another benefit of traffic multiplexing at AP is to alleviate the MAC congestion; with N mobile nodes considered, the number of contending queues can be reduced to $N + 1$ instead of $2N$ queues in the case of separate downlink queues. With traffic multiplexing, the service rate of the downlink queue should be properly provisioned to guarantee the QoS requirements of all the downlink flows, which is to be achieved by properly configured CW-differentiated multiclass DCF MAC. In addition, the MAC should be maintained operating at the OOP for maximum resource utilization and valid G/D/1 queueing analysis.

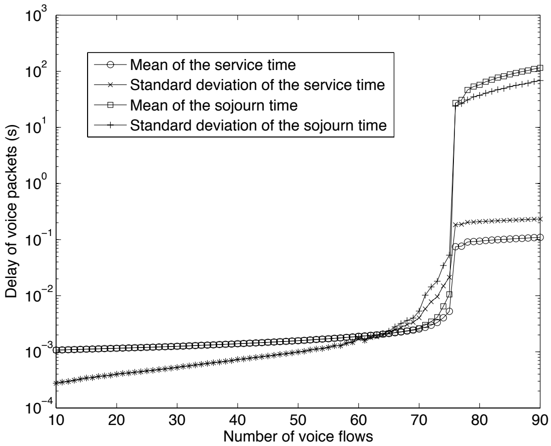


Fig. 1. Performance in a WLAN supporting on/off voice flows.

For each voice flow, the on and off periods are exponentially distributed with average durations of t_{on} and t_{off}, respectively; the activity factor is $p_{on} = t_{on}/(t_{on} + t_{off})$. At the on state, voice packets are generated at a constant rate of R_p packets/slot. Considering that $N(\geq 1)$ on/off flows are multiplexed at the AP, the QoS is to guarantee a stochastic delay bound d, i.e.,

$$P\{D > d\} \leq \epsilon \tag{12}$$

where D is the queueing delay and ϵ the delay bound violation probability. If the queue service rate is μ packets/slot, the delay violation probability can be equivalently mapped to buffer overflow probability at $d\mu$,

$$P\{Q > d\mu\} \leq \epsilon \tag{13}$$

where Q is the queue length of a node.

It is well-known that a conservative approximation of the overflow probability for on/off sources takes an exponential expression (equations (3-42) and (3-43) in [5]), i.e.,

$$P\{Q > x\} \approx \exp\left[-\frac{N(1-\rho)(\alpha+\beta)}{NR_p - \mu}x\right] \tag{14}$$

where $\rho = p_{on}NR_p/\mu$ is the utilization ratio, and $\alpha = 1/t_{on}$, $\beta = 1/t_{off}$ are the transition rates between the on and off states, respectively. After some manipulations of (14), we have

$$P\{Q > d\mu\} \approx \exp\left[-\frac{Nd(\mu/p_{on} - NR_p)}{t_{off}(NR_p - \mu)}\right] \tag{15}$$

Combining (13) and (15), the minimum service rate required to guarantee the QoS is given by

$$\mu = \frac{NR_p(t_{off}\log\epsilon - Nd)}{t_{off}\log\epsilon - Nd/p_{on}}. \tag{16}$$

[3] The results in Fig. 1 are slightly different from those presented in Fig. 3 in [1], because we use an infinite buffer in the simulations, and do not consider background data traffic.

C. Cross-Layer Analysis

The network layer bandwidth requirement should be satisfied by the MAC layer. In addition, the multiclass DCF MAC is used to provision the service differentiation between the AP and the mobile nodes; 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. Consider a constant voice packet size, and use μ_1 and μ_2 to denote the average packet service rate achieved by the MAC for the AP and a mobile node, respectively. In a WLAN supporting N mobile nodes (i.e. $2N$ on/off voice flows), we can form a cross-layer analytical framework including the following set of equations:

$$\mu_1 = \frac{NR_p(t_{off}\log\epsilon - Nd)}{t_{off}\log\epsilon - Nd/p_{on}} \tag{17}$$

$$p_1 = 1-(1-\tau_2\frac{p_{on}R_p}{\mu_2})^N \tag{18}$$

$$p_2 = 1-(1-\tau_1\frac{p_{on}NR_p}{\mu_1})(1-\tau_2\frac{p_{on}R_p}{\mu_2})^{(N-1)} \tag{19}$$

$$\frac{1}{\mu_1} = T_S + N\frac{p_{on}R_p}{\mu_1}T_S + \frac{1}{2}\left[\overline{T}_C + N\frac{p_{on}R_p}{\mu_1}\overline{T}_C\right] + \overline{W}_1 \tag{20}$$

$$\frac{1}{\mu_2} = \left[1+(N-1)\frac{p_{on}R_p}{\mu_2}\right]T_S + \frac{p_{on}NR_p}{\mu_2}T_S + \frac{1}{2}\left[\left(1+(N-1)\frac{p_{on}R_p}{\mu_2}\right)\overline{T}_C + \frac{p_{on}NR_p}{\mu_2}\overline{T}_C\right] + \overline{W}_2 \tag{21}$$

$$\mu_1(\frac{1}{\mu_1} - \overline{W}_1) = \mu_2(\frac{1}{\mu_2} - \overline{W}_2) \tag{22}$$

$$\mu_2(\frac{1}{\mu_2} - \overline{W}_2) = 0.9 \tag{23}$$

In the cross-layer framework, (18) to (21) are the simplifications of the multiclass DCF model, i.e. the equation sets of (7) and (9). Equation (17) indicates the required serving capacity that needs to be provisioned by the MAC to guarantee the network-layer QoS for the downlink traffic. Equation (22) indicates that an uplink packet observes the same CBRO as that observed by an downlink packet, when the MAC guarantees the balanced uplink and downlink throughput. Equation (23) indicates that the maximum network capacity is achieved at the optimal operation point [4]. Note that we calculate the CBRO by $\mu(\frac{1}{\mu} - \overline{W})$, which is in fact a conservative approximation as we ignore the periods where no packet is available for transmission. Such an approximation is preferred due to three reasons: 1) the calculation is reasonably accurate, as the channel utilization is close to 1 at the OOP; 2) the conservative CBRO estimation can lead to a conservative admission region, which is commonly adopted in the CAC to guarantee QoS; 3) the calculation is very simple.

In the framework, we do not explicitly set the QoS requirement of the uplink traffic, which is implicitly guaranteed by keeping the MAC from exceeding the optimal operation point. Note that in the equation set, the average backoff time $\overline{W}_i$ $(i = 1, 2)$ is a function of p_i and $CW_{i,min}$, and μ_1 is a function of N shown by (17). Therefore, we can jointly determine $(p_1,\ p_2,\ CW_{1,min},\ CW_{2,min},\ N,\ \mu_2)$ [5] from the equation set (18) to (23).

It should be emphasized that the cross-layer framework can be applied to WLANs for supporting heterogeneous multimedia applications. For example, in a voice/video integrated WLAN, some nodes may use VoIP services, other nodes may associate with on-line movies via video-streaming applications. As each video stream has a much larger traffic source rate than a voice flow and correspondingly requires a much higher MAC service rate to guarantee the video QoS, we can set a separate queue for each downlink video stream at the AP in addition to the queue aggregating all the downlink voice flows. The admission regions and the CW configurations for both voice and video flows can then be found from the cross-layer framework integrating the network-layer analysis, multiclass DCF modeling, and the OOP control. An in-depth investigation is necessary for future work on the multi-service WLAN.

D. Rate Control

In addition to the admission region or network capacity planning, the voice codec rate can also be specified under the cross-layer framework. Let R_{cd} denote the codec rate and T_{smp} the voice sampling period (or voice packetization interval), which result in a voice packet payload of $R_{cd}T_{smp}$. The transmission time of a MAC DATA frame over the wireless channel of capacity C is

$$T_{DATA} = T_{header} + R_{cd}T_{smp}/C \tag{24}$$

where T_{header} is the overhead time to transmit the headers attached in different layers to the voice data. According to 802.11b, the packet transmission time and the packet collision time are given by

$$T_S = T_{DATA} + SIFS + T_{ACK} + DIFS \tag{25}$$

$$T_C = T_{DATA} + ACK_{timeout} + DIFS = T_S. \tag{26}$$

With a given T_{smp} value and other parameters defined according to the 802.11b standard, T_S and T_C are a function of R_{cd}. In (18) – (23), if the admission region N is predetermined to satisfy a target call blocking probability by the Erlang-B formula, we can then jointly find $(p_1, p_2, CW_{1,min}, CW_{2,min}, R_{cd}, \mu_2)$ from the six equations.

VI. Performance Evaluation

In the following, we present the numerical analysis and computer simulation results from some case studies to illustrate the applications and the associated performance of the proposed cross-layer analytical framework. In all the examples, we use Maple [25] to obtain the numerical results. Considering that the popular simulation tool ns2 has many known bugs in simulating WLANs [26], we develop our

[4] It should be noted that as the optimal operation point of around 0.9 is approximately identified only from computer simulations, a rigorous theoretical investigation of its accurate value and its independence on traffic conditions is a challenging and interesting open issue.

[5] If we write $CW_{2,min} = rCW_{1,min}$, the parameters $(p_1,\ p_2,\ CW_{1,min},\ r,\ N,\ \mu_2)$ will then be solved from the equation set.

own simulation tool in C language using the event-driven simulation technique to implement the 802.11b DCF MAC protocol. Before giving the performance evaluation, we first present a head-of-line outage dropping (HOD) scheme that can effectively improve the channel utilization in serving the realtime applications.

A. Head-of-Line Outage Dropping

In the DCF MAC, a collided packet will be retransmitted; the packet will not be dropped until the retry limit is reached. In a realtime application, when a packet exceeds the delay bound (i.e. delay outage happens), it becomes useless and will be discarded by the receiver even if it is received at last. Continuously serving the outage packets is not only a waste of resources, but also brings a further delay to the subsequent packets in the queue and aggravates the heavy channel collision. In fact, such a collision aggravating procedure due to retransmission is one of the main reasons that lead to the sharp-turning behavior of the DCF MAC.

The channel utilization can be improved and the sharp-turning behavior can be alleviated by a HOD scheme, where the DCF is slightly modified to include the following procedure:

- At each node or each separate queue, a packet is time-stamped upon its arrival. At the end of a backoff stage, before the transmission, the residing time (i.e., current time - arrival time) of the HOL packet is checked. If the delay bound has been exceeded, the packet is dropped; otherwise, the packet is transmitted.
- After each successful packet transmission or dropping (due to delay outage or retry limit), the MAC continues to check the subsequent queued packets until the first packet still within the delay bound is identified for service; all the other packets before the identified packet are dropped due to delay outage.

B. Experiment Results

1) Analysis accuracy and statistical multiplexing gain: In the first example, we analyze an 802.11 DCF network with uplink only voice flows (or a DCF WLAN working at an ad hoc mode). The analytical results are compared with computer simulation results to demonstrate the accuracy of the analysis.

We consider a WLAN with the same configuration as that used in [1]. Table I gives system parameters. In addition, $CW_{2,min} = 32$, $m_b = 5$, $m_r = 7$, and $t_{on} = t_{off} = 300$ ms. During the on periods, voice traffic 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 T_S and the packet collision time T_C can be calculated as $707.27\mu s$ according to (25) and (26). The RTS/CTS mechanism is not considered.

Without traffic from the AP, only class-2 mobile nodes exist in the analytical framework described by (18) – (23). The framework can then be simplified to an equation set including (19), (21), and (23), considering that the traffic arrival rate at AP is now 0 instead of $Np_{on}R_P$. We solve the simplified equation set and obtain $p_2 = 0.2011$, $N = 76.07$ ($\lfloor N \rfloor = 76$), and $1/\mu_2 = 5.21$ ms. In [1], an analytical admission region of 52 is obtained by peak-rate bandwidth allocation at the optimal operation point; however, the simulation results in [1] show that the WLAN can in fact support up to 76 on/off voice flows with QoS satisfaction. Our analytical framework generates an accurate admission region of 76. The observation also means that the proposed analytical model can effectively exploit the statistical multiplexing among the on/off flows over the DCF MAC channel.

TABLE I
IEEE 802.11 DCF PARAMETERS

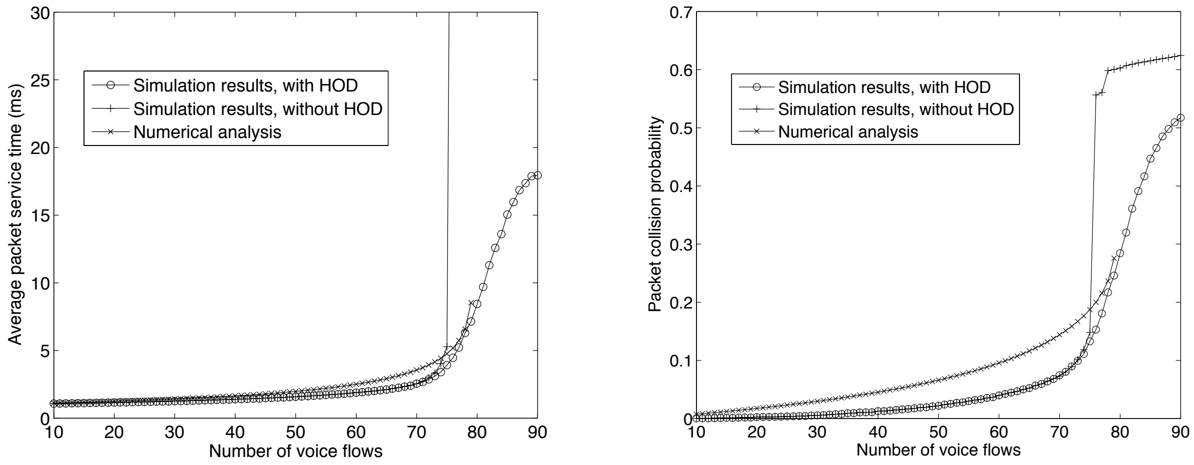
Bit rate for DATA frames	11 Mbps
Bit rate for ACK frames	1 Mbps
Bit rate for PLCP & Preamble	1 Mbps
Slot Time	20 μs
SIFS	10 μs
DIFS	50 μs
PLCP & Preamble	24 bytes
MAC header	28 bytes
IP header	20 bytes
DATA frame	PLCP & Preamble + MAC header + IP header+ payload
ACK fame	PLCP & Preamble + 14 bytes

We also simulate the DCF with the HOD scheme, where packets exceeding the delay bound of 150 ms are dropped according to the procedure described in Section VI-A. We use the further simplified DCF MAC model, including only (19) and (21), to solve for $1/\mu_2$ and p_2 with N varying from 10 to 90, as plotted in Fig. 2(a) and Fig. 2(b), respectively, and compare them with the simulation results. It is observed that the analytical results closely match the simulation results within the admission region ($N < 76$). While the original 802.11b DCF shows a sharp-turning behavior when the admission region is exceeded, the DCF with HOD does exhibit a gradual performance degradation. Within the admission region, the two cases with and without HOD have very similar performance.

Performance of the HOD scheme is demonstrated more clearly in Fig. 3, with respect to the simulation results of the packet delay outage probability at 150 ms. In the original DCF, the over-delayed packets are transmitted by the MAC and dropped by the receivers, while such packets are dropped at the transmitter by the MAC in the HOD scheme. We use delay outage probability (DOP) and outage dropping probability (ODP) to differentiate these two cases, respectively. From the simulation results in Fig. 3, it can be observed that the original DCF in fact only supports an admission region of 74 [6], under the constraint of a stochastic delay bound of 150 ms with the target DOP of 0.01. However, the DCF with HOD can support an admission region of 77, showing an improved channel utilization. In the following, we only present the simulation results with HOD applied.

2) Multiplexing at AP and Contention Window Optimization: In this example, we demonstrate that the analytical framework can generate an accurate admission region in the scenario of downlink multiplexing at the AP; moreover, the network-layer QoS adaption and the CW differentiation are

[6] Our simulation result is slightly conservative than that presented in [1] by ns2 simulation. As we do not know the simulation codes used to generate the results in [1], we can not identify the reasons leading to the discrepancy between the two simulation results.



(a)

(b)

Fig. 2. Comparison between the analytical results and the computer simulation results. (a) Average packet service time. (b) Packet collision probability.

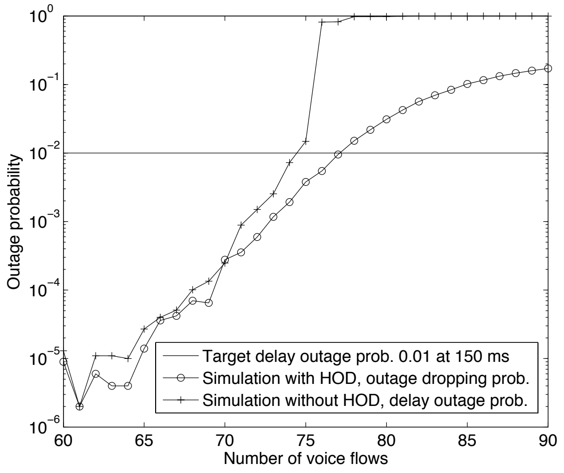


Fig. 3. Efficiency of the HOD scheme compared to the original 802.11b DCF.

TABLE II

ADMISSION REGIONS UNDER DIFFERENT p_{on} AND STOCHASTIC DELAY BOUND, WITH DOWNLINK MULTIPLEXING AT AP

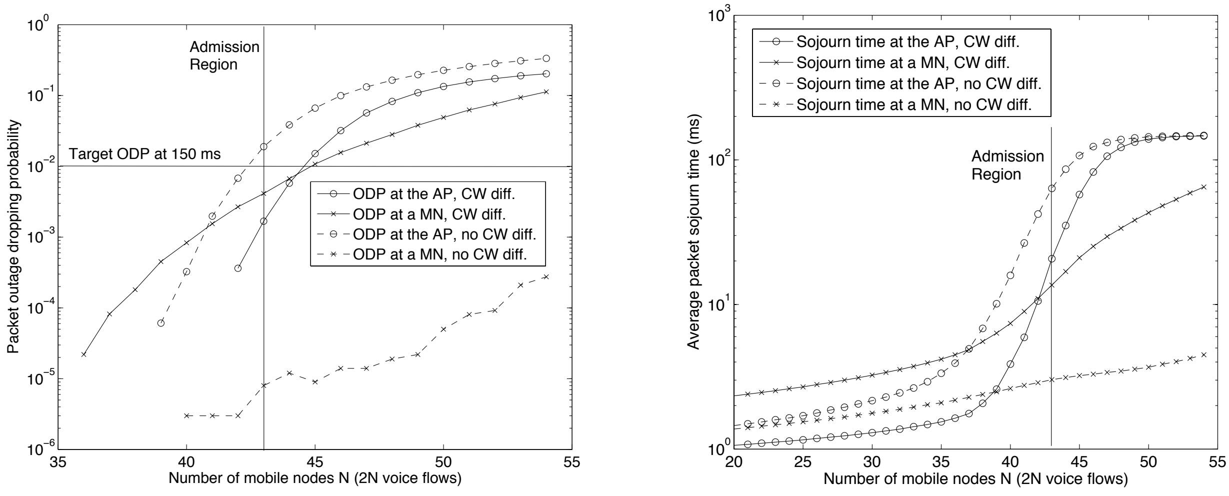
Delay bound			75 ms	150 ms	300 ms
p_{on} (0.5)	Analy.	$\frac{1}{\mu_1}$ (ms)	1.60	1.67	1.71
		N	42.35	43.69	44.46
		$CW_{1,min}$	11	11	12
		$CW_{2,min}$	48	75	118
	Simu.	$\frac{1}{\mu_1}$ (ms)	1.33	1.39	1.50
		$ODP_1/10^{-2}$	0.30	0.20	0.13
		$ODP_2/10^{-2}$	0.64	0.64	0.47
p_{on} (0.3)	Analy.	$\frac{1}{\mu_1}$ (ms)	1.47	1.59	1.67
		N	65.50	70.08	72.67
		$CW_{1,min}$	11	11	12
		$CW_{2,min}$	29	47	79
	Simu.	$\frac{1}{\mu_1}$ (ms)	1.31	1.39	1.56
		$ODP_1/10^{-2}$	0.23	0.21	0.37
		$ODP_2/10^{-2}$	0.39	0.41	0.41

efficient in improving the resource utilization. Specifically, each mobile node has a two-way conversation through the AP with a CN outside the WLAN, and all the downlink flows are aggregated at the AP. Therefore, for N mobile nodes, there are $2N$ activated on/off flows. For the downlink queue, the MAC needs to allocate a sufficient service rate to guarantee a stochastic delay bound d with outage probability not larger than ϵ. The QoS of the mobile nodes is implicitly guaranteed by the OOP control. The DCF parameters are in Table I.

We use the cross-layer analytical framework to determine the capacity regions for the WLAN under different configurations, where the on/off parameters and the network-layer stochastic delay bound are set different. For each configuration, the admission region and the MAC contention windows for the AP and the mobile nodes are jointly solved, under the network-layer QoS constraint. The analysis results are presented in Table II. We use simulations to estimate the QoS under the admission region; the simulation results are also presented in Table II. In our experiments, the target outage probability ϵ is always set as 0.01 for different delay bounds, and t_{on} of the on/off voice flow is fixed as 300 ms with p_{on} changed. The case with a delay bound of 75 ms is considered with the objective to guarantee an end-to-end delay of 150ms when the correspondence node is in another WLAN, assuming a negligible delay incurred over the wireline links. From Table II, we have the following observations:

(1) When the traffic is more bursty (i.e., p_{on} becomes smaller) or the network-layer delay bound is relaxed, the network can support more voice flows due to a higher statistical multiplexing gain and therefore higher resource utilization. Particularly, in the case with $p_{on} = 0.5$, $d = 150$ ms, and $\epsilon = 0.01$, the admission region in terms of the voice flow number $\lfloor 2N \rfloor = 87$, shows an extra increase of $\frac{87-76}{76} \approx 14.47\%$ in the admission region, compared to the corresponding case without AP multiplexing. Such an extra increase is due to the extra statistical multiplexing gain at the AP and the alleviation of channel congestion by aggregating the downlink queues.

(2) In all the scenarios, the achieved average packet service rate is smaller than but close to the analytical result, and the ODPs at both the AP and the mobile nodes are smaller than



(a)

(b)

Fig. 4. The impact of the contention window on the QoS performance when downlink flows are aggregated at the AP, with $p_{on} = 0.5$, $d = 300$ ms, and $\epsilon = 0.01$. (a) Simulation results of the packet outage dropping probability. (b) Simulation results of the average packet sojourn time.

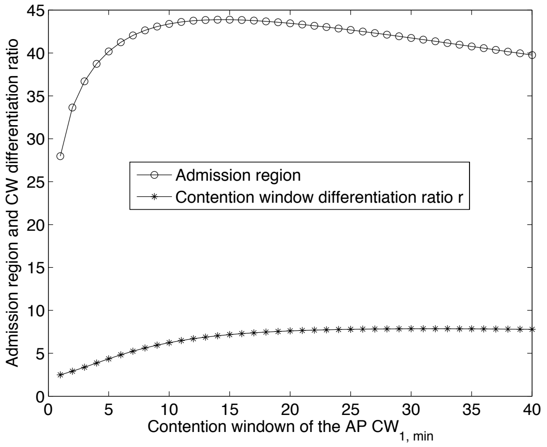


Fig. 5. The impact of the contention window on the admission region when downlink flows are aggregated at the AP.

but close to the target outage probability. That is, the analytical framework can give a moderately conservative admission region to ensure the QoS with efficient resource utilization.

(3) The CW differentiation effectively balances the QoS between the AP and the mobile nodes (i.e. they have the ODPs with the same order of magnitude), although their input traffic loads are largely different.

We here present more experiment results to illustrate the efficiency of the CW differentiation. Under the configuration of $p_{on} = 0.5$, $d = 150$ ms, and $\epsilon = 0.01$, we run simulations to check the QoS performance under the admission region $\lfloor N \rfloor = 43$ for the case with differentiated contention windows $CW_{1,min} = 11$, $CW_{2,min} = 75$, and for the case with homogeneous contention windows $CW_{1,min} = CW_{2,min} = 32$. The simulation results of the ODP and the average packet sojourn time are plotted in Fig. 4(a) and Fig. 4(b), respectively. We observe that the QoS at the AP is far worse than that in the mobiles hosts when using the homogeneous contention windows, due to the large aggregate traffic arrival rate at the AP. In fact, the WLAN with homogeneous contention windows can only support an admission region of 42 due to the bottleneck effect at the AP. Nevertheless, such a QoS discrepancy can be effectively balanced by the CW differentiation, and the admission region is improved to 44.

TABLE III

CAPACITY PLANNING FOR COMMONLY USED CODECS, $t_{on} = 300\ ms$, $p_{on} = 0.5$, $d = 150\ ms$, AND $\epsilon = 0.01$

Codec	G.723.1	GSM 6.10	G.711	G.726-32	G.729
Bit rate (kbps)	5.3	13.2	64	32	8
Packetizing (ms)	30	20	20	20	10
N	37	24	21	23	12
$CW_{1,min}$	9	9	11	10	9
$CW_{2,min}$	51	37	43	40	23

The admission region versus contention window curve, shown in Fig. 5, further demonstrates the impact of the contention window on the resource utilization. From Fig. 5, we can observe that an optimal $CW_{1,min}$ and the corresponding $CW_{2,min}$ exist to maximize the admission region. Due to the discrete effect that the admission region can only be an integer number of mobile nodes, the optimal windows can in fact be selected from a range, 9 to 23 in this example, to achieve the maximum $\lfloor N \rfloor = 43$. The cross-layer analytical framework does give out a contention window (i.e. $CW_{1,min} = 11$) falling in the optimal range.

3) Capacity/Rate Planning: In the third example, we apply the cross-layer analytical framework in network capacity planning for some commonly used codecs, and in rate planning to fit a predetermined call-level capacity.

In Table III, the analytical network capacities and the contention window sizes for the AP and the mobile nodes are listed, with the on/off traffic generated from different codecs. In all the cases, the downlink flows are aggregated at the AP, and the WLAN is configured according to Table I. From Table III, it can be observed that the packetizing interval has a

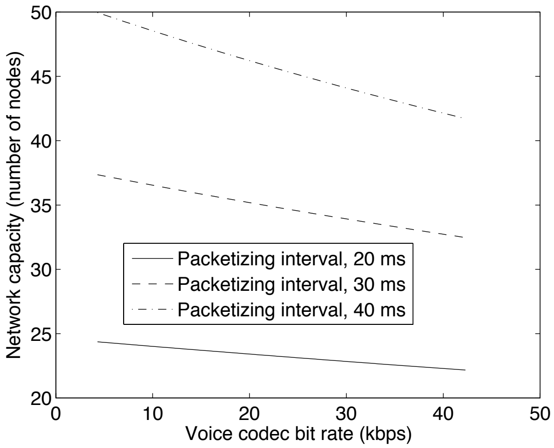


Fig. 6. Curves for rate planning, $t_{on} = 300$ ms, $p_{on} = 0.5$, $d = 150$ ms, and $\epsilon = 0.01$.

more significant impact on the network capacity than the traffic generating rate, which implies that the MAC channel capacity is more sensitive to the packet arrival rate than to the packet size. Note that when transmission errors (which are ignored in our experiments) are considered, the impact of different factors on the network capacity will be more complex [27], which is out of the scope of this paper. In addition, we plot the network capacity versus bit rate curve in Fig. 6 which can facilitate the rate planning, where the network capacity may be predetermined according to a target call blocking probability.

VII. Conclusions

In this paper, we present a cross-layer analytical framework for WLAN network capacity planning. Specifically, the on/off voice flows are considered, for which a stochastic delay bound is guaranteed via the cross-layer design. We investigate a multiclass DCF MAC supporting service differentiation, where the AP is assigned a smaller CW than the mobile nodes, to facilitate downlink traffic multiplexing. A nonsaturated multiclass DCF model is developed, which is combined with the network-layer queueing analysis to form the cross-layer analytical framework. In addition, the channel busyness ratio control is integrated into the framework to guarantee the analysis accuracy. We use extensive numerical analysis and computer simulation results to demonstrate that the framework can be exploited for statistical multiplexing gain analysis, network capacity planning, contention window optimization, and voice traffic rate design.

In practice, in order to apply the network capacity for admission control or the optimal contention window configuration for maximum resource utilization, a certain application-layer negotiation protocol needs to be implemented when a mobile node is connected to an AP. We are investigating the design of such a protocol. For further work, we also plan to investigate the network planning issues in a multi-application WLAN, where a node may have multiple queues to support voice/video/data integrated applications.

References

[1] 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.*, vol. 13, no. 1, pp. 36–43, Feb. 2006

[2] H. Zhai, X. Chen and Y. Fang, "How well can the IEEE 802.11 wireless LAN support quality of service?" *IEEE Trans. Wireless Commun.*, vol. 4, no. 6, pp. 3084–3094, Nov. 2005.

[3] L. Cai, X. Shen, J.W. Mark, L. Cai and Y. Xiao, "Voice capacity analysis of WLAN with unbalanced traffic," *IEEE Trans. Veh. Technol.*, vol. 55, no. 3, pp. 752–761, May 2006.

[4] W. Wang, S. C. Liew, and V. O. K. Li, "Solutions to performance problems in VoIP over a 802.11 wireless LAN," *IEEE Trans. Veh. Technol.*, vol. 54, no. 1, pp. 366–384, Jan. 2005.

[5] M. Schwartz, *Broadband Integrated Networks*, New Jersey: Prentice Hall, 1996.

[6] U. K. Sarkar, S. Ramakrishnan, and D. Sarkar, "Modeling full-length video using Markov-modulated Gamma-based framework," *IEEE/ACM Trans. Networking*, vol. 11, pp. 638-649, Aug. 2003.

[7] F. P. Kelly, "Notes on effective bandwidth," in *Stochastic Networks: Theory and Applications*, F. P. Kelly, S. Zachary, and I. Ziedins, Eds. Oxford, U.K.: Oxford Univ. Press, 1996, pp. 141-168.

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

[9] 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*, vol. 8, no. 6, pp. 785–799, Dec. 2000.

[10] Y. Tay and K. Chua, "A capacity analysis for the IEEE 802.11 MAC protocol," *Wireless Networks*, vol. 7, no. 2, pp. 159–171, Mar. 2001

[11] 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.

[12] 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.

[13] M. van Der Schaar and S. Shankar N,"Cross-layer wireless multimedia transmission: challenges, principles, and new paradigms", *IEEE Wireless Commun.*, vol. 12, no. 4, Aug. 2005, pp. 50–58.

[14] V. T. Raisinghani and S. Iyer, "Cross-layer feedback architecture for mobile device protocol stacks", *IEEE Commun. Mag.*, vol. 44, no. 1, Jan. 2006, pp. 85–92.

[15] S. Khan, Y. Peng, E. Steinbach, M. Sgroi and W. Kellerer, "Application-driven cross-layer optimization for video streaming over wireless networks", *IEEE Commun. Mag.*, vol. 44, no. 1, Jan. 2006, pp. 122–130.

[16] 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", *Proc. IEEE INFOCOM*, 2002, pp. 599–607

[17] Y. Xiao, "Performance analysis of priority schemes for IEEE 802.11 and IEEE 802.11e wireless LANs," *IEEE Trans. Wireless Commun.*, vol. 4, no. 4, pp. 1506–1515, Jul. 2005.

[18] G. Bianchi, I. Tinnirello, and L. Scalia, "Understanding 802.11e contention-based prioritization mechanisms and their coexistence with legacy 802.11 stations," *IEEE Network*, vol. 19, no. 4, pp. 28–34, Jul./Aug. 2005

[19] 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.

[20] 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.

[21] S. Garg and M. Kappes, "Can I add a VoIP call?," in *Proc. IEEE ICC*, 2003, pp. 779–783.

[22] L. B. Jiang and S. C. Liew, "An adaptive round robin scheduler for head-ofline-blocking problem in Wireless LANs", *Proc. IEEE WCNC*, 2005, pp. 1219–1224.

[23] C. Zhu, O. W. W. Yang, J. Aweya, M. ouellette, and D. Y. Montuno, "A comparison of active queue management algorithms using the OPNET modeler," *IEEE Commun. Mag.*, vol. 40, no.6, pp.158–167, Jun. 2002.

[24] S. Floyd and V. Jacobson, "Radom early detection gateways for congestion avoidance," *IEEE/ACM Trans. Networking*, vol. 1, no. 4, pp. 397–413, Aug. 1993.

[25] A. Heck, *Introduction to Maple (3rd ed.)*. Springer-Verlag, New York, 2003.

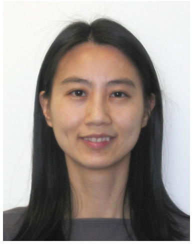
[26] H. L. Vu and T. Sakurai, "Accurate delay distribution for IEEE 802.11 DCF," *IEEE Commun. Lett.*, vol. 10, no. 4, pp. 317–319, Apr. 2006.

[27] Y. Zheng, K. Lu, D. Wu, and Y. Fang, "Performance analysis of IEEE 802.11 DCF in binary symmetric channels," in *Proc. IEEE GLOBECOM*, 2005, pp. 3144–3148.

Yu Cheng (S'01-M'04) 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.

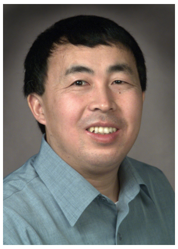
Xinhua Ling (S'03) received the B.Eng. degree in Radio Engineering from Southeast University, Nanjing, China in 1993 and the M.Eng. degree in Electrical Engineering from the National University of Singapore, Singapore in 2001. He is currently pursuing the Ph.D. degree in the Department of Electrical 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.

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).



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.

Weihua Zhuang (M'93-SM'01) received the Ph.D. degree from the University of New Brunswick, Canada, in electrical engineering. 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 (PREA) in 2001 from the Ontario Government for demonstrated excellence of scientific and academic contributions. She is the Editor of *IEEE Transactions on Vehicular Technology* and an Editor of *IEEE Transactions on Wireless Communications*.



Xuemin Shen (M'97-SM'02) has been with the Department of Electrical and Computer Engineering, University of Waterloo, Canada, since October 1993, where he is a professor and the Associate Chair for Graduate Studies. His research focuses on mobility and resource management in interconnected wireless/wireline networks, UWB wireless communications systems, wireless security, and ad hoc and sensor networks. He is a co-author of two books, and has published more than 200 papers and book chapters in wireless communications and networks, control, and filtering. He was Technical Co-Chair for the *IEEE GLOBECOM'03, ISPAN'04, QShine'05, IEEE Broadnets'05*, and *WirelessCom'05*, and is Special Track Chair of the *2005 IFIP Networking Conference*. He serves as Associate Editor for *IEEE Transactions on Wireless Communications*; *IEEE Transactions on Vehicular Technology*; *Computer Networks*; *ACM/Wireless Networks*; *Wireless Communications and Mobile Computing* (Wiley); and *International Journal Computer and applications*. He has also served as Guest Editor for *IEEE JSAC*, *IEEE Wireless Communications*, and *IEEE Communications Magazine*. He received the Premier's Research Excellence Award (PREA) in 2003 from the Province of Ontario, Canada, for demonstrated excellence of scientific and academic contributions, and the Distinguished Performance Award in 2002 from the Faculty of Engineering, University of Waterloo, for outstanding contributions in teaching, scholarship, and service.