Qiang Ni, et al. Journal of Wireless Communications and Mobile Computing, Wiley. 2004: Volume 4, Issue 5: pp.547-566

A Survey of QoS Enhancements for IEEE 802.11 Wireless LAN

Qiang Ni*, Lamia Romdhani, Thierry Turletti

Planete Group, INRIA, 2004 route des Lucioles BP93, 06902 Sophia Antipolis, France

Summary

Quality of service (QoS) is a key problem of today's IP networks. Many frameworks (IntServ, DiffServ, MPLS, etc.) have been proposed to provide service differentiation in the Internet. At the same time, the Internet is becoming more and more heterogeneous due to the recent explosion of wireless networks. In wireless environments, bandwidth is scarce and channel conditions are time-varying and sometimes highly lossy. Many previous research works show that what works well in a wired network cannot be directly applied in the wireless environment. Although IEEE 802.11 wireless LAN (WLAN) is the most widely used WLAN standard today, it cannot provide QoS support for the increasing number of multimedia applications. Thus, a large number of 802.11 QoS enhancement schemes have been proposed, each one focusing on a particular mode. This paper summarizes all these schemes and presents a survey of current research activities. First, we analyze the QoS limitations of IEEE 802.11 wireless MAC layers. Then, different QoS enhancement techniques proposed for 802.11 WLAN are described and classified along with their advantages/drawbacks. Finally, the upcoming IEEE 802.11e QoS enhancement standard is introduced and studied in detail.

Keywords: IEEE 802.11, Medium Access Control (MAC), Quality of Service (QoS), Distributed Coordination Function (DCF), Point Coordination Function (PCF), IEEE 802.11e, Hybrid Coordination Function (HCF)

1. Introduction

IEEE 802.11 wireless LAN (WLAN) [1] is one of the most deployed wireless technologies all over the world and is likely to play a major role in next-generation wireless communication networks. The main characteristics of the 802.11 WLAN technology are simplicity, flexibility and cost effectiveness. This technology provides people with a ubiquitous communication and computing environment in offices, hospitals, campuses, factories, airports, stock markets, etc. Simultaneously, multimedia applications have experienced an explosive growth. People are now requiring to receive high-speed video, audio, voice and Web services even when they are moving in offices or travelling around campuses. However, multimedia applications require some quality of service (QoS) support such as guaranteed bandwidth, delay, jitter and error rate. Guaranteeing those QoS requirements in 802.11 WLAN is very challenging due to the QoS-unaware functions of its medium access control (MAC) layer and the noisy and variable physical (PHY) layer characteristics. In this paper we mainly focus on QoS issues at 802.11 MAC layer.

The primary objectives of this paper are to:

- Introduce an overview of IEEE 802.11 WLAN standard,
- Analyze the QoS problems of IEEE 802.11 MAC layer functions,
- Survey the main QoS enhancement schemes that have been proposed for 802.11 WLAN,
- Describe and study the new IEEE 802.11e QoS enhancement WLAN standard [2].

Some previous work has been done to survey the QoS support for IEEE 802.11 WLAN or mobile ad-hoc networks [5, 11, 37,42], but they do not introduce the latest research work including the new QoS standard, IEEE 802.11e. Moreover, a broad survey is needed to summarize all the current research efforts on QoS support for IEEE 802.11 WLAN. This paper surveys all these efforts and intends to provide a

* Correspondence to: Qiang Ni, Hamilton Institute, National University of Ireland, Maynooth, Co. Kildare, Ireland.
E-mail: Qiang.Ni@ieee.org
This work was done while he was with INRIA. Qiang Ni is currently with the Hamilton Institute.

comprehensive view of the various works. Then, some hints for future research works on QoS enhancement are provided in this paper.

The rest of this paper is organized as follows. Section 2 introduces an overview of IEEE 802.11 WLAN. The limitations of QoS support in 802.11 MAC functions are addressed in Section 3. Section 4 surveys the different QoS enhancement schemes that have been proposed for 802.11 MAC layer. In Section 5, the upcoming QoS enhancement standard 802.11e is introduced and analyzed. Section 6 presents simulation-based evaluations of different QoS-enhanced schemes and Section 7 concludes the survey and presents possible future areas of research.

2. An overview of IEEE 802.11 WLAN

2.1 Standard activities of IEEE 802.11

The IEEE 802.11 WLAN standard covers the MAC sub-layer and the physical (PHY) layer of the open system interconnection (OSI) network reference model [1]. Logical link control (LLC) sub-layer is specified in the IEEE 802.2 standard. This architecture provides a transparent interface to the higher layer users: stations (STAs) may move, roam through an 802.11 WLAN and still appear as stationary to 802.2 LLC sub-layer and above. This allows existing TCP/IP protocols to run over IEEE 802.11 WLAN just like wired Ethernet deployed. Figure 1 shows different standardization activities done at IEEE 802.11 PHY and MAC layers. In 1997, IEEE provided three kinds of options in the PHY layer, which are an InfraRed (IR) baseband PHY, a frequency hopping spread spectrum (FHSS) radio and a direct sequence spread spectrum (DSSS) radio. All these options support both 1 and 2Mbps PHY rate. In 1999, the IEEE defined two high rate extensions: 802.11b in the 2.4GHz band with data rates up to 11Mbps, based on DSSS technology; and 802.11a in the 5GHz band with data rates up to 54Mbps, based on orthogonal frequency division multiplexing (OFDM) technology. Recently, 802.11g is finalized. It extends 802.11b PHY layer to support data rates up to 54Mbps in the 2.4GHz band. Moreover, ongoing 802.11h will enhance 802.11a with adding indoor and outdoor license regulations for the 5GHz band in Europe.

At the MAC layer, 802.11e is the first supplement to enhance the QoS performance of 802.11 WLAN; an Inter-Access Point protocol is defined in 802.11f to allow STAs roaming between multi-vendor access points; 802.11i aims to enhance security and authentication mechanisms for 802.11 MAC.

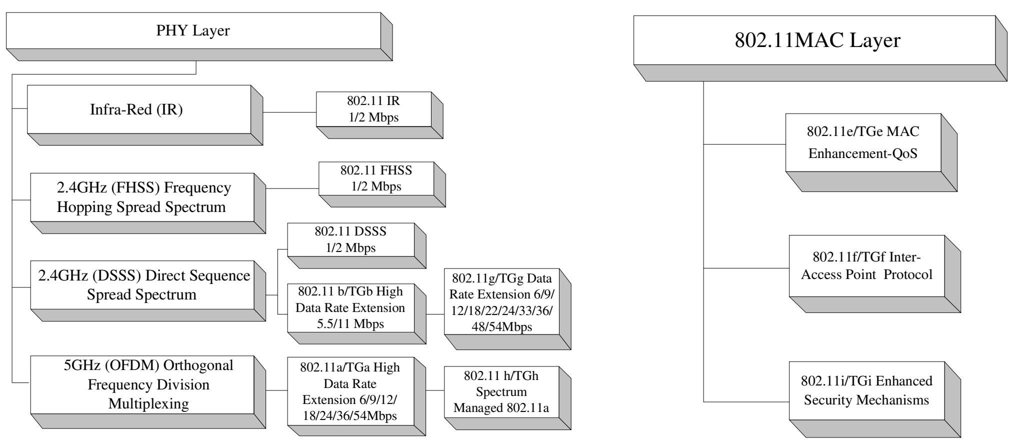


Fig.1 Snapshot of 802.11 PHY and MAC standard activities

2.2 IEEE 802.11 MAC

The IEEE 802.11 MAC sub-layer defines two medium access coordination functions, the basic *Distributed Coordination Function* (DCF) and the optional *Point Coordination Function* (PCF) [1]. 802.11 can operate both in contention-based DCF mode and contention-free PCF mode, and supports two types of transmissions: *asynchronous* and *synchronous*. Asynchronous transmission is provided by DCF whose implementation is mandatory in all 802.11 STAs. Synchronous service is provided by PCF that basically implements a polling-based access. Unlike DCF, the implementation of PCF is not mandatory.

The reason is that the hardware implementation of PCF is thought to be too complex at that time. Furthermore, PCF itself relies on the asynchronous service provided by DCF. As specified in the standard, a group of STAs coordinated by DCF or PCF is formally called a basic service set (BSS). The area covered by the BSS is known as the basic service area (BSA), which is similar to a cell in a cellular mobile network. There are two different modes to configure an 802.11 wireless network: ad-hoc mode and infrastructure mode. In ad-hoc mode, the mobile STAs can directly communicate with each other to form an Independent BSS (IBSS) without connectivity to any wired backbone. In infrastructure mode, the mobile STAs can communicate with the wired backbone through the bridge of access point (AP). Note that the DCF can be used both in ad-hoc and infrastructure modes, while PCF is only used in infrastructure mode.

2.2.1 DCF: Distributed Coordination Function

DCF is a distributed medium access scheme based on *carrier sense multiple access with collision avoidance* (CSMA/CA) protocol. In this mode, an STA must sense the medium before initiating a packet transmission. Two carrier sensing mechanisms are possible: PHY carrier sensing at air interface and virtual carrier sensing at PHY MAC layer. PHY carrier sensing detects the presence of other STAs by analyzing all detected packets and channel activity via relative signal strength from other STAs. Virtual carrier sensing can be used by an STA to inform all other STAs in the same BSS how long the channel will be reserved for its frame transmission. On this purpose, the sender can set a duration field in the MAC header of data frames, or in the RequestToSend (RTS) and ClearToSend (CTS) control frames. Then, other STAs can update their local timers of network allocation vectors (NAVs) to indicate this duration. As shown in Figure 2, if a packet arrives at an empty queue and the medium has been found idle for an interval of time longer than a Distributed InterFrame Space (DIFS), the source STA can transmit the packet immediately [1]. Meanwhile, other STAs defer their transmission while adjusting their NAVs, and then the backoff process starts. In this process, the STA computes a random time interval, called *Backoff_timer*, selected from the contention window (CW): *Backoff_timer= rand* [*0,CW*] · *slot time*, where CW_{min}<CW <CW_{max} and *slot time* depends on the PHY layer type. The backoff timer is decreased only when the medium is idle; it is frozen when another STA is transmitting. Each time the medium becomes idle, the STA waits for a DIFS and continuously decrements the backoff timer. As soon as the backoff timer expires, the STA is authorized to access the medium. Obviously, a collision occurs if two or more STAs start transmission simultaneously. Unlike a wired network, collision detection in a wireless environment is impossible due to significant difference between transmitted and received power levels. Hence, a positive acknowledgement is used to notify the sender that the transmitted frame has been successfully received, see Figure 2. If the acknowledgement is not received, the sender assumes that the transmitted frame was collided, so it schedules a retransmission and enters the backoff process again. To reduce the probability of collisions, after each unsuccessful transmission attempt, the CW is doubled until a predefined maximum value CW_{max} is reached. After each successful transmission, the CW is reset to a fixed minimum value CW_{min}.

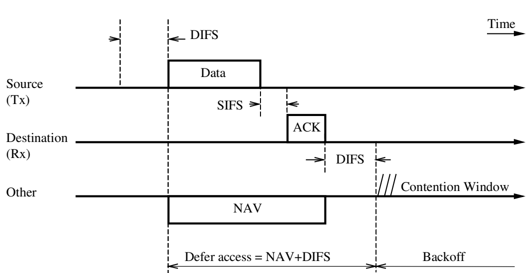


Fig. 2 Basic DCF CSMA/CA

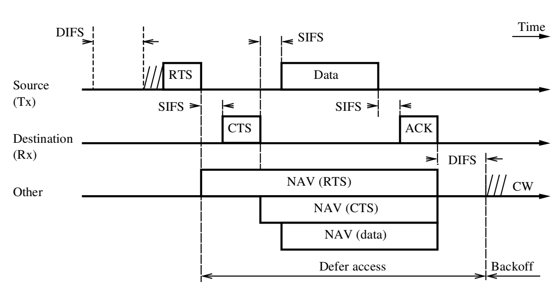


Fig. 3 RTS/CTS access scheme

Hidden terminals are STAs that the receiver can hear but that cannot be detected by other senders. Consequently, the packets from different senders will collide at the same receiver. In order to solve the hidden terminal problem, an optional RTS/CTS scheme is introduced. The source sends a short RTS frame (20 bytes) before each data frame transmission, see Figure 3, and the receiver replies with a CTS frame (14 bytes) if it is ready to receive. After the source receives the CTS frame, it starts transmitting its frame. So, all other STAs hearing an RTS, a CTS or a data frame in the BSS can update their NAVs, and will not start transmissions before the updated NAV timers reach zero. Since a collision of a short RTS or CTS frame is less severe than a collision of data frame (up to 2346 bytes), the RTS/CTS scheme improves the performance of basic DCF scheme considerably in many cases. The overhead of sending RTS/CTS frames becomes considerable when data frame sizes are small, thus the channel is used sub-optimally. Moreover, an uncorrectable error in a larger frame leads to wasting more bandwidth and more transmission time as compared with an error in a smaller frame. So an optimization parameter of fragmentation_threshold is used. That means, when data frame size exceeds this threshold, the data frame will be partitioned into several smaller MAC level frames.

2.2.2 PCF: Point Coordination Function

PCF uses a centralized polling scheme, which requires the AP as a point coordinator (PC). If a BSS is set up with PCF-enabled, the channel access time is divided into periodic intervals named beacon intervals, see Figure 4. The beacon interval is composed of a contention-free period (CFP) and a contention period (CP). During the CFP, the PC maintains a list of registered STAs and polls each STA according to its list. Then, when an STA is polled, it gets the permission to transmit data frame. Since every STA is permitted a maximum length of frame to transmit, the maximum CFP duration for all the STAs can be known and decided by the PC, which is called *CFP_max_duration*. The time used by the PC to generate beacon frames is called target beacon transmission time (TBTT). In the beacon, the PC denotes the next TBTT and broadcasts it to all the other STAs in the BSS. In order to ensure that no DCF STAs are able to interrupt the operation of the PCF, a PC waits for a PCF InterFrame Space (PIFS), which is shorter than DIFS, to start the PCF. Then, all the other STAs set their NAVs to the values of *CFP_max_duration* time, or the remaining duration of CFP in case of delayed beacon. During the CP, the DCF scheme is used, and the beacon interval must allow at least one DCF data frame to be transmitted.

A typical medium access sequence during PCF is shown in Figure 4. When a PC polls an STA, it can piggyback the data frames to the STA together with the CF-Poll, then the STA sends back data frame piggybacked with an ACK after a SIFS interval. When the PC polls the next STA, it piggybacks not only the data frame to the destination, but also an ACK to the previous successful transmission. Note that almost all packet transmissions are separated by the SIFS except for one scenario: if the polled STA does not respond the PC within a PIFS period, the PC will poll the following STA. Silent STAs are removed from the polling list after several periods and may be polled again at the beginning of the next CFP. At any time, the PC can terminate the CFP by transmitting a CF-End packet, then all the STAs in the BSS should reset their NAVs and attempt to transmit during the CP. Normally, PCF uses a round-robin scheduler to poll each STA sequentially in the order of polling list, but priority-based polling mechanisms can also be used if different QoS levels are requested by different STAs.

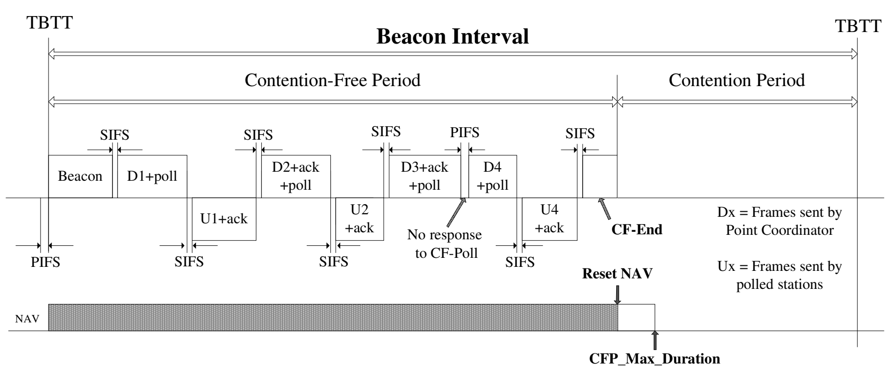


Fig. 4 PCF and DCF cycles

3. QoS limitations of 802.11 MAC

The most important functions of a wireless MAC layer include controlling channel access, maintaining QoS, and providing security. Wireless links have specific characteristics such as high loss rate, bursts of frame loss, packet re-ordering, large packet delay and jitter. Furthermore, the wireless link characteristics are not constant and vary over time and place. Mobility of users may cause the end-to-end path to change when users are roaming. Users expect to receive the same QoS once changing their point of attachment.

This implies that the new path should also support the existing QoS, and problems may arise when the new path cannot support such requirements.

There are several ways to characterize QoS in WLAN such as *parameterized* or *prioritized* QoS [2]. Generally, QoS is the ability of a network element (e.g. an application, a host or a router) to provide some levels of assurance for consistent network data delivery. *Parameterized QoS* is a strict QoS requirement that is expressed in terms of quantitative values, such as data rate, delay bound, and jitter bound. In a Traffic Specification (TSPEC), these values are expected to be met within the MAC data service in the transfer of data frames between peer STAs. *Prioritized QoS* is expressed in terms of relative delivery priority, which is to be used within the MAC data service in the transfer of data frames between peer STAs. In prioritized QoS scheme, the values of QoS parameters such as data rate, delay bound, and jitter bound, may vary in the transfer of data frames, without the need to reserve the required resources by negotiating the TSPEC between the STA and the AP. According to the definitions of QoS above, this section presents the QoS limitations of IEEE 802.11 MAC functions.

3.1 QoS limitations of DCF

DCF can only support best-effort services, not any QoS guarantees. Typically, time-bounded services such as Voice over IP, or audio/video conferencing require specified bandwidth, delay and jitter, but can tolerate some losses. However, in DCF mode, all the STAs in one BSS compete for the resources and channel with the same priorities. There is no differentiation mechanism to guarantee bandwidth, packet delay and jitter for high-priority STAs or multimedia flows. We have made the following simulation to evaluate the performance of DCF in ad-hoc mode using *ns*-2 [12]. The simulation topology is shown in Figure 5 and there is no mobility in the system. Each STA operates at IEEE 802.11a PHY mode-6 [23] and transmits three types of traffic (audio, video and background traffic) to each other. The packet size of audio is equal to 160 bytes and the inter-packet arrival interval is 20ms, which corresponds to 8KB/s PCM audio flow. The video sending rate is 80KB/s with a packet size equal to 1280 bytes. The sending rate of background traffic is 128 KB/s, using a 1600 bytes packet size. All traffic are CBR/UDP sources and the simulation parameters are summarized in Table 1. We vary the load rate from 9.6% to 90% by increasing the number of STAs from 2 to 18. Figure 6 shows the simulation results for the throughput and delay. We can see that average throughput of three kinds of flows per STA are almost stable when the channel load rate is less than 70% (i.e. the number of STAs is up to 10). For example, throughput of audio is about 7.8 KB/s; throughput of video is about 78KB/s; throughput of background is about 125KB/s; and delay is lower than 4ms. When the number of STAs is larger than 10, the throughput of all three traffic decreases very fast, e.g., the throughput is around 60% when the number of STAs is 18 (90% load rate). Moreover, the mean delays of the three flows increase up to 420ms and almost the same for the three flows. This simulation clearly shows that there is no throughput or delay differentiation between different flows since only one queue is shared by all the three flows, thus they all experience the same delay. So, there is no way to guarantee the QoS requirements for high-priority audio and video traffic unless admission control is used.

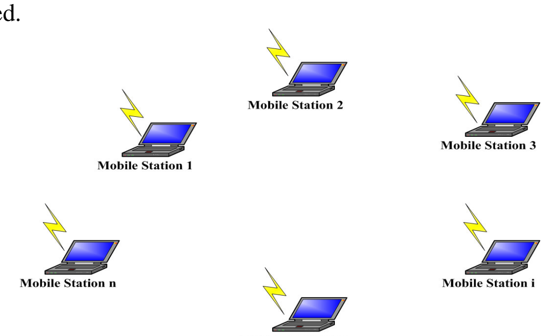
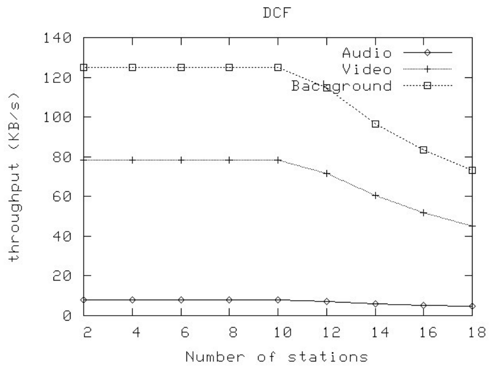


Fig. 5 Simulation topology of DCF in ad-hoc mode

SIFS	16µs	MAC header	28bytes
DIFS	34µs	PLCP header length	4µs
ACK size	14bytes	Preamble length	20µs
PHY rate	36Mbps	CWmin	15
slot time	9µs	CWmax	1023

Table 1 Simulation parameters for 802.11a mode 6



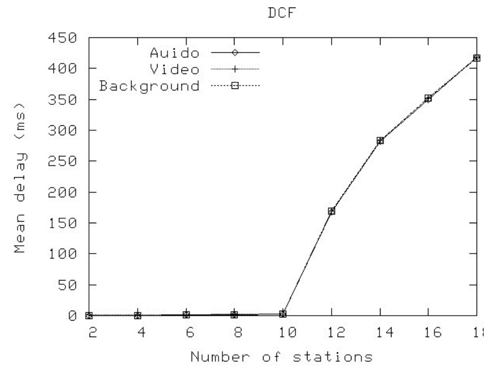


Fig. 6 Throughput and delay performance for DCF

3.2 QoS limitations of PCF

Although PCF has been designed to support time-bounded multimedia applications, this mode has three main problems that lead to poor QoS performances [1,6,7,34]:

First, the central polling scheme is questionable. All the communications between two STAs in the same BSS have to go through the AP, thus some channel bandwidth is wasted. When this kind of traffic increases, a lot of channel resources are wasted.

Second, the cooperation between CP and CFP modes may lead to unpredictable beacon delays [7,34]. The PC schedules the beacon at TBTT for the CFP interval, and then the beacon can be transmitted when the medium has been found idle for an interval of time longer than a PIFS. Hence, depending on whether the wireless medium is idle or busy around the TBTT, the beacon frame may be delayed. In the current 802.11 legacy standard, STAs are allowed to start their transmissions even if the frame transmission cannot terminate before the upcoming TBTT [1]. The duration of the beacon to be sent after the TBTT defers the transmission of time-bounded frames, which may severely impact the QoS performance in each CFP. In the worst case, the maximum delay for beacon frame can be 4.9ms in IEEE 802.11a [7], and the average beacon frame delay can reach up to 250μs [7].

Third, the transmission time of a polled STA is difficult to control. A polled STA is allowed to send a frame of any length between 0 and 2346 bytes, which introduce the varation of transmission time. Furthermore, the PHY rate of the polled STA can be changed according to the varying channel status, so the transmission time is hard to be predicted by the AP. This makes a barrier for the AP to provide guaranteed QoS service for other STAs in the polling list during the rest of the CFP.

All these limitations for both DCF and PCF led to a large number of research activities to enhance the performance of 802.11 MAC (as discussed in Section 4).

4. QoS enhancement schemes for 802.11 MAC

Normally, QoS issues in wired LAN are neglected since the physical layer bandwidth of wired LAN is high enough (1Gbps is now a common link speed between switches in enterprise LANs while 10Gbps 802.3ae Ethernet will appear soon). However, wireless LAN has some distinct features from wired LAN: high bit error rate, high delay and low bandwidth. The characteristics of the wireless channel make high data rate very difficult to achieve. IEEE 802.11 WLAN is originally designed for best-effort services. The error rate at physical layer is more than three orders of magnitude larger than that of wired LAN. Moreover, high collision rate and frequent retransmissions cause unpredictable delays and jitters, which degrade the quality of real-time voice and video transmission. Enhanced QoS-aware coordination can reduce overhead, prioritize frames, and prevent collisions to meet delay and jitter requirements in mobile environment.

Currently, there are two main architectural approaches to add QoS support in the Internet: integrated services (IntServ) [25] and differentiated services (DiffServ) [26]. IntServ provides fine-grained service guarantees to individual flows. It requires a module in every hop IP router along the path that reserves resources for each session. However, IntServ is not deployed since its requirement of setting states in all routers along a path is not scalable. On the contrary, DiffServ only provides a framework offering coarse-grained controls to aggregates of flows. DiffServ attempts to address the scaling issues associated with

IntServ by requiring state awareness only at the edge of DiffServ domains. At the edge, packets are classified into flows, and the flows are conditioned (marked, policed and possibly shaped) to a traffic conditioning specification (TCS). In this way, more simple and effective QoS support can be built from the components during early deployments, and Internet-wide QoS can evolve into a more sophisticated structure. But until now, DiffServ has not been widely deployed, mainly because it is difficult to map between different service domains or subnetworks such as 802.11 WLAN. The problems of both IntServ and DiffServ schemes led to the activities of Integrated Services over Specific Link Layers (ISSLL) Working Group at the IETF to provide IntServ over specific link technologies [35]. One of the key ideas is to provide IntServ QoS by using DiffServ network segments. This solution maintains the IntServ signalling, delay-based admission and the IntServ service definitions. The edge of the network consists of pure IntServ regions. However, the core of the network is considered as a DiffServ region, and all flows are mapped into one of the few DiffServ classes at the boundary. So, in order to support both kinds of IP QoS approaches in 802.11 WLAN links, different kinds of QoS enhancement schemes for both infrastructure and ad-hoc modes have been proposed for 802.11 WLAN. In this section, we classify and evaluate the performances of the main proposed schemes. Because the services required by multimedia applications are based on parameters such as bandwidth, delay, jitter and loss (or bit error) rate, we introduce bandwidth, delay and jitter based service differentiation in Section 4.1 and error control based enhancement schemes in Section 4.2.

4.1 Service (bandwidth, delay, jitter) differentiation based enhancement schemes

4.1.1 Classification of service differentiation based schemes

First of all, QoS enhancement can be supported by adding service differentiation into the MAC layer. This can be achieved by modifying the parameters that define how an STA or a flow should access the wireless medium. Current service differentiation based schemes can be classified with respect to a multitude of characteristics. For example, a possible classification criterion is whether the schemes base the differentiation on per-STA or per-queue (per-priority) parameters. Another classification depends on whether they are DCF-based (distributed control) or PCF-based (centralized control) enhancements. Figure 7 shows a classification in two levels. We distinguish between station-based schemes and queue-based schemes at the top-level and DCF-based versus PCF-based enhancement at the second level. Previous research works mainly focus on the station-based DCF enhancement schemes [3,9,10,11]. Other recent works mainly focus on queue-based hybrid coordination (combined PCF and DCF) enhancement schemes [2,4,7,13-18,41] since queue-based schemes perform more efficiently.

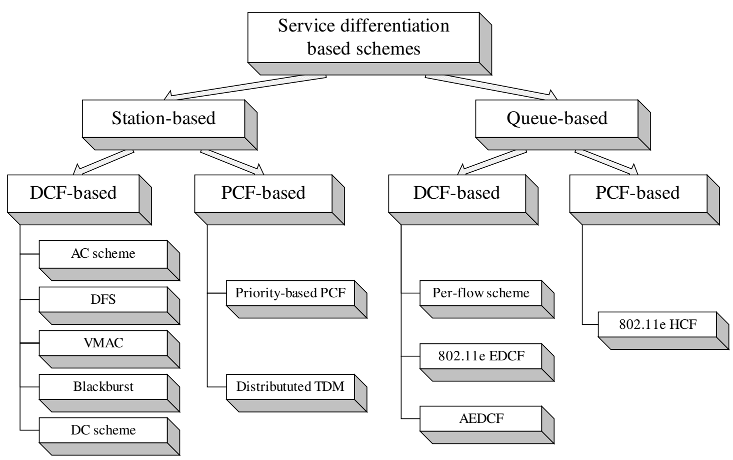


Fig. 7 Classification of service differentiation based schemes

4.1.2 Proposed schemes for service differentiation

(A) Station-based service differentiation using DCF enhancement

AC scheme: In this paper, we denote the scheme proposed in [3] the AC scheme. To introduce priorities for the IEEE 802.11 standard under the DCF access method, Aad and Castelluccia propose three techniques [3]:

(a) Different backoff increase function: Each priority level has a different backoff increment function. Assigning a short contention window to those higher priority STAs ensures that in most (although not all) cases, high-priority STAs are more likely to access the channel than low-priority ones. This method modifies the contention window of the priority level j after i transmission attempts as follows: $CW_{new} = P_j^{2+i} \cdot CW_{old}$, where P_j is a factor used to achieve service differentiation which has the higher value for lower priority STAs. Experiments show that this scheme performs well with UDP traffic but does not perform well with TCP traffic because all TCP ACKs are set the same priorities, which affect the differentiation mechanism.
(b) Different DIFS: Each STA has a different DIFS according to its priority level. In IEEE 802.11, ACK packets have higher priorities than data packets. An ACK packet is sent after sensing the medium for a time of SIFS, whereas the medium has to be sensed for a longer time (equal to DIFS) to send an data packet. The same idea is used to give priorities to data frames when using the DCF access scheme. In this approach, each priority level has a different DIFS, for example, $DIFS_{j+1} < DIFS_j$. So before transmitting a packet, the STAs having priority j+1 will wait for an idle period of length $DIFS_{j+1}$ slot time, which is shorter than that of STA with priority j. To avoid collision between frames with the same priority, the backoff mechanism is maintained in a way that the maximum contention window size added to $DIFS_j$ is $DIFS_{j-1}$-$DIFS_j$. This ensures that no STA of priority j has queued frames when STA of priority j-1 starts transmission.
The main problem of this scheme is that low priority traffic suffers as long as high priority frames are queued. Therefore, the maximum random range (RR_j) after $DIFS_j$ can be made greater than $DIFS_{j-1}$-$DIFS_j$, so the previous issue becomes less severe. In this case, a packet which failed to access the channel at the first attempt is likely to have its priority reduced after several consecutive attempts, depending on the DIFS and Random Range values. Experiments show that there is no backoff problem with TCP, but TCP ACKs also reduce the effects of service differentiation since all ACKs have the same priorities.
(c) Different maximum frame lengths: Each STA has a different maximum frame length according to its priority level, therefore, a high priority STA can transmit more information per medium access than a low priority STA. For the implementation, two possibilities should be distinguished: one is either to drop packets that exceed the maximum frame length assigned to a given STA (or simply configure it to limit its packet lengths), the other is to fragment packets that exceed the maximum frame length. This mechanism is used to increase both transmission reliability and differentiation, and works well for TCP and UDP flows. However, in a noisy environment, long packets are more likely to be corrupted than short ones, which decreases the service differentiation efficiency.

DFS scheme: In order to introduce both priority and fairness, Vaidya et al. [8] propose an access scheme called distributed fair scheduling (DFS) which utilizes the ideas of self-clocked fair queueing (SCFQ) [31] in the wireless domain. In DFS, the backoff process is always initiated before transmitting a frame. Different from 802.11 DCF, backoff interval is computed as a function of packet size and weight of the station, which can be linear, exponential or square-root function [8]. For example, in linear scheme, the backoff interval is set proportional to the packet size (l) and inversely proportional to the weight of the station (ϕ): $B = \lfloor \rho \cdot \lfloor Scaling\_factor \cdot l / \phi \rfloor \rfloor$, where *Scaling_factor* is used to scale the backoff interval to a suitable value and ρ is a uniform random variable in the interval [0.9,1.1]. This causes STAs with low weights to generate longer backoff intervals than those with high weights, thus getting lower priority. Fairness is achieved by considering the packet size in the calculation of the backoff interval, causing flows with smaller packet size to be sent more often. If a collision occurs, a new backoff interval is calculated using the original backoff algorithm of the IEEE 802.11 DCF. However, the implementation complexity of this scheme limits its deployment.
VMAC scheme: Based on DCF, Campbell et al. [11] propose a fully distributed service quality estimation, radio monitoring, and admission control approach to support service differentiation. A virtual MAC (VMAC) algorithm monitors the radio channel and estimates locally achievable service levels. The VMAC estimates MAC level statistics related to service quality such as delay, jitter, packet collision, and packet loss. The VMAC algorithm operates in parallel to the MAC in the mobile host but does not handle real packet transmission like in MAC. This is why it is called virtual MAC. The advantage of virtual MAC is that it can estimate higher order statistics than first-order performance statistics without too much overheads. By this way, more sophisticated analysis and traffic control methods can be applied. Moreover, a virtual source (VS) algorithm can utilize the VMAC to estimate application-level service quality. The VS allows application parameters to be tuned in response to dynamic channel conditions based on "virtual delay curves". The goal of the VMAC is to estimate QoS parameters in the radio channel accurately since relative service differentiation is not enough for real-time services. Moreover, this

scheme uses the following backoff timer differentiation: $CW_{\min}^{high\_pri} < CW_{\min}^{low\_pri}$, $CW_{\max}^{high\_pri} < CW_{\max}^{low\_pri}$. Simulation results show that: (a) When these distributed virtual algorithms are applied to the admission control of the radio channel, then a globally stable state can be maintained without the need for complex centralized radio resource management. (b) Delay differentiation can be increased by increasing the gap between $CW_{\min}^{high\_pri}$ and $CW_{\min}^{low\_pri}$, i.e., decreasing $CW_{\min}^{high\_pri}$ and increasing $CW_{\min}^{low\_pri}$ provide high priority traffic lower delay than before, and low priority traffic higher delay than before [11]. However, one drawback of the VMAC scheme is that the interactions between application and MAC layers introduce complexities.

Blackburst scheme: Sobriho and Krishnakumar propose the Blackburst scheme in [10]. The main goal of Blackburst is to minimize the delay of real-time traffic. Unlike other schemes, it imposes certain requirements on high priority STAs: 1) All high priority STAs try to access the medium with equal and constant intervals, t_{sch}, and 2) The ability to jam the medium for a period of time. When a high priority STA wants to send a frame, it senses the medium to verify if it has been idle for an interval of time PIFS and then sends its frame. If the medium is busy, the STA waits for the medium to be idle for a PIFS and then enters a black burst contention period: the STA sends a so-called black burst to jam the channel. The length of the black burst is determined by the time the STA has waited to access the medium, and is calculated as a number of *black slots*. After transmitting the black burst, the STA listens to the medium for a short period of time (less than a black slot) to see if some other STA is sending a longer black burst which would imply that the other STA has waited longer and thus should access the medium first. If the medium is idle, the STA will send its frame, otherwise it will wait until the medium becomes idle again and enters another black burst contention period. After the successful transmission of a frame, the STA schedules the next transmission attempt t_{sch} seconds in the future. This has the nice effect that real-time flows will synchronize, and share the medium in a time division multiple access (TDMA) fashion [10]. In Blackburst scheme, low priority STAs use the ordinary CSMA/CA access method of IEEE 802.11. This means that unless some low priority traffic comes and disturbs the order, very few blackburst contention periods will have to be initiated once the STAs have been synchronized. Simulation results show that Blackburst can support more real-time nodes than CSMA/CA, with stable data and real-time traffic operation, due to the absence of collisions. From the delay point of view, the Blackburst offers very low delay and jitter, even at high traffic load. The main drawback of Blackburst is that it requires constant access intervals for high-priority traffic, otherwise the performance degrades considerably.

DC scheme: Deng and Chang propose a service differentiation scheme [9], which requires minimal modifications of the basic 802.11 DCF. In this paper, we denote the scheme proposed in [9] as the DC scheme. The DC scheme uses two parameters of IEEE 802.11 MAC, the backoff interval and IFS between each data transmission, to provide the differentiation. Thus, the backoff time is divided into two parts and each interval of time is combined with two different IFS lengths PIFS and DIFS, as shown in Table 2.

Priority	IFS	Backoff algorithm
0	DIFS	$B = 2^{2+i}/2 + \lfloor rd \times 2^{2+i}/2 \rfloor$
1	DIFS	$B = \lfloor rd \times 2^{2+i}/2 \rfloor$
2	PIFS	$B = 2^{2+i}/2 + \lfloor rd \times 2^{2+i}/2 \rfloor$
3	PIFS	$B = \lfloor rd \times 2^{2+i}/2 \rfloor$

Table 2 DC scheme's priority classes
(rd is a uniform random variable in (0,1), and $\lfloor x \rfloor$ represents the largest integer less than or equal to x.)

As mentioned in Table 2, four classes of priorities can be supported. An STA that uses PIFS gets higher priority than an STA using DIFS. Using the DC scheme, higher priority STAs have shorter waiting time when accessing the medium. Moreover, when a collision occurs, higher priority STAs could have more chances to access the medium than the lower priority ones. On the other hand, when there are no high priority STAs which want to transmit packets, the low priority ones still generate a long backoff time. Thus, an additional delay is imposed by long backoff time.

Table 3 summarizes the comparison of different station-based schemes using DCF enhancement:

MAC scheme	Main features	Strength	Weakness
AC scheme	Differentiation is based on different backoff increase function, DIFS and maximum frame length	Good service differentiation is achieved	This scheme works very well with UDP traffic, but does not perform well with TCP traffic
DFS scheme	A fair scheduling algoritm is defined according to packet size and flow weight	Fairness is achieved. The performance of high priority flows is enhanced	The implementation complexity of this scheme is high
VMAC scheme	A virtual MAC is introduced to estimate delay, jitter, packet collisions and packet losses, then VS tunes the application parameters based on estimations	Channel conditions are taken into consideration. “Virtual delay curves” can be used by the applications to tune their parameters	Interactions between application and MAC layers introduce complexities
Blackburst scheme	A black burst contention period is used. It indicates the time that the station has waited to access the medium	Delay of high real time traffic is minimized. Very nice synchronization between high prioprity flows is achieved when there are no low priority ones	If the two requirements on high priority flows can not be supported, the performance will degrade
DC Scheme	The backoff interval is divided into two parts. Then each part is combined with two IFSs yielding four priorities	The service differentiation is achieved that ensures a good performance for high priority traffic	Starvation of low priority stations when there are no high priority ones because they generate a long backoff

Table 3 Comparison of station-based service differentiation schemes using DCF enhancement

(B) Station-based service differentiation using PCF enhancement

Priority-based PCF [6, 33]: Since PCF is optional in 802.11 standard, few research works extend station-based PCF to support service differentiation. However, PCF can be used to provide service differentiation support using a priority-based polling scheme instead of the default round-robin polling algorithm. Indeed, the AP sends priority-based polling packets to a succession of STAs in the wireless BSA, which can give STAs different priorities.

Distributed TDMA: This mechanism does not modify the polling scheme of PCF, but rather sets up TDMA-like slot time periods, and specifies which STA gets which slot time to provide differentiation. Once the slot time has been assigned, each STA knows when it can transmit, and packet transmissions can take place with very little intervention from the AP (in contrast with PCF, where the AP has to use its polling capability to direct the transfer of every frame to be sent).

(C) Queue-based service differentiation using DCF enhancement

Per-flow scheme [4]: The motivation to use queue-based differentiation scheme comes from the following observations: (1) In station-based schemes, when several TCP senders with different priorities share the same receiver, they all receive the TCP-ACKs with the same priority (limited to the same receiver priority). This tends to reduce the differentiation effect. Furthermore, if the shared receiver is slow, the observed relative priority is also reduced. (2) Moreover, differentiation effect may also be reduced if one sender sends two flows to two receivers. Since there is a multi-path fading effect in wireless channel, one receiver may stay in a good channel condition (e.g. low error rate), another receiver may stay in a bad channel condition (e.g. high error rate), which causes both receiers' frames to wait long in the sender's queue. These two issues motivate the use of per-flow and per-queue differentiation where the shared node uses different priorities for different flows. The authors in [4] introduce a per-flow differentiation, and all packets are put in the same queue, independent of their priorities. But this scheme introduces mutual interferences between priorities: when the AP serves a low priority and slow flow, the global speed and efficiency of AP depends on the occupation time of the slow flow. If most of the time the flow occupies the AP, even if there are other high-priority fast flows, the AP has to be slow, and service differentiation gets lower. A possible solution is to assign different queues to different flows in the AP. Simulations [4] show that with this solution there is a total independence between flows: even if a low priority flow passes through the AP, it does not slow down the AP (the shared node), and differentiation effect is much better than one-queue per-flow scheme. Note that when using this approach with CW_{min} differentiation, the collision avoidance at the shared node (e.g. the AP) is lower than that of wireless STAs. In fact, when a single queue per MAC sub-layer is used, we just have one packet per STA contending to access the channel. However, when an STA has several queues for multiple TCP connections, there are multiple packets per CW period. Internal collisions in one STA will increase when there is an increase in the number of connections. This tends to use the admission control and transmission opportunity (TXOP) scheduler in EDCF scheme [2].

EDCF [2]: To introduce better differentiation performance than per-flow scheme, IEEE 802.11e EDCF extends the basic DCF to support up to four EDCF queues in one STA and each queue contends for Transmission Opportunity (TXOP) in one STA to send the packets. We will discuss EDCF in detail in section 5.

AEDCF [41]: One problem of the basic EDCF ad-hoc mode is that the values of CW_{min}, CW_{max} and backoff function of each queue are static and does not take into account dynamicity of wireless channel conditions [2]. In adaptive EDCF (AEDCF) scheme, relative priorities are provisioned by adjusting the size of the contention window of each traffic class taking into account both application requirements and network conditions. After each successful transmission, AEDCF does not reset the contention window to CW_{min}. Instead, the scheme takes into account the estimated collision rate in each STA noted by f_{curr}^{j}. A multiplicative factor for each class i is introduced by: $MF[i] = \min((1+2i)\cdot f_{avg}^{j}, 0.8)$, where $f_{avg}^{j} = (1-\alpha) f_{curr}^{j} + \alpha f_{avg}^{j-1}$, α is a factor to smooth the estimation (α = *0.8*). Then, the contention window is updated as follows: $CW_{new}[i] = \max(CW_{\min}[i], CW_{old}[i]\cdot MF[i])$. After each collision, a persistence factor *PF*[i] is introduced in AEDCF for further differentiation. Performance comparisons between AEDCF and 802.11e EDCF scheme show that AEDCF outperforms the EDCF, especially at high traffic load conditions: AEDCF increases the medium utilization ratio and reduces more than 50% of the collision rate. While achieving delay differentiation, the overall goodput obtained is up to 25% higher than EDCF.

(D) Queue-based service differentiation using PCF enhancement

HCF [2, 18]: Hybrid Coordination Function (HCF) is a queue-based service differentiation scheme proposed by IEEE 802.11e working group using both PCF and DCF enhancements. It combines the advantages of distributed contention access (EDCF) and centralized polling access (PCF) methods. HCF uses QoS-enhanced access point (QAP) as a traffic director for different queues. HCF is discussed in detail in Section 5.

4.2 Error control based enhancement schemes

In parallel, QoS enhancement can also be obtained by error control enhancements. In the Internet architecture, the end-to-end reliability should be entirely provided by the end nodes. The Internet may occasionally drop, corrupt, duplicate or reorder packets. So, the transport protocol (e.g., TCP) or the application itself (e.g., if UDP is used as the transport protocol) must recover from these errors on an end-to-end basis. Error recovery in the subnetwork is justified only to the extent that it can enhance overall performance. However, some subnetworks like wireless links require link layer error recovery mechanisms to enhance the performance. These enhancements should be lightweight. For example, wireless links normally require link-layer error recovery (such as 802.2 LLC) and MAC-level error recovery in subnetwork. There are two basic categories of error recovery schemes: ARQ (Automatic Repeat reQuest) [19,21-22,27-28] and FEC (Forward Error Correction) [20,40]. Which are described in the following subsection.

4.2.1 Automatic Repeat reQuest (ARQ)

ARQ is an error control protocol, which is mostly implemented at both link and transport layers. It is efficient on a high speed WLAN link when round trip delay is small. But it may cause a large delay when a lot of retransmissions are introduced on a slow link.

(A) Stop and Wait ARQ (SW-ARQ) [20]

SW-ARQ is a simple and very efficient technique for data communications. Basically, in SW-ARQ, a sender transmits a single packet and then waits for the response. The receiver sends an Acknowledgement (ACK) for each packet correctly received. If there is no response after a time out, the sender retransmits the packet. Under normal transmission, the sender receives an ACK for the data and then starts transmission of the following data packet. The sender may have to wait a considerable time for this response. While it is waiting, the sender is unable to send another packet. In fact, the current MAC mechanism of IEEE 802.11 WLAN uses this error control mechanism because it is more efficient and simpler than FEC [28,1].

(B) Selective Repeat ARQ (SR-ARQ) [21]

Unlike SW-ARQ, when using SR-ARQ, packets are transmitted continuously by the Data Link Control (DLC) layer. The receiver acknowledges each successfully received packet. If the acknowledgment is not received for a packet after the expiration of a timeout, the packet is retransmitted. Once a packet has been retransmitted, the sender resumes transmission of packets from where it left off, i.e., if j is the packet with the largest sequence number that has been transmitted, packet with sequence number j+1 is transmitted next (assuming that no other timers have expired in the meantime). Note that with SW-ARQ mechanism, a lot of idle time is wasted for waiting the sequenced ACKs. When the SR-ARQ protocol is used, packets are continuously transmitted, which removes the idle time associated with SW-ARQ. In fact, when SR-ARQ is used, packets can be accepted out of sequence. Hence, packets received out of sequence have to be buffered and re-sequenced before they can be delivered to the application layer. Indeed, SR-ARQ is the most efficient scheme for saving end-to-end delay. However, it is a very complex error recovery mechanism.

(C) Go-Back-N ARQ (GBN-ARQ) [21]

When GBN-ARQ is used, packets are transmitted continuously as in SR-ARQ. However, the receiver accepts packets only in the order in which they have been transmitted. Packets received out of sequence are discarded and not acknowledged. Since the receiver accepts packets only in-sequence, the sender retransmits the packet that timed out and also all the following packets. Hence, each time a timeout occurs, all the packets that have not been acknowledged are retransmitted. It is important to observe that GBN-ARQ attempts to combine the desirable features of SR-ARQ and SW-ARQ, i.e., packets are transmitted continuously, as in SR-ARQ scheme, without the need to buffer out of sequence packets and there is no re-sequencing overhead.

4.2.2 Forward Error Correction (FEC)

FEC involves addition of redundant bits that help to recover erroneous bits. It has been suggested for real-time applications due to the strict delay requirements and semi-reliable nature of media streams. However, FEC incurs constant transmission overhead even when the channel is error free [20,40].

With ARQ, the receiver requests retransmission when it detects an error, but ARQ leads to variable delays that are unacceptable for real-time services. FEC schemes help maintaining homogeneous throughput and bounded time delay. However, the decoding error rate of FEC increases rapidly with the increase of channel error rate. So, when channel error rate is high, a long FEC code is necessary. This makes the coder-decoder pair complex and also imposes a high transmission overhead. Furthermore, the wireless channel is non-stationary and the channel bit error rate varies over time. Only FEC or ARQ is not very efficient for high-speed error-prone WLAN channel. In order to overcome their individual drawbacks, hybrid FEC-ARQ schemes have been developed [19,22,27,29]. In the next subsection we present two kinds of error control mechanisms that combine FEC with ARQ: Type-I and Type-II Hybrid FEC-ARQ schemes.

4.2.3 Hybrid FEC-ARQ

(A) Type-I Hybrid FEC-ARQ

The Type-I Hybrid FEC-ARQ uses parity bits for both error detection and error correction in every packet. If the number of erroneous bits in a received packet is within the error correction capability of the code, the errors are corrected. If an uncorrectable error pattern is detected, the packet is rejected and a retransmission is requested. In Type-I Hybrid ARQ schemes, information can be recovered from each transmitted packet. The sender retransmits the same codeword. When the retransmitted codeword is received, the decoder attempts to correct the errors within the error capability of the code. If the packet arrives with detectable and uncorrectable errors, the receiver discards the received codeword and a retransmission is requested again. This process continues until the packet is successfully received or the maximum number of retransmissions has been reached. A disadvantage of Type-I hybrid ARQ schemes is that the uncorrectable packets are discarded by the decoder even if they might contain some useful information.

(B) Type-II Hybrid FEC-ARQ

In Type-II Hybrid FEC-ARQ, the uncorrectable packet is kept for future use instead of being discarded [19,22]. In case the packet cannot be successfully decoded at the destination, the receiver requests for a retransmission and uses the saved packet to help the decoding processor to correct the detected errors. This process is repeated until the packet can be successfully decoded.

SW-ARQ is the current mechanism used in the 802.11 MAC layer because it is simple and easy to be deployed. MAC-level FEC has been introduced in the early version of IEEE 802.11e draft and has been finally removed because of considerable overheads. The other proposed techniques such as Type-I Hybrid FEC-ARQ and Type-II Hybrid FEC-ARQ schemes are able to reduce delay for high priority traffic. They are planned to be used in the next-generation high-throughput wireless networks [43]. But their implementation is relatively complex.

5. Upcoming IEEE 802.11e QoS enhancement standard

There are many new features in the 802.11e draft 4.2 [2]. In this section, we will briefly describe three of them: Hybrid Coordination Function (HCF), Direct Link Protocol (DLP) and block acknowledgement.

5.1 HCF: Hybrid Coordination Function

In order to support both IntServ and DiffServ QoS approaches in 802.11 WLAN, 802.11e has defined a new mechanism, namely, Hybrid Coordination Function (HCF). HCF is composed of two access methods: contention-based channel access (also called EDCF) and controlled channel access mechanisms, which are respectively detailed in sub-sections 5.1.1 and 5.1.2.

One main feature of HCF is to introduce four access category (AC) queues and eight traffic stream (TS) queues at MAC layer. When a frame arrives at MAC layer, it is tagged with a traffic priority identifier (TID) according to its QoS requirement, which can take the values from 0 to 15. The frames with TID values from 0 to 7 are mapped into four AC queues using EDCF access rule. On the other hand, frames with TID values from 8 to 15 are mapped into eight TS queues using HCF controlled channel access rule. The reason of separating TS queues from AC queues is to support strict parameterized QoS at TS queues while prioritized QoS is supported at AC queues.

Another main feature of the HCF is the concept of transmission opportunity (TXOP), which is the time interval permitted for a particular STA to transmit packets. During the TXOP, there can be a series of frames transmitted by an STA separated by SIFS. The TXOP is called either *EDCF-TXOP*, when it is obtained by winning a successful EDCF contention; or *polled-TXOP*, when it is obtained by receiving a QoS CF-poll frame from the QoS-enhanced AP (QAP). The maximum value of TXOP is called *TXOPLimit*, which is determined by the QAP.

5.1.1 Enhanced Distributed Coordination Function (EDCF)

The EDCF is designed for the contention-based prioritized QoS support. Figure 8 shows that in EDCF, each QoS-enhanced STA (QSTA) has 4 queues (ACs), to support 8 user priorities (UPs) as defined in IEEE 802.1D [39]. Therefore, one or more UPs are mapped to the same AC queue, see Table 4. This comes from the observation that usually eight kinds of applications do not transmit frames simultaneously, and using less ACs than UPs reduces the MAC layer overheads. Each AC queue works as an independent DCF STA and uses its own backoff parameters.

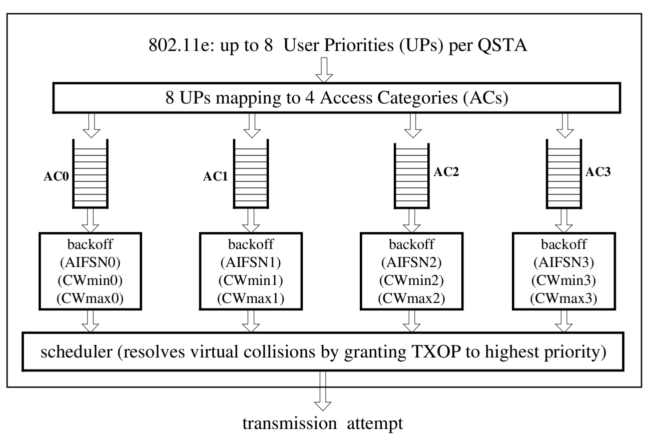


Fig. 8 EDCF proposed by 802.11e

UP, User Priority (Same as 802.1D)	802.1D Designation	802.11e AC (Access Category)	Service type
2	Not defined	0	Best Effort
1	Background (BK)	0	Best Effort
0	Best Effort (BE)	0	Best Effort
3	Excellent Effort (EE)	1	Video Probe
4	Controlled Load (CL)	2	Video
5	VI (Video <100ms latency and jitter)	2	Video
6	VO (Video <10ms latency and jitter)	3	Voice
7	Network Control (NC)	3	Voice

Table 4. Mapping between User Priority (UP) and Access Category (AC)

In EDCF, two main methods are introduced to support service differentiation:

The first one is to use different InterFrame Space (IFS) sizes for different ACs. Figure 9 shows the detailed timing diagram of the EDCF scheme. A new kind of IFS called Arbitration IFS (AIFS) is used in EDCF, instead of DIFS in DCF. The AIFS [AC] is determined by

$$\text{AIFS [AC]} = \text{AIFSN [AC]} \cdot \text{SlotTime} + \text{SIFS},$$

where the default value of the arbitration inter frame spacing number (AIFSN) is defined as either 1 or 2 [2]. When AIFSN = 1, high priority queues AC1, AC2 and AC3 have AIFS value equal to PIFS. When AIFSN = 2, the low priority queue AC0 has AIFS value of DIFS. When a frame arrives at an empty AC queue and the medium has been idle longer than AIFS [AC]+SlotTime, the frame is transmitted immediately. If the channel is busy, the arriving packet in each AC has to wait until the medium becomes idle and then defer for AIFS+SlotTime. So the AC with the smaller AIFS has the higher priority. For example, the earliest transmission time for high priority queue is to wait for PIFS+SlotTime = DIFS, while the earliest transmission time for best effort queue is to wait for DIFS + SlotTime.

The second method consists in allocating different CW sizes for different ACs. Assigning a short CW size to a high priority AC ensures that in most cases, high-priority AC is able to transmit packets ahead of low-priority one. If the backoff counters of two or more parallel ACs in one QSTA reach zero at the same time, a scheduler inside the QSTA will avoid the virtual collision by granting the EDCF-TXOP to the highest priority AC. At the same time, the other colliding ACs will enter a backoff process and double the CW sizes as if there is an external collision. In this way, EDCF is supposed to improve the performance of DCF under congested conditions. However, our simulation results show that although internal collision rates are low for EDCF, external collisions between the same priorities in different QSTAs are still high [41].

The default values of AIFSN [AC], CWmin [AC], CWmax [AC] and TXOPLimit [AC] are announced by the QAP in beacon frames, and the 802.11e standard also allows the QAP to adapt these parameters dynamically depending on network conditions [2]. But how to adapt to the channel has not been defined by the standard and remains an open research issue.

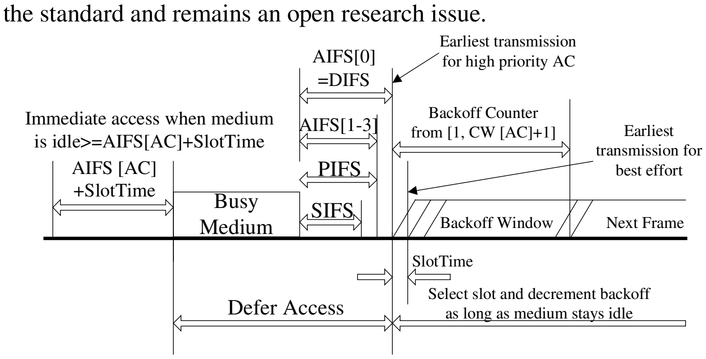


Fig. 9 EDCF channel access IFS Relationships

To improve the throughput performance, EDCF packet bursting can be used in 802.11e [2], meaning that once a QSTA has gained an EDCF-TXOP, it can be allowed to send more than one frame without contending for the medium again. After getting access to the medium, the QSTA can send multiple frames as long as the total access time does not exceed the TXOPLimit bound determined by QAP. To ensure that no other QSTA interrupts the packet bursting, SIFS is used between packet bursts. If a collision occurs, the EDCF bursting is terminated. This mechanism can reduce the network overhead and increase throughput by multiple transmissions using SIFS and burst acknowledgements. However, bursting may also increase the jitter, so TXOPLimit should not be longer than the time required for the transmission of the largest data frame.

5.1.2 HCF controlled channel access

The HCF controlled channel access mechanism is designed for the parameterized QoS support, which combines the advantages of PCF and DCF. HCF can start the controlled channel access mechanism in both CFP and CP intervals, whereas PCF is only allowed in CFP. Figure 10 is an example of typical 802.11e beacon interval, which is composed of alternated modes of optional CFP and CP. During the CP, a new contention-free period named controlled access phase (CAP) is introduced. CAPs are several intervals during which frames are transmitted using HCF-controlled channel access mechanisms. HCF can start a CAP by sending downlink QoS-frames or QoS CF-Poll frames to allocate polled-TXOP to different QSTAs after the medium remains idle for at least PIFS interval. Then the remaining time of the CP can be used by EDCF. This flexible contention-free scheme makes PCF and CFP useless and thus optional in the 802.11e standard. By using CAP, the HCF beacon interval size can be independent of targeted delay bounds of multimedia applications. For example, in order to support audio traffic with a maximum latency of 20 millisecond (ms) using PCF, the beacon interval should be no more than 20 ms since the fixed portion of CP forces the audio traffic to wait for the next poll. On the other hand, the HCF controlled channel access can increase the polling frequency by initiating CAP at any time, thus guarantee the delay bound with any size of beacon interval. So there is no need to reduce the beacon interval size that increases the overheads. Moreover, the problem of beacon delay in PCF is solved, because in HCF, a QSTA is not allowed to transmit a frame if the transmission cannot be finished before the next TBTT.

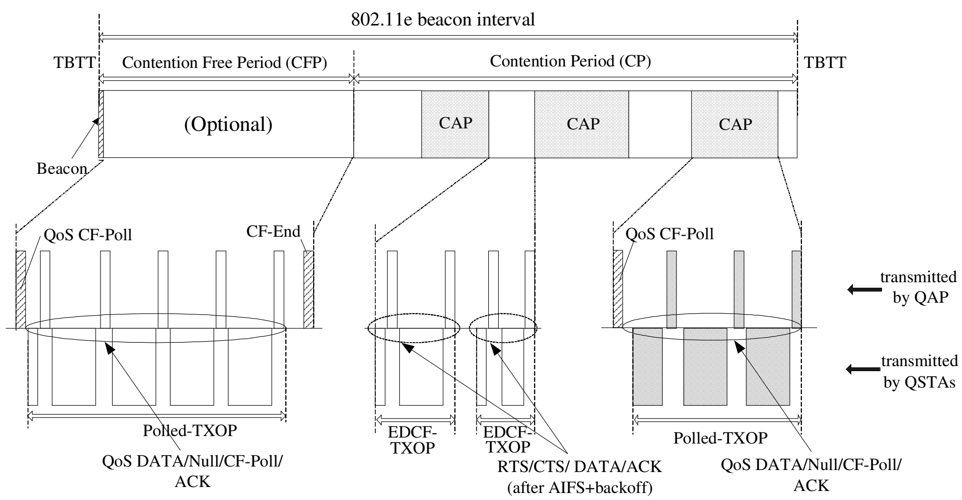


Fig.10 A typical 802.11e HCF beacon interval

In HCF controlled channel access mechanism, QoS guarantee is based on the traffic specification (TSPEC) negotiation between the QAP and the QSTA(s). Before transmitting any frame that requires the parameterized QoS, a virtual connection called traffic stream (TS) is established. In order to set up a TS, a set of TSPEC parameters (such as mean data rate, nominal frame size, maximum service interval, delay bound, etc.) are exchanged between the QAP and the corresponding QSTA(s). Based on these TSPEC parameters, the QAP scheduler computes the duration of polled-TXOP for each QSTA, and allocates the polled-TXOP to each QSTA. Then the scheduler in each QSTA allocates the TXOP for different TS queue according to the priority order. A simple round-robin scheduler is proposed in the IEEE 802.11e draft 4.2 [2]. The simple scheduler uses the following mandatory TSPEC parameters: mean data rate, nominal MAC frame size and maximum service interval or delay bound. Note that the maximum service interval requirement of each TS corresponds to the maximum time interval between the start of two successive TXOPs. If this value is small, it can provide low delay but introduce more CF-Poll frames. If different TS has different maximum service interval requirements, the scheduler will select the minimum value of all maximum service interval requests of all admitted streams for scheduling. Moreover, the QAP is allowed to use an admission control algorithm to determine whether or not to allow new TS into its BSS. When a TS is set up, the QAP attempts to provide QoS by allocating the required bandwidth to the TS. During a CFP, the medium is fully controlled by the QAP. During a CP, it can also grab the medium whenever it wants (after a PIFS idle time). After receiving a QoS CF-poll frame, a polled QSTA is allowed to transmit multiple MAC frames denoted by contention-free burst (CFB), with the total access time not exceeding the TXOPLimit. All the other QSTAs set their NAVs with the TXOPLimit plus a slot time. By this way, they will not contend for the medium during that period. If there are no frames to be sent to the QAP, the polled QSTA will send a QoS-Null frame to the QAP which can poll another QSTA.

5.2 Direct link protocol

The legacy standard [1] does not allow an STA to directly transmit frames to another STA within the infrastructure BSS: All the communications between two STAs have to go through the AP. On the other hand, in the IEEE 802.11e, direct link protocol (DLP) is introduced for QSTAs to setup the direct communication between each other in the infrastructure mode [2], which can significantly increase the bandwidth. With DLP, the sender first sends a direct link request message including its supported rates and some other information to the receiver through the QAP. Once the receiver acknowledges the request, the direct link between two QSTAs is established. When there are no frame transmissions between two QSTAs for duration of DLPIdleTimeout, the direct link is disabled. In this case, frames between two QSTAs will be sent via the QAP again. However, with DLP, the traffic between two QSTAs are not buffered at QAP for forwarding, which may wake up QSTAs in power-saving mode frequently and reduce the efficiency of power-saving when DLP is not used.

5.3 Block acknowledgement

The legacy MAC is based on the simple SW-ARQ scheme. This involves a lot of overheads due to the immediate transmissions of ACKs. In 802.11e, a new SR-ARQ mechanism named block acknowledgement (BlockAck) is introduced. In this mechanism, a group of data frames can be transmitted one by one with SIFS interval between them. Then, a single BlockAck frame is sent back to the sender to inform it how many packets have been received correctly. Obviously, this scheme can improve the channel efficiency. There are two kinds of BlockAck mechanisms used in 802.11e: *immediate* and *delayed*. In case of immediate BlockAck, the sender transmits a BlockAck-request frame after transmitting a group of data frames; the receiver has to send back the BlockAck after a SIFS interval. If the sender receives the BlockAck frame, it retransmits frames that are not acknowledged in the BlockAck frame, either in another group or individually. Immediate BlockAck is very useful for applications that require high-bandwidth and low-latency. But it is very difficult for implementations to generate the BlockAck in SIFS interval. On the other hand, the delayed BlockAck does not require the strict timing limit. In delayed BlockAck mechanism, the receiver is allowed to send a normal ACK frame first to acknowledge the BlockAck-request. Then the receiver can send back the BlockAck at any other time less than a delayed BlockAckTimeout. Delayed BlockAck scheme is useful for applications that can tolerate moderate latency. If the sender does not receive the BlockAck or ACK frame from the receiver, it will retransmit the BlockAck-request frame. When the maximum BlockAck-request retransmission limit number is reached, the whole group of data frames will be deleted.

6. Simulation-based evaluations of QoS-enhanced schemes

In [6], different simulations have been conducted to compare the performances of an early version of EDCF [14], Blackburst, PCF, and DFS schemes. In their simulations, each STA has its own priority and generates one kind of traffic type. Blackburst shows to be the best choice for high-priority traffic; on the other hand, it starves the low-priority traffic in case of high load. EDCF has the similar performance as Blackburst, but leads to higher collision rates compared with other schemes. With DFS, the performance is not so satisfying as the former two schemes for high-priority traffic, but the good point is that DFS does not starve the low-priority traffic. The simulation results [6] show the worst performance for high-priority traffic are obtained with the PCF scheme, but that PCF does not starve the low-priority traffic.

To evaluate the performance of the 802.11e EDCF scheme [2], we take the same simulation topology as used by DCF in Section 3.1, but the difference is that we separate the three flows (audio, video and background) into three queues. The EDCF MAC parameters for three queues are summarized in Table 5. Figure 11 shows the throughput and delay performances for EDCF. As compared with DCF in Figure 6, EDCF can support service differentiation for different types of flows. In Figure 11, the throughput of EDCF audio, video traffic is stable when the number of STAs is less than or equal to 16 (76% load rate). However, the throughput of background traffic decreases dramatically when the number of STAs is larger than 10 (48% load rate) and almost drops to 10KB/s when the number of STAs is 18 (90% load rate). This means that EDCF maintains the throughput of high-priority audio and video flows by punishing the background traffic. Furthermore, when the channel is 90% loaded, the throughput of audio and video start to decrease, which means that admission control for audio and video is required during very high load. On the other hand, the mean delay of background flows increases very fast when the number of STAs is more than 10 (48% traffic load) and up to 4.5s in case of 90% traffic load. When traffic load is less than or equal to 76%, the mean delays of audio and video flows keep low. One interesting remark, which can be seen in Figure 12, is that the total throughput of EDCF is lower than that of DCF when the traffic load is larger than 48%: EDCF reduces the throughput of low-priority flows considerably and therefore results in decreasing the total throughput.

Parameters	Audio	Video	Background
CW_{min}	7	15	31
CW_{max}	15	31	1023
AIFSN	1	1	2
Packet Size (bytes)	160	1280	1600
Packet Interval (ms)	20	16	12.5
Sending Rate (KB/s)	8	80	128

Table 5. EDCF parameters for three queues

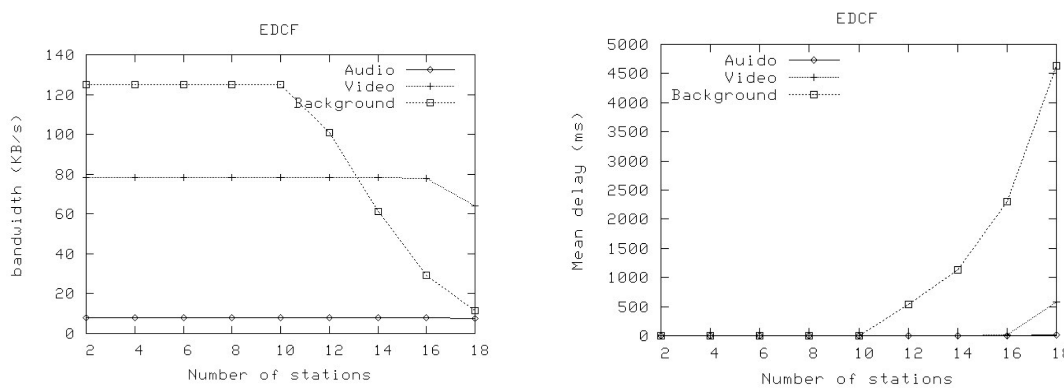


Fig. 11 Throughput and delay performance for EDCF

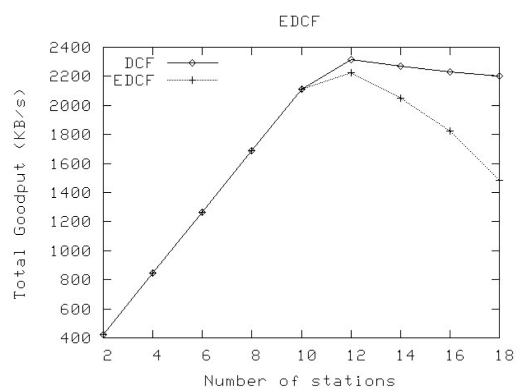


Fig. 12 Comparison of total goodput between EDCF and DCF

In order to compare HCF controlled channel access mechanism with EDCF, we run the following simulations: 6 QSTAs send audio (8KB/s On-Off traffic), high-priority CBR (Constant Bit Rate) and low-priority CBR video traffic to QAP at the same time. The sending rate of high-priority CBR is set to 25.4KB/s with a packet size of 660bytes and an interarrival time of 26ms. We vary the channel load rate by increasing the packet size of low-priority CBR video from 900bytes (0.3MB/s) to 1500bytes (0.5MB/s). The PHY and MAC parameters are the same as those in Table 1 and 5. In this simulation, we use the simple HCF scheduler explained in Section 5.1.2. For audio traffic, peak sending rate (8KB/s) and maximum service interval (50ms) are selected as QoS requirement; while for CBR traffic, constant sending rate and maximum service interval (100ms) are selected as QoS requirement. We map the high-priority and low-priority CBR video to different queues with the same EDCF parameters. Figure 13 shows the mean delays of audio, and low-priority CBR video versus the channel load rate. We can see that EDCF provides very low delays for audio traffic. When we increase the load rate to 80% by increasing the sending rate of CBR video, the mean delay of audio still keeps low and the mean delay of CBR increases to about 185ms. While in HCF controlled channel access scheme, the delay keeps at almost 20ms for both audio and low-priority video. The simulation results show that the HCF controlled channel access mechanism can guarantee the minimum delay requirement (50ms) for all the admitted flows in different load rate. On the other hand, EDCF works very well under low load conditions but suffers from delay degradation in high-load condition.

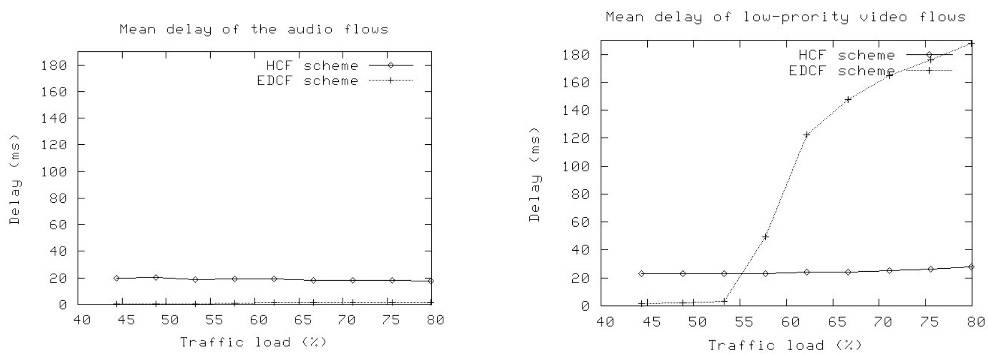


Fig. 13 Mean delay of audio, CBR video versus channel load for EDCF and HCF

7. Conclusions and future research areas

This survey analyzes the QoS limitations of the original IEEE 802.11 wireless LAN MAC layer. We evaluate and classify different QoS enhancement techniques proposed for IEEE 802.11 wireless LAN and study their advantages and drawbacks. Research activities and performance evaluations of the upcoming IEEE 802.11e QoS enhancement standard are also introduced and analyzed. As described, many QoS enhancement schemes have been proposed to improve the performance of original 802.11 wireless LAN. Among them the upcoming queue-based 802.11e standard offers some improvements. But it has not been finalized yet and needs to be analyzed more. There are still many research topics and open issues for QoS enhancement in IEEE 802.11 WLAN, among them we cite:

- Adapt the parameters to the traffic load and channel condition efficiently in ad-hoc EDCF mode,
- Optimize the tradeoff between channel efficiency, priority and fairness,
- Evaluate the efficiency and performance of EDCF packet bursting and Contention-Free Burst (CFB),
- Compare different scheduling schemes for HCF controlled channel access mechanism,
- Map between IP DiffServ (AF, EF), IntServ priorities and IEEE 802.11e MAC priorities,
- Standardize good 802.11e simulation models and tools,
- Evaluate IEEE 802.11e with different QoS requirements under different scenarios.

8. Acknowledgements

The authors would like to thank Dr. Imad Aad (INRIA) and the anonymous referees for providing valuable comments and suggestions that improved the quality of this paper. We also thank the IEEE 802.11 TGe Working Group members Dr. Ho Jin-meng (Texas Instruments Corp. USA), Dr. Sunghyun Choi (Philips Research USA) and Dr. Stefan Mangold (Aachen Univ. of Tech., Germany) for early discussions. This work has been supported by the French ministry of industry in the context of the national project RNRT-VTHD++.

References

[1] IEEE 802.11 WG, Reference number ISO/IEC 8802-11:1999(E) IEEE Std 802.11, 1999 edition. International Standard [for] Information Technology - Telecommunications and information exchange between systems-Local and metropolitan area networks-Specific Requirements - Part 11:Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications, 1999.

[2] IEEE 802.11 WG, Draft Supplement to STANDARD FOR Telecommunications and Information Exchange Between Systems-LAN/MAN Specific Requirements - Part 11: Wireless Medium Access Control (MAC) and Physical Layer (PHY) specifications: Medium Access Control (MAC) Enhancements for Quality of Service (QoS), IEEE 802.11e/Draft 4.2, Februray 2003.

[3] Aad I, Castelluccia C. Differentiation mechanisms for IEEE 802.11, *Proc. IEEE Infocom 2001*, Anchorage, Alaska, USA, April 2001; 1:209-218.

[4] Aad I, Castelluccia C. Remarks on per-flow differentiation in IEEE 802.11, *Proc. of European Wireless (EW2002)*, Florence, Italy, February 2002.

[5] Perkins DD, Hughes HD. A survey on quality-of-service support for mobile ad hoc networks. *Wireless Communications and Mobile Computing* 2002; 2:503-513.

[6] Lindgren A, Almquist A and Schelen O. Evaluation of quality of service schemes for IEEE 802.11 wireless LANs. *Proc. of the 26th Annual IEEE Conference on Local Computer Networks (LCN 2001)*, Tampa, Florida, USA, November 15-16, 2001; 348-351.

[7] Mangold S, Choi S, May P, Klein O, Hiertz G, and Stibor L. IEEE 802.11e wireless LAN for quality of service, *Proc. of European Wireless (EW2002)*, Florence, Italy, February 2002.

[8] Vaidya NH, Bahl P, and Gupa S. Distributed fair scheduling in a wireless LAN, *Proc. of the Sixth Annual International Conference on Mobile Computing and Networking (Mobicom 2000)*, Boston, USA, August 2000; 167-178.

[9] Deng J and Chang RS. A priority scheme for IEEE 802.11 DCF access method, *IEICE Transactions in Communications* 1999; 82-B(1): 96-102.

[10] Sobrinho JL and Krishnakumar AS. Real-time traffic over the IEEE 802.11 medium access control layer, *Bell Labs Technical Journal* 1996; 172-187.

[11] Veres A, Campbell AT, Barry M, and Sun LH. Supporting service differentiation in wireless packet networks using distributed control, *IEEE Journal of Selected Areas in Communications (JSAC), Special Issue on Mobility and Resource Management in Next-Generation Wireless Systems* 2001; 19(10):2094-2104.

[12] NS-2 simulator. http://www.isi.edu/nsnam/ns/, Access date: April 2002.

[13] Singla A, Chesson G. HCF and EDCF simulations, *IEEE 802.11e working document* 802.11-01/525r0, October 2001.

[14] Benveniste M, Chesson G, Hoeben M, Singla A, et. al. EDCF proposed draft text. *IEEE 802.11e working document* 802.11-01/131r1, March 2001.

[15] Choi S, Prado JD, Grag A, Hoeben M, Mangold S, et. al. Multiple frame exchanges during EDCF TXOP, *IEEE 802.11e working document* 802.11-01/566r3, January 2002.

[16] Hoeben M and Wentink M. Suggested 802.11 PCF enhancements and contention free bursts, *IEEE 802.11e working document* 802.11-00/113, 2000.

[17] Chesson G, Diepstraten W, Kitchin D, Teunissen H, and Wentink M. Baseline D-QoS proposal, *IEEE 802.11e working document* 802.11-00/399, 2000.

[18] Fischer M. Hybrid Coordination Function (HCF) frame exchange and NAV details, *IEEE 802.11e working document* 802.11-01/109r2, 2001.

[19] Liu H and Zarki MEI. Delay bounded type-II hybrid ARQ for video transmission over wireless channel, *Proc. Conference on Information Sciences and Systems*, March 1996.

[20] Liu H, Ma H, Zarki MEI and Gupta S. Error control schemes for networks: An overview. *Baltzer Journal of Mobile Networks and Applications* 1997; 2:167-182.

[21] Zorzi M, Rao RR, and Milstein LB. ARQ error control for fading mobile radio channels, *IEEE Transactions on Vehicular Technology*, 1997; 46(2):445-455.

[22] Kallel S and Haccoun D. Generalized type-II hybrid ARQ scheme using punctured convolutional coding, *IEEE Transactions on Communications 1990*; 38(11):1938-1946.

[23] IEEE WG, 802.11a, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: High-speed Physical Layer in the 5Ghz Band, Sep. 1999.

[24] O'Hara B and Petrick A. IEEE 802.11 handbook. A designer's companion, *IEEE Press*, 1999.

[25] Braden R.et. al. Integrated services in the Internet architecture: an overview, *IETF Standard RFC* 1633.

[26] Blake S, Baker F, and Black D. An architecture for differentiated services, *IETF Standard RFC* 2475.
[27] Wang Y and Lin S. A modified selective-repeat type-II hybrid ARQ system and its performance analysis, *IEEE Transactions on Communications* 1983; 31(5):583-589.
[28] Khansari M, Jalali A, Dubois E, and Mermelstein P. Robust low bit-rate video transmission over wireless access system, *Proc. of International Conference on Communication ICC*' 1994.
[29] Hagenauer J. Rate-compatible punctured convolutional codes (RCPC codes) and their applications, *IEEE Transactions on Communications* 1988; 36:389-400.
[30] Crow BP, Widjaja I, Kim JG, and Sakai PT. IEEE 802.11 wireless local area networks, *IEEE Communication Magazine*, September 1997; 35(9):116-126.
[31] Golestani SJ. A self-clocked fair queueing scheme for broadband applications. *Proc. IEEE Infocom* 1994.
[32] Wolisz A, Fitzek FHP. QoS support in wireless networks using Simultaneous MAC packet transmission (SMPT), *Applied Telecommunication Symposium* (ATS), San Diego, USA, April 1999.
[33] Ganz A, Phonphoem A, and Ganz Z. Robust superpoll with chaining protocol for IEEE 802.11 wireless LANs in support of multimedia applications, *Wireless Networks* 2001; 7: 65-73.
[34] Visser MA and Zarki ME. Voice and data transmission over an 802.11 wireless network, *Proc. of PIMRC*, Toronto, Canada, Sept. 1995.
[35] http://www.ietf.org/html.charters/issll-charter.html. Access date: June 2002.
[36] Chesson G, Sherman M, Choi S, and et al. A simulation framework for evaluation of 802.11 QoS MAC enhancements: progress report, *IEEE 802.11e working document* 802.11-00/265r1, Sept. 2000.
[37] Drabu Y. A survey of QoS techniques in 802.11. http://www.cs.kent.edu/~ydrabu/research/wmac/mac.pdf. Access date: June 2002.
[38] Choi S. Emerging IEEE 802.11e WLAN for Quality-of-Service (QoS) Provisioning, *SK Telecom Telecommunications Review* 2002, 12 (6): 894-906.
[39] IEEE 802.1D-1998, Part3: Media Access Control (MAC) Bridges, ANSI/IEEE Std. 802.1D, 1998 edition, 1998.
[40] Aikawa S et al. Forward error correction schemes for wireless ATM systems, *Proc. IEEE ICC* 1996.
[41] Romdhani L, Ni Q, and Turletti T. Adaptive EDCF: Enhanced Service Differentiation for IEEE 802.11 Wireless Ad Hoc Networks. *IEEE WCNC'03 (Wireless Communications and Networking Conference)*, New Orleans, Louisiana, March 16-20, 2003.
[42] Chalmers D and Sloman M. A survey of Quality of Service in mobile computing environments, *IEEE Communications Surveys*, April 1999; 2:2-10.
[43] Ware C, Simoens S, et al. HTSG (High Throughput Study Group) requirements - scope and purpose. *IEEE 802.11 HTSG working document* 802.11-02/567r0, Sept. 2002.

Authors' Biographies

Qiang Ni received the B.S. degree, the M.S degree and the Ph.D. degree from Hua Zhong University of Science and Technology (HUST), Wuhan, China in 1993, 1996 and 1999 respectively. From 1999 to 2001, he was a postdoctoral research fellow in the Multimedia Communication Laboratory, Department of Electronics and Information Engineering, HUST. In 2001, he joined INRIA Sophia Antipolis, where he is now an expert research engineer at INRIA, Planète group. Since 2002 he has been a voting member for IEEE 802.11 wireless LAN Standard Working Group. His current research interests include QoS enhancements for IEEE 802.11 wireless LAN, next-generation high-speed wireless networks and multimedia transmission over hybrid wireless and wired networks.

Lamia Romdhani graduated from the Faculty of Sciences of the Tunis I University in Tunisia 2001 (Computer Science Department). She obtained her M.S. Networking and Distributed Systems from the University of Nice Sophia-Antipolis in 2002. From March 2002 to September 2002 she did her M.S. internship at INRIA, Planète Group. In October 2002, she joined the Mobile Communications Department of the Eurécom Institute in Sophia-Antipolis to prepare a Ph.D on computer science. Her main research topic is cross-layer QoS support for wireless ad-hoc networks.

Thierry Turletti received the M.S. (1990) and the Ph.D. (1995) degrees in computer science both from the University of Nice - Sophia Antipolis, France. He has done his PhD studies in the RODEO group at INRIA Sophia Antipolis. During the year 1995-96, he was a postdoctoral fellow in the Telemedia, Networks and Systems group at LCS, MIT. He is currently a research scientist at the Planète group at INRIA Sophia Antipolis. His research interests include multimedia applications, congestion control and wireless networking. Dr. Turletti currently serves on the Editorial Board *of Wireless Communications and Mobile Computing*.

Appendix

Abbreviations and acronyms

AC	Access Category
ACK	Acknowledgement
AIFS	Arbitration Inter Frame Spacing
AIFSN	Arbitration Inter Frame Spacing Number
AP	Access Point
CA	Collision Avoidance
CAP	Controlled Access Period
CFB	Contention Free Burst
CFP	Contention Free Period
CF-Poll	Contention Free – Poll
CF-End	Contention Free – End
CP	Contention Period
CSMA	Carrier Sense Multiple Access
CW	Contention Window
CW_{max}	Contention Window Maximum
CW_{min}	Contention Window Minimum
DCF	Distributed Coordination Function
EDCF	Enhanced Distributed Coordination Function
FEC	Forward Error Correction
HC	Hybrid Coordinator
HCF	Hybrid Coordination Function
IEEE	Institute of Electrical and Electronics Engineers
ISM	Industrial, Science, and Medical
MAC	Medium Access Control
MSDU	MAC Service Data Unit
NAV	Network Allocation Vector
PC	Point Coordinator
PCF	Point Coordination Function
PHY mode	Physical Layer mode, coding and modulation scheme
PIFS	PCF Inter Frame Space
PSDU	Physical (layer) Service Data Unit
QAP	QoS access point
QBSS	Quality of Service Basic Service Set
QIBSS	Quality of Service Independent Basic Service Set
QoS	Quality of Service
QSTA	QoS station
RS	Reed-Solomon
RTS/CTS	Request to Send/Clear to Send
SIFS	Short Inter Frame Space
TC	Traffic Category
TBTT	Target Beacon Transmission Time
TCID	Traffic Category Identifier
TID	Traffic Identifier
TS	Traffic Stream
TSID	Traffic Stream Identifier
TSPEC	Traffic Specification
TXOP	Transmission Opportunity
WLAN	Wireless Local Area Network