On Sustained QoS Guarantees in Operated IEEE 802.11 Wireless LANs

Abdelhamid Nafaa and Adlen Ksentini, *Member*, *IEEE*

Abstract—Most of quality of service (QoS)-capable IEEE 802.11 MAC protocols fall short to deliver sustained QoS guarantees while maintaining high network utilization, particularly under congested network conditions. The problem often resides in the fact that flows belonging to the same service class are assigned the same MAC parameters, regardless their respective bit rate, which leads to throughput fairness rather than perceived QoS fairness. Harmonizing the MAC parameters of traffic classes' flows may further lead to suboptimal situations, since certain network configurations (in terms of per-class traffic load) cannot be accommodated without readjusting the basic MAC parameters. In this paper, we propose a new cross-layer MAC design featuring a delay-sensitive backoff range adaptation, along with a distributed flow admission control. By monitoring both the MAC queue dynamics of each traffic class and the overall network contention level, our MAC adaption scheme reacts based on the degree to which application QoS metrics (delay) are satisfied. Aside from that, we use a distributed admission control mechanism to accept new flows while protecting the active one. Simulation results show that compared to the Enhanced Distributed Coordination Function (802.11e Enhanced Distributed Control Access (EDCA)) and AEDCF (Adaptive EDCF), our protocol consistently excels in terms of network utilization, bounded delays, and service-level fairness.

Index Terms—IEEE 802.11 MAC protocol, cross-layer QoS performance metrics guarantees.

1 Introduction

IEEE 802.11 [1] has been widely accepted as the de facto standard for Wireless Local Area Networks (WLANs). Currently offering nominal data rates up to 54 megabits per second (Mbps), IEEE 802.11 can provide a serious alternative to existing wired LAN technologies. However, the open shared-air medium places an additional burden on the Media Access Control (MAC) protocol. In its basic form, the IEEE 802.11 Distributed Coordination Function (DCF) provides a simple and flexible mechanism for sharing the medium but lacks the ability to guarantee service levels to meet the demands of multimedia applications. As a consequence, there has been a considerable effort to improve the MAC's ability to serve and interact with higher level QoS mechanisms. The IEEE 802.11e Task Group has worked toward designing and developing a framework for QoS support. Based on the basic DCF, the 802.11e proposals [2] focus primarily on providing differentiated access to individual traffic classes (TCs). In particular, the Enhanced Distributed Control Access (EDCA) uses priority concepts to alter the existing MAC scheme. During initialization, EDCA assigns static MAC parameters for each TC. Based on these parameters, the MAC protocol provides different service levels to different TCs. It is readily realized that EDCA parameters do not accommodate all network configurations in terms of relative (per-class) network load [3]. Particularly, EDCA is unable to absorb a large number of multimedia flows due to a too narrow backoff range (0, 31) assigned to high-priority (HP) flows, which lead to high intra-TC contention level. This situation entails high collision rate, poor medium utilization, and increased medium access delays.

There are many proposals of priority schemes [4], [5], [6], [7] that utilize a variety of mechanisms for differentiating between TCs, including adjusting interframe spaces (IFSs), minimum/maximum values for the contention window (CW), Transmission Opportunity (TXOP) duration, and the CW increase/decrease functions. Nonetheless, there is no previous work that clearly addressed the QoS interaction between the MAC layer and upper layers. From the Network Operator (NO)'s point of view [20], it is indeed crucial to be able to translate (enforce) common application-level QoS metrics into medium access mechanisms (that is, reflecting the application-level QoS notions at the MAC layer) [8]. WLANs should allow a variable number of users with heterogeneous QoS requirements to share a common radio channel. By extending WLAN's MAC protocol to provide several TCs, the NO will be able to fragment its QoS offer into several levels of guarantees so as to accommodate any network configuration in terms of per-class network load. In this context, the network resources may be fully utilized by flows from a single TC, or in contrast, the load may be differently distributed among the supported TCs according to the instantaneous per-class offered load. Depending on the nature of content coding and the targeted applications (for example, videoconferencing, IP telephony, and media streaming), the multimedia streams may be mapped into different TCs characterized by different guaranteed QoS metric performance thresholds.

The random nature of the EDCA scheme makes it difficult to maintain high channel utilization and fair channel utilization. As the network becomes congested, backoff intervals must increase in order to keep the

- *A. Nafaa is with the School of Computer Science and Informatics, University College of Dublin, Dublin 4, Ireland. E-mail: nafaa@ieee.org.*
- *A. Ksentini is with the IRISA Laboratory, University of Rennes 1, Campus Universitaire de Beaulieu, 35042 Rennes, France. E-mail: adlen.ksentini@irisa.fr.*

Manuscript received 7 Feb. 2007; revised 3 Aug. 2007; accepted 11 Sept. 2007; published online 26 Sept. 2007.
Recommended for acceptance by M. Ould-Khaoua.
For information on obtaining reprints of this article, please send e-mail to: tpds@computer.org, and reference IEEECS Log Number TPDS-0045-0207.
Digital Object Identifier no. 10.1109/TPDS.2007.70785.

probability of collision relatively low. However, this can also mean that the medium has an increased chance to remain idle, wasting valuable bandwidth. In contrast, when a transmission is successful, the sending station will reduce its CW in order to try fully filling the medium. This can result in considerable unfairness, as a node can dominate the channel with repeated transmissions. Although some work has proposed adaptive CW schemes (see [4], [5], [6], [7], and [18]) designed to coordinate MAC parameters between TC[i]'s flows in different stations, they still provide access opportunities fairness rather than service-level fairness. In fact, if traffic was balanced between nodes, achieving fairness between flows within the same TC would require the MAC parameters on different nodes to remain harmonized. The fact is that traffic load is typically unbalanced in a real WLAN deployment, with a large variation in TC volume from one node to the next.

In this work, we focus on guaranteeing the same QoS metrics (for example, loss rate, mean delay, and mean jitter) for all flows belonging to the same TC. That is, we aim at maintaining a sustained application-level perceived QoS, regardless the bit rate of each single TC flow. This is an imperative in most of existing and forthcoming operated 802.11 networks (Hotspots) [19], [20]. We also study the cross-layer interaction between the TC QoS requirements and the network dynamics behavior to derive an accurate model that estimates the achievable delays at the application level. One particular issue inherent to 802.11-based networks is that in certain circumstances, for the same overall offered load, the network may exhibit widely different performances (that is, availability levels), depending on the number of competing flows and their respective bit rates. It is therefore difficult for a *NO* to a priori figure out whether a new service (based on its requirements) can be admitted or not based only on monitoring the overall network load.

Although existing work focuses on studying throughput limits in WLANs [10], [11], [12], [14], [15], [16], less emphasis is put on modeling end-to-end delay distribution (see [18] and the references therein). Nevertheless, there is a key trade-off between fully filling the network capacity and maintaining acceptable network access delays. Delay is a particularly important metric in the statistically shared environment such as 802.11 networks, where medium access delays may widely vary with the network load, producing different enqueuing delays at different stations, depending on the offered load in each station. For each network configuration, in terms of the number of flows, it indeed exists an optimal operation point, in which the network cannot carry additional traffic load without violating delay constraints at certain stations. Typically, for a fixed delay budget, the optimal operation point would determine the allowed load per flow, although the load may be unbalanced between network flows. In this paper, we aim at gaining insight into the trade-off between achieving bounded delays and maximizing the network utilization in order to design an effective admission control (AC) mechanism. By analyzing all factors that influence the medium access delay, we derive a distributed model able to accurately predict the achievable delay at each network flow using different network measurements. Delays bounds associated with each TC are assumed to be communicated to the MAC layer through a top-down cross-layer interaction. Based on the latter model, we derive the AC algorithm, which allows us to a priori assess the achievable throughput before admitting new incoming streams, taking into considerations their QoS requirements. This could contribute in improving network utilization, the objective being to preserve the QoS of already-active flows while maximizing the volume of QoS-enabled services, providing to NOs an improved resource control mechanism (that is, allows for generating more revenues).

The remainder of this paper is organized as follows: The next section provides background material on QoS provisioning at the IEEE 802.11 MAC layer, highlighting the motivating factors that lead us to propose a new adaptive MAC protocol. Section 3 describes the design of a delay-sensitive service differentiation algorithm based on a delay modeling at the MAC level. The model is validated through simulation and further compared to EDCA and AEDCF [5]. Based on this latter adaptation model, Section 4 develops a fully distributed AC protocol. Detailed simulations of our proposal have been constructed in the Network Simulator (ns-2) in order to evaluate its performance under a variety of conditions. We describe these simulations in Section 5. Finally, we have drawn several key conclusions from this work, and these are stated in Section 6.

2 Sustained Service-Level Guarantees in 802.11-Based Networks

In WLANs, it is crucial to restrict the volume of traffic in order to maintain the QoS of current serving traffic. If there are no restrictions to limit the volume of traffic being introduced to the service set, performance degradation will result due to higher backoff time and collision rate. An effective resource allocation in IEEE 802.11 is difficult to achieve due to the intrinsic nature of the Carrier Sense with Multiple Access with Collision Avoidance (CSMA/CA) scheme. Unlike traditional wired networks (or point-coordinated wireless networks), where bandwidth provision can be managed by only using bandwidth availability information, flow AC in distributed 802.11 networks asks for additional parameters and more advanced models. Actually, for the same overall offered load, the network may exhibit widely different performances (that is, availability levels), depending on the number of competing flows and their respective bit rates. For instance, the network contention level (collision) involved by 10 active flows with a rate of 100 kilobits per second (Kbps) would be different from the one involved by two 500-Kbps-rate active flows. The difficulty with distributed 802.11 networks lies in estimating the achievable QoS performance in the WLAN. This estimation depends on several time-varying factors, including the number of active stations and the offered traffic volume for each TC.

Many recent work on 802.11 network dimensioning ([10], [11], [21], and references therein) has mainly focused on the analysis of throughput and delay in saturated conditions. Aside from considering a single TC, the work derived models by assuming balanced traffic distribution between active wireless stations. If these analyses are to be used for AC, flow admission in the network would be achieved in terms of the number of active stations rather than in terms of single flows.

The Distributed Bandwidth Allocation/Sharing/Extension (DBASE) protocol [9] addresses the problem on resource control in the DCF-based mode by splitting the contention period into two subperiods: a period for contention between real-time stations and another for contention between nonreal-time stations. This protocol allows the voice station to a priori reserve bandwidth by using specific messages and an updated network reservation table at each station to coordinate between competing stations. The differentiation between these two contention periods is based on different AIFSs for real-time and nonreal-time traffic. Aside from leading to substantial traffic overhead during the reservation process, DBASE is unable to effectively separate between different TCs when the network gets fairly loaded, since both TCs still use the exponential backoff algorithm: non-real-time traffic can draw small backoff intervals and frequently access the network, wasting valuable bandwidth.

Based on local network measurements, Zhai et al. [15] propose controlling the arrival rate at each station to achieve a given objective such as the maximum throughput, maximum delay, jitter, or loss rate in the network. The developed analytical model is able to assess the capability of 802.11 for supporting major QoS metrics. The model is further extended in [16] to control the admission of network flows based on a new metric (channel busyness ratio) as a good indicator of the network state. The channel busyness ratio is used to derive the rate control algorithm, that is, Call Admission and Rate Control (CARC). Aside from not being applicable to 802.11e-like protocols, where several TCs (having several requirements) may simultaneously operate in the network and even coexist at a single station, CARC tries finding the optimal network utilization (maximize the throughput) while barely considering delays fluctuations.

A common drawback of the above introduced techniques [9], [11], [15], [18], resides in the fact that it is not possible to provision different TCs at a given station, as a common admission criterion should be enforced by all stations. This severely limits the flexibility for realistic deployment of multimedia streams with different requirements. That is, the network stations should be either voice stations or best effort stations. Designing advanced AC mechanisms is clearly very important to operate future value-added services in WLANs, where the NO is able to fragment its quality-of-service (QoS) offer into different service classes.

The DAC and Two-level Protection and Guarantee Mechanisms [12], [13] are combined to address the above-mentioned issues. DAC is a measurement-based AC mechanism that was considered by the 802.11e Working Group. In this algorithm, the resource budget for each TC is periodically announced by the AP in the beacon frame so that each station may decide whether or not to accept new flows. A new stream to be admitted first tries accessing to the network, and it rejects itself after a certain period if its requirements are not met. With this algorithm, the residual network resources are fairly distributed among the competing streams (streams seeking for acceptation) at different stations in the sense that different TCs (in different stations) compete to accommodate their new entering streams. The stream is then locally accepted if it reaches its targeted throughput. This situation may cause spectrum waste, because there may be enough resources to admit one additional stream, but due to the algorithm fairness and absence of coordination, none of the competing streams is accepted, and the available bandwidth remains unfilled.

Another shortcoming of the DAC algorithm resides in the lack of protection to existing flows only when the network load is not too heavy. If the network resources are not sufficient to admit the new stream, the performance degradation will affect all of the TC's streams (as much as it does for other TCs active in the network). This is due to the fact that entering streams are aggregated with other active streams in the same TC queue. The above-mentioned phenomenon is usually referred to as the "*spill-over*" effect in WLANs: when traffic is overloaded in a TC, performance in other TCs will also be affected. Nonetheless, the major problem with DAC-based approaches consists of the fact that the overall network bandwidth is statically allocated among different TCs, so each TC receives a fixed share of bandwidth that cannot be exceeded. This may severely affect the flexibility of the AC mechanism, since it is very difficult to beforehand forecast the per-TC traffic volume in realistic multimedia-dedicated WLANs. Therefore, streams from a given TC may be rejected, whereas some bandwidth is still unfilled in other TCs, which means bandwidth wasting or additional revenue loss for NO. Another side effect is that the admission decision depends only on the local measurements collected at the admitting station level. However, the stream admission may have different impacts at different stations (respectively, flows), depending on the load of each active station. The stream admission may actually cause QoS violation at certain stations while not at all affecting other active stations in the network. This is particularly prevalent for high-bit-rate stations, which usually cannot carry the load in a sufficiently timely manner as the load (respectively, the medium access delay) increases.

It is readily realized that it is essential for an AC model to be able to a priori estimate the achievable application-level QoS metrics. This way, the admission decision does not affect the existing flows. Pong and Moors [21] propose estimating the achievable throughput under saturation to ultimately control the flow admission in EDCA-based 802.11e networks. Aside from the limitations entailed by the saturation assumption, this scheme delivers only throughput guarantees without considering multimedia delay requirements. Furthermore, due to the static per-TC parameters used in 802.11e, it is not possible to accommodate an important number of multimedia flows [3]. More specifically, HP flows use a too-narrow backoff range, which provokes a high intraclass contention.

The Virtual MAC (VMAC) and Virtual Source (VS) Algorithms [17], [18] propose a fully distributed VMAC algorithm that operates in parallel to the real MAC in the mobile host, but the VMAC does not handle real packets. Rather, it handles "*virtual packets*." Each station runs a VMAC instance that monitors the capability of the wireless channel and passively estimates whether the channel can support new service demands (for example, delay and loss). Unlike

the case of real packets, VMAC does not transmit anything but estimates the probability of collision. When a collision is "*detected*," the VMAC enters a backoff procedure just as a real MAC would do. The VS algorithm consists of a virtual application, an interface queue, and the VMAC. The virtual application generates virtual packets like a real application. Packets are time-stamped and placed in a virtual buffer. After a virtual packet has been processed in the VMAC, the total delay is calculated.

VMAC's main criterion to make an AC decision is based only on delay and collision estimates. It does not provide any achievable throughout information, which is also useful to multimedia applications. The achievable QoS is estimated only at the admitting station, although flow admission may unevenly affect the different backlogged flows, provoking delay violation at certain flows, whereas other flows in the network still experience acceptable delays. As mentioned earlier, the outcome of stream admission should be beforehand assessed at all active stations. In fact, flows belonging to the same TC use roughly the same CWs' ranges, and thus, they more or less experience the same packet service times (PSTs; that is, the time needed to successfully transmit the frame located at the front of the queue). Hence, depending on the volume of their offered load, different flows may suffer from widely different enqueuing delays. In other words, the admission of a new flow means a slightly increased PST, with different outcomes on different active flows. The impact of a stream admission should therefore be assessed at all active stations.

Dynamic Multiple-Threshold Reservation [22] proposes an algorithm that is capable of granting differential priorities to different TCs in wireless multimedia network with cellular infrastructure. DMTBR generalizes the concept of relative priority and hence gives the NO more flexibility to adjust the AC policy by taking into account the offered load. However, as highlighted earlier, cellular networks are point coordinated and present widely different characteristics compared to contention-based distributed WLANs. In the cellular architecture, the network offered load is not correlated with delays, since the resources are centrally managed by the BS, and each flow receives fixed transmission slots when admitted in the network.

Obviously, the candidate AC mechanism should be distributed and able to manage different TCs at each single station while providing high flexibility with respect to the relative (per-class) network load configurations (that is, the AC mechanism that enables all possible per-class load distributions, as long as the QoS metrics are not violated). Admission decision may be made based on the estimates of the achievable QoS at different active stations rather than only at the admitting station.

3 Delay-Sensitive Media Access Control Adaptation Model

Conventional IEEE 802.11 backoff schemes have many shortcomings that make it difficult to provide deterministic guarantees. The exponential CW increasing more likely produces probabilistic service assurances and high oscillations in delays (throughput), since the CW is reinitialized to its minimum value (CWmin) with each successful transmission. In order to limit the effect of high inter-TC contention, different AIFS[i]s may be assigned to different TCs TC[i]. This would differ from transmissions of low-priority flows only when their respective transmission attempts coincide with HP flow transmission. At this point, managing the contending flows through an appropriate CW scheme is a key component to effectively maintain an acceptable QoS level for multimedia flows.

In this section, we present a measurement-based CW adaptation scheme. The objective is to guarantee the same QoS metrics (for example, loss rate, mean delay, and mean jitter) for all flows belonging to the same TC. That is, we aim at maintaining a sustained application-level perceived QoS: this is an imperative in most of existing and forthcoming operated networks [20]. In this respect, we set a predefined QoS metric (MAC-level transmission delay) threshold for each supported TC. Based on distributed measurements, our protocol is able to guarantee multimedia streams requirements (MaxDelay, MaxLoss, and ensured bit rate) in different network configurations. A key point to enforce predictable QoS performances resides in the ability of our scheme to accurately model the achievable QoS metric performances. In the next section, we will generalize our achievable QoS assessment model to derive the AC algorithm.

3.1 Delay-Driven Dynamic Contention Window Adjustment

At the MAC layer, packets are serviced with a variable latency that depends on the current CW size, the mean frame size E[P], and the mean number of transmission attempts before effectively gaining access to the medium. In addition, the network load (that is, transmission volume from other nodes) may strongly affect the end-to-end communication latency as a substantial amount of time slots is occupied, which ends up provoking frequent backoff freezing. Actually, each new packet selects a random backoff interval E[CW] that is more or less quickly decremented, depending on the number of time slots where the medium was observed as busy. The packet transmission deferring period depends on the selected backoff interval as much as it depends on the degree of network load.

We define PST as the time needed to successfully transmit a packet. This delay is defined as the time interval elapsed between the time when a packet arrives in front of the queue and the time when it is received by the receiver. The delay considers only channel access delay, transmission delay, and associated overhead (that is, the queuing delay is not included).

Let $\mathrm{B(T)} = \mathrm{B/I}$ be the number B of busy time slots over the number I of idle slots observed during the last T time slots $(\mathrm{T} = \mathrm{B} + \mathrm{I})$. The total deferring time for a packet can be approximated by $\mathrm{E(CW)} \cdot (1 + \mathrm{B/I})$. This delay takes into account both the backoff interval and the freezing period. Compared to the technique that achieves direct measurement of the freezing period at each flow [11], our technique is based on continuous monitoring of the overall network load, which could be better exploited to predict network load trends. Measuring the freezing period for each transmitted packet may exhibit high oscillations, not to

mention the involved complexity. Using the overall network occupancy (B/I) to estimate the access delay leads to inherent measurement coordination between different active flows, as they observe the same network activities.

We define E[P] as the mean number of time slots occupied by a single packet transmission, including the PHY/MAC overhead, SIFSs, and ACK when considering the DCF basic mode. It is worth mentioning that within the DCF basic method (without RTS/CTS handshaking), each failing transmission (due to frame collision or bit alteration) occupies roughly the same number of slots as a successful transmission [3]. In the following, we assume a DCF MAC protocol operating without RTS/CTS handshaking. In order to better assess the accuracy of our model with simulations, we assume that packet loss provoked by wireless link interferences (BER) is negligible.

The overall PST may be quantitatively estimated as follows:

$$PST = [E(CW) \cdot (1 + B(T) + E(P))] \cdot E[TransAtt]. \quad (1)$$

Here, E[TransAtt] is the mean number of transmission attempts needed to successfully access the medium. This parameter depends on the Packet Error Rate (PER) and the automatic retransmission (ARQ) scheme being used at the MAC layer. Generally speaking, a packet is kept in the transmitter queue until either a timer times out (that is, after seven failed transmission attempts) or the packet is successfully received and acknowledged by the receiver. Since the backoff process has a geometric distribution with probability of success p, the mean number of transmission attempts E[TransAtt] would be $1/p$. At this point, the probability of transmission success p can be approximated as the fraction of the number of transmitted frames over the number of transmission attempts. Thus, the mean number of transmission attempts E[TransAtt] can be estimated as

$$\begin{aligned} E[TransAtt] &= \frac{1}{1 - \frac{Collisions}{TransmissAttempts}} \\ &= \frac{TransAttempts}{SucceedTransmissions}. \end{aligned} \quad (2)$$

Note that E[TransAtt] may return different values, depending on the flow's TC and its associated AIFS. Obviously, inter-TC collisions are most of the time avoided, since the flows with the highest priority seizes the medium, whereas other flows enter in differing states.

As B(T) is calculated based on the overall network load, it is inherently coordinated between stations. Each station averages the measurements over the period T required to sense "*CWmax*" idle time slots. By choosing the frequency of measuring B/I this way, we are ensured that all backlogged flows (regardless of the priority) would have attempted to access the medium at least once within this period. Thus, B/I measurements are more accurate by considering all active flows and are also more stable as they are averaged over a long-enough period. Throughout this paper, the value of T is set to 1,024 "*idle*" slots. For the same reasons, E[TransAtt] values are also averaged over the period needed to sense 1,024 idle time slots.

As apparent in Fig. 1, each station in the network may have different TCs with different requirements in terms of the QoS metric performances. Several LLC/MAC queues are indeed implemented within a single station. Each queue supports one TC, behaving similar to a single DCF entity within the 802.11 standard. In this context, the last packet in the queue (packet N) should not exceed the maximum delay tolerated by the TC to which it belongs. By considering that both the arrival λ and the service μ are exponential, the PST will be therefore constrained by

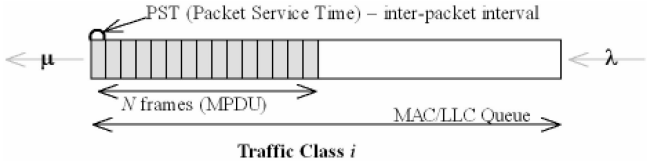

Fig. 1. MAC layer queue for TC i.

$$PST \leq \frac{MaxDelay}{N}. \quad (3)$$

The formula above generalize our PST estimation model to estimate the enqueuing time by taking into account the number of packets N currently in the MAC queue (the N packets ahead of the last packet entering the queue). From (1) and (3), given the queue length N, the appropriate maximum CW size CWmax that would satisfy the delay constraints associated with each service class (regardless of its bit rate) is obtained as follows:

$$CW_{\max} = 2.E[CW] \leq \frac{2.(MaxDelay - N.E[TranAtt].E[P])}{N.(1 + B(T)).E[TransAtt]}. \quad (4)$$

It is commonly accepted [11] that the WLAN capacity (that is, channel utilization) decreases with an increasing number M of active flows. This is caused by a high contention level, in which case the medium is often occupied by collisions. In this situation, the mean number of attempts to successfully transmit a frame would grow, resulting in additional delays at active flows. The CW size should be continuously adopted, thereby reacting to changing network conditions while meeting QoS constraints. Actually, when M increases, the CW size is increased to absorb the increasing number of contending flows, hence minimizing the collision probability for these flows. On the other hand, when M becomes small, the CW size is decreased, which reduces the spacing between successive frame transmissions. Large values of the CW size may indeed strongly limit the throughput of fewer backlogged flows. As a matter of fact, the current CWsize in use should always be larger than a certain variable threshold CWmin to avoid the network performance from collapsing. From [7] and [23], the minimum CW size that maximizes network performances with M contending flows is given by

$$CW_{\min} \geq \left[M \cdot \sqrt{2Tc}\right], M \approx \frac{E[O(T)] \cdot (E[oldCW] + 1)}{T}. \quad (5)$$

Here, Tc is the average time (in time slots) of channel unavailability upon a collision. Tc is dependent on the PHY layer and is equal to $\mathrm{PHYhdr} + \mathrm{E[F]} + \mathrm{DIFS}$ when the RTS/CTS mode is disabled. E[oldCW] is the current

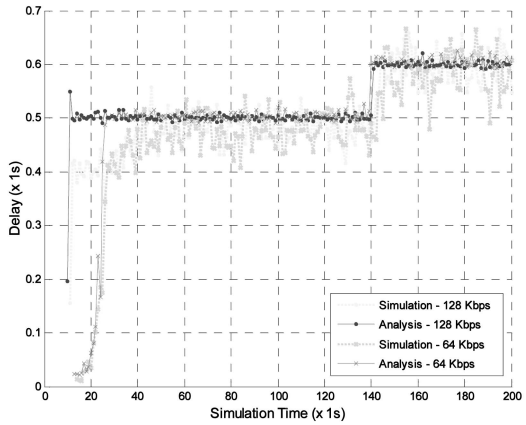

Fig. 2. HP flows' instantaneous delays.

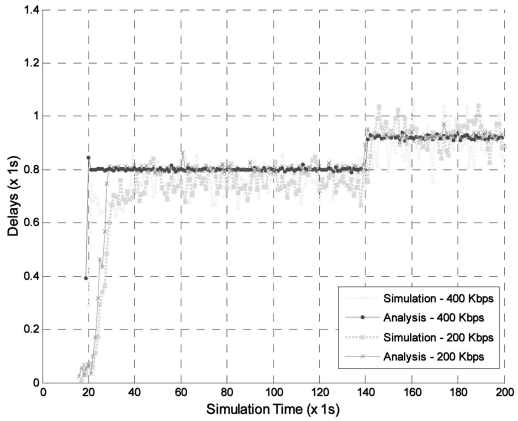

Fig. 3. MP flows' instantaneous delays.

mean backoff value. O(T) is the number of slots where the medium was observed as busy out of the previous T slots B. Like all other network measured parameters (that is, E[TransAtt] and B(T)), O(T) is weighted with respect to past measures by using the Exponential Weighted Mean Average (EWMA).

Although this is not accurate (that is, much incertitude still exists due to different flows' priorities and bit rates), the estimate of the number M of active flows is quite pertinent, since it still precisely reflects the overall trend of the network contention level, which allows readjusting the CW to optimize the network performance. In fact, constraining the CW by CWmin helps keep a low collision rate and, hence, acceptable mean transmission attempts (that is, E[TransAtt] lower than 1.5, which means three transmission attempts for two successful transmissions). The new CW to be maintained by each TC is given by

$$newCW_{size} = \frac{CW_{\min} + CW_{\max}}{2} \quad \text{with} \quad CW_{max} \geq CW_{min}. \tag{6}$$

If CWmax is smaller than CWmin, we assign CWmin to CWmax. In this case, newCWsize is simply reinitialized with a CWmin value. This situation does not guarantee MaxDelay. Instead, it keeps network collisions within an acceptable level. Using the above-introduced CWsize adjustment model, a given flow would use the interval [0, newCWsize] to randomly draw a backoff interval. Note that parameter CWmin is not necessarily coordinated between flows, since its value is, in part, based on the current CW size that is maintained by the flow. Accordingly, flows calculate different CWmin values, depending on their class of service (MaxDelay constraints) and their offered load as well.

3.2 Model Validation

In this section, we highlight the abilities of our MAC protocol to sustain certain QoS guarantees when the network is under high contention levels. The aim is to evaluate the accuracy of CW adaptation in maintaining bounded MAC queuing delays, regardless of changes in the network load. Throughout our experiment, the relative (per-class) network load is deliberately changed to evaluate the performance of our protocol to suit different network configurations. In order to assess the accuracy of our analysis in terms of the estimated achievable delay, Figs. 2 and 3 compare the model-predicted enqueuing delays $(\mathrm{N} \cdot \mathrm{E}[\mathrm{PST}])$ with the delays effectively experienced during simulations. All network configurations, flows characteristics, and simulation scenarios used in the simulations are thoroughly discussed in the performance evaluation section (Section 5.2).

Fig. 2 illustrates the instantaneous delays experienced by two HP flows having different bit rates (128 and 64 Kbps). Fig. 3 gives the instantaneous delays experienced by two medium-priority (MP) flows with bit rates of 200 and 400 Kbps, respectively. We give, in each figure, the model-based delays estimated by the four involved flows. The given delays are each time averaged over 1 second. The maximum delay bound to not violate MaxDelay is fixed to 0.5 second for HP flows, whereas it is set to 0.8 second for MP flows. Note that the model-estimated delays are calculated at each TC by using different MAC parameters, as explained in the previous section.

Globally, when the network is sufficiently relaxed, there are no violations of delay thresholds, except for some brief spikes that are rather due to 1) short-term fluctuations in collision rate measurements and 2) disparity between the successively drawn backoff intervals.

As clearly apparent in Figs. 2 and 3, our protocol ensures roughly the same delays to TC's flows, regardless of their respective bit rates. The negligible disparity between delays experienced by different TC flows is mainly due to slightly different short-term network-measurements (for example, E[TransAtt] and B(T), O(T)). In fact, for a longer measurement averaging period or resolution T, the different stations would be more coordinated although with a seriously reduced responsiveness in the face of network load variations.

As apparent from the above figures, the service-level fairness among TC flows is achieved, even when the MaxDelay threshold is violated (between $\mathrm{t} = 140$ seconds and $\mathrm{t} = 200$ seconds). Although there is sufficient bandwidth available in the network to carry additional offered load, flows experience higher latencies due to higher enqueuing delays induced by an increase in PST (frequent network occupations cause an increase in the mean number of transmission attempts). At this point, the model-calculated

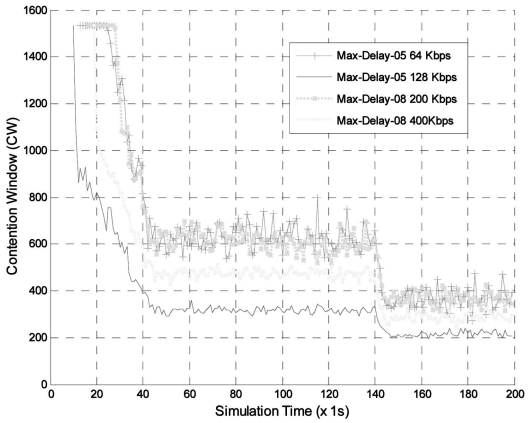

Fig. 4. Variation of the CW size (newCWsize).

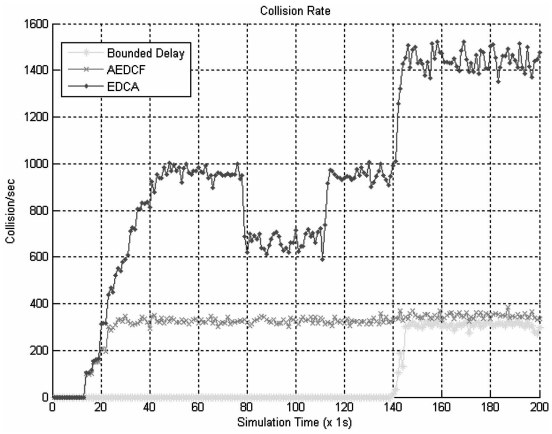

Fig. 6. Collision rate experienced by an HP flow.

CWmax that would accommodate the delay constraints is actually too low (that is, CWmax lower than CWmin), which causes the flows to use CWmin as the maximum CW size (see (6)). Since CWmin calculation is mostly based on the currently used CW size, its value is roughly proportional to previous CWmax values. As a consequence, the fairness between the achieved delays is maintained, since different flows belonging to the same TC use different CWmin values. Consider that flows with a higher offered load usually maintain lower CWmin in order for them to carry the load during high-contention situations.

It is worth mentioning that within our protocol, the network may be fully utilized by MP (respectively, HP) flows. There are no constraints regarding the relative (per-class) network load, since the backoff ranges of supported TCs are not limited by static CWmin/CWmax values as in EDCA-based protocols.

As discussed above, when the network can no longer guarantee the delay exigencies (between $t = 140$ seconds and $t = 200$ seconds), the CW values of different flows, with different priorities, tend to use quite stable values (that is, CWmin). This explains, as shown in Fig. 4, the results between $t = 140$ seconds and $t = 200$ seconds, where all stations are using CWmin as the final CW (newCWsize) to be used in the backoff process for each packet transmission. This fact causes more important delay violations at HP flows (see Figs. 2 and 3), since it is more difficult to ensure delays below 0.5 second under stressed conditions.

In Figs. 5, 6, and 7, we compare the performance of our scheme (Bounded Delays) with the performances of both 80211e EDCA and AEDCF, which was recently proposed to overcome some of EDCA's shortcomings. We used the same network setup and traffic backlogging scenario, with the same number of HP and MP flows as in the previous experiment (see Section 5 for further details).

Fig. 5 depicts the achieved goodput during the simulation. In relaxed situations (between $t = 0$ second to $t = 140$ seconds), both AEDCF and our scheme carry the load, as it is offered without any performance degradation. This can be confirmed in Figs. 6 and 7, where both collision and loss measurements reveal quite a stable behavior with both AEDCF and our scheme. This is not the case with EDCA, which experiences serious performance degradation during the period [$t = 80$ seconds to $t = 110$ seconds]. This situation is the consequence of a high-contention period. As shown in Fig. 6, the flows enter in a high-collision period ($t = 40$ seconds to $t = 80$ seconds) that finishes by causing all the flows to use very large CW sizes, limiting the utilization of the network.

Note that AEDCF performs better than EDCA, because it uses a smoothing mechanism when varying the CW size upon successful/unsuccessful transmissions. The successively drawn backoff intervals are therefore more stable and

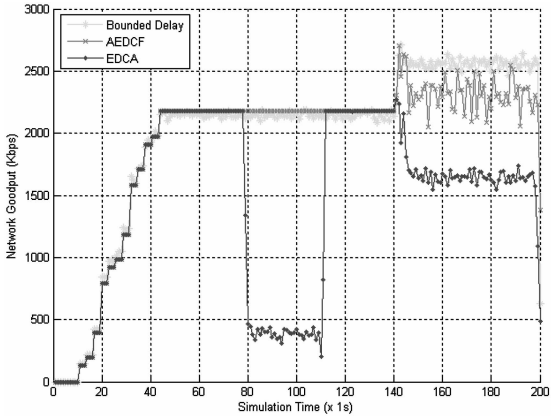

Fig. 5. Goodput measured for an HP flow.

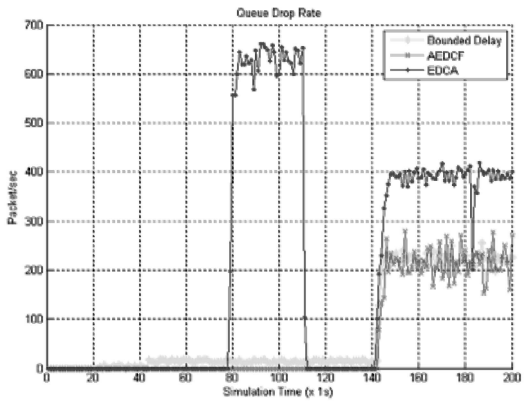

Fig. 7. Drop rate experienced by an HP flow.

prevent situations where all flows are maintaining high CW sizes.

During high network contention, our scheme outperforms both EDCA and AEDCF, thanks to a more careful CW size adjustment at the different backlogged flows, which considerably reduce the collision rate (see Fig. 6) and improve the network utilization (see Fig. 5). In particular, using an additional constraint on the number of contending flows (see (5)), the different active flows use the appropriate CW sizes that maintain an acceptable spacing between successive transmission attempts. It is also worth mentioning that our scheme provides more stable (nonoscillating) performances, as shown in Fig. 7, where the measured loss ratio is quite stable compared to AEDCF.

3.3 Discussion

An important observation that came out from the above results is that there is a critical trade-off between the achieved network throughput and delay guarantees for certain flows. Obviously, it is not possible to fully fill the network capacity while still satisfying strict delay requirements. From a practical point of view, increasing a flow's throughput, beyond a certain extent, means increasing the enqueuing delays, thus probably violating delay constraints. The instantaneous transmission delay at a given flow F is a multifaced problem that depends on the network configuration, that is, depends on many factors such as the bit rate of F, the maximum tolerated delay by F, the overall network load, the network contention level (which itself depends on the overall offered load distribution over the different active flows), and, finally, the delay constraints on the other active flows. Different network configurations (different combinations of the above-mentioned parameters) may result in the same overall achieved throughput, although with different achieved delays. From the NO point of view, this situation poses a major problem.

In fact, it is essential to each time find out the optimal network's operation point by maximizing the number of QoS-enabled services in the network, regardless of the network configuration. This requires a distributed model able to a priori (before admitting new services) predict network performances in terms of the achievable QoS metrics. The AC mechanism should allow for various per-class traffic load distributions to allow NOs to optimize their underlying resources and increase their revenues. The difficulty of implementing this approach in 802.11 lies in estimating the consequences at different active network flows provoked by streams' admission.

4 Multimedia Services Admission and Protection

As apparent from the results presented earlier, our model delivers a fairly good estimation of the achievable delay. Since delay estimation is based on interpacket interval assessment, the achievable throughput, together with potential degradations (mean loss rate), may be predictable as well. Using the packet arrival rate λ, which is a priori known for a given TC, it is possible to capture the queue dynamics based on instantaneous network activities. The packet arrival rate may be, for example, provided by preestablished Service-Level Agreements (SLAs). The objective is to predict the impact of a new stream's acceptation on the overall network performance. In other words, we assess the consequences resulting from increasing the arrival rate of a given TC/station (that is, stream admission) before actually admitting any new entering service.

TABLE 1
A Cross-Layer QoS Mapping

	Conversational services	Streaming services	Best Effort services
Type of application	Interactive voice and video gaming	Streamibg audio/video, Multimedi broadcasting	Web browsing, E-mail, Telnet
IP Diffserv class	EF	AF11	Best effort
IP transmission Delay	600 ms	800 ms	Unspecified
Loss Percentage Resiliency (α_i)	1%	2%	Unspecified

This section introduces a multimedia stream AC algorithm for IEEE 802.11 networks. A flow, in the context of this discussion, is defined as a set of packets belonging to the same TC of a station and uses the same set of MAC parameters. A flow can be seen as an aggregation of several applications' streams belonging to the same TC.

4.1 Multimedia Flow Admission and Protection

As illustrated in Fig. 1, we consider a MAC/LLC queue with a buffer size k. Service is exponential with parameter μ, and interarrival times are exponential with parameter λ. A loss occurs whenever an arriving packet finds the queue full. The queue occupation rate is thus

$$\rho = \frac{\lambda}{\mu} = \lambda \cdot E[PST]. \tag{7}$$

The queue model is assumed to be a single-server queue with finite waiting room (M/M/1/K). Certainly, the Poisson assumption for the arrivals of packets is not the most realistic, but considering the exponential case, this reveals essential features of the system and is a fairly appropriate assumption for an aggregate of different streams (TC). The mean loss rate Lr of an M/M/1/K queue is given by

$$Lr_i = \frac{(1-\rho)\rho^k}{1-\rho^{k+1}}. \tag{8}$$

Since the maximum tolerated loss rate $(\mathrm{MaxDrop} = \mathrm{Lr})$ is a priori known for each TC i, we can numerically fix ρ, since the MAC queue size K is as well known. In fact, the NO may propose different levels of QoS guarantees, where each level is characterized by maximum QoS metric performance bounds (MaxDelay and MaxLoss). Table 1 illustrates an example of TCs when using DiffServ classes mapping.

For instance, assuming a queue length of $\mathrm{k} = 30$ packets and with a maximum tolerated loss rate of $\mathrm{MaxDrop} = 1\ \mathrm{percent}$, the queue occupation rate ρ should be lower than 0.935. In the same manner, $\rho = 0.97$ for a maximum

tolerated loss rate of $\mathrm{MaxDrop} = 2\ \mathrm{percent}$. In this paper, we aim at categorizing the traffic into service classes, where each service class has a maximum delay and a maximum loss rate to not violate.

Based on the delay analysis (that is, PST) and the mean tolerated loss rate, we can now determine the appropriate μ (that is, 1/E[PST]) that satisfies (8). Thus, we analytically figure out the appropriate CW that provides a mean interpacket transmission interval E[PST] necessary to maintain a queue occupation rate at the desired level ρ. By combining (1) and (7), we obtain the appropriate CW size that satisfies the loss requirements associated to a given TC:

$$NewCW_{size} = 2 \cdot E[CW] = 2 \cdot \frac{\frac{\rho}{\lambda} - E[P] \cdot E[TransAtt]}{(1 + B(T)) \cdot E[TransAtt]}, \tag{9}$$

with $CW_{min} \leq newCWsize$, and $newCW_{min} \leq CW_{max}$.

Although the CW (CWsize) given by (6) ensures an acceptable delay with regard to TC requirements, (9) allows for avoiding TC queue overflow by each time checking if the current PST (that is, NewCWsize) is able to absorb the packet arrival rate λ. More precisely, the new CW size ensures that the TC flow in which the entering stream will be aggregated will not violate its maximum tolerated loss rate. The new calculated CW size (NewCWsize) should be also larger than CWmin. This means that the network is able to accommodate the new stream's offered load while still meeting delay guarantees (NewCWsize $<$ CWmax) and keeping an acceptable contention level (NewCWsize $>$ CWmin) to avoid network performance collapse.

Combined with the delay-driven CW adjustment introduced in (6), the above formula may be used to accept new streams in the network. This consists of assessing if a new stream may be serviced while not interfering with the already-active flows. As highlighted already, an overadmission will unavoidably affect all currently serviced flows, as the medium is shared, and an increase in the contention level affects all flows, regardless of their bit rates or priorities. As revealed in Fig. 4, on the other hand, different active flows may simultaneously maintain widely different CW sizes due to different values of CWmin and CWmax. The maintained CW depends actually as much on the flow's offered load as it does on the flow's TC. In certain circumstances, overadmission may cause a certain flow to violate its CWmin limit, whereas other flows still use CW sizes larger than their calculated CWmin. Flows with high bit rates are generally the first flows to reach their CWmin limits. At this point, it is readily realized that the impact of new stream admission should be estimated at all stations.

At new stream admission, each flow in the network recalculates the values of CWmin, CWmax, and NewCWsize according to (5), (6), and (9). The new values of these parameters should take into account changes in network availability entailed by admitting a new stream. Accordingly, certain determinant measurement-based parameters such as B(T), O(T), and E[TransAtt] should be reconsidered. Although E[TransAtt] fluctuations are limited by using an appropriate CWmin, both B(T) and O(T) exhibit significant changes that should be considered to accurately reestimate the new achievable QoS performances.

Again, it is worth mentioning that λ is actually the arrival rate of streams' aggregate belonging to the same TC. At new stream admission, the overall arrival rate at the TC queue would increase as $\overline{\lambda} = \lambda + \Delta\lambda$, where $\Delta\lambda$ is the packet arrival rate of the new entering stream. In this case, the network load should be updated to reflect the additional load induced by the new stream:

$$\overline{B}(T) = \frac{\overline{B}}{\overline{I}} = \frac{B + \beta}{I - \beta}, \quad \text{with} \quad \beta = \overline{\lambda} \cdot T \cdot (20 \cdot 10^{-6}) \cdot L. \tag{10}$$

Here, L is the mean number of time slots occupied by a MAC packet of a given flow, including the overhead involved by acknowledgment. O(T) should be as well updated with the new flow arrival as follows:

$$\overline{O}(T) = \frac{\overline{B}}{T} = \frac{B + \beta}{T}. \tag{11}$$

Given the above-mentioned parameters, all active stations calculate the new values of CW[i]min, CW[i]max, and NewCW[i]size for each TC i. If the new values satisfy all QoS constraints $(\mathrm{CW[i]\,min} < \mathrm{NewCW[i]size} < \mathrm{CW[i]\,max})$ associated with each TC i, then the station concludes that the entering stream will not affect its already-serviced streams. If all stations will not be affected by the entering stream, the AC algorithm may then proceed with stream admission. Otherwise, it means that the stream admission may severely degrade the quality of currently servicing flows, which should lead to the rejection of the entering stream.

In the following section, we further investigate the design of a distributed AC protocol to implement the above-mentioned procedure while allowing for scalability, bandwidth efficiency (low overhead), and accurate coordination between active stations.

4.2 Admission Control Coordination

The first issue to tackle when designing a distributed AC mechanism is the coordination between competing nodes. In fact, aside from necessitating a unified admission model for all stations, we further require harmonizing the estimation of achievable QoS at different stations in order to achieve a coordinated AC decision. In particular, multiple new real-time streams may be simultaneously admitted by individual nodes if not coordinated, causing "*overadmission*." To mitigate this problem while keeping the distributed feature of our protocol, we divide the time into admission cycles (epochs), where only one single stream may be accepted in an admission cycle. The network is assumed to operate on "*slotted*" synchronization epochs, where each epoch is actually equal to a beacon period. This way, the admission cycle is long enough to allow network measurements E[TransAtt] at different stations to converge toward accurate values reflecting the real network conditions before admitting a new stream in the next synchronization epoch.

To completely avoid the overadmission problem, we adopt a coordinator-aided AC scheme. In other words, all admission decisions are made by a coordinating node (CN),

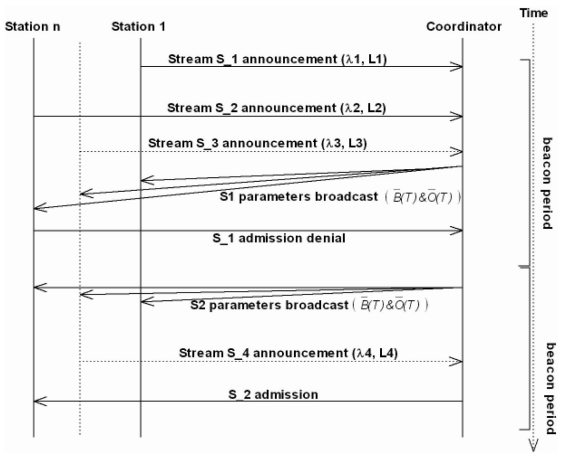

Fig. 8. AC message exchange.

which can record the current number of admitted real-time flows and their occupied channel bandwidth in the network. Clearly, this will prevent overadmission situations. The coordinator node is also in charge of other responsibilities related to the SLA. These additional CN's responsibilities are further discussed in the next sections.

It is important to note that a coordinator is available whether the wireless LAN is working in the infrastructure mode or in the ad hoc mode. If the network is working in the infrastructure mode, the access point is inherently the coordinator. Otherwise, a mobile node can be elected to act as the coordinator in the network by using one of the many algorithms in the literature (see [24] and references therein). A natural solution would be to appoint the node in charge of sending the MAC-level beacon as the CN. As in the 802.11 ad hoc mode, in case of failure, a distributed backoff-based mechanism would design a new node to periodically send the beacon. Further details on the election process are beyond the scope of this paper.

Each time a station S has a new stream to admit, it should beforehand evaluate locally its impact by using new values of B(T) and O(T), as given by (10) and (11). Using (9), the station S should as well assess the risk of having overflows by calculating NewCWsize, where λ is replaced by $\lambda + \Delta\lambda$. In Fig. 8, λ_i $(i = 1 \text{ to } 3)$ stands for the rate $\Delta\lambda$ of a new entering stream.

If the new entering stream does not affect the locally active TC flows, the station S announces the stream's bit rate λ and nominal MSDU size (in terms of time slots) to the CN, which, in turn, recalculates the new values of network occupancy parameters (B(T) and O(T)) to be broadcast. Then, all active stations evaluate the impact of new stream admission (that is, with new B(T) and O(T) changes) on their TC flows and eventually deny the admission if the QoS of one of its TCs may degrade. Note that each TC[i] flow in the network calculate NewCWsize by using its own packet arrival rate λ and maximum queue occupation ratio ρ_i corresponding to its TC.

Fig. 8 illustrates a scenario where in the first beacon period, the coordinator receives three new-stream announcements. The coordinator calculates and broadcasts parameters associated with the first stream S_1. The admission is then aborted by station n when the admission of S_1 interferes with its QoS constraints. In the second beacon period, the coordinator broadcasts S_2 parameters and finishes by accepting the stream, as no active station has denied the acceptation within the current beacon period. Typically, here, S_2 should have a lower packet rate than S_1.

For scalability reasons, AC message handshakes are kept to a minimum by broadcasting CN messages (that is, parameter broadcast and admission messages). Furthermore, response messages (that is, admission denial messages) are sent by an active station only if one of their QoS thresholds associated with TC flows would be violated with the new stream admission. A single denial message suffices to abort the whole stream admission process, so other stations no longer need to send denial messages; that is, all stations overhear AC messages.

To increase the reliability of a CN's broadcasted messages, we use an efficient basic data rate (1 Mbps) usually employed to transmit the beacon, RTS/CTS, and ACK messages. On the other hand, during the AC process, all directed messages exchanged between the coordinator node and other stations are fully persistent in the sense that they are retransmitted until successful reception.

Upon the first admission in a given beacon period, the other flows seeking admission in the network should differ the announcement in the next beacon period, and additional network measurements are carried out before final admission. This allows all stations to take into account the changes in network availability before accepting new streams (that is, allows the different competing stations to have a coherent perception of the network availability by carrying out measurements during a long-enough period such as a beacon period).

5 Performance Evaluation

In order to evaluate the advantages of the proposed protocol, we have constructed a simulation using ns-2. We compare our distributed AC protocol scheme by using the last IEEE 802.11e Standard [2]. Our AC protocol was implemented atop the last ns-2 implementation of IEEE 802.11e that uses a more realistic MAC implementation, where 802.11 nodes are more synchronized, thanks to a considerably improved backoff freezing process. We further improved this implementation with a more accurate MAC Timer for better synchronization between flows with respect to network load measurement (that is, B(T) and O(T) measurements). In this section, we highlight various aspects entailed by deploying effective AC mechanisms in WLAN, with a special focus on the appropriate *brokering strategies*[1] to be adopted by NOs.

5.1 Simulation Model

For the simulations, we have created a network consisting of 16 wireless terminals WT[i], $\mathrm{i} = 1, .., 16$. A single CN is arbitrarily chosen among the 16 nodes. The CN is actually the node that periodically sends the beacon frame in the 802.11 ad hoc mode.

Each WT may generate up to two different TC flows at the same time, representing two uniquely prioritized TCs: HP with a MaxDelay of 500 ms and MP with a MaxDelay of

1. By brokering strategies, we mean the pricing strategies of network operators in offering QoS-enabled services on the basis of preestablished SLAs.

TABLE 2
Traffic Characteristics

Traffic features	Packet size (Bytes)	Generation Interval (second)	Bit rate (bps)
Max_delay_0.5 (HP)	160	0.02	64000
Max_delay_0.5 (HP)	160	0.01	128000
Max_delay_0.5 (MP)	500	0.02	200000
Max_delay_0.5 (MP)	500	0.01	400000

800 ms. In our simulation, we choose to generate only one flow per station so as to make the contention for seizing the medium worse. In fact, if the backoff counters of two or more TCs collocated in one station elapse at the same time, a scheduler inside the station treats the event as a "*virtual*" collision without causing waste of the network's time slots. In this case, the medium is seized by the TC with the highest priority among the colliding TCs, whereas other colliding TCs defer their transmissions as if the collision occurred in the real medium.

Constant bit rate (CBR) sources are used for all traffic. The properties of these flows are specified in Table 2. CBR sources put more stringent exigencies (for example, packet rate and enqueuing delay) on the network than VBR sources. In fact, multiple CBR sources would require that the network sustains the overall offered load (the summation of CBR sources' bit rates) throughout the simulation period, which may provoke MAC queue overflows after a fairly long run. In contrast, with multiple VBR sources, the peaks of bit rates unlikely occur at the same time, which allows the network to absorb the brief offered load bursts exhibited by different traffic sources at different time scales. In practice, the SLA between the service costumer and the service provider specifies the service characteristics in terms of the fixed mean data rate (and, eventually peak, data rate) with associated QoS metric performance bounds. It is extremely difficult, in practice, to precisely characterize the burstiness of a VBR stream.

We performed several simulations runs in order to evaluate the performance of our AC scheme with respect to different QoS metrics (loss and delay). We also give the evolution of the CW size at each TC flow type as the network configuration changes over the time. Each run consists of 200 seconds of simulated network lifetime with a fixed scenario in terms of per-TC traffic load variation and the order of single flow backlogging. From time $t = 0$ second to $t = 10$ seconds, the channel is empty, whereas from $t = 10$ seconds, new flows of each class are started at 3-seconds intervals and begin competing for the channel. Byt $= 37$ seconds, each class has five active flows: two 64-Kbps and three 128-Kbps HP flows and three 200-Kbps and two 400-Kbps MP flows. From $t = 37$ seconds to $t = 140$ seconds, the network remains in this state in order for us to asses to what extent our protocol can sustain the QoS. At $t = 140$ seconds, four new flows are started at 1-second intervals as follows: a 128-Kbps HP flow, a 400-Kbps MP flow, a 64-Kbps HP flow, and, finally, a 400-Kbps MP flow. At this point, the network is exhibiting a high contention level, which means an increased mean number of unsuccessful transmissions attempts. From $t = 140$ seconds to $t = 200$ seconds, the simulation is completed with 16 flows backlogged in 16 different stations.

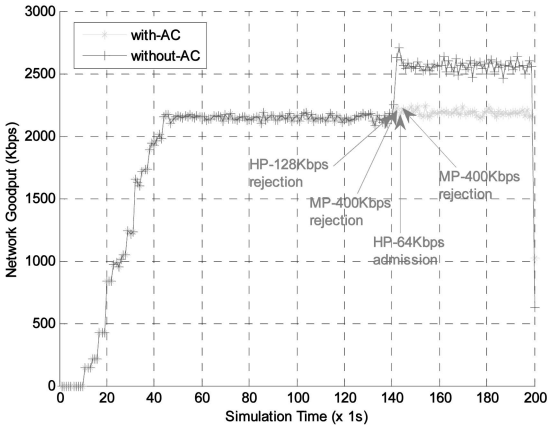

Fig. 9. Overall network utilization.

The objective of the above-described network dimensioning is to assess if our model is able to react to frequent changes in the elative (per-class) network load while still meeting the QoS requirements. Indeed, we keep changing the per-class traffic load and increasing the number of backlogged flows and, with it, the network contention level. We use flows with different bit rates in order to thoroughly study the service-level fairness among flows belonging to the same TC but with unbalanced offered load.

5.2 Experimental Results

Since the ability of our scheme to maintain sustained QoS guarantees was validated in Section 3.2, in this section, we will focus on evaluating the AC part. We specially assess to what extent our AC scheme is able to protect already-active flows. Another important aspect highlighted in this section is the ability of our scheme to keep admitting new entering flows based on a careful evaluation of their impact on the already active flows.

In this section, we compare the performance of our scheme, that is, the bonded-delay (BD) scheme, when using the AC mechanism and without using AC. We refer to these two operation modes as with AC and without AC, respectively.

The overall network utilization is shown in Fig. 9 in terms of the total achieved throughput (goodput) during the simulation. Clearly, when the network is sufficiently relaxed (before $t = 140$ seconds), there is sufficient bandwidth available, and both BDS-AC and BDS achieve similar throughputs, carrying the load as it is offered. However, under stressed conditions, BDS gains a significant advantage over BDS-AC. The goodput gain reaches about 20 percent when the load is around 2.4 Mbps (between $t = 140$ seconds and $t = 200$ seconds). At this point, the AC mechanism in BDS-AC rejects three entering flows: the 128-Kbps HP flow at $t = 140$ seconds, the 400-Kbps MP flow at $t = 141$ seconds, and, finally, the 400-Kbps MP flow at $t = 143$ seconds. Meanwhile, a 64-Kbps HP flow was accepted at $t = 141$ seconds. This bandwidth gain comes, however, with a serious degradation in the QoS of all active multimedia flows, as clearly revealed by the delay measurements of single flows.

As shown in Figs. 10, 11, 12, and 13, the four flow types experience high delays as from $t = 140$ seconds when no AC is applied. The performance degradation starts at $t = 140$ seconds with the overadmission of a 128-Kbps HP flow.

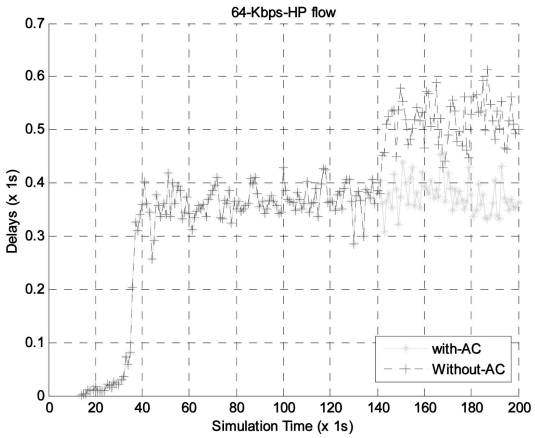

Fig. 10. End-to-end delays for 64-Kbps HP flows.

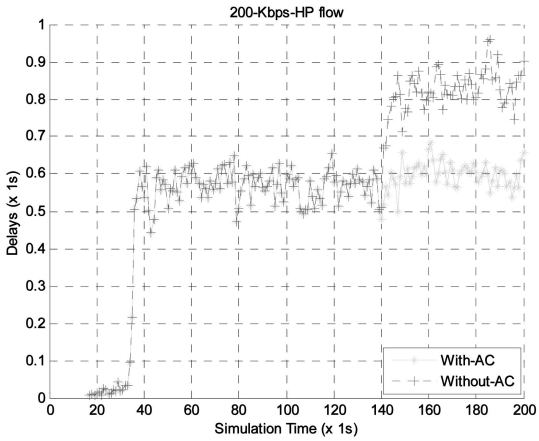

Fig. 12. End-to-end delays for 200-Kbps MP flows.

The performance is further degraded with the acceptance of three other flows. Depending on their respective offered load, the different TC flows are differently affected by this increase in the network contention level.

Although high bit rate flows maintain quite small CW sizes (see Figs. 14 and 15) compared to other flows, they are still unable to overcome the increasing network offered load and the entailed high PSTs. This decrease in CW sizes is driven by the delay constraint presented in (6), without taking into account the queue overflow risk. As a matter of fact, throughput degradation is mostly caused by excessive packet dropping due to overloaded MAC queues. Here, the advantage of the AC becomes essential by clearly establishing relation between the packet arrival rate and the PST and ultimately assessing the achievable QoS before actually admitting any new entering flow. This beforehand flow admission assessment at each active station is done by deriving the ideal CW size (that is, NewCWsize, which ensures that the loss rate constraint will be respected) to be used and comparing it with both 1) CWmin to make sure that the contention is still controlled and 2) CWmax to make sure that the delay constraints will be respected at each flow active in the network.

It is worth mentioning that the entering streams were each time rejected by high-bit rate TC flows (that is, 400-Kbps MP flows). In other words, during the AC process, stations carrying high-data-rate load reject the new entering flow. Based on network-based PST measurements, it is much more difficult to maintain an acceptable loss rate if the TC flow is handling a high packet arrival rate λ.

An important observation to point out is that the results given in the model validation section and those presented in the performance evaluation section are slightly different. For instance, in the above-presented results, the achieved delays of different TCs are far below their respective MaxDelay thresholds. This is due to the fact that with the AC mechanism, the CW size effectively maintained by each flow is generally smaller than the one that would be maintained if AC is not used. In fact, with an additional constraint to avoid MAC queue overflows (that is, NewCWsize calculation; see (9)), the actually used CW size is smaller than the one given by (6).

5.3 Discussion

It is essential for NOs to determine the per-service cost according to the bit rate and the TC of that service. This is indeed a complex task, since the cost (price) is not necessarily expressed in terms of data volume per TC (that is, a fixed price for 1 Kbps/TC). The service cost should be fairly established based on 1) real network resources committed by the NO to accommodate the new entering

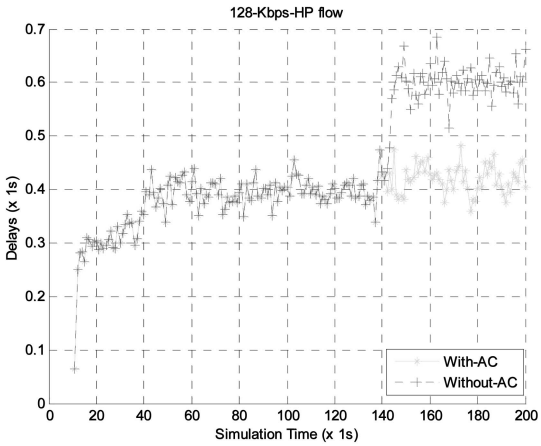

Fig. 11. End-to-end delays for 128-Kbps HP flows.

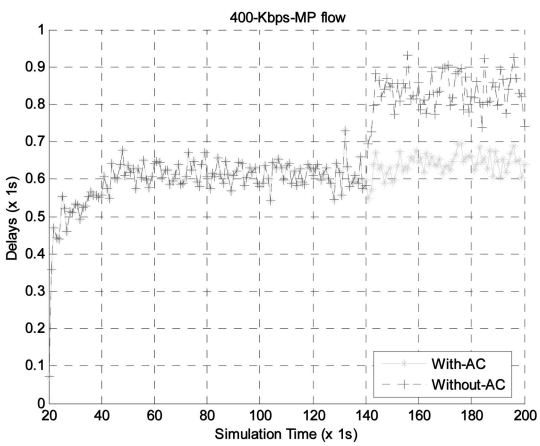

Fig. 13. End-to-end delays for 400-Kbps MP flows.

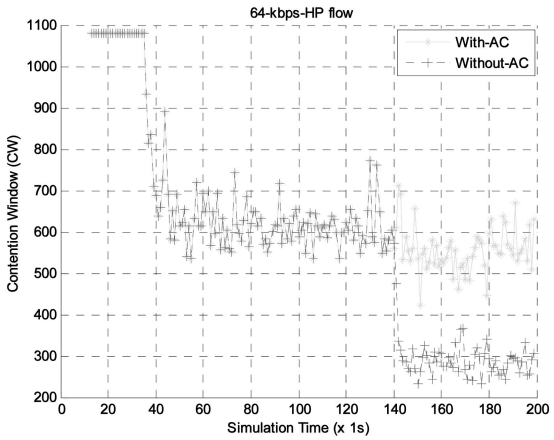

Fig. 14. CW for 64-Kbps HP flows.

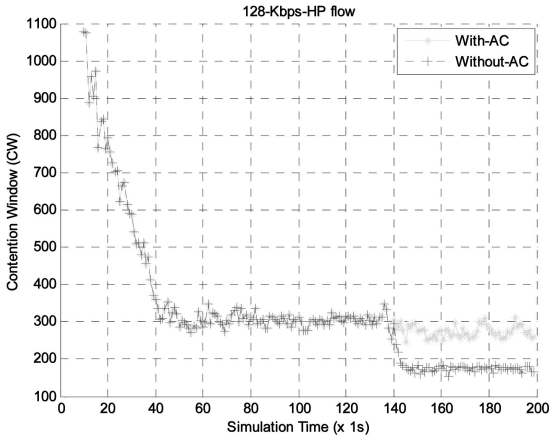

Fig. 15. CW for 128-Kbps HP flows.

flow and 2) the impact of the service admission on the overall network availability.

As revealed earlier, in certain circumstances, compared to a single flow with a 100-Kbps bit rate, two flows with a 50-Kbps bit rate would have a stronger impact on the network resources availability (due to higher collision rates). The optimal pricing strategy could consist of proposing a certain fixed price per service and an additional charging for the data volume (Kbps/TC). Another approach that is worth investigating consists of dynamically changing the per-service price based on the network load and ultimately designing some sort of auction-based pricing that weight the resource demand with the resource availability in an online basis before fixing the per-service price.

In conclusion, the cost of a given service may widely vary, depending on the instantaneous network configuration, as the impact on network availability of that service admission is different. Achieving such a resource management strategy suggests the presence of a central entity that controls and keeps track of the SLAs contracted by the service provider and the service subscribers. Further discussion on the pricing strategy issues is beyond the scope of this paper.

6 Conclusion

In this paper, we presented a delay-sensitive scheme combined with an AC mechanism that is based on thorough analysis of the trade-off between high network utilization and achieving bounded QoS metrics in operated 802.11-based networks. First, we derive an accurate delay estimation model to adjust the CW size in a real-time basis by considering key network factors, MAC queue dynamics, and application-level QoS requirements. Second, we use the above-mentioned delay-based CW size adaptation model to derive a fully distributed AC model that provides protection for existing flows in terms of QoS guarantees. Compared to existing QoS-capable protocols, the proposed scheme offers an improved ability to guarantee cross-layer QoS while allowing for different network configurations in terms of relative (per-class) network load. Common application-level QoS metric performance thresholds may be henceforth guaranteed at the MAC level.

As briefly discussed throughout this paper, the SLA appears as an important aspect to deal with when deploying the AC mechanism in operated/commercial 802.11 networks. Translation of a QoS-enabled service request into a quantifiable resource commitment (and, thus, a specific cost) is indeed a challenging task. Therefore, our future work will focus on deriving a viable model to effectively establish the service cost in terms of network resource commitment, depending on its impact on the network availability.

References

[1] *IEEE Standard 802.11, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications,* IEEE, June 1999.

[2] *Final 802.11e Standard, Wireless LAN Medium Access Control (MAC) Enhancements for Quality of Service (QoS),* IEEE, July 2005.

[3] A. Nafaa et al., "Sliding Contention Window (SCW): Towards Backoff Range-Based Service Differentiation over IEEE 802.11 Wireless LAN Networks," *IEEE Network Magazine,* special issue on wireless local area networking: QoS provision and resource management, vol. 19, no. 4, July/Aug. 2005.

[4] I. Aad and C. Castelluccia, "Differentiation Mechanisms for IEEE 802.11," *Proc. IEEE INFOCOM '01,* vol. 1, pp. 209-218, Apr. 2001.

[5] L. Romdhani et al., "Adaptive EDCA: Enhanced Service Differentiation for IEEE 802.11 Wireless Ad Hoc Networks," *Proc. IEEE Wireless Comm. and Networking Conf. (WCNC '03),* vol. 2, pp. 1373-1378, Mar. 2003.

[6] M. Malli et al., "Adaptive Fair Channel Allocation for QoS Enhancement in IEEE 802.11 Wireless LAN," *Proc. IEEE Int'l Conf. Comm. (ICC '04),* pp. 347-3475, July 2004.

[7] Z.J. Haas and J. Deng, "On Optimizing the Backoff Interval for Random Access Schemes," *IEEE Trans. Comm.,* vol. 51, no. 12, pp. 2081-2090, Dec. 2003.

[8] A. Ksentini et al., "Toward an Improvement of H.264 Video Transmission over IEEE 802.11e through a Cross-Layer Architecture," *IEEE Comm. Magazine,* special issue on cross-layer protocol Eng., vol. 44, no. 1, pp. 107-114, Jan. 2006.

[9] S.-T. Sheu and T.-F. Sheu, "A Bandwidth Allocation/Sharing/Extension Protocol for Multimedia over IEEE 802.11 Ad Hoc Wireless LANs," *IEEE J. Selected Areas in Comm.,* vol. 19, pp. 2065-2080, Oct. 2001.

[10] G. Bianchi, "Performance Analysis of the IEEE 802.11 Distributed Coordination Function," *IEEE J. Selected Areas in Comm.,* vol. 18, pp. 535-547, Mar. 2000.

[11] E. Ziouva and T. Antonakopoulos, "CSMA/CA Performance under High Traffic Conditions: Throughput and Delay Analysis," *Elsevier Computer Comm. J.,* vol. 25, no. 3, pp. 313-321, 2002.

[12] Y. Xiao and H. Li, "Evaluation of Distributed Admission Control for the IEEE 802.11e EDCA," *IEEE Comm. Magazine,* vol. 42, no. 9, pp. S20-S24, 2004.

[13] Y. Xiao et al., "Protection and Guarantee for Voice and Video Traffic in IEEE 802.11e Wireless LANs," *Proc. IEEE INFOCOM,* 2004.
[14] F. Cali et al., "Dynamic Tuning of the IEEE 802.11 Protocol to Achieve a Theoretical Throughput Limit," *IEEE/ACM Trans. Networking,* vol. 8, no. 6, pp. 785-790, Dec. 2000.
[15] H. Zhai, X. Chen, and Y. Fang, "How Well Can the IEEE 802.11 Wireless LAN Support Quality of Service?" *IEEE Trans. Wireless Comm.,* vol. 4, no. 6, pp. 3084-3094, 2004.
[16] H. Zhai, X. Chen, and Y. Fang, "A Call Admission and Rate Control Scheme for Multimedia Support over IEEE 802.11 Wireless LANs," *Proc. First Int'l Conf. Quality of Service in Heterogeneous Wired/Wireless Networks (QSHINE),* 2004.
[17] M. Barry, A.T. Campbell, and A. Veres, "Distributed Control Algorithms for Service Differentiation in Wireless Packet Networks," *Proc. IEEE INFOCOM '01,* vol. 1, pp. 582-590, 2001.
[18] A. Veres et al., "Supporting Service Differentiation in Wireless Packet Networks Using Distributed Control," *IEEE J. Selected Areas in Comm.,* vol. 19, no. 10, pp. 2081-2093, Oct. 2001.
[19] A. Nafaa, "Provisioning of Multimedia Services in IEEE 802.11 Networks: Facts and Challenges," *IEEE Wireless Comm.,* 2006.
[20] *End-to-End QoS through Integrated Management of Content, Networks and Terminals,* EU-IST Integrated Project IST-2003-507637/ENTHRONE, Sixth EU Framework Program, 2003-2007.
[21] D. Pong and T. Moors, "Call Admission Control for IEEE 802.11 Contention Access Mechanism," *Proc. IEEE Global Telecomm. Conf. (GLOBECOM '03),* pp. 174-178, Dec. 2003.
[22] X. Chen, B. Li, and Y. Fang, "A Dynamic Multiple-Threshold Bandwidth Reservation (DMTBR) Scheme for QoS Provisioning in Multimedia Wireless Networks," *IEEE Trans. Wireless Comm.,* vol. 4, no. 2, Mar. 2005.
[23] G. Bianchi et al., "Performance Evaluation and Enhancement of the CSMA/CA MAC Protocol for 802.11 Wireless LANs," *Proc. Seventh IEEE Int'l Symp. Personal, Indoor and Mobil Radio Comm. (PIMRC '96),* pp. 392-396, Oct. 1996.
[24] H. Garcia-Molina, "Elections in a Distributed Computing System," *IEEE Trans. Computers,* vol. 31, no. 1, Jan. 1982.

Abdelhamid Nafaa received the master's and PhD degrees from Versailles Saint Quentin en Yvelines University in 2001 and 2005, respectively. He is a Marie Curie research fellow under the EU-FP6 EIF Marie Curie Action, which seeks broader synergy in the European research space. He is currently with the School of Computer Science and Informatics, University College of Dublin (UCD), undertaking independent research work in multimedia service distribution over carrier-grade networks under the Marie Curie Award. Before joining UCD, he was a professor assistant at Versailles Saint Quentin en Yvelines University, where he was involved in several national and European projects such as NMS, IST-ENTHRONE1, IST-ATHENA, and IST-IMOSAN. He was also a technology consultant for a US-based and a European-based companies in the area of reliable multimedia communication over Wi-Fi technology and IMS-based multicasting in DVB-S2 satellite networks, respectively. He is now involved in a successful FP7 proposal CARMEN, which aims at developing a mixed Wi-Fi/WiMax wireless mesh networks to support carrier-grade services. He is a coauthor of more than 25 technical journal papers and international conference proceedings on multimedia communications.

Adlen Ksentini received the MS degree in telecommunications and multimedia networking from the University of Versailles in 2001, and the PhD degree in computer science in 2005 from the University of Cergy-Pontoise. His PhD dissertation focused on QoS provisioning in IEEE 802.11-based networks. He is an associate professor at the University of Rennes 1, Rennes, France, where he is also a member of the IRISA Laboratory. He is involved in several industrial projects and European projects such as FP6 IST-ANEMONE, which aim at realizing a large-scale testbed supporting mobile user on heterogeneous wireless technologies. His research interests include mobility and QoS support in IEEE 802.16, QoS support in the newly IEEE 802.11s mesh networks, and multimedia transmission over WLAN. He is a coauthor of more than 20 technical journal papers and international conference proceedings. He is a member of the IEEE.