QoS Enhancements in IEEE 802.11 Wireless LANs to Support Real-Time Services

by

Irshad Ali Qaimkhani

A Thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
in partial fulfilment of the requirements of the degree of

MASTER OF SCIENCE

Department of Electrical and Computer Engineering
University of Manitoba
Winnipeg

THE UNIVERSITY OF MANITOBA

FACULTY OF GRADUATE STUDIES

*****

COPYRIGHT PERMISSION

QoS Enhancements in IEEE 802.11 Wireless LANs to Support Real-Time Services

BY

Irshad Ali Qaimkhani

A Thesis/Practicum submitted to the Faculty of Graduate Studies of The University of Manitoba in partial fulfillment of the requirement of the degree

MASTER OF SCIENCE

ABSTRACT

We address the research problems relating to channel scheduling and call admission control for real-time voice services in IEEE 802.11 Wireless LANs. Here, scheduling refers to the efficient allocation of the available resources among the admitted voice calls, and the call admission control (CAC) refers to the method for deciding whether to admit or to drop an incoming voice call depending on some decisive performance metrics. We carry out a background study of IEEE 802.11 MAC standard with special reference to its architectural aspects, quality of service (QoS) limitations on real time services, QoS enhancement aspects and efforts done so far within the standard. We specify the core issues related to the QoS provisioning and efficient bandwidth utilization. We also present an extensive research survey and critically analyze the major research works on QoS-aware wireless medium access control (MAC) protocols in the recent literature.

Analytical performance evaluation of two well-known contention-free schemes that suppress the silence periods in voice calls is carried out. Numerical results verify that these schemes are not suited to delay sensitive voice traffic. To this end, we developed a new QoS-aware wireless MAC protocol, called Hybrid Contention-Free Access (H-CFA) protocol, and a measurement-based call admission control technique, called Traffic Stream Admission Control (TS-AC) algorithm. The H-CFA combines the two contention-free wireless medium access approaches, i.e., the round-robin polling and the time-division multiple access (TDMA)-like time slot assignment, to suppress the idle periods of voice calls efficiently at no cost of increased waiting time delays.

In order to provide the acceptable level of packet loss to already admitted traffic streams, the proposed TS-AC algorithm keeps the number of admitted stations below some measured maximum count. The TS-AC algorithm maximizes the capacity based on the maximum negotiated QoS level in terms of the percentage of times an admitted voice user may undergo non-consecutive packet loss.

Performance analyses of the proposed H-CFA MAC protocol and the TS-AC scheme are carried out through simulation models in comparison to the related well-known schemes in the literature, including the round-robin polling scheme. The performance results show that the H-CFA protocol out-performs all its counterparts and that the TS-AC scheme provides consistent delay satisfaction and increases the capacity gain significantly at the cost of different accepted QoS levels.

Examiners:

Prof. E. Hossain, Supervisor, Dept. of Electrical & Computer Engineering

---

Prof. J. Cai, Dept. of Electrical & Computer Engineering

Prof. R. Eskicioglu, External Examiner, Dept. of Computer Science

Table of Contents

List of Figures

List of Tables

Acknowledgement

I would like to thank

- Professor Ekram Hossain for his guidance and support,
- my research group colleagues for their time and discussion,
- and my family: Zareena and Yaqoob (my parents), Farhana (my better half), and others, for their ever-lasting love and prayers that made me strong enough to have such a success.

Dedication

I dedicate this thesis to my beloved deceased grandma, Halima, whose love, affections, devotion to her family, righteousness, and simple but reality-based approach towards life, are always the source of inspiration and motivation to me.

Chapter 1

Introduction

Over the last few years, there is a rapidly growing trend of migrating customers from traditional telephone companies to mobile phones and Voice-over-IP (VoIP) services. Until recently, voice user demand for mobility has been fulfilled by the circuit-switched based cellular technology only. Although, the dedicated bandwidth allocation in circuit-switched networks makes them very efficient in terms of end-to-end (ETE) delays and jitters (frames inter-arrival time), but, at the same time, it also makes them bandwidth-inefficient. This is due to the fact that voice calls are bursty in nature with alternating talk-spurt and silence periods.

The packet-switched telephony technology (i.e., VoIP) is being increasingly deployed worldwide on commercial grounds on wired-IP networks through the Internet. The ease of statistical multiplexing of IP packets in the packet-switched networks makes it highly bandwidth efficient and cost-effective. By 2009, VoIP systems are expected to represent 91% of all enterprise phone systems worldwide [1]. Although, VoIP technology suppresses the alternating silence periods of voice calls and increases the bandwidth utilization by multiplexing those with other voice calls, it does not fulfill the increasing mobility demand of the user.

The demand for mobile VoIP-telephony has been increasing with exponential pace which has attracted tremendous research work in this area. There are two major grounds of research for this purpose. One is the IEEE 802.16 standard-based WiMAX (Worldwide Interoperability for Microwave Access) technology, which ranges to several miles to cover a metropolitan city area through wireless access. The other is IEEE 802.11 standard-based wireless LAN or WiFi (Wireless Fidelity) technology, which provides wireless medium access within around hundred meters to cover small public places such as university or college campuses, hospitals, airports, and restau-

rants. According to a research estimate, the enterprise market for VoIP over WiFi equipment is to be worth around $2bn$ in 2007 and growing to $15bn$ in 2012 [2]. Our focus in this research thesis is the wireless VoIP using WiFi.

Pervasive commercial deployment of VoIP over wireline networks and the unprecedented mobility, flexibility and scalability provided by the WiFi hot-spots in the recent years have attracted great research efforts in the area of WiFi VoIP. Real-time voice traffic requires some parametric type of quality of service (QoS) such as maximum end-to-end delay bound, acceptable jitter, and acceptable inconsecutive packet loss. For example, the interactive voice can normally tolerate end-to-end delay up to 25 $msec$ without echo canceller and 150 $msec$ with echo canceller [6]. This stringent parametric QoS requirement can only be fulfilled through end-to-end connection oriented service. However, voice traffic can tolerate inconsecutive packet loss and can provide the opportunity of multiplexing gains through the suppression of its alternating silence periods.

Medium access control (MAC) protocol design plays a key role in provisioning QoS for packet-switched services in a WLAN. Real-time voice traffic poses stringent QoS requirements, such as maximum delay bound, acceptable jitter, and acceptable inconsecutive packet loss. The distributed coordination function (DCF) which is mandatory in the IEEE 802.11 MAC standard uses a contention-based medium access protocol, called carrier sense multiple access with collision avoidance (CSMA/CA). The binary exponential back-off mechanism used in CSMA/CA causes non-deterministic channel access delays which grows even faster under heavy and unbalanced traffic conditions [23]. Therefore, DCF is not suitable for real-time voice traffic. Whereas, the point coordination function (PCF), which is optional in the IEEE 802.11 MAC standard, offers a contention-free and connection-oriented service through a round-robin polling scheduler. Although, PCF is suited to real-time traffic but its round-robin polling scheduler is inefficient with respect to the bursty nature of voice that has alternating periods of idleness (silence) and activeness (talk-spurt). Round-robin polling scheduler causes wastage of bandwidth and incurs unnecessary waiting delays to other stations having frames to send when the polled station does not have any frame during its idle periods. Suppression of idle periods in voice sessions can greatly reduce such unnecessary waiting time delays and can bring multiplexing gains to increase

the network capacity significantly.

In the QoS-enhanced version, i.e., 802.11e, the contention-based channel access mechanism, called enhanced distributed channel access (EDCA), provides relative-priority type QoS which is not enough for real-time voice traffic. However, its contention-free hybrid coordination function controlled channel access (HCCA), provides parameterized QoS through a central coordinator, called hybrid coordinator (HC). The HC does so by setting up traffic streams (TS) after negotiating the traffic specifications (TSPEC), such as nominal MAC frame size, maximum service interval, delay bounds, jitters, mean data rates between the stations and the access point (AP). However, the same round-robin type polling scheduler is used in the HCCA as in the PCF. The difference is that the HCCA uses the mandatory TSPEC parameters to prioritize different TSs in its polling list and to determine the respective length of the polled transmission opportunities (polled-TXOP) for all the poll-able stations. For guaranteeing the negotiated delay bounds, 802.11e also defines the call admission control outlines but does not specify any technique for the admission control. In a nut shell, the inefficiency with respect to the bursty nature of voice traffic is still there in HCCA as, like PCF, it is also unable to suppress silence periods and provide multiplexing gains. Therefore, the research challenge still remains the same.

1.1 Motivation

The motivation for this research comes from an observation that the HCCA/PCF mode offers a "packet-switched connection-oriented" service, which is well suited for telephony traffic. Telephony traffic has been shown to have alternating periods of talk-spurts and silences [22]. In packet-switched solutions, silent periods of a voice call can be multiplexed with voice data from other calls. In this way, packet-switched solutions are more bandwidth-efficient than circuit-switched solutions.

This has been one of the primary reasons for the ongoing movement in the telecommunications industry towards moving telephony traffic from circuit-switched networks on to packet-switched networks. In wireless networks, where bandwidth is more constrained, the use of packet-switched techniques, which is more bandwidth efficient in carrying voice traffic, is needed. Currently, most of the wireless customers use

cellular or cordless telephones within buildings, where the availability of an 802.11 wireless LAN enables a more efficient usage of overall wireless bandwidth. With the increased number of customers using wireless LAN access for their voice calls inside the building, more of the cellular resources can be made available for outdoor users and buildings which are without 802.11 LANs. Therefore, our motivation is to take advantage of the packet-switched aspect of IEEE 802.11 standard to support bursty telephony traffic and achieve better overall wireless bandwidth utilization, and thus QoS.

Further, polling by the hybrid coordinator (HC) in HCCA means that the HC designates the channel to the stations alternatively for specific time periods (i.e., polled-TXOPs in case of HCCA) in round-robin fashion giving them the opportunity to send their voice packets in the uplink voice transmission period (UVTP), a sub-division of the contention-free period (CFP) or of the controlled access phase (CAP). Voice has the characteristics of alternating periods of silences (while hearing or not communicating at all) and talk-spurts (while talking). During the silent period, a station need not be polled because it has nothing to send. In other words, silent periods can be suppressed by not polling a silent station in order to reduce the polling over-head and, consequently, increasing the performance in terms of end-to-end throughputs and queuing delays. HCCA/PCF polling mechanism does not care about the alternating silent periods of voice sessions and polls all the stations during the downlink voice transmission period (DVTP), a sub-division of CFP or of CAP.

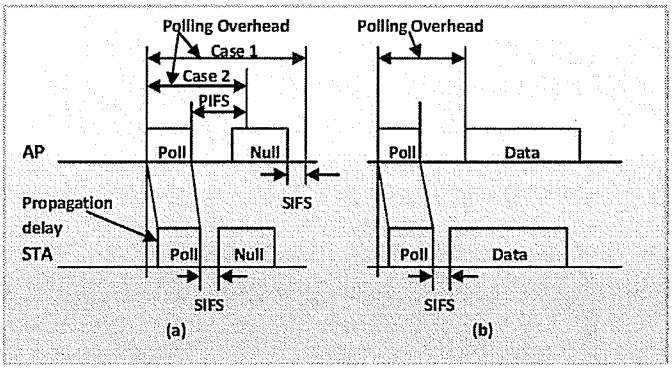

Figure 1.1. *Polling Overhead in IEEE 802.11 HCCA/PCF mode: (a) Station has no data to send, (b) station has data to send.*

Polling of each station incurs some cost in terms of spending some time called polling over-head (POH). Fig. 1.1(a) is the case when polled node does not have data to send and Fig. 1.1(b) is the case when polled node does have data to send. From these two figures, it can be seen that the POH is larger in the case when polled node has no data to send than in the case when the polled node does have data to send. This means that the POH can be reduced significantly by not polling a node during its silent periods.

The "connection-oriented" aspect of the HCCA/PCF mode would allow the network to provide delay guarantees necessary for interactive voice. The end-to-end delay requirement for interactive voice is 25ms without echo cancellers, 150ms with echo cancellers for excellent quality voice, and 400ms with echo cancellers for acceptable quality voice [6], and for round trip time (RTT), these delay values are doubled. The HCCA/PCF mode would allow for delay guarantees to be made for voice calls.

So, there is a motivating need to develop intelligent polling MAC schemes that can avoid polling all nodes during their silent periods in order to improve the bandwidth efficiency and QoS.

1.2 Objectives and Scope of the Thesis

The main objective of this thesis is to study comprehensively the wireless medium access scheduling and call admission control problems relating to the real-time voice and video services in the current and next generation wireless IP networks, and to develop improvement schemes or protocols for efficient bandwidth utilization and QoS provisioning for real-time applications. In this context, we attempt to develop analytical models for the critical study of performance metrics in the important related research works proposed recently in the literature in comparison to the IEEE standard in order to ascertain their viability for wide-scale commercial deployment. By doing this analytical performance evaluation of the current state of the art, our objective is to develop improvement schemes and to carry out their performance evaluation in comparison to the current state of the art.

In order to achieve the objectives of this thesis, we would need comprehensive analytical and simulation models to study the interdependencies among the different

system parameters and the performance metrics. However, analytical models can be used to obtain better performance metrics as compared to computer simulations. Based on these analytical models, optimization formulation can be developed and solved for obtaining the optimal system parameter setting. In nut shell, both analytical and simulation frameworks would be required for design and engineering of the wireless medium access scheduling and call admission control. In this thesis, Matlab is used in obtaining both the analytical and the simulative results. The methodology used for simulations is the discrete-event system modeling with the process interaction approach.

The scope and major contributions of this thesis are described as follows:

- It provides a comprehensive study of the wireless MAC architecture and various QoS aspects with special reference to that which can guarantee QoS-support and efficient use of the bandwidth to real-time voice services in the IEEE 802.11 WLANs. Besides, it explores the QoS-related features in the legacy standard and its enhanced version, it identifies the limitations and shortcomings within the standard as open research and improvement problems for the researchers.
- It puts forth a consolidated research survey of the related research work in the literature that claims to enhance QoS and/or efficient bandwidth utilization for time-sensitive voice services. In this research survey, various schemes are analyzed critically in the context of our objectives, and a qualitative comparison of these schemes along with the schemes that we propose in this thesis is also presented.
- It carries out the analytical performance evaluation of some well-known recent schemes in the literature that are closely related to our objectives. To this end, mathematical models are developed to evaluate important performance metrics like the average polling overhead time saved at the scheduler for each voice station during its each voice activity cycle and the average wireless medium access waiting time delay that a typical first talk-spurt packet would experience in these schemes when the voice station changes its state from silence to talk-spurt. The numerical results of this analysis reveals that, although these schemes suggest considerable reduction in the polling overhead time for efficient utilization of the bandwidth, these are not viable in case of delay sensitive real-time voice

services.

- It introduces a novel wireless MAC protocol called Hybrid Contention-Free Access (H-CFA) protocol for providing efficient bandwidth utilization and QoS to voice services in WiFi networks. This protocol provides purely contention-free medium access mainly through an intelligent round-robin type polling algorithm adjoining it with a TDMA-like time slot algorithm for Contention-Free Activity Detection from idle to active state change. In this way, it suppresses the idle periods of real-time voice efficiently at no cost of increased waiting time delays. A computer simulation model is developed to carry out the performance evaluation of the H-CFA protocol in comparison to the related well-known schemes in the literature including the round-robin polling scheme used in the legacy IEEE 802.11 standard. The performance results show that the H-CFA protocol out-performs all its counterparts.
- It also introduces a measurement-based call admission control scheme called Traffic Stream Admission Control (TS-AC) scheme for QoS provisioning to delay sensitive voice services. In order to provide a consistent level of parametric QoS and the acceptable level of packet loss, to already admitted traffic streams, the proposed TS-AC algorithm keeps the number of admitted stations below some measured maximum count. The TS-AC algorithm maximizes the capacity by exploiting one characteristic of voice service that it can tolerate inconsecutive packet loss to some acceptable level. This capacity enhancement is based on the maximum negotiated QoS level in terms of the percentage of times an admitted voice user may undergo inconsecutive packet loss. As an important performance measure, we evaluate the maximum number of voice stations that can be satisfied for different levels of maximum negotiated QoS through a simulation model. The performance results of the TS-AC algorithm shows that it provides consistent delay satisfaction on one hand and increases the capacity gain significantly at the cost of different accepted QoS levels on the other hand.

1.3 Organization of the Thesis

This thesis is organized as follows:

- **Chapter 2** presents a consolidated background study of IEEE 802.11 MAC standard with special reference to its important architectural aspects, its limitations on QoS for real-time services, QoS requirements and a basic survey of QoS enhancement aspects and efforts done so far within the standard and out of the standard.
- **Chapter 3** discusses and elaborates the core QoS-related design problem that is common in all versions of IEEE 802.11 MAC standard (whether 802.11a/b/g or 802.11e). That is the inefficiency of the WiFi MAC standard in exploiting a very important characteristic of packet-switched networks, i.e., multiplexing gains and quality of service provisioning in case of bursty type of traffic. This chapter also presents a consolidated survey and critically analyzes the major research works on QoS-aware wireless MAC protocols in the recent literature. For end-to-end delay requirement of the real-time sensitive services, it also discusses a few schemes that deal with another very important QoS-provisioning aspect that is call admission control.
- **Chapter 4** provides the analytical evaluation of two well-known fully contention-free schemes that suppress the silence periods in voice calls. Their short-comings are elaborated. Mathematical models are developed to evaluate the two important performance metrics, i.e., the average polling overhead time that the scheduler saves for each voice station during its one uplink voice activity cycle, and the unnecessary average waiting time delay that the first talk-spurt frame at each voice station suffers.
- **Chapter 5** presents and evaluates a novel wireless MAC protocol called Hybrid Contention-Free Access (H-CFA) protocol. The H-CFA protocol provides substantial multiplexing gains through silence suppression. This protocol provides purely contention-free medium access mainly through an intelligent round-robin type polling algorithm adjoining it with a TDMA-like time slot algorithm for Contention-Free Activity Detection from idle to active state change. It suppresses the idle periods of real-time voice efficiently at no cost of increased waiting time delays.
- **Chapter 6** introduces a measurement based call admission control scheme called Traffic Stream Admission Control (TS-AC) algorithm that keeps the

number of admitted stations below some measured maximum count in order to provide a consistent level of parametric QoS, specially, the delay bounds and the acceptable level of packet loss, to already admitted traffic streams. The TS-AC algorithm maximizes the capacity by exploiting one characteristic of voice service that it can tolerate inconsecutive packet loss to some acceptable level.

- **Chapter 7** concludes this thesis by summarizing its contributions, and it also outlines the ongoing and a few future research directions.

Chapter 2

IEEE 802.11 Wireless LANs: MAC Protocol Architectures

2.1 IEEE 802.11 Wireless LANs

IEEE 802.11 WLAN standard defines MAC sub-layer and the physical (PHY) layer. Logical link control (LLC) sub-layer is defined in the IEEE 802.2 standard which provides a transparent interface to the higher layer users. Stations (STAs) while roaming through 802.11 WLAN still appear as stationary to 802.2 LLC sub-layer and higher layers enabling existing TCP/IP protocols to run over IEEE 802.11 WLAN just like wired Ethernet. IEEE provided three PHY layer options in 1997, i.e., InfraRed (IR) baseband PHY, frequency hopping spread spectrum (FHSS) radio, and direct sequence spread spectrum (DSSS) radio. All these options support both 1 and 2 Mbps PHY rate. In 1999, IEEE extended 802.11 to 802.11b and 802.11a, where 802.11b works in 2.4 GHz band with data rates up to 11 Mbps, based on DSSS PHY, and 802.11a works in the 5 GHz band with data rates up to 54 Mbps based on OFDM PHY. 802.11b has been further extended to 802.11g to support high data rates up to 54 Mbps in 2.4 GHz band, and 802.11h is the further extension of 802.11a to support indoor and outdoor license regulations for the 5GHz band in Europe. The version 802.11f defines an Inter-Access Point protocol that allows STAs to roam among multi-vendor access points. Security and authentication issues are addressed in 802.11i. And the version 802.11e has targeted to enhance quality of service (QoS) performance of the legacy IEEE 802.11 standard. Figs. 2.1-2.2 show different standardization activities done at IEEE 802.11PHY and MAC layers.

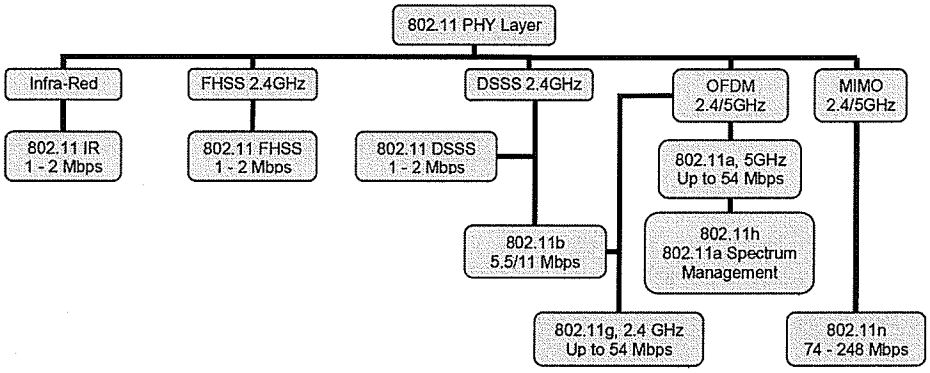

Figure 2.1. *Different IEEE 802.11 versions classified on PHY layers.*

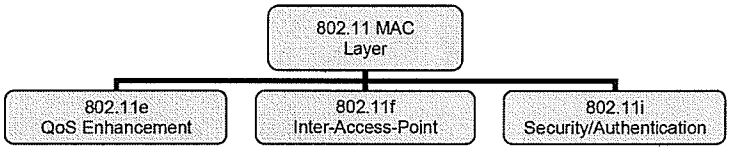

Figure 2.2. *Different IEEE 802.11 versions defining only MAC layer.*

2.2 IEEE 802.11 MAC Architecture

The basic architecture of IEEE 802.11 standard includes the definitions of both medium access control (MAC) sub-layer and physical layer (PHY). WLAN MAC architecture is based on either of the following two well-known approaches or a combination of the both.

- *Contention-based medium access (also called non-deterministic medium access):* It uses the carrier-sense multiple-access with collision avoidance (CSMA/CA) protocol in which the non-deterministic contention window (CW) back-off mechanism causes non-deterministic channel access delays. CSMA/CA is incorporated in the DCF mode of IEEE 802.11a/b/g and in the enhanced distributed coordination function (EDCF) of IEEE 802.11e with some modifications for prioritizing services. DCF and EDCF work at the user end. Due to its non-deterministic nature, contention-based medium access architecture is suited for asynchronous data transmission and for best-effort service.
- *Contention-free medium access (also called deterministic medium access):* Since contention-free medium access is deterministic, medium access delays are also

deterministic. There are two sub-approaches which are used in contention-free wireless MAC design. One is the round-robin type polling method where each poll-able station is individually polled by the central controller at its turn. Round-robin type polling method is used in the point coordination function (PCF) mode of IEEE 802.11a/b/g [3] and the hybrid coordination function controlled channel access (HCCA) mode of IEEE 802.11e [4]. These two modes work at the central controller (i.e., the access point) and are combined in the hybrid co-ordination function (HCF). HCF can work concurrently with the DCF and the PCF for backward compatibility combining functions from both DCF and PCF with some enhanced QoS-specific mechanisms and frame subtypes (see Figs. 2.3-2.4). The other approach is the TDMA-like time slots assignment to stations under some defined order which is being proposed in the litrature for reducing the polling over-head. Both the sub-opproaches are taking-turn type, suited for synchronous transmission, and are deployed through a central controller, however, one has advantages or disadvantages over the other in different respects.

In TDMA-like time slot assignment, polling overhead is reduced greatly by broadcasting only one poll frame to all the admitted users during one contention-free period assigning them a specific time slot for uplink transmission when their turn comes. One major drawback of this scheme is the wastage of the entire time slot if the users do not have packets to send. Time synchronization problem of the TDMA-like schemes poses complexities in accommodating asynchronous service time requirements of individual admitted services. However, in round-robin polling schemes there is no time synchronization problem as each admitted user is assigned the channel for its uplink transmission only at the time of its own turn in a defined order through a contention-free poll. In this case, the asynchronous service time requirements of individual admitted services become easy to address. In case there is no packet to send, the central controller has to wait for comparatively much smaller time (e.g., PIFS time in case of 802.11 PCF-mode) before it proceeds to poll the next admitted user. In this way, the bandwidth wastage in case of the TDMA-like schemes is much larger than that in polling schemes, but in terms of polling overhead reduction, the former has

edge over the latter.

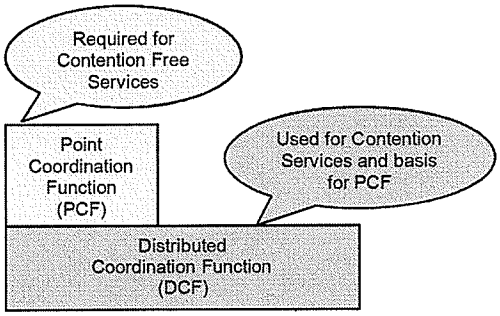

Figure 2.3. *Basic WiFi MAC (IEEE 802.11a/b/g) architecture.*

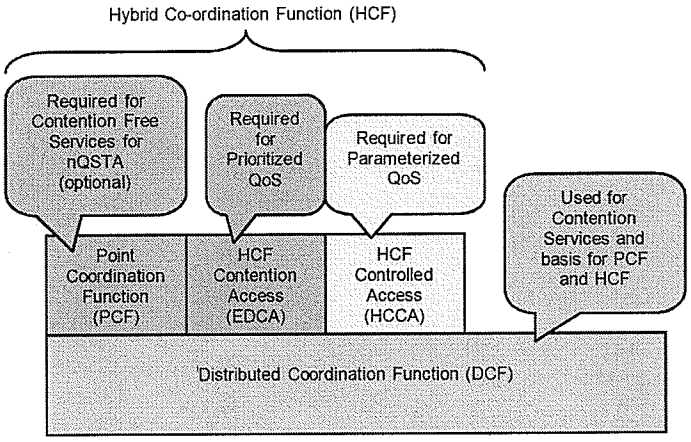

Figure 2.4. *Enhanced WiFi MAC (IEEE 802.11e) architecture.*

In the enhanced MAC (i.e., IEEE 802.11e) architecture which supports quality of service (QoS), all the terminologies are prefixed with Q for QoS. For example, STA (station) is represented by QSTA, a station not supporting QoS is represented by nQSTA, a basic services set (BSS) by QBSS, an infrastructure BSS (IBSS) by QIBSS etc. A nQSTA does not has HCF, and a QSTA has both DCF and HCF. In any case, PCF is always optional.

According to 802.11 Standard the group of STAs coordinated by DCF or PCF is formally called a basic service set (BSS). The area containing this BSS is called basic service area (BSA) which is like a cell in a cellular mobile network. IEEE 802.11 wireless networks can be configured into two different modes: ad-hoc and infrastructure

modes. In ad-hoc mode, all wireless STAs within the BSA can communicate directly with each other to form an independent BSS (IBSS) with out connecting to a wired back-bone, i.e., distribution system (DS), whereas in infrastructure mode, an access point (AP) is needed to connect all stations to a distribution system (DS), and each station can communicate with others through the bridge of AP within the same BSS or through the wired distribution system to stations in other BSSs. DCF is the basic medium access mechanism for both ad-hoc (IBSS) and infrastructure modes. But PCF can only be used in infrastructure mode [5].

The transmitting frames have in-between time gaps called inter frame spacing (IFS). The idleness of the wireless medium is ascertained by the carrier sense function if the medium is sensed idle for the specified IFS. Five following IFS with different time lengths are defined to prioritize the wireless medium access.

Small IFS (SIFS): SIFS is the shortest of all types of IFS (top-most priority) and a STA that has seized the wireless medium (i.e., in transmission sate) uses SIFS to keep the medium seized while it performs frame exchange sequence. SIFS is used for acknowledgement (ACK) frame, clear to send (CTS) frame, subsequent MPDU (MAC Protocol Data Unit) of a long fragmented burst/frame (i.e., MAC service data unit (MSDU) or a MAC management protocol data unit (MMPDU)), responding to a poll frame in the PCF mode or it may be used for any type frame during the Contention Free Period (CFP).

PCF IFS (PIFS): It is the second shortest time frame used to give priority to capture the idle wireless medium at the start of each PCF mode cycle, and also by the access point (AP) to wait for the response from a polled STA.

DCF IFS (DIFS): DIFS is the third smallest time frame and is only used in DCF mode. A STA operating in DCF mode waits for DIFS time period sensing an idle medium to remain idle so that it can seize the idle medium for frame transmission or if the STA is in back-off mode, it waits for DIFS time period to start decrementing its back-off counter just after the busy medium is sensed idle.

Extended IFS (EIFS): EIFS is the largest time frame used in DCF mode only to provide enough time to a STA to acknowledge an incorrectly received frame with an incorrect FCS (frame check sequence) value. EIFS period begins just after the indication of incorrect reception by the PHY carrier sensing without regard to the

virtual carrier sensing.

Arbitration IFS (AIFS): AIFS is used by the QoS facility of 802.11e to transmit MPDUs, all MMPDUs, PS-Poll, RTS, CTS (when transmitted as a response to the RTS), BlockAckReq, and BlockAck (when not transmitted as response to the BlockAckReq). A QSTA gets a TXOP for an access category (AC) using the EDCAF if its carrier sense mechanism finds the wireless medium idle at the AIFS slot boundary, after a correctly received frame, and the back-off time for the AC has expired.

2.2.1 Distributed Coordination Function (DCF) Mechanism

The basic DCF is CSMA/CA (Carrier Sense Multiple Access with Collision Avoidance) and is implemented in all STAs for use during both ad-hoc and infrastructure mode. Carrier sensing is performed using physical carrier sensing (by air interface) as well as virtual carrier sensing at PHY MAC layer. In PHY carrier sensing all detected packets and channel activity is analyzed through comparing the relative signal strength from other STAs in order to detect if any of the present STAs is transmitting. Virtual carrier sensing uses the duration of the packet transmission which is included in the headers of control packets (RTS and CTS) and in DATA frames. The channel is considered to be busy if either physical or virtual carrier sensing indicates that the channel is busy. Collision avoidance (CA) is implemented to reduce the collision probability through non-deterministic bi-exponential back-off mechanism. Otherwise, the collision probability, just after the channel is sensed idle after it has been busy, is very high because many STAs might be having frames to transmit at the same time.

As soon as the STA has a packet at the head of its buffer, it starts contending for the wireless channel through CSMA/CA mechanism. If it finds:

The Channel is busy (on vacation): The STA defers to contend for the wireless medium immediately for the specified time slots as set in its network allocation vector (NAV) according to the envisaged transmission time slots of the on-going busy medium. After this deferral is over (i.e., the NAV has become zero), the STA backs off for a randomly chosen back off time from 0 to W discrete time slots (where $W = CW_{min}$ is called minimum contention window) by initiating a back off counter which initially remains inactive as far as the channel is busy and does not become idle for DIFS time period. In the back off state the STA continuously senses the channel

for if it is busy or idle. As soon as the channel is sensed idle for a DIFS or EIFS period (as the case may be), the back off counter starts decrementing by one for each elapsing time slot and keeps on decrementing as far as the channel remains idle. If, in the mean while, the channel becomes busy, the back off counter stops decrementing the back off time and remains stopped as far as the channel is busy and does not becomes idle for a DIFS period. As soon as the channel becomes idle again for a DIFS or EIFS period for an incorrect reception of frame, the back off counter starts decrementing the counter from where it had stopped it last time. This process of back off counter decrementing while the channel is idle and back off counter stopping while the channel is busy is repeated as many times as the channel becomes idle and busy respectively during the whole randomly chosen back of time (from 0 to W) till the back off counter reaches to 0 ultimately.

The Channel is idle (ready to serve): The STA waits for DIFS time period and keeps on sensing the channel for if it remains idle or becomes busy during the DIFS time.

In case if the channel becomes busy before the DIFS time is over, the STA immediately defers contending for the medium for time slots set in its NAV, and, then after, the above stated random back mechanism is carried out. In case if the channel remains idle for the whole period of DIFS, the STA enters the transmission state immediately for transmission of its initial frame.

In the above stated both cases, the contending STA has to wait for at least a fixed DIFS time if the STA finds the channel idle at the start and the channel remains idle for the DIFS time period. Otherwise, the contending STA has to wait for time slots which comprise of randomly chosen back off time slots and the random time slots for which the channel may go on single or multiple vacations during the back off time period. In both cases when DIFS time slots or the randomly chosen CW time slots reaches to zero the STA enters the transmission state from the contending state.

In the transmission state, packet transmission may result in success or in collision. Contrary to the wired-medium packet transmission where the collision/packet loss can be detected while transmitting the packet by sensing the received packet signal strength at the same time and thus the transmission bandwidth can be saved by aborting the transmission of rest of the transmitting packet, the collision in the

wireless medium transmission can not be sensed at the same time because of the large difference between the transmitting and receiving packet signal strengths. The only way that the wireless transmission is considered successful is the reception of ACK frame by the sending STA from the receiving STA within a time-out period. Also, every STA in the BSS/IBSS has to maintain a time vector called network allocation vector (NAV). Each time a transmitting STA transmits a control/data frame, the NAV is updated among all the STAs that are within the transmission range of the transmitting STA according to the estimated time slots of transmission indicated in the headers of the transmitting control/data frame so that all other STAs can become idle for that time slots by setting the NAV timer accordingly. In this way the virtual carrier sensing is basically implemented by NAV, when the NAV counter is non-zero the medium is busy, and when the NAV counter is zero the wireless medium is idle.

In case of successful transmission, the STA enters the contention state afresh if it has another packet arrived at the head of its buffer in the mean while, and it repeats the same random back-off mechanism as described above with the CW reset to CW_{min} after it has sensed the medium idle for DIFS time slots. Here the successful transmission is determined with the reception of the ACK frame within the ACK time out period. In case of packet collision (or unsuccessful transmission) which is determined by the non-reception of the ACK frame within the ACK time out period, the STA backs off again for a randomly chosen back off time from 0 to $2W$. In this way, the contending STA undergoes a bi-exponential back off process for each successive unsuccessful transmission before it transmits the packet successfully.

For each unsuccessful transmission the back off window is increased bi-exponentially up to a maximum of CW_{max}. The backoff time is computed as follows [3]:

$$Backoff\ Time = Random() \times SlotTime\,(bit\ time) \tag{2.1}$$

Where $Random()$ is a pseudorandom integer drawn from a uniform distribution over the interval $[0, CW]$. CW is an integer within the range of values of the PHY characteristics CW_{min} and CW_{max}, i.e., $CW_{min} \leq CW \leq CW_{max}$. SlotTime ($bit\ time = \sigma\mu s$) equals the value of the corresponding PHY characteristics. So, CW_{min} is equal to $32 \times \sigma\mu s$(σ is the duration of a time slot, i.e., bit time) and CW_{max} is equal to $1024 \times \sigma\mu s$. CW parameter shall take an initial value of CW_{min}.

So, at m^{th} unsuccessful transmission attempt, CW is randomly chosen from CW_{min} to $2^m \times CW_{min} - 1$ where $0 \leq m \leq 5$ and m is the parameter that denotes the stage of the back-off window and is incremented every time a station fails to transmit the packet successfully. In default mode, it is not incremented further when $m = 5$. And once the transmission has become successful, the contention window CW is reset to C_{Wmin} for wireless medium access for the next packet transmission.

DCF has two modes of operation after the STA has accessed the wireless channel (transmission state). One is basic DCF (2-way hand-shake) and the other is RTS/CTS DCF (4-way hand-shake). In Basic DCF mode the data packet is transmitted immediately after the channel has been sensed idle for DIFS time period in the start or if in the back-off mode, the back-off counter has become zero. In RTS/CTS DCF mode, the nodes, first, exchange control packets (RTS/CTS) after the channel has been sensed idle for DIFS time period in the start or, if in the back-off mode, the back-off counter has become zero (see Fig. 2.5).

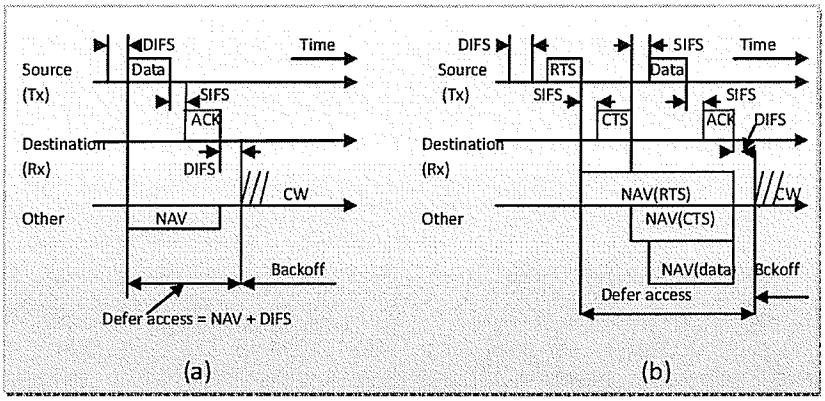

Figure 2.5. *(a) Basic DCF (2-way hand-shake), (b) RTS/CTS DCF (4-way hand-shake).*

2.2.1.1 Characteristics of CSMA/CA

Traditional contention resolution methods like CSMA used in wired network MAC protocols are not effective in wireless networks because the propagation path losses in the wireless medium cause signal power to vary with the distance. In wireless networks, the signal is not heard equally by all STAs which is the requirement of

CSMA. In wireless networks it is difficult to say from the measured carrier power alone whether the channel is usable or not. The matching of the total received interference with a predefined threshold value determines whether the channel is busy (unusable) or idle (usable). Therefore, the carrier sensed at the source can not correctly assess the interference level at the receiver side. The fading strength of the wireless medium creates so-called hidden and exposed terminal problems in the wireless CSMA that causes throughput degradation. However, the propagation path loss in wireless medium facilitates frequency reuse enabling the same channel to be simultaneously used at two well apart locations having no effect of interference from each other.

Hidden terminal: As we can see in Fig. 2.6(a), STA's A and C are out of transmission range of each other, therefore, they are hidden to each other and can transmit to STA B at the same time causing collision. However, RTS/CTS DCF mode resolves this problem to some extent.

Exposed terminal: STA B wants to transmit to STA A but backs off due to carrier sensing of STA C's transmission. STA A being not within the transmission range of STA C could be able to receive B's packet because B's back off may be unnecessary, as there is a chance that C's interference might be too weak to affect reception at A as shown in Fig. 2.6(b).

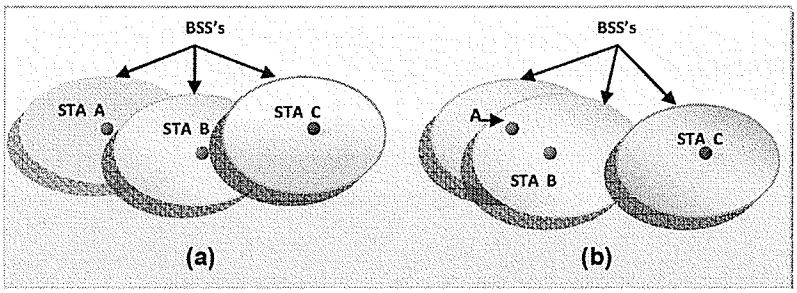

Figure 2.6. *(a) Hidden terminal and (b) exposed terminal.*

The basic DCF may cause substantial throughput degradation, especially in the case when size of the data frame is large, because of collision/loss due to hidden terminals transmission and/or bandwidth wastage due to not transmitting un-necessarily because of the exposed terminals. This problem is overcome to some extent with the RTS/CTS Mode of DCF (i.e., 4-way hand-shake mode). In RTS/CTS mode as soon

as the STA enters the transmission state, it sends a short request to send (RTS = 20 bytes) frame to the intended receiver and waits for SIFS time period. All other STAs that are in the transmission range this sending STA also senses the RTS frame and updates their NAV vector accordingly to defer their transmission for the estimated time slots as envisaged in the RTS frame. In reply, the intended receiver sends a further shorter clear to send (CTS = 14 bytes) frame to the sender STA notifying that it may transmit the data frame. CTS frame also has the estimated time of channel capture, so the STAs which are not in the transmission range of the sender and has not updated their NAV vector but are in the transmission range of the receiving STA updates their NAV vectors accordingly. As soon as the sending STA receives that CTS frame, it transmits the comparatively large data frame.

Though, RTS/CTS frames also face the hidden/exposed terminal problem, but as these are very short frames, the performance of DCF in RTS/CTS mode is not affected severely as compared to the basic DCF mode where the collisions of large size data frames (up to 2346 bytes) degrades the performance substantially.

However, the RTS/CTS mode is not always beneficial, especially in case when data frame size is small like in case of dot11Fragmentation of large data frame into small size fragments. In this case the over-head of permanently using RTS/CTS is large enough to be compared with the collision/loss over-head of small fragment frames in case of the basic DCF mode. So, there is always a trade-off between using basic DCF and RTS/CTS DCF which is controlled by the dot11RTSThreshold attribute. According to this, each STA can be configured to use RTS/CTS either always, or never, or only in the case of a data frame/fragment whose size exceeds an optimally defined data frame threshold size.

A STA that is not configured to initiate RTS/CTS mechanism will still sense the RTS from other STAs, and will update its NAV vector according to the duration/ID field of the sensed RTS frame and will reply by transmitting CTS frame if the RTS is destined to it.

2.2.2 Point Coordination Function (PCF)

PCF is implemented on top of DCF and uses a centralized contention-free polling access method to facilitate real-time services. It is carried out by a funtion called

point coordinator (PC) incorporated in the access point (AP). It performs polling for stations that are capable of being polled. Before a PCF polling cycle starts, AP contends with other stations in DCF, but AP has to wait for a PIFS period, which is shorter than DIFS. In order to prevent stations from starvation that are not allowed to send during the contention-free period (CFP), the contention period is long enough to transmit one maximum MAC protocol data unit (MMPDU). To fully utilize the polling mechanism in CFP, the length of CFP is kept as long as possible. The 802.11 standard does not specify a mechanism for determining the relative lengths of CFP and CP. However, while in PCF mode, the beacon interval must allow at least one DCF data frame to be transmitted in the CP.

When a station associates to an AP, the AP assigns an association id (AID) to it and puts it in the polling ready queue (polling queue) in the order of its AID. The PC maintains the polling list that specifies the order in which stations are polled. The PC polls the stations in a round-robin fashion. If there is no pending data for transmission, the station either responds with a null frame containing no payload or does not respond at all. If the AP does not get a response from the station, then the AP may re-poll the station after PIFS interval instead of SIFS interval, which is the normal interval between any two polls. If the CF terminates before all stations have been polled, the polling queue is resumed at the next station in the following CF cycle. Fig. 2.7 shows a super-frame (beacon interval) that includes CFP and CP. In CFP, the AP polls the stations in the polling queue; the down-link and up-link (i.e., from PC to STA and STA to PC) transmissions take place alternately. In the down-link, the PC piggybacks the data frames, if any, along with the CF-Poll frame, and in the up-link, the polled STA piggybacks the data frame, if it has to send, with an ACK frame after a SIFS interval. Before AP can start CFP, it must catch an idle medium in PIFS interval by sending a beacon frame. After the beacon frame, the communication between AP and stations starts. Before the CFP ends, AP sends a CF-End frame, and then CP starts.

All STAs in the BSS (other than the PC) set their NAVs to the $CFPMaxDuration$ value at the nominal start time of each CFP. The maximum value for CFPMaxDuration, in microseconds, when operating with a contention window of a $CWmin$, is calculated as under.

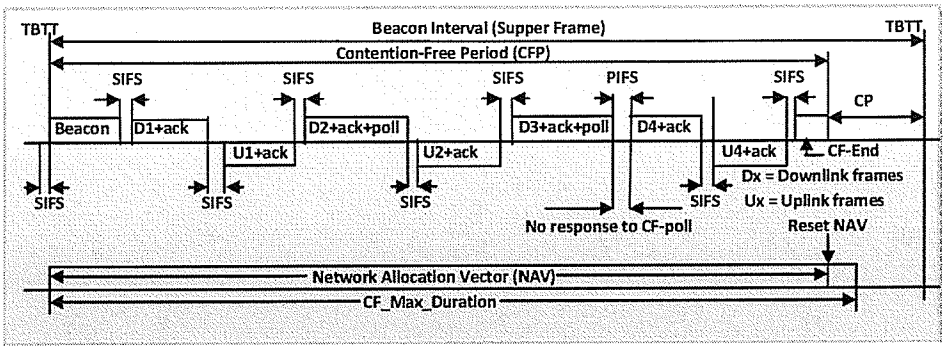

Figure 2.7. *PCF mode: Structure of the super-frame cycle.*

$$\begin{aligned} CFPMaxDuration &= (BeaconPeriod \times DTIMPeriod \times CFPRate) \\ &\quad - [MaxMPDUTime + (2 \times aSIFSTime) \\ &\quad + (2 \times aSlotTime) + (8 \times ACKSize)] \end{aligned} \tag{2.2}$$

$MaxMPDUTime$ is the time to transmit the maximum sized MAC frame, expanded by the WEP, plus the time to transmit the PHY preamble, header, trailer, and expansion bits, if any. This allows sufficient time to send at least one data frame during the CP. Non-CF-Pollable or unpolled CF-Pollable STAs acknowledge frames during the CFP using the DCF ACK procedure.

2.2.2.1 CFP Structure and Timing

A CFP starts with a Beacon frame that contains a DTIM element. The CFPs occurs at a defined repetition rate in a synchronized fashion with the beacon interval. The PC generates CFPs at the contention-free repetition rate (CFPRate) defined as a number of DTIM intervals and is communicated to other STAs in the BSS in the $CFPPeriod$ field of the CF parameter set element of Beacon frames as shown in the Fig. 2.8.

CFP is controlled by the PC, and its maximum or the actual duration is not constrained to be the multiple of the beacon interval. $CFPDuration$ field in the CF Parameter set element is non-zero in case of CFP, and it will be zero in case of CP. The PC may terminate any CFP at or before the $CFPMaxDuration$ on the basis of

the available traffic and the polling list size. Since the beacon transmission may be delayed due to busy medium, the CFP is shortened accordingly.

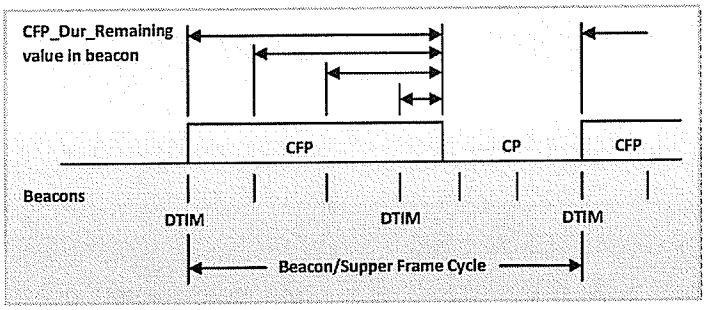

Figure 2.8. *Beacons and CFPs.*

2.2.2.2 PCF Fundamental Access

At the beginning of each CFP, the PC senses the medium to be idle for a PIFS period and broadcast a beacon frame to all the associated STAs in the BSS containing the CF parameter set element and a DTIM element. After the first beacon frame, the PC waits for a SIFS period, and then may transmit a data frame, or a CF-Poll frame, or a Data+CF-Poll frame, or a CF-End frame. In case if there is nothing to send in the CFP, the PC transmits a CF-End frame immediately after the initial beacon. If an error-free frame receiving STA is pollable during the CFP, it responds after a SIFS period, and if not CF-pollable, it always responds with an ACK frame.

2.2.2.3 NAV Operation during the CFP

Each non-PC STA, at each target beacon transmission time (TBTT), preset its NAV to the $CFPMaxDuration$ value which is obtained from the CF parameter set element in beacons from the PC. And each non-PC STA updates its NAV using the $CFPDurRemaining$ value in the error-free CF parameter set element of the beacon frame including the $CFPDurRemaining$ values that are received from other overlapping BSSs. A STA joining a BSS operating with a PC shall use the information in the $CFPDurRemaining$ element of the CF parameter set of any received beacon or probe response frames to update its NAV prior to initiating any transmissions. The

PC shall transmit a CF-End or CF-End+ACK frame at the end of each CFP. A STA that receives either of these frames, from any BSS, shall reset its NAV.

2.2.2.4 Contention-Free Polling List

The PC maintains a polling list to enroll all CF-Pollable STAs whether or not the PC or the CF-Pollable STA has any data frame to send. The polling list is logical which is not exposed outside of the PC. During the CFP, the PC issues polls to a subset of the STAs on the poling list in ascending order of AID values. A STA indicates its CF-Pollability using the CF-Pollable subfield of the capability information field of association request and re-association request frames. If a CF-Pollable STA does not want to be placed in the polling list, like in the power-save mode, it performs association with both the CF-Pollable subfield false and the CF-Poll request subfield true. For removal from the polling list, the STA performs a re-association.

2.2.2.5 Multi-Rate Support

IEEE 802.11 supports multi-rate PHYs capabilities for dynamic rate switching to improve the performance. Under this provision, all the control frames, and multi-cast and broadcast frames are transmitted at one of the *BSSBasicRateSet* or at one of the rates in the PHY mandatory rates set so they will be understood by all STAs. However, the data and/or management MPDUs with a unicast immediate address are sent on any data rate selected by the rate switching mechanism, not defined by the legacy 802.11 standard, but is supported by the destination STA.

2.3 QoS Limitations in IEEE 802.11 MAC Protocol

The wireless MAC layer should provide a good control over the wireless medium access, QoS, and security. Real-time services such as voice and video are time-sensitive and can not tolerate end-to-end delays and jitters beyond certain upper bounds. For example, the interactive voice can normally tolerate end-to-end delay up to 25 msec without echo canceller and 150 msec with echo canceller [6]. Generally, the QoS in

the context of a wireless MAC is characterized either as parameterized QoS or as prioritized QoS as being the ability of a wireless network element to provide some levels of assurance for consistent network data delivery. In parameterized QoS, some quantitative values, such as delay bound, mean data rate, jitter, nominal frame size, maximum service interval etc., are guaranteed at negotiated levels to the admitting service. While prioritized QoS simply provides relative delivery priority among different services without guaranteeing any level of parametric values (e.g., data rate, delay bound, jitter bound). Peculiar characteristics of wireless channels, such as high loss rate, packet disordering, bursts of packet loss, large packet delays and jitter, which also vary with time and space (mobility), put forth great challenges on how to accommodate time-sensitive services that require some minimum level of parameterized-QoS in the wireless packet-switched networks.

2.3.1 QoS Limitations of DCF

Due to bi-exponential back-off mechanism of the CSMA/CA used in the DCF mode of IEEE 802.11a/b/g for wireless medium access, the queuing delays are non-deterministic, and since there is no service priority categorized in the DCF mode, DCF does not promise parameterized or prioritized QoS.

2.3.2 QoS Limitations of PCF

Since the PCF mode is designed with the contention-free wireless medium access approach based on round-robin type polling scheme [3], it can support QoS but the following inherent problems in the PCF degrade its ability to support QoS.

- The centralized communication between any two peers of the same basic services set (BSS) through the access point causes bandwidth wastage, specially, when such traffic increases.
- The non-deterministic beacon delays are barriers in guaranteeing strict delay bounds. Since the PC schedules beacons at TBTT for the CFP, beacons can only be transmitted if the channel becomes idle around the TBTT for at least PIFS time period. PIFS is shorter than the DIFS (used by stations to wait after the channel is idle in the DCF mode) giving priority to the AP over

stations. But, if a station has not finished its transmission at the arrival of TBTT, it will continue its transmission according to the 802.11 legacy standard causing delays in the beacon transmission, and subsequently, causing delays to the priority traffic in the CFP.

- Since PCF can only work in the designated CFP of the beacon interval, the size of the beacon interval becomes dependent on the delay bounds imposed by a real-time service.
- The AP is unable to predict the transmission time of a polled station due to the physical (PHY) rate variation in time and space and frame size variation (between 0 and 2346 bytes), and therefore, guaranteeing strict delay-bound services to other stations in the polling list during the rest of the CFP is not a trivial job.
- The round-robin scheduler polls all the stations in its polling list whether a station has frame(s) to send or not causing substantial bandwidth wastage, especially in case of bursty type of real-time traffic such as voice that always has alternating active and idle periods.

2.4 QoS Enhancements in 802.11 MAC Standard

Some high layer applications such as video, audio, email, and data transfer have different service requirements in terms of bandwidth, delay, jitter, and packet loss guarantees. Since the basic 802.11 MAC standard does not promise strict QoS requirements of QoS-hungry services, QoS support for IEEE 802.11 becomes critical for its success, specially, in real-time applications. For this purpose, a QoS-enhanced version, named IEEE 802.11e, has been developed that defines a new hybrid coordination function (HCF). HCF concurrently exists with basic DCF/PCF for backward compatibility and exploits both controlled contention-free and contention-based channel access approaches in the single channel access protocol.

At present, there are two service based approaches, integrated services ($IntServ$) and differentiated services ($DiffServ$), to enhance QoS in 802.11 MAC. Both have implementation problems [7][8], the solution to which is given in [9] providing $IntServ$ QoS by using $DiffServ$ network segments. Since real-time services need bounds on

bandwidth, delay, jitter, and bit error, we take into account service differentiation schemes that are based on these parameters except the error control schemes which are discussed separately. By manipulating the parameters that define how to access the wireless medium, $DiffServ$ schemes can be supported at MAC level. $DiffServ$ schemes can be mainly based on either per-STA or per-queue, and each can further be sub-based on DCF or PCF. Since queue-based schemes perform more efficiently, we will focus on such schemes.

However, in per-STA DCF-based schemes, there are different approaches like: using different size of CW according to the priority STAs; using different size of DIFS according to the priority STAs; using different size of maximum frame length for STAs according to their priority level. But these are not much efficient as the TCP ACKs have same priorities [10]. Others are: Blackburst scheme [11] requiring a constant access interval for each priority STA and the ability to jam the wireless medium; Virtual MAC (VMAC) scheme [12] that monitors and estimates the QoS DiffServ parameters working virtually in parallel to actual MAC but causes complexities for application and MAC layers interaction; Distributed Fair Scheduling (DIFS) scheme [13] that introduces an initial back-off mechanism before each transmission based on the packet size and the weight/priority of the STA which is different from the actual DCF back-off mechanism; Deng and Chang scheme (DC scheme) that suggests four STA priorities by bifurcating the actual DCF back-off interval each part combined with two IFSs. All these schemes require modification in the actual IEEE 802.11 MAC Standard along with the inherent problems/draw-backs in each scheme.

In per-STA PCF-based approach there is a little work as PCF is not used normally being an optional feature of IEEE 802.11 MAC. However two main works are worth-mentioning: one is Priority-based PCF scheme [14][15] that suggests modification in the actual round-robin type polling scheduler with priority-based one; the other is the Distributed TDMA mechanism which suggests TDMA-like slot times and assignments of slots according to the priority of STAs rather than the polling mechanism of the PCF.

In the DCF-based under Queue-based service differentiation approach QoS enhancement schemes are: Per-flow scheme [16] where the shared node (AP) uses different priorities for different flows but all types of packets are put in the same queue

which causes interferences between priorities slowing down the AP. In [16], solution to such interferences is proposed assigning different queues to different flows in the AP. In case of several queues in one STA, multiple TCP connections are there causing increased internal collisions in one STA.

This problem is resolved in the IEEE 802.11e EDCF [4] (Enhanced-DCF) that supports four queues in one STA and each queue contends for a transmission opportunity (TXOP) to send packets. The problem of EDCF is the static values of CW_{min}, CW_{max}, and back-off function that doe not conform to the dynamicity of the wireless channel conditions. Adaptive EDCF (AEDCF) [17] resolves this problem by providing relative priorities by adjusting the size of the CW of each traffic class taking in to account both application requirements and network conditions. Comparison of AEDCF with EDCF shows that AEDCF reduces more than 50% of collision rate and gives 25% higher good-put than EDCF. PCF sub-based $DiffServ$ based on the Queue-based approach is mainly proposed in the hybrid coordination function (HCF) of the IEEE 802.11e MAC which is implemented in the QoS-enhanced access point (QAP). Among many features of QoS enhancement in IEEE 802.11e, three are worth mentioning: the hybrid coordination function (HCF), direct link protocol (DLP), and block acknowledgement.

2.4.1 Hybrid Coordination Function (HCF) Mechanism

Each QSTA is incorporated with HCF that can work concurrently with the DCF and the PCF for backward compatibility combining functions from both DCF and PCF with some enhanced QoS-specific mechanisms and frame subtypes. Thereby, it allows a uniform set of frame exchange sequences for use in both $IntServ$ and $DiffServ$ QoS-frame transfers during both contention-free and contention periods.

HCF has two sub-functions. One is the Enhanced Distributed Channel Access (EDCA) which is contention-based for prioritized QoS to four queue-based differentiated services called access categories (AC). The other is HCF controlled channel access (HCCA) which provides contention-free medium access for parameterized QoS to eight queue-based differentiated services called traffic streams (TS). A frame arriving at the MAC layer is tagged with a traffic priority identifier (TID) that may have a value from 0 to 15 according to the QoS requirement. If the MAC frame's

TID value is from 0 to 7, it is mapped into any of the four ACs accordingly. If the MAC frame's TID value is from 8 to 15, it is mapped into any of the eight TSs accordingly. Another feature is that HCF allots transmission opportunity (TXOP) time units to stations in both EDCA (called EDCA-TXOP or EDCF-TXOP) and HCCA (called polled-TXOP or HCCA-TXOP) modes according to the ACs priority or TSs requirement. The QAP defines the maximum value of TXOP called $TXOPLimit$.In this way, frame-bursting and block-acknowledgment are possible for efficient use of the bandwidth. However, bursting may result in delay jitter [4].

2.4.1.1 Contention-Based Enhanced Distributed Channel Access (EDCA)

In EDCA, two methods are defined for providing prioritized QoS. One is that the EDCA uses four Access Categories (ACs = four queues) to support eight different User Priorities (UPs) and is designed for contention-based prioritized QoS support. One or more UPs are mapped to the same AC queue as shown in Fig. 2.9

Priority	UP (same as 802.1D user priority)	802.1D designation	802.11e AC	Service Type
Lowest	1	BK (Background)	AC-BK = 0	Background (Best Effort)
	2	Not defined	AC-BK = 0	Background (Best Effort)
	0	BE (Best Effort)	AC-BE = 0	Best Effort
	3	EE (Excellent Effort)	AC-BE = 1	Video Probe
	4	CL (Controlled Load)	AC-VI = 2	Video
	5	VI (Video < 100ms latency and jitter)	AC-VI = 2	Video
Highest	6	VO (Video < 10ms latency and jitter)	AC-VO = 3	Voice
	7	NC (Network Control)	AC-VO = 3	Voice

Figure 2.9. *UP to AC mappings.*

Each AC queue works as independent DCF STA and uses its own back-off parameters as shown in the Fig. 2.10. Instead of DIFS, as in DCF, A new kind of IFS, Arbitration IFS (AIFS), is used in EDCF for different ACs. AIFS is determined by as:

$$AIFS[AC] = AIFSN[AC] \times SlotTime + SIFS \tag{2.3}$$

Here, AIFSN (arbitration inter frame spacing number) is determined (by default)

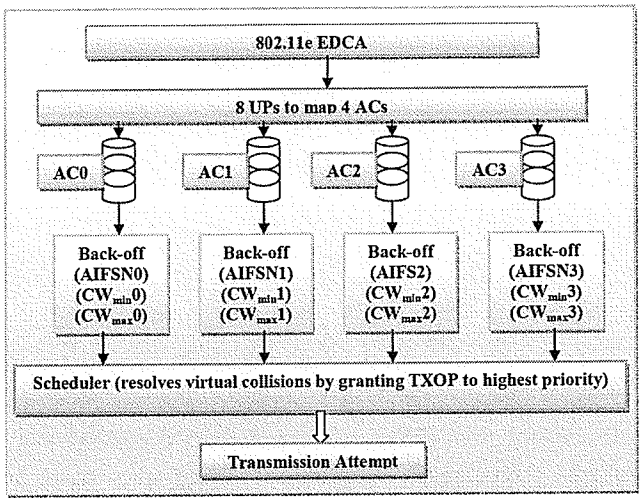

Figure 2.10. *EDCA access mechanism.*

either as 1 for priority queues AC1, AC2, and AC3 to have AIFS value equal to PIFS or as 2 for low priority queue AC0 to have AIFS equal to DIFS. Therefore, in the start, if the channel is found idle for longer than $AIFS + SlotTime$, the frame is transmitted straightaway, or if the channel is found busy, the transmission is deferred till the channel becomes idle and remains idle for a $AIFS+SlotTime$ period. Fig. 2.11 shows the detailed timing diagram of the EDCF scheme. Since the PIFS is smaller than the DIFS, AC1, AC2 and AC3 have higher priority than the AC0 in contending for the wireless channel. In this way, the waiting time at the head of the queue for the high priority queue is to wait for $PIFS + SlotTime$, and for the best effort queue is to wait for $DIFS + SlotTime$.

The other method for providing prioritized-QoS is to use different sizes of CW for different ACs according to the priority, i.e., shorter CW for higher priority AC and vice versa. Internally, in the same QSTA, if the back-off counter of two or more ACs becomes zero at the same time, the scheduler avoids the virtual collision by granting the EDCA-TXOP to the higher priority AC queue, and the other colliding ACs backs off by doubling their CW sizes. In this way, the internal collision rates in the same QSTA are reduced. However, external collision rates between the same-priority ACs in different QSTAs remain still high. $AIFSN[AC]$, $CW\_min[AC]$, $CW\_max[AC]$ and $TXOPLimit[AC]$ are the default values which are determined by the QAP and are announced in the beacon frames, however, the IEEE 802.11e allows to dynam-

ically change these parameters depending on the network conditions but does not define how to change these dynamically. EDCA also has the provision to improve the throughput performance by packet bursting, i.e., transmitting more than one frame in the same EDCF-TXOP with SIFS in between them and being under the limit of $TXOPLimit$ bound that is determined by the QAP. Along with the packet-bursting, burst-acknowledgements are also allowed to reduce the network overhead and increase the throughput. However bursting may result jitter.

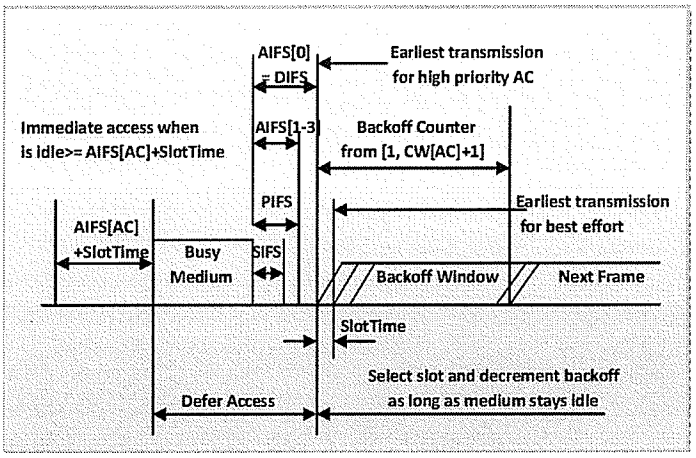

Figure 2.11. *EDCF channel access and IFSs relationship.*

EDCA back-off procedure: Each EDCA function (related to each access category AC) maintains a state variable from 0 to $CW[AC]$ where $CW[AC]$ varies between $CW\_min[AC]$ and $CW\_max[AC]$. On successful transmission of a frame by a specific EDCAF, which is evident from successful reception of a CTS, ACK, or a *BlockAck* or *ACK* frame in response to an RTS, a unicast MPDU or *BlockAck*, a *BlockAckReq* frame respectively, or the successful transmission of a multicast frame or a frame with no Ack policy, the $CW[AC]$ is set to $CW\_min[AC]$.

If any of the following events occurs, EDCAF of a specific AC will provoke backoff.

1. There is a frame in that AC to be transmitted but the channel is sensed busy and the back-off timer for that AC is zero.
2. The final transmission in that specific AC was successful.
3. There is a transmission failure of a RTS/CTS, MPDU/ACK, or a BlockAck-Req/BlockAck/Ack frame.

4. Two or more EDCAFs in the same QSTA are granted a TXOP at the same time and thus their transmission attempts are collided with each other.

In case of item 1. above, $CW[AC]$ is unchanged, and in case of item 2., $CW[AC]$ is reset to $CW\_min[AC]$. And in case of item 3. and item 4., $CW[AC]$ is updated as:

- If the short retry counter, $QSRC[AC]$, or the long retry counter, $QLRC[AC]$, for the QSTA has reached to $dot11ShortRetryLimit$ or $dot11LongRetryLimit$ respectively, $CW[AC]$ is reset to $CW\_min[AC]$.
- If $CW[AC]$ is less than $CW\_max[AC]$, it is set to a the value $(CW[AC] + 1) \times 2 - 1$, and if it is equals to $CWmax[AC]$, $CW[AC]$ remains unchanged for the remainder of any retries.

All backoff slots follow an $AIFS[AC]$ period during which the medium is idle for the $CW\_min[AC]$ period, or these follow an $EIFS - DIFS + AIFS[AC]$ period during which the medium is idle for the duration of $EIFS - DIFS + AIFS[AC]$ period following detection of a frame that was not received correctly.

EDCA retransmit procedure: A QSTA maintains a short retry counter, $QSRC[AC]$, and a long retry counter, $QLRC[AC]$, for each AC which are initiated with a zero value. The QSTA also maintains short and long retry counters, initially set to zero, for each MSDU or MMPDU that belongs to a transmission category (TC) requiring ACK. For each transmission failure of a MSDU or MMPDU frame less than or equal to the $dot11RTSThreshold$, the short retry count (for TC) and the $QSRC[AC]$ is incremented and are reset when such a MAC frame succeeds. And for each transmission failure of a MSDU or MMPDU frame greater than $dot11RTSThreshold$, the longt retry count (for TC) and the QLRC[AC] is incremented and are reset when such a MAC frame succeeds. Retries for failed transmissions continue until the short retry count is equal to $dot11ShortRetryLimit$ or until the long retry count is equal $dot11LongRetryLimit$. MSDU or MMPDU is discarded if either of these limits is reached. Each QSTA maintains a transmit MSDU timer that is initiated at the time when the MSDU is passed to the MAC, and if this timer exceeds the appropriate value in the $dot11EDCATableMSDULifetime$ attribute of MIB, the relevant MSDU or any remaining un-transmitted part thereof will be discarded.

2.4.1.2 Contention-Free HCF Controlled Channel Access (HCCA)

HCCA provides parameterized QoS support and uses a central coordinator, called hybrid coordinator (HC), like the point coordinator (PC) in the PCF mode of nQoS. But HCCA operates under quite different rules than under PC, although it may also implement PCF optionally. The main difference is that the HCF frame exchange sequence may be used among QSTAs during both CP and CFP, i.e., HCF can start controlled channel access mechanism in both CFP and CP intervals whereas PCF is allowed only in the CFP. The IEEE 802.11e beacon interval is composed of alternating modes of optional CFP and CP as shown in the Fig. 2.12. During the CP, a new contention-free period, called controlled access phase (CAP), is used for HCF-controlled channel access. 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. And the remaining time of the CP can be used by the EDCF. It is a flexible contention-free scheme that makes CFP and PCF useless, and thus, optional in the IEEE 802.11e standard. The major benefit of CAP is that the HCF beacon interval size can be independent of targeted delay bounds of real-time services. For example, in audio traffic, the maximum latency should not be more than 20 msec using PCF mode, so the beacon interval in this case should not be more than 20 msec because the fixed portion of the CP forces the audio traffic to wait longer for the next poll. However, the HCF-controlled channel access can increase the polling frequency, and, hence, the polling overhead (POH), by initiating the CAP at any time guaranteeing the delay bound with any size of beacon interval that can be used to decrease the increased overheads of shorter beacon intervals. Also, the beacon delay problem is resolved in PCF as in HCF a QSTA is not allowed to transmit a frame if the transmission can not be finished before the next TBTT.

The HCCA guarantees parameterized-QoS through setting up virtual connection called traffic stream (TS) after negotiation of traffic specification parameters (TSPEC) such as mean data rate, nominal frame size, maximum service interval, delay bound, delay jitter etc., between the AP and the mobile station. This negotiation of TSPECs determines the length of the polled-TXOP. The scheduler in the station allocates the polled-TXOP to its TSs queues as per priority. A simple round-robin scheduler is proposed in the IEEE 802.11e that uses these mandatory TSPEC param-

eters, i.e., nominal MAC frame size, maximum service interval or delay bound and mean data rate. Here, the maximum service interval requirement of each TS is based on the maximum time interval between the start of two successive TXOPs, and if it is small, it can provide low delay but more CF-Poll frames. In case if different TS has different maximum service interval requirements, the scheduler selects the minimum value of all maximum service interval requests of all allowed streams for scheduling. IEEE 802.11e does have the provision of using an admission control algorithm to accept new TS into the QBSS. When TS is setup, the QAP provides QoS by allocating the required bandwidth to the TS.

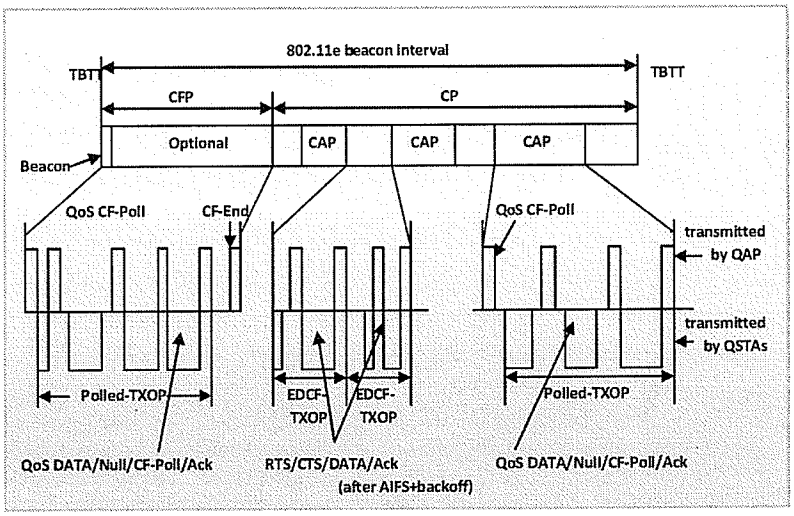

Figure 2.12. *A typical 802.11e HCF beacon interval.*

The other difference of HCCA with the PCF is that the single polled TXOP granted by the HC to a non-AP QSTA may be used for transmission of multiple frame exchange sequences denoted by the contention-free burst (CFB) subject to the limit on $TXOPLimit$. All other QSTAs set their NAVs with the $TXOPLimit$ plus a slot time. All STAs inherently observe the NAV rules of the HCF and all QSTAs respond to QoS(+)CF-Poll frames received from HC with the address-1 field matching with their own addresses. If a polled-QSTA does not have frames to send to QAP, it just sends a QoS-Null frame to the QAP enabling it to poll to another QSTA.

CFP generation: The HC may work as PC using CFP with the restriction that such CFP will always end with a CF-end frame. Since the HC may also issue QoS(+)CF-

Poll frames to the associated QSTAs during the CFP and it can also issue polled-TXOPs, by sending QoS(+)CF-Poll frames during the CP, the HC is not bound to use CFP only for QoS data transfers.

CAP generation: The HC senses the wireless medium idle for a PIFS time period whenever it needs to start a CFP or a TXOP in CP, and it transmits the first frame of any permitted frame exchange sequence with the duration value set to cover the CFP or the TXOP. A beacon frame is the first in a CFP after a TBTT. After the last frame of all other non-final frame exchange sequences during a TXOP, the holder of the TXOP shall wait for a SIFS period before transmitting the first frame of the next frame exchange sequence, and, in the meanwhile, the HC may sense the channel idle for PIFS duration to capture it after the TXOP, otherwise a CAP ends automatically as shown in the Fig. 2.13.

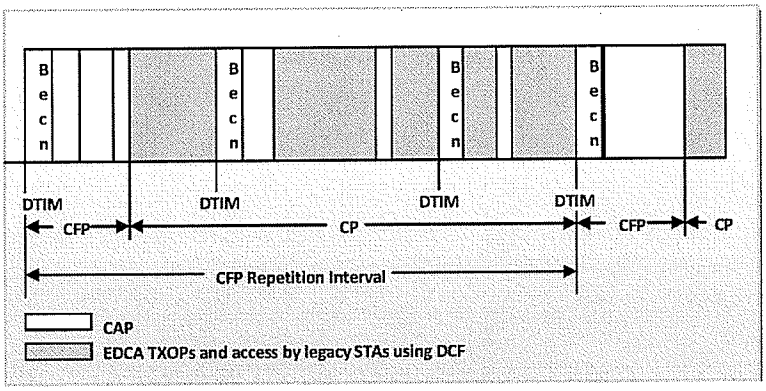

Figure 2.13. *CAP/CFP/CP periods.*

Recovery from the absence of an expected reception: Since the NAVs of all the associated QSTAs are already set up by the HC transmissions, recovery rules from the absence of an expected reception are different from EDCA. The recovery rules for the multiple frame transmission are different because a non-AP STA may always be hidden and may have not set its NAV due to the transmission by another non-AP STA. However, since the HC is collocated with the QAP, QAP may recover using these rules even if the recovery is from the absence of an expected reception.

If the beginning of reception of an expected response, as detected by the occurrence of PHY-CCA.indication(busy) primitive at the QSTA that is expecting the response, does not occur during the first slot time following SIFS, then:

- If the transmitting STA is the HC, it may initiate recovery by transmitting at a PIFS after the end of the HC's last transmission only if PHY-CCA.indication primitive is clear.
- If the transmitting STA is a non-AP QSTA, it shall initiate recovery by transmitting at a PIFS after the end of the last transmission, if the polled TXOP limit is greater than 0 and at least one frame (re)transmissions can be completed within the remaining duration of a nonzero polled TXOP limit.

If the transmitted frame is not of type QoS (+)CF-Poll and the expected response frame is not received correctly, regardless of the occurrence of the PHY-RXSTART.indication primitive, the QSTA may initiate recovery following the occurrence of PHY CCA.indication(idle) primitive so that a SIFS time interval occurs between the last energy on the air and the transmission of the recovery frame. If the QSTA sends a frame to HC with a duration/ID covering only the response frame, the HC assumes that QSTA is terminating its TXOP. In this case the HC may initiate other transmissions or allow channel to go into the CAP.

The polled QSTA sends a QoS(+)Null frame if it has no queued frames or if the MPDUs it has are too long to send within the allotted TXOP limit. In the former case, the QoS control field of the null frame is set to size 0 for any TID with the duration/ID set to one ACK frame transmission time plus one SIFS interval, and in the later case, the QoS control field of the null frame contains the TID and TXOP duration or a nonzero queue size required to send the ready MPDU. And the HC combines this queue size information with the rate of the received null frame to determine the required size of the requested TXOP.

The unused portion of the TXOPs after the transmission of the final frame and its expected ACK response frame is returned back to the HC as final frame with the Ack policy subfield set to normal Ack, and in case of no frame to send to the HC, the QSTA sends a QoS null frame to the HC with the queue size subfield in the QoS control field set to 0. If there is no enough time in the unused TXOP to transmit either QoS null frame or the frame with the duration/ID field covering only the response frame, then the QSTA will cease the control of the channel. If the beginning of the reception of an expected ACK response frame to the final frame does not occur, detected as the nonoccurrence of PHYCCA.indication(busy) primitive at

the non-AP QSTA that is expecting the response during the first slot time following SIFS, the non-AP QSTA shall retransmit the frame or transmit a QoS null frame, with the Ack policy subfield set to normal Ack and the Queue size subfield set to 0, after PIFS from the end of last transmission, until such time that it receives an acknowledgment or when there is not enough time remaining in the TXOP for sending such a frame, and if PHYCCA.indication(busy) primitive occurs, it is assumed as an indication that the channel control has successfully transferred and no further frames shall be transmitted by the QSTA in that TXOP, even though the ACK frame from the HC is incorrectly received. This is to avoid the situation where the HC may not receive the frame and may result in an inefficient use of the channel. Here, it should be noted that the use of the PHY-CCA.indication(busy) primitive is used to determine the control of the channel, and not to determine the success or failure of the transmission.

TXOP structure and timing: A QoS data frame of subtype that includes CF-Poll contains a TXOP limit in its control field which is protected by the NAV set by the duration field of the frame containing the QoS(+)CF-Poll function as shown in Fig. 2.14. Within a polled TXOP, a QSTA may initiate transmissions of one or more frame exchange sequences each separated by a SIFS. The QSTA shall not initiate transmission of a frame unless the transmission and any acknowledgment or other immediate response expected from the peer MAC entity are able to complete prior to the end of the remaining TXOP duration. TXOP includes all transmissions including the response frames and the HC accounts for these while setting the TXOP limit.

A TXOP or transmission within a TXOP shall not extend across TBTT, $dot11CAP\text{-}Limit$, $dot11MaxDwellTime$ (if using an FH PHY), or $dot11CFPMaxDuration$ (if during CFP). The HC ensures that the full duration of any granted TXOP meets these requirements so that non-AP QSTAs may use the time prior to the TXOP limit of a polled TXOP without checking for these constraints. Thus, all decisions concerning what MSDUs and/or MMPDUs are transmitted during any given TXOP are made by the QSTA that holds the TXOP.

NAV operation during a TXOP: The HC sets its own NAV during the TXOP of a QSTA, however, it may reclaim it if the QSTA is not using it at all or it has ended earlier. Whether in CFP or CP, and if there is no more QSTAs to poll and no more

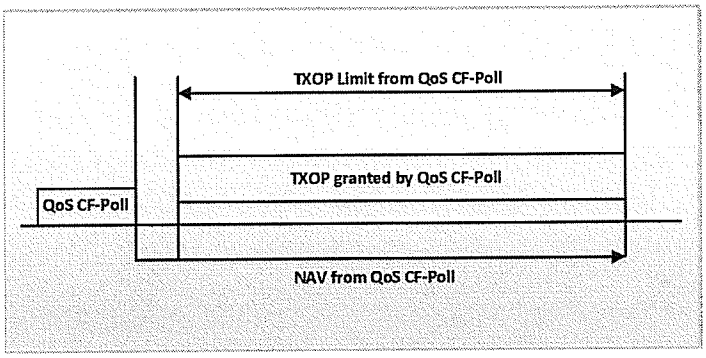

Figure 2.14. *Polled TXOP.*

data, management, *BlockAckReq*, or *BlockAck* frames to send, the HC may reset the NAVs of all QSTAs in the QBSS by sending a QoS CF-Poll frame with the RA matching its own MAC address and with the duration/ID field set to 0. When the QAP contains a PC, during the CFP, it may reset the NAVs of all STAs within the QBSS by sending a CF-End frame regardless of how the NAVs have been originally set. Subsequently, the QSTA receiving the CF-End frame containing the BSSID shall reset its NAV to 0. A QSTA that receives the QoS(+)CF-Poll frame with the MAC address in the address 1 field matching the HC's MAC address and the duration/ID field value equal to zero, shall also reset its NAV to 0. However, before resetting, if the QSTA receives a RTS frame with the RA address matching the MAC address of the QSTA and the MAC address in the TA field in the RTS frame matches the saved TXOP holder address, the QSTA, without regard/resetting its NAV to 0, shall send the CTS frame after a SIFS period. A QSTA responds with an ACK or QoS+CF-Ack frame without regard of its NAV, and it accepts a polled-TXOP by initiating a frame exchange sequence without regarding to its NAV.

HCCA transfer rules: A QSTA, after obtaining a TXOP, is bound to finish its transmissions within the TXOP limit, however it may transmit or receive any kind of frame, i.e., QoS data frames, management frames etc. but MSDUs may be fragmented in order to fit within TXOPs. Only the HC can send QoS(+)CF-Poll frame. And QoS(+)Data frame can only be sent in response to a QoS(+)CF-Poll frame. A non-AP QSTA that wishes to inform the HC of queue status may use the QoS null frame indicating the TID and the queue size or TXOP duration request.

In both CFP and CP, QSTAs always responds to QoS data frames having the Ack policy subfield in the QoS control field set to normal Ack with an ACK frame, unless the acknowledgment is piggybacked in which case it shall use a QoS +CF-Ack frame. Piggybacked frames are allowed only in CFP or within TXOPs initiated by the HC. ***TXOP requests:*** Non-AP QSTAs may send TXOP requests to the HC, with the TXOP duration requested or queue size subfield value and TID subfield value indicated therein, during the polled-TXOPs or EDCA-TXOPs using the QoS control field in the QoS data or QoS null frame directed to the HC. Even if the value of TXOP duration requested subfield or queue size subfield in a QoS data frame is zero, the HC shall continue to poll according to the negotiated schedule.

2.4.2 Admission Control at the HC

Admission control is used to regulate the available bandwidth resources. In case of IEEE 802.11 networks, admission control is required when QSTA needs a guaranteed/deterministic channel access time. The HC administers the admission control. For two access modes of IEEE 802.11e, i.e., EDCA and HCCA, there are two distinct admission control mechanisms. Admission control, in general, depends on vendors' implementation of the scheduler, available channel capacity, link conditions, retransmission limits, and the scheduling requirements of a given stream.

2.4.2.1 Admission Control in EDCA

QAPs supports admission control procedures, at least to the minimal extent of advertising that admission is not mandatory on its ACs by using ACM (admission control mandatory) subfields in the EDCA parameter set element to indicate whether admission control is required for each of the ACs. Even though the limit parameters of the CWmin, CWmax, AIFS, and TXOP may be adjusted over time by the QAP, the ACM bit remain static throughout the life time of the BSS. The QSTA sends ADDTS request frame to the HC in order to request admission of traffic in any direction, i.e., uplink, downlink, direct, or bidirectional, using an AC that requires admission control. The ADDTS request frame shall contain the UP associated with the traffic and shall indicate EDCA as the access policy, and QAP, in turn, associates the received UP with the appropriate AC as per UP-to-AC mapping rules. The QSTA may trans-

mit un-admitted traffic for such ACs for which the QAP does not require admission control, and if a QSTA desires to send data with out admission control using an AC that mandates admission control, it uses the EDCA parameters corresponding to lower priority requiring no admission control.

Procedures at the QAP: The QAP, on receipt of an ADDTS request from a QSTA, determines locally whether to accept or deny the request using any appropriate algorithm. In case if the QAP determines to accept the request, it will derive the medium time from the information conveyed in the TSPEC element in the ADDTS request frame using any algorithm like K.2.2, and will transmit to the concerned QSTA a TSPEC element contained in the ADDTS response frame specifying therein the medium time.

Procedure at non-AP QSTAs: The EDCAF at the QSTA maintains two variables: admitted-time and used-time, which are set to 0 at the time of (re)association, and these can be used to explicitly determine the medium time for the AC associated with the specified UP. To make such request, the QSTA shall transmit a TSPEC element contained in an ADDTS request frame with the following fields specified (i.e., nonzero): nominal MSDU size, mean data rate, minimum PHY rate, inactivity interval, and surplus bandwidth allowance. However, the medium time field is not used in the request frame and shall be set to 0. The QSTA, on receipt of the TSPEC element contained in the ADDTS response frame, computes the *admitted_time* for the specified EDCAF as:

$$admitted\_time = admitted\_time + dot11EDCAAveragingPeriod \times (medium\ time\ of\ TSPEC) \tag{2.4}$$

Two parameters, used_time and admitted_time in units of $32microsec$, are defined to describe the behavior of the non-AP QSTA. The QSTA updates the value of the usded_time as:

At $dot11EDCAAveragingPeriod$ second intervals:

$$used\_time = max\left((used\_time - admitted_time), 0\right) \tag{2.5}$$

After each successful or unsuccessful MPDU (re)transmission attempt:

$$used\_time = used\_time + MPDUExchangeTime \tag{2.6}$$

The $MPDUExchangeTime$ equals the time required to transmit the MPDU sequence. For the case of an MPDU transmitted with normal Ack policy and without RTS/CTS protection, this equals the time required to transmit the MPDU plus the time required to transmit the expected response frame plus one SIFS. If the *used_time* value reaches or exceeds the *admitted_time* value, the corresponding EDCAF shall no longer transmit using the EDCA parameters for that AC as specified in the QoS parameter set element. However, a non-AP QSTA may choose to temporarily replace the EDCA parameters for that EDCAF with those specified for an AC of lower priority, if no admission control is required for those ACs. In case if the QSTA is running short of the *admitted_time* due to the slowed down data rates caused, possibly, due to worsening PHY conditions, it may request to the QAP for more *admitted_time*, and, at the same time, it may downgrade the EDCA parameters for that AC for short intervals for sending some of the traffic at the admitted priority and some at the un-admitted priority, while waiting for the response to and additional admission request.

2.4.2.2 Admission Control in HCCA

The HC, while providing controlled channel access to non-AP QSTAs, is responsible for granting or denying polling service to a TS based on the parameters in the associated TSPEC. The HC should not tear down a TS, except at the expiry of the inactivity timer, or initiate a modification of TSPEC parameters of an admitted TS unless requested by the STA. The polling service based on admitted TS provides a *guaranteed channel access* from the scheduler in order to have its QoS requirements met. Though, it is not so easy to guarantee QoS due to specific wireless medium characteristics, the behavior of the scheduler, under controlled environment (e.g., no interference), can be observed and verified to be compliant to meet the service schedule , that is:

- The scheduler is implemented so that, under controlled operating conditions, all STAs with admitted TS are offered TXOPs that satisfy the service schedule.

- If a TS is admitted by the HC, then the scheduler shall service the non-AP QSTA during an SP. An SP is a contiguous time during which a set of one or more downlink unicast frames and/or one or more polled TXOPs are granted to the QSTA. An SP starts at fixed intervals of time specified in service interval field. The first SP starts when the lower order 4 bytes of the TSF timer equals the value specified in service start time field. Additionally, the minimum TXOP duration shall be at least the time to transmit one maximum MSDU size successfully at the minimum PHY rate specified in the TSPEC. If maximum MSDU size is not specified in the TSPEC, then the minimum TXOP duration shall be at least the time to transmit one nominal MSDU size successfully at the minimum PHY rate. The vendors are free to implement any optimized algorithms, such as reducing the polling overheads, increasing the TXOP duration, etc., within the parameters of the transmitted schedule.

The QAP schedules the transmissions in HCCA TXOPs and communicate the service schedule to QSTA, and, in case if the QSTA had set the Aggregation field in its TSPEC request, the HC provides an aggregate service schedule. Anyhow, the service schedule is communicated to the QSTA in a schedule element contained in the ADDTS response frame and the modified service start time in it does not exceed the requested service start time by more than one maximum service interval (SI). The HC uses the maximum SI for the initial scheduling only as there may be situations that HC may not be able to service the TS at the scheduled timing, due to an EDCA or DCF transmission or other interferences interrupting the schedule. The Service Interval field value in the Schedule element shall be greater than the minimum SI. The service schedule may be updated at any time by the HC by sending a schedule element in a schedule frame and it is considered as in effect when the HC receives the ACK frame for the schedule frame. The cumulative TXOP duration, during any time interval $[t1, t2]$ including the interval that is greater than the specification interval, is greater than the time required to transmit all MSDUs of nominal size arriving at the mean data rate for the stream over the period $[t1, t2 - D]$ where D is the parameter set to the specified SI in the TSPEC, and if maximum SI is not specified, then D is set to delay bound in the TSPEC.

If the minimum PHY rate is present in the TSPEC field in the ADDTS response, the HC uses it in calculating TXOPs, otherwise the HC uses an observed PHY rate. QSTAs may have an operational rate lower than the minimum PHY rate due to varying conditions on the channel for a short time, and it may be still able to sustain the TS without changing the minimum PHY rate in the TSPEC.

During the TSPEC negotiation, the specification of a minimum set of parameters, i.e., mean data rate, nominal MSDU Size, minimum PHY rate, surplus bandwidth allowance, and at least one maximum service interval and delay bound in the ADDTS request frame, is needed for the scheduler to determine a schedule for the stream to be admitted. And in the ADDTS response frame, these parameters are mean data rate, nominal MSDU size, minimum PHY rate, surplus bandwidth allowance, and maximum service interval and shall be nonzero when a stream is admitted, and if not non-zero in the ADDTS request frame, the HC may replace these unspecified parameters with non-zero values and admit the stream indicating about it to the QSTA through the ADDTS response frame, or it may reject it. If both maximum SI and delay bound are specified, the HC may use only the maximum SI. If any other parameter is specified in the TSPEC element, the scheduler may use it when calculating the schedule for the stream. The HC may also use the UP value in the TS Info field for admission control or scheduling purposes.

2.4.3 IEEE 802.11e Direct Link Protocol (DLP)

IEEE 802.11e MAC protocol introduces a direct link protocol (DLP) for setting up direct communication between the QSTAs [16] to increase the bandwidth utilization. The sender first sends its DLP request to the receiver through the QAP including its supported rates and some other information. Once the receiver acknowledges this request, a direct link is setup. If there is no frame transmission between these two directly linked QSTAs for a duration of $DLPIdleTimeout$, the direct link is disabled.

2.4.4 IEEE 802.11e Block Acknowledgement

Instead of simple SW-ARQ, as in 802.11b, more effective mechanism called selective repeat ARQ (SR-ARQ) is used in the IEEE 802.11e. In this scheme a group of data

frames can be transmitted one by one with SIFS interval between each group, but the receiver sends single ACK frame acknowledging whole the block and indicating how many packets have been received correctly. Two sub-kinds of block acknowledgements are proposed in the IEEE 802.11e. One is immediate *BlockAck* where the sender sends a BlockAck-request frame after transmitting a group of data, and the receiver has to send back the *BlockAck* after a SIFS interval, and if the sender receives the *BlockAck* frame, it retransmits the un-acknowledged frames mentioned in the *BlockAck* frame either in an other group or separately. Immediate *BlockAck* is very useful for applications requiring high bandwidth and low latency but very difficult to generate *BlockAck* in SIFS interval. The other is the delayed *BlockAck* which does not need strict timing limit and useful for applications that can tolerate moderate latency. In delayed *BlockAck* scheme the receiver sends a normal ACK frame first in response to a BlockAck-request and then can send the *BlockAck* at any other time less than the delayed *BlockAckTimeout*.

2.4.5 Error Control in IEEE 802.11 Wireless LANs

Error control based enhancement schemes are implemented at both transport and link layers. Since wireless links are more prone to data loss, error recovery / error control at the link level becomes a must. Two error recovery schemes are basically used. One is ARQ (Automatic Repeat request) and the other is FEC (Forward Error Correction). In IEEE 802.11 MAC stop and wait ARQ (SW-ARQ) scheme is used in which the sender, before starting to transmit the next packet, stops and wait for the ACK from the receiver for the last packet it sent to the receiver. SW-ARQ, though easy to implement, is not QoS supportive as it causes bandwidth wastage in waiting for the ACK. Solution to this problem is the selective repeat ARQ (SR-ARQ) scheme in which packets are transmitted continuously without waiting for their ACKs. The receiver acknowledges each successfully received packet, and the sender retransmits only that packet the ACK of which is not received in the time limit and then resumes transmission from where it had left this scheme is very much bandwidth effective but very complicated to implement and needs buffer resources and re-sequencing overhead. Go-Back-N ARQ (GBN-ARQ) scheme combines the features of SW-ARQ and SR-ARQ. The sender transmits packets continuously but the receiver acknowledges on

that packet which it receives in sequence and discards all others which are not in sequence, thus requiring no buffer re-sequencing arrangement. However, the sender needs to re-transmit all the packets it sent before the un-acknowledged packet which is also an extra overhead.

Since error correction with the ARQ is done at the sender's end, it leads to unacceptable variable delays to real-time services. Forward error correction scheme (FEC) helps recover/add the redundant/erroneous bits at the receiver's end that leads in maintaining a homogenous throughput and bounded time delays. Since wireless channel is high speed error prone media, FEC decoding error rate increases with channel high error rate and needs a long FEC code creating high transmission overhead. So, in order to overcome the individual drawbacks of ARQ and FEC, hybrid FEC-ARQ schemes are developed combining their features. One is Type-I Hybrid FEC-ARQ scheme which uses parity bits for both error detection and error correction in each packet. Error is corrected only if the number of erroneous bits is within the error correction capability of the code, otherwise retransmission of the packet is requested. This process continues until the packet is successfully received or the maximum number of retransmissions has been reached. The drawback of Type-I Hybrid FEC-ARQ is that the uncorrectable packets are discarded straightaway without regard that they might contain useful information, and this aspect is covered in the Type-II Hybrid FEC-ARQ. Any how, so far, the error correction scheme implemented in the IEEE 802.11 is SW-ARQ, and these hybrid error correction schemes are planned to be used in the next-generation high-speed wireless networks.

2.5 Chapter Summary

In chapter 2 we have presented a comprehensive study of IEEE 802.11 Wireless LAN MAC standard with special reference to its important architectural aspects, its limitations on QoS for real-time services, QoS requirements and a basic survey of QoS enhancement aspects (for a detailed survey, see next chapters) and efforts done so far within the standard and out of the standard. After discussing the fundamental architecture, special and detailed discussion is focused on the new version IEEE 802.11e that has outlined rules for both prioritized-QoS and parameterized-QoS en-

hancement. Since real-time services need bounds on bandwidth, delay, jitter, and bit error, specially, in wireless medium where the bandwidth is always scarce and more error and delay prone depending on time and space, we took into account in detail the service differentiation schemes that are based on these parameters and are called per-queue based $ServDiff$ schemes.

One of such scheme is the EDCF or EDCA that provides prioritized QoS by defining four access category queues (AC) to support eight different user priorities (UPs) in the contention environment, using the CSMA/CA. Although, the EDCA scheme improves performance by providing prioritized service differentiation to important data but, as it uses the non-deterministic channel access mechanism, CSMA/CA, it can not guarantee bounds on bandwidth and latencies which are pre-requisites of time-sensitive services like voice and multimedia. The other scheme is the HCCA that provides parameterized QoS differentiation by providing a controlled channel access during both contention and contention-free periods. HCCA guarantees mean data rate, nominal frame size, maximum service interval, delay bound etc. by setting up a virtual connection called traffic stream (TS) between the QAP and QSTA and letting them to negotiate these parameters as traffic specification (TSPEC) before actually transmitting any frame. A simple round-robin type scheduler is proposed in the IEEE 802.11e that uses these mandatory TSPEC parameters. In order to regulate the available bandwidth resources and to assure the guaranteed/deterministic channel access time, HC administers the admission control both at EDCA and HCCA levels that along with the other measures like: direct link protocol (DLP) enabling the peer QSTA to communicate directly without QAP, and the block acknowledgement for efficient utilization of the bandwidth, are also discussed and explored as has been provisioned in the IEEE 802.11e MAC standard. To deal with the high speed error prone wireless medium, the reverse error correction techniques (stop and wait ARQ, selective-repeat ARQ, Go-back-N ARQ etc.) and forward error correction technique (FEC), and their limitations, are discussed along with a joint-venture of these two approaches that is called hybrid FEC-ARQ error correction technique. FEC-ARQ is also discussed with reference to its implication on the high speed error prone wireless channel.

In view of the discussions in this chapter, the issues regarding the QoS enhance-

ment in IEEE 802.11 MAC standard can be summarized as follows:

- In case of the contention-based channel access mechanism EDCA, the AIFSN[AC], CWmin[AC], CW_max[AC] and TXOPLimit[AC] are the default values which are determined by the QAP and are announced in the beacon frames, however, the IEEE 802.11e allows to dynamically change these parameters depending on the network conditions but does not define how to change these dynamically.
- IEEE 802.11 standard supports multi-rate PHYs capabilities for dynamic rate switching that can improve the throughput. However it does not define the mechanism or algorithm that can decide the best rate channel momentarily and then can efficiently or opportunistically switch to that channel momentarily.
- The parameterized QoS that guarantees real-time service requirements is provided in the HCCA of the HCF which is contention free channel access mechanism in the IEEE 802.11e. However, IEEE 802.11e suggests the same round-robin type polling scheduler for HCCA as is provided in the IEEE 802.11a/b/g. The vendors are set free to implement any optimized algorithms, such as reducing the polling overheads specially in case of bursty traffic that has alternating periods of talk-spurts and silence, increasing the TXOP duration, etc., within the parameters of the transmitted schedule.
- The reservation information, the round-robin type scheduler uses, include average values of traffic characteristics of the flows, like mean packet size, mean required throughput etc., that restricts the HCF to allocate fixed polling schedule which is only suitable for constant bit rate (CBR) traffic. But what about the variable bit rate (VBR) type of traffic, such as quality-controlled MPEG4 or video conferencing. Increasing the TXOP by a fixed amount would be a solution but at a probable cost of decreased total network capacity. In the literature [18], there are three proposed approaches described, i.e., TXOP, service interval, and polling schedule. Reference works at [19][20][21] have attempted to improve for VBR traffic.
- One issue is the optimization of the tradeoff between the channel efficiency, priority and fairness.
- Another important issue is analytical modeling to evaluate the efficiency and performance of EDCF packet bursting and contention-free burst (CFB).

Chapter 3

QoS Provisioning in IEEE 802.11 Wireless LANs

3.1 Introduction

With the pervasive deployment of IEEE 802.11-based WLAN technologies world-wide such as in private sites and WiFi commercial hot-spots, there has been a growing demand for WLANs to put forth QoS for packet-switched real-time services (e.g., voice, audio, and video). Voice-over-IP (VoIP) is an example of packet-based voice service which is being increasingly deployed commercially over the wired networks through the Ethernet technologies. As the benefits of having integrated wireless and wired networks for data, voice, and video applications are being unraveled, more and more research works are being focused to make WLANs to support real-time multimedia services. Since WLANs hold the promise of providing unprecedented mobility, flexibility and scalability than its wire-line counterpart, much research effort has been attracted in this area.

In this chapter we discuss and dig out the core QoS-related design problem that is common in all versions of IEEE 802.11 MAC standard (whether 802.11a/b/g or 802.11e). Also, we discuss the motivation and gravity for solving this QoS design problem that causes in-efficient utilization of the scarce wireless bandwidth.

Besides the QoS provisioning in IEEE 802.11e discussed in detail in the previous chapter, we carry out a consolidated survey of the major research works on QoS-aware wireless MAC protocols in the recent literature and summarize in this chapter.

3.2 QoS Provisioning for Real-Time Voice Traffic in IEEE 802.11 MAC

The legacy WiFi standard, i.e., IEEE 802.11a/b/g, does not provide parametric QoS required by voice traffic. Out of its two modes of medium access, i.e., the distributed coordinated function (DCF) and the point co-ordination function (PCF), the DCF uses the carrier sense multiple access with collision avoidance (CSMA/CA) protocol which is contention based and causes non-deterministic medium access delays, and the PCF is contention-free that uses the round-robin polling scheduler that provides connection-oriented medium access. However, round-robin polling scheduler does not suppress the silence periods of voice calls, and also, non-deterministic beacon delays in the PCF mode becomes barriers in guaranteeing strict delay bounds. The hybrid coordination function controlled channel access (HCCA) mode of the QoS-enhanced version, i.e., IEEE 802.11e, addresses most of the QoS-related problems of the PCF, and it also defines the call admission control outlines. However, the HCCA deploys the same inefficient round-robin scheduler and does not specify any technique for the admission control.

The round-robin type polling scheduler is kept common in both HCCA and PCF with the following difference. The HCCA uses the mandatory TSPEC parameters to prioritize different TSs in its polling list and to determine the respective length of the polled transmission opportunities (polled-TXOP) to all the poll-able stations. But it polls in the same round-robin fashion as the PCF does regardless of whether the polled station has data to send or not. Therefore, the research challenge still remains as it is.

From the perspective of real-time voice traffic, IEEE 802.11a/b/g/e MACs are lacking at following two corners.

- The round-robin style scheduler of HCCA or PCF in IEEE 802.11 causes bandwidth wastes and incurs unnecessary delay to other stations having packets to send when the polled station does not have packets in its local queue.
- Voice traffic exhibits bursty nature with alternating periods of talk-spurts and silences where silence suppression is needed for QoS assurance, but in IEEE 802.11 the HCCA or the PCF assumes that each station in its polling list

always has packet(s) to transmit without taking into account the alternating silence periods between two consecutive bursty voice packets.

The performance of HCCA/PCF due to these two major shortcomings is degraded as opposed to the VoIP technology that is being widely used commercially through the wired Ethernet. The VoIP suppresses the alternating silence periods of voice calls and achieves multiplexing gains and provides QoS to the users. To make VoIP compatible with the IEEE 802.11 MAC so that WiFi technology can efficiently support real-time voice services, it is imperative to develop an efficient algorithm or protocol. This protocol should make the round-robin scheduler intelligent enough so that it can suppress alternating silence periods of voice calls using the bandwidth efficiently with increased throughput and minimized queuing delays.

3.3 Research Advancements for QoS-Provisioning in WLANs: Current State of the Art

In this section we discuss the major research works that have been proposed in the literature to enhance the QoS in WLANs. Since voice traffic has bursts of packets to travel with in-between periodic silent periods, there is great motivation to suppress these silent periods by somehow not attending the user during its idle periods and multiplexing other traffic with it to reduce the bandwidth wastage. In the polling type wireless medium access, this job is done by the central controller (i.e., AP in IEEE 802.11 MAC) through polling list management function that should exclude a voice user from the polling list during its silent periods and include it back in the polling list when the station becomes active (talking). Since the central controller does not communicate with the user during its silent state, the hardest part of this process for AP is to know when the silent to talk-spurt state change occurs at the user side so that the user can be included again in the polling list.

For completeness of the topic, we categorize and summarize the selected research works in the following sub-sections. The first sub-section includes schemes which address the priority and/or fairness type of QoS issues. The second sub-section includes the schemes which either partially deploy contention-free medium access approach or fully deploy contention-free approach, to address the bursty nature of real-time

voice traffic for suppressing idle periods. We disregard the former sub-category since its partially contention-based medium access causes non-deterministic waiting time delays and/or consecutive frame losses that worsen exponentially even if the traffic load increases only linearly [23]. This is not tolerable to real-time voice traffic. The schemes in the latter category comprises of those protocols which suggest contention-free medium access through adaptive type of polling schemes to suppress the idle periods of voice calls. However, they do so at the cost of increased waiting time delays to stations at the time when they change their state from idle to active contributing significantly to the end-to-end delays. The third sub-section includes schemes which address QoS issues by suggesting some call admission control strategies.

3.3.1 Priority and Fairness-Based QoS Schemes

The Priority Effort-Limited Fair (priority-ELF) scheme proposed in [24] checks the priority of the stations defined as type of service (TOS) combined with the effort-limited fair scheduling, wherein the central controller (PC) does not polls the priority user for which a counter shows that there is no more than one data frame received during the last polling session during a threshold time period. The PC does so with a mechanism of checking if the effort (air time spent on a flow) is higher than a predefined threshold of its power factor. This scheme, at one hand, deters the PC from polling in an error-prone wireless link, and, at the other hand, provides prioritized-QoS like the HCF of IEEE 802.11e does.

The Priority schemes proposed in [25] provide prioritized-QoS to voice and video services over the data service with the PCF mode by giving the priority control to AP or to both AP and the stations. The AP decides whether to poll to a user during the contention-free period (CFP) based on the priority of the station.

Both of these schemes are sufficient to maintain the prioritized-QoS of the multimedia traffic. However, they do not consider the effect of the system conditions in an unstable WLAN and the overall performance of PCF.

Distributed Deficit Round-Robin (DDRR) scheme in [26] proposed a fairness-based polling scheme instead of basic round-robin polling scheme. This scheme requires an additional state variable called the deficit counter (DC) in the PCF for each associated station and schedules either to poll or not to poll to the poll-able user

based on the positive or negative value of the DC, thereby providing a sort of fairness of opportunity to all poll-able stations.

These schemes do not propose any measure to guarantee bounded end-to-end delays and jitters that are needed to voice traffic, especially under high load situations.

3.3.2 QoS Schemes Based on Silence Suppression in Voice Services

PCF with implicit signaling scheme proposed in [27] uses the contention-based DCF mode to notify the talk-spurt arrival to the AP. Active and idle lists are managed by the PCF for active and idle voice stations, respectively. When a voice activity transits from idle state to active state, the first frame is transmitted to the AP according to the contention-based DCF. This causes non-deterministic access delay which increases exponentially as the traffic load increases. If the frame is received successfully, the PC includes the station to its active polling list.

The uplink notification scheme in [23] proposes a contention-based uplink status notification mechanism which is quite different from the WiFi DCF mode. This scheme compared polling with contention mechanisms under the general tree-based topology quantifying the range of parameters, like round-trip delays and number of active stations, where the proposed contention mechanism can give better delay performance than a polling scheme. However, the quantitative results of this work stamp on the fact that contention-based wireless medium access approach can only be deployed to a limited small number of delay sensitive load and is not appropriate for large scale commercial deployment.

Simultaneous Transmit Response Polling (STRP) scheme in [28] exploits capture phenomenon to notify the uplink status by sending a weak jamming signal to the AP in response to a special control frame sent by the AP. Theoretically, it succeeds in reducing the polling overhead that would have been caused by stations during their idle states. In this scheme, the AP transmits a special control frame to two stations simultaneously, one in the active state (i.e., included in the polling list) and the other in the idle state (i.e., excluded from the polling list). Then, it lets the idle station to respond with a weak jamming signal with an extended delay than the active station with the presumption that this weak jamming signal will notify the

uplink status to the AP successfully through the capture effect. However, the capture effect may not work in a harsh fading environment, and, hence, it may not guarantee the required delay bound.

All the above schemes use the round-robin type polling scheduler to poll stations in the active list but the uplink talk-spurt notification mechanism they use is contention-based. To avoid the contention-based unpredictable delays, there are schemes such as those proposed in [29] and [30] which proposed to spare separate time zone from the contention period for addressing only the reconnection requests. A problem with such a scheme is the wastage of bandwidth if there is no reconnection request.

There are some recent schemes that suggest TDMA-like time slot assignment for scheduling instead of round-robin scheduler and claim significant silence suppression. But the mechanisms they suggest for uplink activity detection from silence to talk-spurt state change is based on contention-based medium access approach. These are summarized as under.

Deterministic Access Priority of Voice in CP proposed in [31] suggests a modified service interval structure of HCCA in IEEE 802.11e along with a TDMA-like time slots assignment scheme for MAC scheduling to support real-time voice services. Instead of alternate downlink and uplink transmissions at each scheduled user, the whole contention free period (whether in the PCF or CAP) is sub-divided in to downlink and uplink transmission periods. The AP transmits all the down packets to their destined users in FIFO order in the downlink time zone. Then, at the start of the uplink time zone, the AP broadcast the single supper CF-poll frame to all the poll-able users notifying to them a TDMA-like time slots scheme for their respective turns. Since voice can tolerate packet loss to some extent, no acknowledgment (ACK) or retransmission is proposed for voice transmission in order to avoid the retransmission delay. Voice multiplexing gains are also proposed through active and idle polling list management at the AP suppressing the periodic idle periods of voice calls.

For uplink status notification from idle to active state change, this scheme proposes a so-called "deterministically prioritized access". In this scheme, a fraction of the CP is dedicated for those stations whose status has changed from idle to active state so that these can notify their uplink status change to the AP without contending with other low priority access categories (ACs). However, the contention among the same

high-priority voice stations which want to notify their uplink status change is still there. If there is no reconnection request, this dedicated time is wasted. Instead of using the EDCA contention mechanism, this scheme proposes a black burst jamming scheme where each contending high-priority voice station transmits a jamming energy signal whose time length equals to its back off contention window. In this way, the long lasting black burst wins the channel.

Isochronous Coordination Function (ICF) in [32] is a recent idea proposed to transport voice packets over IEEE 802.11 WLANs. The concept of working cycle in ICF is similar to the CAP cycle in the HCF of IEEE 802.11e for parameterized-QoS to delay-sensitive voice service. But unlike CAP, which suggests simple round-robin polling, ICF proposes TDMA-like time slot assignment scheme where the assignment information is broadcast to all the poll-able stations at the start of each ICF cycle through a supper poll frame called ICF-poll frame. Thereby, it reduces the polling overhead significantly. For polling list management, ICF-poll frame includes a status vector (SV) which is a sequential string of logical bits, i.e., 0 for the idle station and 1 for the active station so that only the active stations should avail their assigned time slots for their uplink transmissions. For uplink status notification from idle to active state change, ICF proposes the EDCA mode of IEEE 802.11e which, though, provides prioritized-QoS, is a contention-based medium access scheme. Hence parameterized-QoS requirement of the time-sensitive voice service is not fulfilled in this scheme.

In both of the above schemes, the TDMA-like single polling scheme reduces the polling overhead at the following cost.

- Both of these schemes suffer time synchronization problems of active stations with their respective TDMA time slots.
- And whole the time slot is wasted if the active user does not receive the single broadcast supper-poll frame correctly or it becomes silent in the meanwhile.

Non-uniform Fully Gated Limited (non-uniform FGL) and FGL polling schemes in [33], based on the service limited (time-limited or number-limited) disciplines, are analyzed for managing the distributed queuing systems, like the uplink distributed queues in case of the HCCA/PCF mode of IEEE 802.11 WLANs. The non-uniform FGL polling scheme attempts to reduce the wastage of the bandwidth due to polling empty queues. The basic idea of non-uniform FGL is to poll each

station as frequently as its traffic requirement needs. Though this scheme saves bandwidth wastage but 802.11 MAC protocol does not support it as it needs to piggyback additional information at the MAC layer.

All of the schemes we have discussed above attempt to reduce the polling overhead through suppressing the silence periods of voice calls, but they, one way or the other, do not provide explicit contention-free wireless medium access control that is the prime requirement of the delay sensitive voice traffic. Following are a couple of schemes that have attempted to do this job.

Cyclic Shift and Station Removal (CSSR) polling scheme proposed in [34] reduces the number of polls to voice stations by not polling them for a fixed number of threshold time cycles during their silent periods. A station notifies to the AP about its uplink status change from talk-spurt to silent state through a null frame. The AP replaces this idle station from the talking user list to the silent user list. This scheme increases the capacity of the WLAN by reducing the polling overhead but at the cost of increased waiting time delays to the voice frames that arrive at the silent stations but can not be sent due to the imposition of the threshold time period during which the station is kept in the silent polling list.

Adaptive Polling MAC Scheme proposed in [35] suggests a talk-spurt detection algorithm to reduce the number of polls to voice stations during their silent periods using the round-robin scheduler. The average duration of the silent period (which is assumed to be exponentially distributed) of a voice user is taken as 1.3 seconds. The algorithm divides this average idle period into decreasing-value threshold time intervals. This is opposed to the scheme in [34] that proposed fixed threshold time intervals during the silent state. The uplink status from talk-spurt to silent state change is notified to the AP through in-band signaling by setting the more-data field in the last frame to 0, and, subsequently, the silent station is replaced from the talking user list to the silent user list. If the dwell time in the silent state surpasses the threshold time limit, the PC replaces the station from its silent to talking user list. In the next SF cycle, if this station responds to the poll by sending a frame, it is kept in the active list. If it does not respond or sends a null frame, the PC puts it back to the idle list, however, at this time, its threshold time variable is set to the next comparatively small threshold time, and so on until the threshold time equals

the supper frame time or the silent to talking state change occurs. Again, this scheme also suppresses the silent periods to increase the capacity but at the cost of increased waiting time delays as in [34].

Both of the above schemes provide silence suppression but at the cost of increased waiting time delays to the first talk-spurt frames that are not sent due to the imposition of the threshold time period.

3.3.3 Schemes Based on Admission Control for Voice Services

The connection admission control (CAC) scheme in association to a signaling protocol proposed in [36] quantifies the maximum capacity, delay jitter in relation to the inter-poll periods (super frame time) in the PCF mode, and maximum delay bound requirements of the voice calls. It suggests a call admission control strategy based on the capacity of the PCF mode under specified conditions. However, this scheme suggests the contention-based DCF mode for signaling the arrival of the first talk-spurt packet to the AP. Also, it does not suggest anything to avoid the non-deterministic beacon delays caused by the peculiar design of the PCF mode. Moreover, this scheme does not provide a clear strategy for silence suppression of voice calls.

Measurement-based CAC strategies introduced in [37] and in [38] are based on the channel occupancy analysis. The channel occupancy rate, which is the ratio of the channel busy time to an observation time, is constantly measured and matched to a predetermined threshold value to know the traffic congestion level and to admit new users accordingly. But these schemes only consider the contention-based DCF and EDCA modes, respectively, which are not suited for the high traffic load situations.

A qualitative comparison among the different QoS-aware MAC schemes is shown in Table 3.1. This comparison is carried out among the schemes we discussed above categorizing them into three three sub-groups, i.e., one is that which deploys partially contention-based medium access approach, second is that which deploys fully contention-free medium access approach, and third is that which address QoS issues by suggesting some call admission control strategies.

Table 3.1. *Qualitative performance comparison among different QoS-aware MAC schemes.*

Schemes on contention-based activity detection					
	Delay bound guarantees	Silence Suppression	Waiting time *delays*[2]	Consecutive packet *loss*[4]	Fairness
STRP [28]	*Partial*[1]	Very good	*Small*[3]	*Small*[3]	Moderate
Uplink Notification Scheme [23]	*Partial*[1]	Good	*Moderate*[3]	*Moderate*[3]	Moderate
PCF with implicit signaling [27]	*Partial*[1]	Good	*Moderate*[3]	*Moderate*[3]	Moderate
ICF [32]	*Partial*[1]	Excellent	*Moderate*[3]	No	Good
Deterministic Access Priority [31]	*Partial*[1]	Excellent	*Small*[3]	*Small*[3]	Moderate
Schemes on contention-free activity detection					
	Delay bound guarantees	Silence Suppression	Waiting time *delays*[2]	Consecutive packet loss	Fairness
Adaptive polling [35]	Yes	Very good	Large	Large	Moderate
CSSR [34]	Yes	Very good	Very large	Large	Moderate
Round-Robin (PCF/HCCA) [4][3]	Yes	No	No	No	Moderate
DDRR [26]	Yes	Good	Large	Large	*Excellent*[5]
H-CFA[8]	Yes	Very good	No	No	*Excellent*[6]
Call admission control schemes					
	Delay bound guarantees	Contention-free Access	Waiting time *delays*[2]	Capacity gains	Fairness/ Priority
Measurement Based CAC [37]	No	No	Probabilistic	*Moderate*[3]	Moderate
WLAN CAC [38]	No	No	Probabilistic	*Moderate*[3]	Moderate
CAC [36]	*Partial*[1]	Partial	*Moderate*[3]	*Moderate*[3]	Moderate
TS-AC[8]	Yes	Yes	No	*Large*[7]	Moderate

Note:
1 If number of contending stations is large, delay bound guarantees become increasingly vulnerable.
2 Unnecessary delays to initial frames when idle-to-active state change arrives.
3 Number of contending stations is assumed small, and if large, it will lead to un-acceptable for delay-bounds.
4 It is assumed that the delayed frame is lost.
5 Fairness in terms of bandwidth allocation.
6 Fairness according to the traffic arrival behavior at each admitted station.
7 At the cost of accepted large inconsecutive packet loss.
8 For H-CFA and TS-AC, see chapters 5 and 6, respectively.

3.4 Chapter Summary

In this chapter we have discussed and elaborated the core QoS-related design problem that is common in all versions of IEEE 802.11 MAC standard (whether 802.11a/b/g or 802.11e). That is the inefficiency of the WiFi MAC standard in exploiting a very important characteristic of packet-switched networks, i.e., multiplexing gains and quality of service provisioning in case of bursty type of traffic.

We have carried out a consolidated survey of the major research works on QoS-aware wireless MAC protocols in the recent literature and summarized in this chapter categorically, i.e., those which address the fairness and priority related QoS issues; those which deploys both the well-known contention-based and contention-free medium access approaches for capacity and QoS enhancement; and those which are purely based on the contention-free medium access approach for this purpose. For end-to-end delay requirement of the real-time sensitive services, we have also discussed a few schemes that deal with another very important QoS-provisioning aspect that is call admission control.

We have observed that the fairness and priority-based QoS schemes, the fully or partially contention-based schemes, due to their non-deterministic medium access nature, are not viable to ensure parametric-QoS, like end-to-end delay and maximum acceptable jitters level, that is required by the real-time services like voice and video. The last two categories, i.e., the fully contention-free schemes, and the call admission control-based schemes, are important in dealing with the capacity enhancement and parametric-QoS provisioning to real-time services. However, these schemes, although show significant improvement in reducing the polling overhead, are not efficient. This is due to the fact that these schemes increase waiting time delays to first talk-spurt frames when VoIP stations change their state from idle to active, and therefore, increase the end-to-end delay substantially. The call admission control-based schemes discussed in this chapter are also not suitable in case of parametric-QoS provisioning. This is due to the fact that these schemes use a contention-based medium access approach which causes unacceptable non-deterministic channel access delays.

In the following chapters we will present analytical performance evaluation of these schemes, and subsequently, present our novel protocols which provide superior performance.

Chapter 4

Performance Analysis of Contention-Free Approaches for Silence Suppression in Voice Calls

4.1 Introduction

In this chapter, we carry out performance analysis of two popular schemes which suppress silence periods of bursty voice traffic with fully contention-free medium access approach. We choose these schemes because these schemes use a fully contention-free medium access approach in order to guarantee deterministic medium access time for time-sensitive voice traffic. These schemes are: Adaptive Polling MAC Scheme [35] and Cyclic Shift and Station Removal (CSSR) polling scheme [34].

While doing the analysis, we take into account both pros and cons of these schemes. Our objective of carrying out this performance evaluation is to determine analytically how much of the unnecessary polling overhead time that the round-robin polling scheduler incurs during the alternating silence periods of voice calls can be reduced in these different approaches. As has been mentioned in the previous chapter, these schemes reduce the polling overhead at the cost of increased waiting time delays to the first talk-spurt frames in accessing the wireless medium. We also aim to determine analytically the unnecessary wireless medium access delay time that a typical first talk-spurt frame suffers in these two different approaches. The evaluation of these performance metrics will lead us to the conclusion whether or not and to what extent each of these schemes is viable.

4.2 Analysis of the Adaptive Polling MAC Scheme [35]

In this scheme the CFP is divided into two following sub-periods as illustrated in Fig. 4.1:

4.2.1 Downlink Voice Transmission Period (DVTP)

During DVTP, the enqueued data frames at the AP are transmitted to their respective destined nodes and each node that receives the data frame transmits ACK frame to the AP after a SIFS time period. The transmission order in the voice queue follows FIFO (first in first out) discipline. Fig. 4.1 only describes the voice data (D and U), Ack and CF-poll without any Data + CF-Poll frame with the presumption that echo canceller is used and there is no double talk situation. For double-talk situation (i.e., when two stations try to talk simultaneously), data + CF-poll is also used. In this case the PC, instead of sending data frames in DVTP and CF-poll frame in the UVTP to the same node, sends data frame first and then sends CF-poll frame to the same station consecutively.

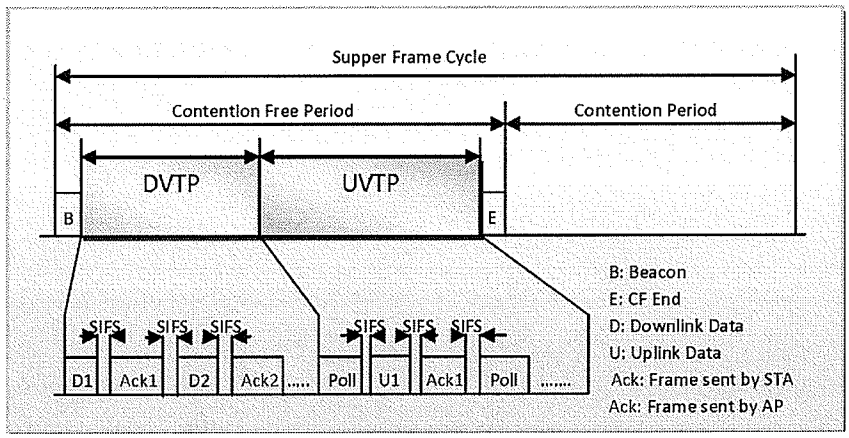

Figure 4.1. *Separated downlink and uplink transmissions in CFP of the beacon interval.*

4.2.2 Uplink Voice Transmission Period (UVTP)

Important jobs, like: polling; polling list management (PLM); and uplink voice activity, are carried out by the PC during the uplink voice transmission period (UVTP).

4.2.2.1 Polling List Management (PLM)

PC manages 4 logical lists during the UVTP, i.e., pollable list (PL), active node list (ANL), inactive node list (INL), and adaptive polling list (APL). The PL enlists the association IDs of all the associated nodes in the BSS. The ANL enlists the AIDs of all such nodes which are currently in the talk-spurt state. The INL enlists the AIDs of all such nodes which are currently in the silent state. The APL enlists the AIDs of all the active nodes plus the AIDs of some nodes from the INL whose dwell time in the silent state has surpassed a threshold time. Initially, all the associated stations are assumed in the talk-spurt state.

4.2.2.2 Uplink Voice Activity

The voice activity that is modeled as a 2-state Markov process (talk-spurt and silence) is handled during the UVTP (see Fig. 4.2). Talk-spurt and silence periods have exponential distributions with means $(\frac{1}{\alpha} =)1sec$ and $(\frac{1}{\beta} =)1.35sec$ respectively. Here α and β are transition probabilities from talk-spurt to silence and vice versa respectively, and $1-\alpha$ and $1-\beta$ are transition probabilities from talk-spurt to talk-spurt and silence to silence respectively. Packet arrival rates in talk-spurt and silence periods are λ and 0, respectively.

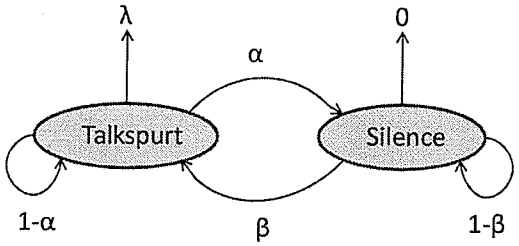

Figure 4.2. *Voice traffic model.*

The uplink voice activity cycle has four following phases as shown in Fig. 4.3.

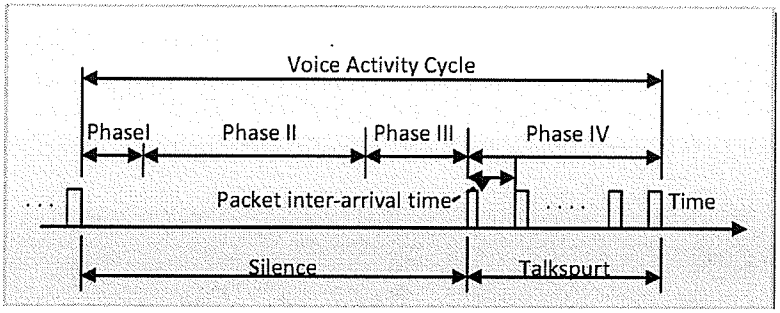

Figure 4.3. *Four phases of voice activity.*

1. Phase I is the detection of talkspurt to silence state change which is performed through the in-band signaling method by setting 'more data field' to 'FALSE' in the current uplink packet. When phase I is detected, the PC replaces the node from its APL and ANL to INL.
2. Phase II is the silence state where the PC does not poll the nodes at all.
3. Phase III is the detection of silence to talk-spurt state change which is performed by the speech activity detection (SAD) function that uses a talk-spurt detection algorithm as illustrated in Fig. 4.4 When phase III is detected, the PC replaces the node from its INL to APL.
4. Phase IV is the talk-spurt state where the PC keeps the node in the ANL and APL as long as 'more data field' in the current uplink packet is set to 'TRUE'.

In phase III of the uplink voice activity cycle, the talk-spurt detection algorithm divides average silent period (r_o) into exponentially decreasing threshold time intervals according to the following formula such that the first threshold time $Th(1) >> T_{SF}$. According to the algorithm if the dwell time in the silent state surpasses the threshold time, the PC replaces the node from its INL to APL for polling. If the node does not respond after polling, the PC does not poll the node anymore until the end of the second threshold time (Tth(2)) and so on until threshold time equals the supper frame time.

$$T_{th}(k) = r_o p(1-p)^{k-1}, \ where\ r_o = 1/\beta,\ T_{th}(k) \geq T_{SF},\ 0 < p < 1 \qquad (4.1)$$

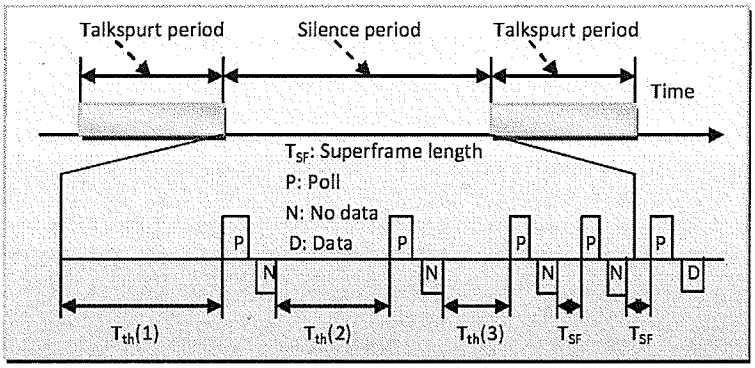

Figure 4.4. *talk-spurt detection algorithm in adaptive polling MAC scheme.*

DVTP and UVTP schemes mentioned above are described below in a combined way and in the flow-chart in Fig. 4.5.

1. Voice packets from far-end side of a mobile station are enqueued in the AP.
2. At the beginning of the nominal SF boundary, the PC seizes the control of the channel by transmitting beacon frame after the channel has been idle for a PIFS time interval.
3. The PC divides the CFP into DVTP and UVTP and starts DVTP first.
4. During the DVTP, a PC transmits each voice packet, enqueued in its voice queue, to the destination node in a FIFO discipline. The maximum number of voice packets transmitted during a DVTP are set to be a half of the number of voice sessions because, on an average, only a half of the VoIP users are listening or speaking to their correspondents.
5. Following DVTP, the PC initiates UVTP by sending a poll frame. The PC sends poll frames to stations in its polling list to receive voice packets. During the UVTP, the PC controls the polling sequence and the polling list by using the uplink transmission scheme and the PLM described earlier in this section.

4.2.3 Discussions on Shortcomings

The shortcomings of this scheme can be summarized as follows:

- The simulations results showed that this scheme achieves a substantial reduction in the polling overhead, but it does so at the cost of increased waiting delays to

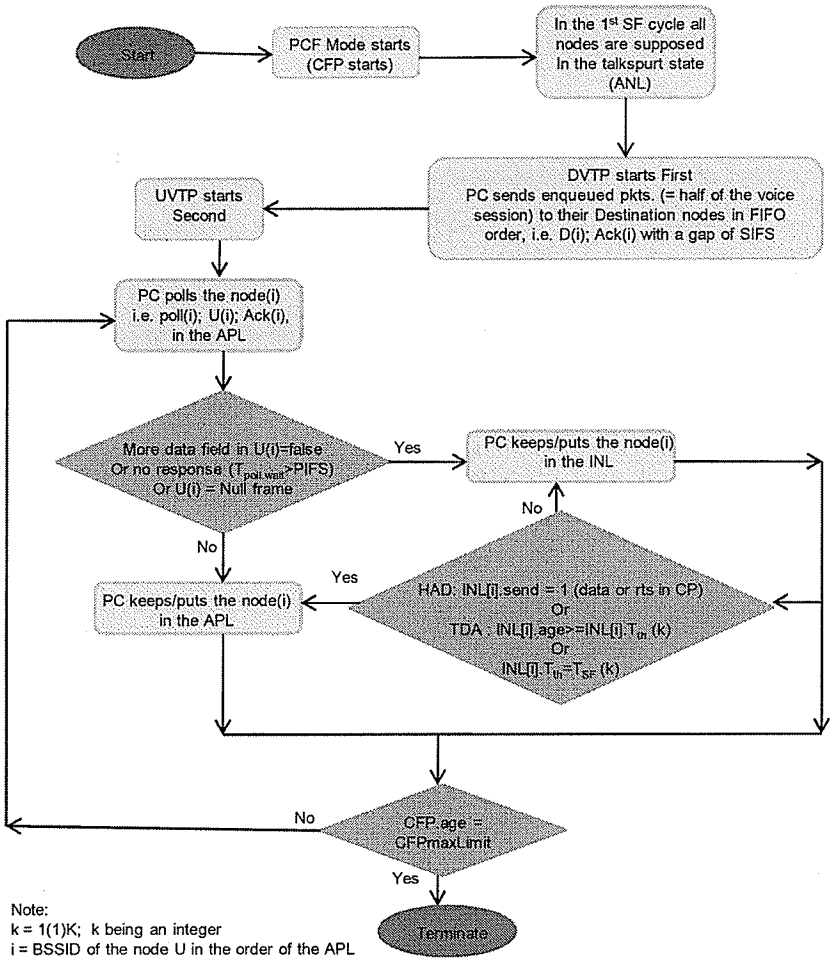

Figure 4.5. *Sequence of actions in DVTP and UVTP of the CFP.*

uplink traffic. This is because, according to the talk-spurt detection algorithm of this scheme, a packet generated at a node during its silence period has to wait excessively for its transmission as this node can not be included in the APL for poling unless its dwell time in the silent state surpasses the threshold time limit.

- The division of the CFP into DVTP and UVTP also causes more waiting delays to the downlink traffic. The downlink voice packets enqueued in PC from the current UVTP has to wait till the next SF cycle starts because DVTP of the current SF cycle has already lapsed. And, if not all, some of these downlink packets destined to such nodes which have not yet been polled in the current

SF cycle, would have been transmitted in the same SF cycle.

4.3 Analysis of the Cyclic Shift and Station Removal (CSSR) Polling Scheme [34]

In this scheme the average silence period is divided into equal threshold time intervals as opposed to the adaptive polling scheme in [35] where the average silence period is divided into exponentially decreasing threshold time intervals(see Fig. 4.6).

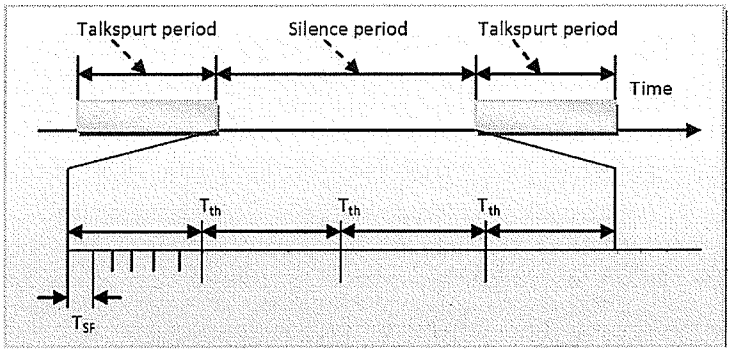

Figure 4.6. *Talk-spurt detection algorithm in CSSR scheme.*

Due to its fixed threshold time intervals, the CSSR scheme would reduce polling overhead more than the adaptive polling scheme, but at the same time it would cause more waiting time delays to the first talk-spurt frames.

4.4 Performance Evaluation

4.4.1 Performance Metrics

The objective is to make the round-robin scheduler of the IEEE 802.11a/b/g/e MAC standard intelligent and efficient so that it can suppress alternating silence periods in the bursty real-time traffic to make the PCF and the HCCA more bandwidth and delay efficient.

The two approaches discussed in the previous sections suggest considerable reduction in the polling overhead. At the same time, these approaches cause excessive waiting time delays to the first talk-spurt frames. In this section these approaches

are analyzed. We develop analytical models for these approaches to evaluate two QoS performance metrics. One is the average polling overhead time saved at the scheduler side (i.e., at PC) during each uplink voice activity cycle of a voice call normalized to that in the round-robin polling . The other is the average waiting time delay caused to the first talk-spurt frame at the station side when the station transits from silence to talk-spurt state.

4.4.2 Modeling and Assumptions

In both the approaches, the talk-spurt and silence periods in the bursty voice traffic are assumed to have exponential distribution [29][26]. Assuming the super frame time interval as discrete unit, we consider the silence period as having geometric distribution. In the adaptive polling MAC approach [35], the average silence period is divided into exponentially decreasing threshold time intervals. In the CSSR approach [34], the average silence period is divided into equal threshold time intervals.

Let p be a parameter that may take a value between 0 and 1. The choice of p defines the length of the threshold time intervals, and the number of polls required before the first talk-spurt arrives, and the channel access delay caused due to not polling during the threshold time interval. Therefore, p should be chosen in correlation with the required delay bounds in order to make an optimal trade off between the reduced number of polls and the unnecessary channel access delays caused due to the talk-spurt detection algorithm.

4.4.3 Formulation of Analytical Models

4.4.3.1 Input Distributions (Random Variables)

Silence and talk-spurt states have uniform geometric distribution. Let x be a geometric random variable representing n independent trials (i.e., n supper frame intervals), each having the probability β of being success (talk-spurt detection occurs) until the success occurs. Then

$$Prob\,[x = n] = (\beta - 1)^{n-1}\,\beta. \tag{4.2}$$

The expected value of x (i.e., $\bar{x}$) at $n = \infty$ is given as follows.

$$E\{x=n\} = \bar{x} = \sum_{n=1}^{\infty} n(1-\beta)^{n-1}\beta = \frac{1}{\beta} \tag{4.3}$$

At the truncated value of n (say $n = l$), $\bar{x}$ is calculated as follows.

$$E\{x=n\} = \bar{x} = \sum_{n=1}^{l} n(1-\beta)^{n-1}\beta = \frac{1-(l+1)(1-\beta)^{l}+l(1-\beta)^{l+1}}{\beta} \tag{4.4}$$

It should be noted that $\bar{x}$ is the average number of supper frame time cycles or the average number of polls, say N_o, in the round-robin scheme before the first success (talk-spurt) occurs.

Let N_o represent the average number of polls in the average silent period r_o according to the actual round-robin scheme. Then

$$r_o = N_o T_{SF} = \bar{x} T_{SF} = T_{SF}\left[\frac{1-(l+1)(1-\beta)^{l}+l(1-\beta)^{l+1}}{\beta}\right]. \tag{4.5}$$

4.4.3.2 Steady State Probability in 2-State Markov Process of Uplink Voice Activity

In Fig. 4.2 that models the uplink voice activity as 2-state Markov process, i.e., talk-spurt and silence, the transition probability matrix, say $\mathbf{P}$, can be expressed as follows.

$$\mathbf{P} = \begin{bmatrix} 1-\beta & \beta \\ \alpha & 1-\alpha \end{bmatrix} \tag{4.6}$$

To evaluate the QoS performance metrics, the steady state probabilities for the system states, i.e., silence and talk-spurt, are required. Since the size of the transition probability matrix $\mathbf{P}$ is 2×2, let $\boldsymbol{\pi} = [\pi_1 \ \pi_2]$ denote the system steady state probability vector. Here π_1 and π_2 respectively denote the steady state probabilities that a voice station is in the silent and talk-spurt states. These steady state probabilities of silence and talk-spurt are calculated from the following two steady state conditions.

$$\boldsymbol{\pi}\mathbf{e} = 1, \quad \text{where } \mathbf{e} \text{ is a } 2\times1 \text{ column vector of ones} \tag{4.7}$$

$$\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi} = [\pi_1 \ \pi_2] \tag{4.8}$$

4.4.3.3 Adaptive Polling MAC Approach [35]

In Fig. 4.4, let the talk-spurt detection algorithm divide the average silence period r_o into k threshold time intervals. Thus the number of polls in the silence period is equal to k.

Average time saved $\left(\overline{T_{saved}(k)}\right)$ ***at the scheduler per uplink voice activity cycle of each voice station in comparison to the round-robin scheme:*** The time saved, say $T_{saved\_r_o}(k)$, by the talk-spurt detection algorithm of this scheme per average silence period in comparison to the round-robin scheme is calculated as follows.

$$T_{saved\_r_o}(k) = T_{poh}\left[\sum_{i=1}^{k} \frac{T_{th}(k)}{T_{SF}} - k\right] \tag{4.9}$$

where $i = 1(1)k$ (i being an integer), $T_{th}(k) = r_o p(1-p)^{k-1}$, T_{poh} is the time (in terms of bit-times) to transmit one poll frame, and T_{SF} is the supper frame time. Here $T_{saved\_r_o}(k)$ is also a function of a geometric random variable k that represents k independent threshold time intervals. The k^{th} time saved ($T_{saved\_r_o}(k)$) is a successive addition with the next, until the first success, i.e., the detection of talk-spurt, occurs. Here k is ultimately co-related with the geometric random variable x. The probability mass function (pmf) of $T_{saved\_r_o}(k)$ is given as follows.

$$p\left(T_{saved\_r_o}(k)\right) = \sum_{x=round\_off\left[N_o\left[1-(1-p)^{k-1}\right]+1\right]}^{round\_off\left[N_o\left[1-(1-p)^{k}\right]\right]} (1-\beta)^{x-1}\beta \tag{4.10}$$

where $k = 1(1)K$ (k being an integer). The expected value of $T_{saved\_r_o}(k)$ is calculated as follows.

$$E\left[T_{saved\_r_o}(k)\right] = \overline{T_{saved\_r_o}(k)} = \sum_{k=1}^{K}\left[\sum_{x=round\_off\left[N_o\left[1-(1-p)^{k-1}\right]+1\right]}^{round\_off\left[N_o\left[1-(1-p)^{k}\right]\right]} (1-\beta)^{x-1}\beta\right] T_{saved\_r_o}(k) \tag{4.11}$$

The average time saved ($\overline{T_{saved}(k)}$) at the scheduler per uplink voice activity cycle of each voice station in comparison to the round-robin scheme can be expressed as follows.

$$E\left[T_{saved}\left(k\right)\right] = \overline{T_{saved}\left(k\right)} = \pi_1 \times \overline{T_{saved\_r_o}\left(k\right)} + \pi_2 \times 0 = \pi_1 \times \overline{T_{saved\_r_o}\left(k\right)} \quad (4.12)$$

It should be noted that the expected value of the time saved during the average talk-spurt period is 0 as the voice station during its talk-spurt period is polled in each beacon cycle.

Average unnecessary channel access delay ($\overline{T_{ud}(k)}$) to the first talk-spurt frame at each voice station: Consider the first voice packet (silence to talk-spurt state change) arrives at the start of k^{th} threshold time ($T_{th}(k)$). Then the unnecessary delay ($T_{ud\_r_o}(k)$) is calculated as follows.

$$T_{ud\_r_o}\left(k\right) = r_o p\left(1-p\right)^{k-1} = T_{SF} N_o p\left(1-p\right)^{k-1} \quad (4.13)$$

The number of the threshold interval, k, is co-related with the geometric random variable x (the number of supper frame cycles). And unnecessary delay $T_{ud\_r_o}(k)$ is basically the function of the G.R.V x, i.e., for different value of x the value of $T_{ud\_r_o}(k)$ is different. The expected value of unnecessary channel access delay $T_{ud\_r_o}(k)$ is calculated as follows.

$$\begin{aligned}\overline{T_{ud\_r_o}(k)} \;=\; & \sum_{k=1}^{K}\Bigg[\sum_{x=round\_off\left[N_o\left[1-(1-p)^{k-1}\right]+1\right]}^{round\_off\left[N_o\left[1-(1-p)^{k}\right]\right]}\Big[r_o p\left(1-p\right)^{k-1} - [x-1 \\ & -round\_off\left[N_o\left[1-\left(1-p\right)^{k-1}\right]\right]\Big] T_{SF}\Bigg]\left(1-\beta\right)^{x-1}\beta\Bigg] \end{aligned} \quad (4.14)$$

where $k = 1\,(1)\,K$, k being an integer. The average unnecessary channel access delay ($\overline{T_{ud}(k)}$) to the first talk-spurt frame at each voice station in comparison to the round-robin scheme can be expressed as follows.

$$E\left[T_{ud}\left(k\right)\right] = \overline{T_{ud}\left(k\right)} = \pi_1 \times \overline{T_{ud\_r_o}\left(k\right)} + \pi_2 \times 0 = \pi_1 \times \overline{T_{ud\_r_o}\left(k\right)} \quad (4.15)$$

It should be noted that the expected value of the unnecessary channel access delay to talk-spurt frames during the average talk-spurt period is 0 as the voice station during its talk-spurt period is polled in each beacon cycle.

4.4.3.4 Cyclic Shift and Station Removal (CSSR) Polling Approach [34]

In Fig. 4.6, let the talk-spurt detection algorithm divides the average silence period r_o into k equal threshold time intervals ($T_{th}(k)$). Thus the number of polls in the silence period is equal to k. $T_{th}(k)$ is calculated as follows.

$$T_{th}(k) = r_o p, \quad \text{where } 0 < p < 1 \tag{4.16}$$

Average time saved $(\overline{T_{saved}(k)})$ ***at the scheduler per uplink voice activity cycle of each voice station in comparison to the round-robin scheme:*** The time saved, say $T_{saved\_r_o}(k)$, by the talk-spurt detection algorithm of this scheme per average silence period in comparison to the round-robin scheme is calculated as follows.

$$T_{saved\_r_o}(k) = kT_{poh}(N_o p - 1) \tag{4.17}$$

where $N_o = r_o/T_{SF}$, T_{poh} is the time (in terms of bit-times) to transmit one poll frame, and T_{SF} is the supper frame time. Here $T_{saved\_r_o}(k)$ is also a function of a geometric random variable k that represents k independent threshold time intervals. The k^{th} time saved ($T_{saved\_r_o}(k)$) is a successive addition with the next, until the first success, i.e., the detection of talk-spurt, occurs. Here k is ultimately co-related with the geometric random variable x. The probability mass function (pmf) of $T_{saved\_r_o}(k)$ is given as follows.

$$p\left(T_{saved\_r_o}(k)\right) = \sum_{x=round\_off[(k-1)N_o p+1]}^{round\_off[kN_o p]} (1-\beta)^{x-1}\beta \tag{4.18}$$

where $k = 1\,(1)\,K$ (k being an integer). The expected value of $T_{saved\_r_o}(k)$ is calculated as follows.

$$E\left[T_{saved\_r_o}(k)\right] = \overline{T_{saved\_r_o}(k)} = \sum_{k=1}^{K}\left[\sum_{x=round\_off[(k-1)N_o p+1]}^{round\_off[kN_o p]} (1-\beta)^{x-1}\beta\right] T_{saved\_r_o}(k) \tag{4.19}$$

As in the case of Adaptive polling MAC approach, the average time saved $(\overline{T_{saved}(k)})$ at the scheduler per uplink voice activity cycle of each voice station in comparison to the round-robin scheme can be expressed as follows.

$E\left[T_{saved}\left(k\right)\right] = \overline{T_{saved}\left(k\right)} = \pi_1 \times \overline{T_{saved\_r_o}\left(k\right)} + \pi_2 \times 0 = \pi_1 \times \overline{T_{saved\_r_o}\left(k\right)}$

It can be noted that the expected value of the time saved during the average talk-spurt period is 0 as the voice station during its talk-spurt period is polled in each beacon cycle.

Average unnecessary channel access delay ($\overline{T_{ud}(k)}$) to the first talk-spurt frame at each voice station: Consider the first voice packet (silence to talk-spurt state change) arrives at the start of k^{th} threshold time ($T_{th}(k)$). Then the unnecessary delay ($T_{ud\_r_o}(k)$) is calculated as follows.

$$T_{ud\_r_o}\left(k\right) = r_o p \tag{4.20}$$

The number of the threshold interval, k, is co-related with the geometric random variable x (the number of supper frame cycles). And unnecessary delay $T_{ud\_r_o}(k)$ is basically the function of the G.R.V x, i.e., for different value of x the value of $T_{ud\_r_o}(k)$ is different. The expected value of unnecessary channel access delay $T_{ud\_r_o}(k)$ is calculated as follows.

$$\overline{T_{ud\_r_o}(k)} = \sum_{k=1}^{K}\left[\sum_{x=round\_off[(k-1)N_o p+1]}^{round\_off[kN_o p]}\left[r_o p - \left[x - 1 - round\_off\left[\left(k-1\right)N_o p\right]\right]T_{SF}\right]\left(1-\beta\right)^{x-1}\beta\right] \tag{4.21}$$

Where $k = 1\left(1\right)K$, k being an integer. As in the case of Adaptive polling MAC approach, the average unnecessary channel access delay ($\overline{T_{ud}(k)}$) to the first talk-spurt frame at each voice station in comparison to the round-robin scheme can be expressed as: $E\left[T_{ud}\left(k\right)\right] = \overline{T_{ud}\left(k\right)} = \pi_1 \times \overline{T_{ud\_r_o}\left(k\right)} + \pi_2 \times 0 = \pi_1 \times \overline{T_{ud\_r_o}\left(k\right)}$. It can be noted that the expected value of the unnecessary channel access delay to talk-spurt frames during the average talk-spurt period is 0 as the voice station during its talk-spurt period is polled in each beacon cycle.

4.4.3.5 Optimal Selection of Parameter p

In this model, p is the parameter that can take any value between 0 to 1. However, while choosing the value of p following constraints should be taken into account.

In the adaptive polling MAC approach [35], the last threshold time interval must be greater than or equal to the supper frame time cycle (i.e $T_{th}(K) \geq T_{SF}$). The

value of K, i.e., the maximum number of threshold intervals, is the function of the value of p, i.e., $T_{th}(K) = r_o p(1-p)^{K-1} = T_{SF}$, and, in case of adaptive polling MAC approach, is calculated as follows.

$$K = round\_off\left[\frac{\ln(1/N_o) - \ln(p) + \ln(1-p)}{\ln(1-p)}\right] \tag{4.22}$$

where $N_o = 1/\beta$ for infinite (i.e., very large) value of n - the number of independent trials of supper frame intervals, and $N_o = \left(1 - (l+1)(1-\beta)^l + l(1-\beta)^{l+1}\right)/\beta$, for the truncated value of $n(= l)$. The value of l can be calculated by the iterative method from the following equation.

$$l_{new} = round\_off\left[\frac{\ln(1 - N_o\beta) - \ln(1 + l_{old}\beta)}{\ln(1-\beta)}\right] \tag{4.23}$$

Numerical example: For $N_o = r_o/T_{SF} = 1.35/0.02$ and $\beta = 0.0148$, l converges to 619.

In the CSSR approach [34], we calculate the number of threshold time intervals for different values of parameter p as follows. Sum of all the threshold time intervals should be equal to the average silence period (r_o), i.e.,

$$Kr_o p = r_o \;\Rightarrow\; K = round\_off\left[\frac{1}{p}\right]. \tag{4.24}$$

In the adaptive polling MAC approach [35], the first threshold time interval that is the largest among all the rest of the threshold intervals and any threshold interval of the CSSR approach [34] must be less than or equal to the maximum accepted delay bounds (say d $msec$), i.e., $r_o p \leq d$. Also, the threshold time should not be less than the T_{SF} ($= 20$ $msec$), i.e., $r_o p \geq T_{SF}$, otherwise, there will be no saving in polling time. Therefore, the range of values of p is defined with the following in-equality constraint.

$$1/N_o \leq p \leq \frac{d}{N_o T_{SF}} \tag{4.25}$$

where $N_o = 1/\beta$ for infinite (very large) value of n, i.e., the number of independent trials of supper frame intervals, and N_o ($\bar{x}$) is given by Equation (4.4) for truncated value of $n(= l)$.

Elimination of errors created in rounding off the summation limits: In case of the CSSR scheme, since the average silence period is divided into equal threshold time intervals, the number of last threshold interval (i.e., K) which is the rounded off number of the fraction $1/p$, may have large variance (additive or subtractive) to the fractional remainder of the actual silence period. The large variance will count for the inclusion of large number of fake super frame cycles in the last threshold time interval as a threshold interval is the multiple of super frame cycle. In order to eliminate such variance when we calculate the performance metrics, we round off K towards its ceiling integer and keep track of the actual fraction of the silence period for the last threshold time interval through the additive variance. With this fraction we calculate the correct upper summation limit for the last threshold time interval. In implementation, we do in this way: the actual fraction (F_r) for the ceiling K is calculated as $F_r = 1 - [K - (1/p)]$. The upper summation limit (x_{upper}) in Equations (4.18, 4.19, 4.21), when $k = K$, is calculated as $x_{upper} = round\_off[(K-1)N_o p + F_r N_o p]$.

4.4.4 Performance Results

For $N_o = r_o/T_{SF} = 1.35/0.02$ and $\beta = 0.0148$, *polling overhead* $(Tph) = 80$ μsec, we present comparative numerical results of the Adaptive Polling MAC and the CSSR approaches with reference to the actual round-robin polling scheme in Figs. 4.7. These figures show the number of polls saved in both the approaches normalized to the actual number of polls in the round-robin scheme versus the number of threshold intervals for different values of parameter p taken within the constrained range as defined in the previous section for maximum acceptable channel access delay $d = 200$ $msec$.

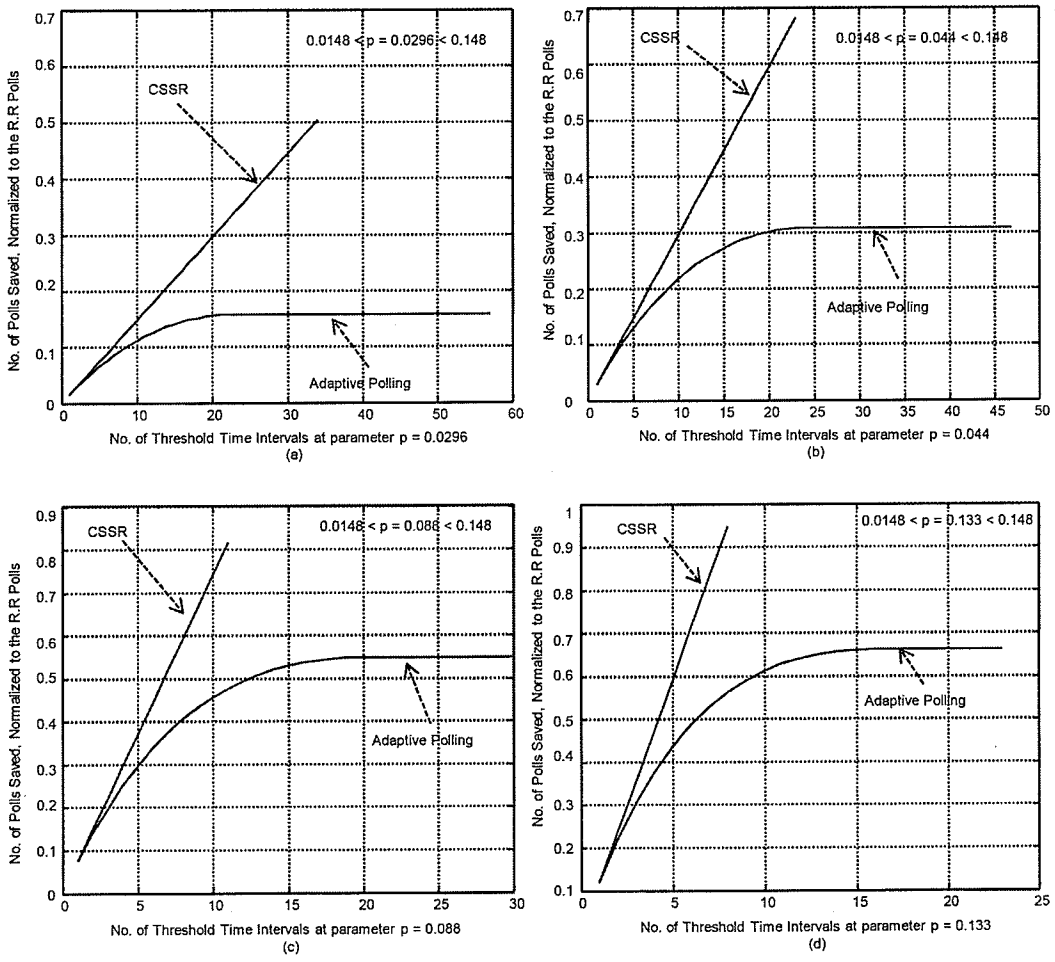

Figure 4.7. *Numerical results for polls saved in the Adaptive Polling MAC and the CSSR approaches vs round-robin approach.*

As we can observe in Fig. 4.7(a), for $p = 0.0296$, the Adaptive Polling MAC scheme

can save up to a maximum of about 16% of the polls of round-robin scheme while the CSSR scheme can save up to a maximum of 50% of the polls of round-robin scheme. As the value of p is increased within the constrained range, see Figs. 4.7(b),(c),(d), the number of polls saved also increases in both the schemes. However, the CSSR scheme saves more polls in all cases with a substantial difference than the Adaptive Polling MAC scheme.

In Fig. 4.8 we discus the numerical results on waiting time delays to the first talk-spurt frames caused due to the talk-spurt detection algorithm, i.e., the division of average silent period into threshold time intervals. This figure represents the comparative numerical results of the number of polls required in both the schemes under discussion normalized to the actual number of polls of the round-robin scheme and the worst-case average channel access delay to the first talk-spurt frame normalized to the average silence period versus different values of parameter p.

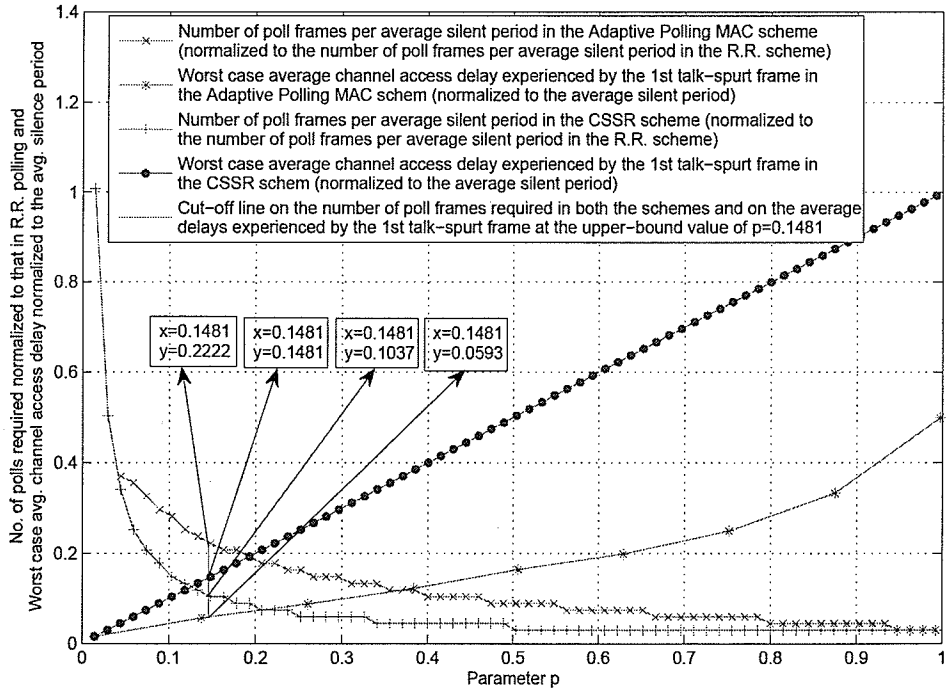

Figure 4.8. *Comparison of numerical results for Adaptive Polling MAC and CSSR approaches vs. round-robin approach: number of poll frames saved vs. the worst-case average channel access delay.*

As we see in Fig. 4.8, at the upper bound constrained value of parameter p, i.e.,

0.1481, the Adaptive Polling MAC scheme requires about 22% of the polls of round-robin scheme while the CSSR scheme needs about 10% of the polls of round-robin scheme to poll to a voice station during its silent period. We can see that with the CSSR approach, polling overhead of the round-robin scheme can be reduced more than with the Adaptive Polling MAC approach by a factor of 12%. However, in the worst-case situation, i.e., the talk-spurt frame arrives at the beginning of the threshold interval, the first talk-spurt frame in the CSSR approach will have to wait for about 14.8%, i.e., 200 *msec*, of the average silence period (1.35 *sec*) before it can be transmitted. In the Adaptive Polling MAC scheme, the first talk-spurt frame will have to wait for about 5.93%, i.e., 80 *msec*, of the average silence period in the worst-case situation. This unnecessary waiting time delay is caused due to the talk-spurt detection algorithm used in these approaches. It can be observed that in terms of the waiting time delays to the first talk-spurt frames, the Adaptive Polling MAC approach is better than the CSSR approach.

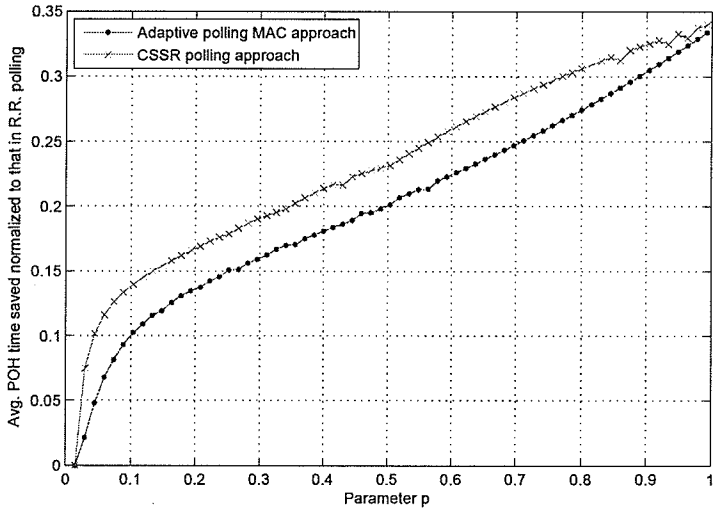

Figure 4.9. *Numerical results of the average polling overhead time saved at the scheduler per station per uplink voice activity cycle normalized to that in the round-robin polling at varying parameter p.*

Now we present numerical results over the two important performance metrics we analytically modeled in the previous sections. Fig. 4.9 shows the comparative numerical results of the average polling overhead time that the scheduler saves for each voice station during its one uplink voice activity cycle at varying parameter p.

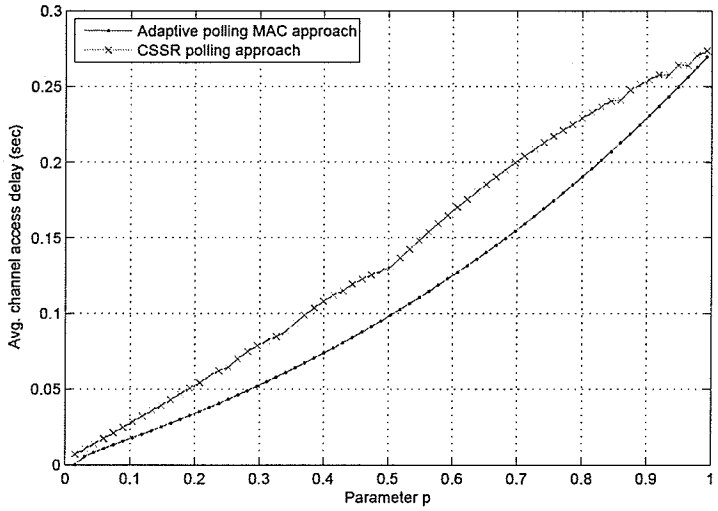

Figure 4.10. *Numerical results of the unnecessary average delay experienced by the first talk-spurt frame at varying parameter p.*

This average time saving is normalized to the average polling overhead time that would have been incurred by the scheduler in the round-robin polling during one uplink voice activity cycle. For example, within the constrained range of parameter $p = 0.1$, the scheduler saves 10% and 14% of the average polling overhead time that would have been incurred in the round-robin polling in case of Adaptive polling MAC and CSSR approaches, respectively.

Fig. 4.10 shows comparative numerical results of the unnecessary average waiting time delay at varying parameter p that the first talk-spurt frame at each voice station has to experience. As an example, we can see that, within the constrained range of parameter $p = 0.1$, the first talk-spurt frame at each voice station suffers waiting time delay of 20 *msec* and 30 *msec* in case of Adaptive polling MAC and CSSR approaches, respectively.

In Fig. 4.11, we present comparative numerical results of the average polling overhead time that the scheduler saves for each voice station during its one uplink voice activity cycle at varying silence periods. The results are obtained at a fixed value of parameter $p = 0.1481$ within its constrained range. This average time saving is normalized to the average polling overhead time that would have been incurred by the scheduler in the round-robin polling during one uplink voice activity cycle. For example, at average silence period of 1.2 *sec*, the scheduler saves 12% and 15% of the average polling overhead time that would have been incurred in the round-robin

polling in case of Adaptive polling MAC and CSSR approaches respectively.

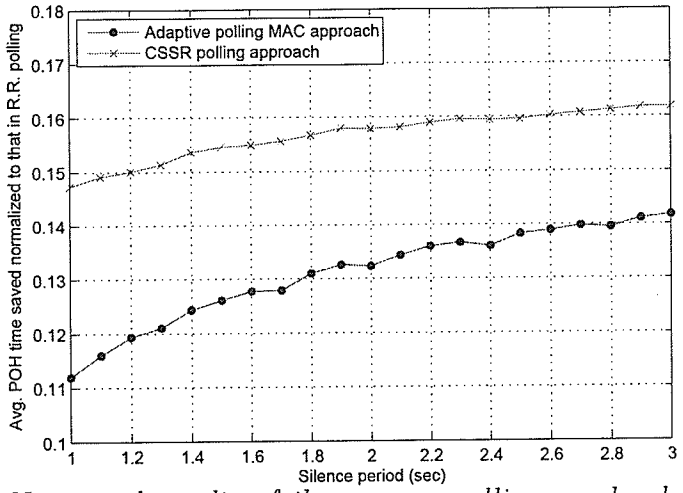

Figure 4.11. *Numerical results of the average polling overhead time saved at the scheduler per station per uplink voice activity cycle normalized to that in the round-robin polling at varying silence period.*

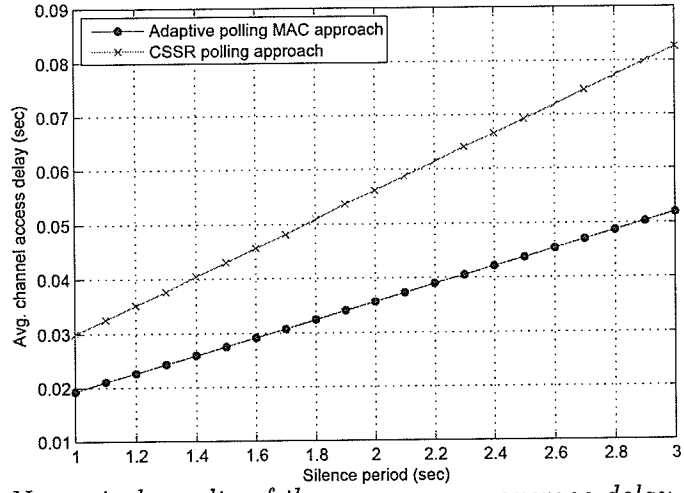

Figure 4.12. *Numerical results of the unnecessary average delay experienced by the first talk-spurt frame at varying silence period.*

Fig. 4.12 shows comparative numerical results of the unnecessary average waiting time delay at varying silence periods that the first talk-spurt frame at each voice station has to experience. As an example, at average silence period of 1.2 *sec*, the first talk-spurt frame at each voice station suffers waiting time delay of 23 *msec* and 36 *msec* in case of Adaptive polling MAC and CSSR approaches, respectively.

4.5 Chapter Summary

Performance of two well-known fully contention-free schemes that suppresses silence periods in voice calls has been analyzed in this chapter. The short-comings of these schemes have been elaborated. Mathematical models have been developed to evaluate two important performance metrics, i.e., the average polling overhead time that the scheduler saves for each voice station during its one uplink voice activity cycle, and the unnecessary average waiting time delay that the first talk-spurt frame at each voice station suffers. In order for the talk-spurt detection algorithm to divide the silent periods into threshold time intervals, constraints are defined for optimal selection of the parameter p.

Performance results have been presented and discussed. These results have shown that the CSSR approach is better than the Adaptive polling MAC approach in terms of the average polling overhead reduction. However, the CSSR approach is the worst in terms of unnecessary average waiting time delays that it causes to the first talk-spurt frames. Both of the schemes reduce the polling overhead time considerably during the silence periods of voice calls. But these schemes cause waiting time delays for medium access in the order of tens of milliseconds to the first talk-spurt frames. This unnecessary delay may not fulfill the end-to-end delay bound requirement of voice calls keeping in view the other contributing delays such as queuing, processing, transmission, and propagation delays within the core of the IP network.

Chapter 5

A Novel QoS-Aware MAC Protocol for Voice Services over IEEE 802.11-Based WLANs

5.1 Introduction

In the previous chapter, we have carried out performance evaluation of such protocols that suggest contention-free medium access through adaptive type of polling schemes to suppress the idle periods of voice calls. The performance results show that these schemes do so at the cost of increased waiting time delays to stations at the time when they change their state from idle to active contributing significantly to the end-to-end delays. In this context, we introduce a novel quality of service (QoS)-aware wireless medium access control (MAC) protocol in this chapter, called Hybrid Contention-Free Access (H-CFA) protocol. This protocol provides purely contention-free medium access mainly through an intelligent round-robin type polling algorithm adjoining it with a TDMA-like time slot algorithm for Contention-Free Activity Detection from idle to active state change.

In the proposed scheme, the bursty nature of the delay-sensitive voice traffic with the characteristic of periodic idle periods is exploited and idle periods are suppressed and multiplexed with other voice traffic efficiently at no cost of increased waiting time delays. This increases the capacity of WLANs significantly as compared to the simple round-robin type mechanism incorporated in the IEEE 802.11 a/b/g/e MAC standard. Also, the proposed scheme outperforms similar other schemes recently

proposed in the literature which generally improve WLAN capacity at the cost of increased waiting time delays. The H-CFA protocol combines different contention-free wireless medium access approaches, i.e., contention-free polling and Time-Division Multiple Access (TDMA)-like time slot assignment, in the single contention-free period (similar to the optional point coordination function (PCF) or the controlled access phase (CAP) in IEEE 802.11a/b/g or 802.11e respectively). The H-CFA protocol is augmented with a Contention-Free Activity Detection (CF-AD) algorithm and an intelligent polling list management algorithm.

We carry out performance analysis of the proposed H-CFA protocol through simulations in comparison to other similar contention-free protocols including the IEEE 802.11 round-robin polling scheme. The results show that the H-CFA protocol outperforms all its counterparts with regard to system capacity and waiting time delays to stations when they change their states from idle to active.

5.2 The Novel H-CFA Protocol

5.2.1 Objectives

The schemes, which we have discussed in the previous chapters, have specific characteristics but no one serves our purpose completely, i.e., to enhance the parameterized-QoS for time-sensitive voice services at one hand and to enhance the system capacity on the other hand. The schemes that provide priority access through the contention-based medium access approach can not guarantee delay bounds especially when the number of users increases. Although, the schemes which suggest silence suppression deploying the contention-based medium access approach enhance the capacity through multiplexing gain, they are unable to guarantee delay bounds especially in high load situations. H-CFA protocol is designed by exploiting some of the useful characteristics of the relevant research works in the literature we have described in the previous chapters. While developing the Hybrid Contention-Free Access (H-CFA) protocol, we set following objectives to achieve:

- It should be viable for very large scale commercial deployment (Fig. 5.1).
- It should increase the capacity of the AP by reducing the polling overhead.

- Capacity enhancement should not be at no cost of extra overhead such as increased waiting time delay to first MAC frame(s).
- It should provide guaranteed delay-bound service.
- Contention-free medium access approach should address the time synchronization issues and ensure efficient utilization of the bandwidth.

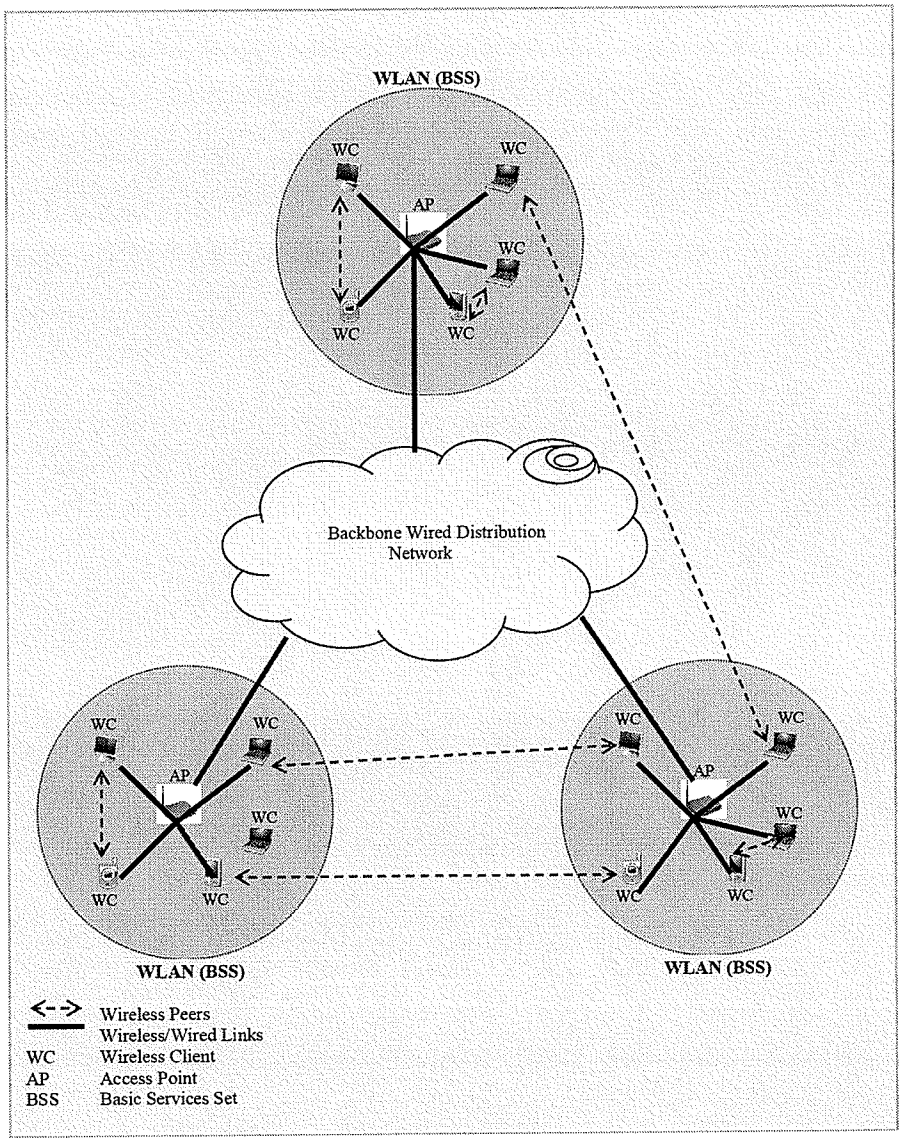

Figure 5.1. *Components of wide-scale heterogeneous WLANs environment.*

5.2.2 Protocol Overview

We achieve the above objectives through the H-CFA protocol. The H-CFA defines a Hybrid Contention-Free Interval (H-CFI) like the CAP in HCCA or the CFI in PCF. In H-CFA, the Dynamic Polling List Management (D-PLM) algorithm incorporated in the central controller eliminates completely the bandwidth wastage that is caused by the round-robin polling scheduler while polling all the admitted stations whether silent or talking. D-PLM function manages three logical lists. The first one is the Admitted Service List (ASL) that enlists the association ID's (BSSID) of all the admitted users (poll-able users) under the traffic stream (TS) setup mechanism in [4] on a first-come first-served basis and under a Call Admission Control (CAC) mechanism described in the next chapter. The second one is the Polling List (PL) that enlists association ID's of those admitted services which are in talking state. The initiation of H-CFI is must as far as the PL is not empty. The third one is the Idle Service List (ISL) that enlists those admitted services which are in the silent state. A comprehensive description and demonstration of the D-PLM algorithm is provided in the following section.

The H-CFA performs uplink status notification job through a Contention-Free Activity Detection (CF-AD) algorithm that works in the Contention-Free Activity Detection Interval (CF-ADI), a small portion of the H-CFI at its beginning. The CF-AD algorithm uses a TDMA-like time slot assignment scheme for this purpose. The remaining portion of H-CFI is further sub-divided into Contention-Free Downlink Voice Interval (CF-DVI) and Contention-Free Polling Interval (CF-PI) for downlink and uplink transmissions, respectively, through the round-robin polling. During the CF-DVI, traffic is transmitted down from AP to QoS-stations in FIFO order. During the CF-PI, the AP allocates polled-TXOPs to the active stations in its updated PL enabling them to send their frames up to the AP using contention-free polling type wireless medium access approach. Since we use both the approaches, i.e., contention-free polling and contention-free TDMA-like slot time assignment, we name it as Hybrid Contention-Free Access (H-CFA) protocol.The structure of the H-CFI is demonstrated in Fig. 5.2.

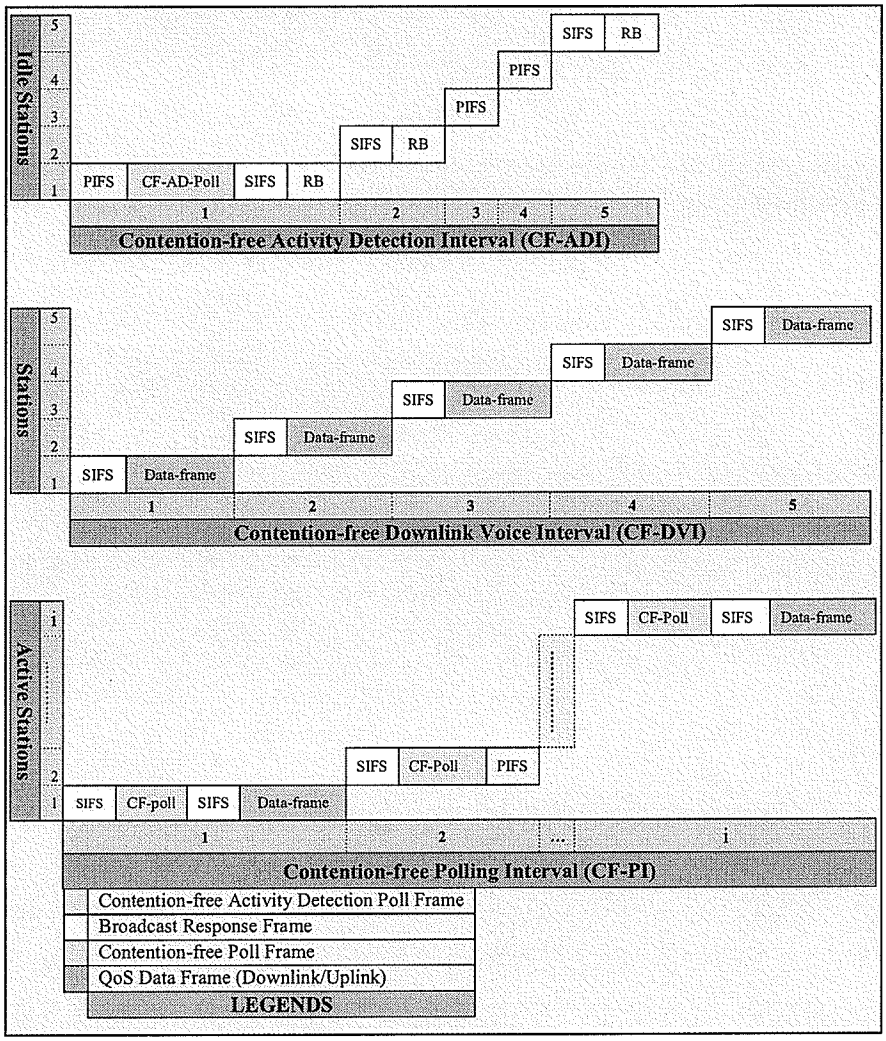

Figure 5.2. *Structure of Hybrid Contention-Free Interval (H-CFI).*

5.2.3 Protocol Description

The AP starts the H-CFI as long as there is at least one active user in the PL, either in the optional CFP of periodic beacon intervals or in the CP like CAP or the ICF as in [4], [32] respectively. However, the internal architecture and working of the H-CFI is quite different, as it comprises of three sub-intervals, i.e., Contention-Free Activity Detection Interval (CF-ADI), contention-free downlink voice interval (CF-DVI), and Contention-Free Polling Interval (CF-PI).

5.2.3.1 Contention-Free Activity Detection (CF-AD) Algorithm

With the start of H-CFI after the channel has been sensed idle for a PIFS time interval, the CF-ADI also starts. In the start of CF-ADI, the AP broadcasts a single Contention-Free Activity Detection Poll (CF-AD-Poll) frame having a field vector (FV) that represents the AIDs of all the idle stations in the ISL in the order as of the ISL. This FV enables idle stations to notify their status change to the AP in contention-free manner according to their respective order in the FV. In this way, a station at the first position in the FV gets the opportunity to respond first. The response from the idle station is also a Broadcast Response (RB) frame. The silent station responds only in the case of activity arrival, otherwise, it does not respond. The next silent station waits for a SIFS time period if it hears the RB frame, otherwise it waits for a PIFS time period before it can avail its own opportunity to respond. The sequences of actions during CF-ADI are shown in Fig. 5.3.

Following this procedure, each idle station in the FV can notify its uplink status change from idle to active state. The AP replaces the idle station from the ISL to the PL on hearing the RB frame from that station. If two consecutive idle stations do not respond, the next station responds after two PIFS time intervals and so will the AP expect from that station. We can see that the length of the CF-ADI (say *AD_Length* in terms of bit times) is a function of the number of activity arrivals and can be calculated as follows.

Let n be the number of idle stations in the FV and m out of these n idle stations do not broadcast RB frames. We denote the lengths of CF-AD-Poll frame, SIFS, RB frame, and PIFS, with *ADPol_Length*, *SIFS_Length*, *RB_Length*, and *PIFS_Length*, respectively, in terms of bit times. According to the CF-AD algorithm as described above,

$$\begin{aligned} AD\_Length &= I_n + (n-m) \times SIFS\_Length + (n-m) \\ &\quad \times RB\_Length + m \times PIFS\_Length. \end{aligned} \tag{5.1}$$

where $I_n = \{(0, ADPol\_Length) \,|I_{n=0} = 0 \; and \; I_{n\geq 1} = ADPol\_Length\}$, and $n = 0 \Rightarrow m = 0$. Following two extreme cases are worth considering when $n \geq 1$.

Case 1: There is no activity arrival (i.e., no user broadcasts RB frame), that means

$m = n$, in which case,

$AD\_Length = ADPol\_Length + n \times PIFS\_Length.$

Case 2: All n users in the ISL broadcast RB frame, i.e., $m = 0$, in which case,

$AD\_Length = ADPol\_Length + n \times SIFS\_Length + n \times RB\_Length.$

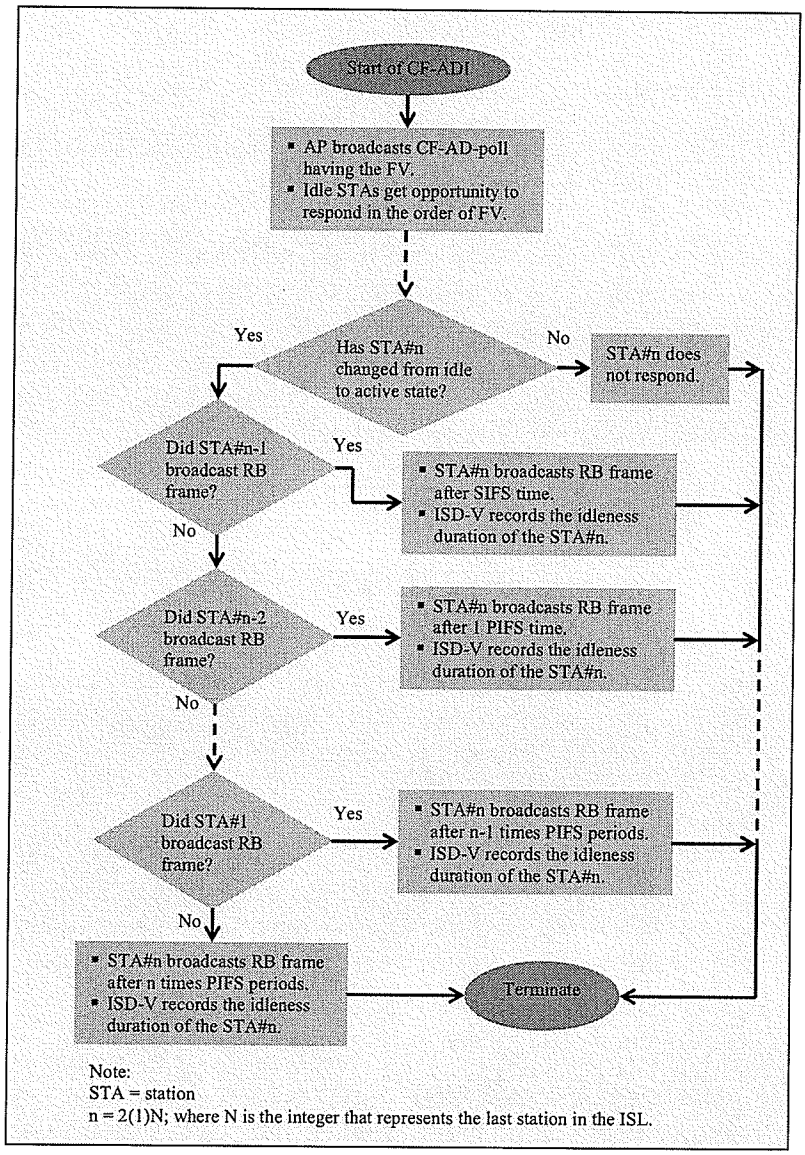

Figure 5.3. *Sequence of actions during contention-free activity detection interval of H-CFA.*

At the end of CF-ADI, the AP starts the contention-free downlink voice interval (CF-DVI) after a SIFS time interval. In CF-DVI, the AP sends frame(s) from its

local queue, separated by SIFS, to their destined stations (active or idle) in the FIFO order. At the end of CF-DVI, the Contention-Free Polling Interval (CF-PI) starts after a SIFS time interval. In the CF-PI, the AP sends the CF-Poll frame to allocate the polled-TXOP to active station according to its order in the PL. After a SIFS time interval, the polled station sends its frame to the AP. If the polled station has more than one frame, it may send multiple frame exchange sequences through the Contention-Free Burst (CFB), each frame sequence separated by a SIFS, as far as the TXOP-limit is not reached. The remaining frame exchange sequence(s) will either be dropped if it is only one or will wait for their transmission in the next CF-PI.

If transmission of the frame(s) is completed before the TXOP-limit, the remaining time of the TXOP is surrendered to the AP. After receiving the last frame, the AP waits for a SIFS time and then sends the CF-Poll frame to the next active station in the PL. If active to idle state change occurs during the allocated TXOP time, the station adds 0 as more data field in its last frame to let the AP know about its status change as in [35] (see D-PLM algorithm for details). In such a case, the AP will replace the station from PL to ISL. Moreover, keeping in view that voice calls can tolerate packet loss to some extent, the frame-acknowledgment scheme can be eliminated. Elimination of such unnecessary poll-frames and subsequent PIFS time intervals or null-frames, and ACK-frames reduces the unnecessary overhead to further increase the system capacity in order to be able to admit more real-time traffic streams (TSs).

5.2.3.2 Dynamic Polling List Management (D-PLM) Algorithm

A newly admitted voice station is presumed active and is placed at the top of the PL above the current active stations in the order of first-come first-served. The order of the polling list (PL) may not be the same in each H-CFI and may not be in the order (ascending/descending) of AID's of the active stations. However, the order of the Admitted Service List (ASL) is kept constant in the ascending order of the AID's of all admitted stations. The order of the Idle Service List (ISL) is maintained according to the FIFO principle, i.e., the station whose status has changed from active to idle state prior to another station will get prior position in the ISL. During the Contention-Free Activity Detection Interval (CF-ADI), an idle station avails its

respective opportunity to respond according to its order in the ISL.

The D-PLM function initiates a timer called Idle State Timer (IS-T) for each idle station as soon as it becomes idle. The IS-T calculates the silence duration of the silent station. With this timer, the D-PLM function calculates the exact duration of idle period of each station at the time when the idle station notifies its status change from idle to active state through the RB frame. In this way, a station for which the duration of idle period is shorter than any other station will get prior position in the PL. However, stations that notify their status change from idle to active state get their respective position at the head of the PL before the old active stations in the PL.

The D-PLM function maintains an Idle State Duration Vector (ISD-V). Based on the size of the silence durations, ISD-V records the positions of stations that notify their state change from idle to active during the current H-CFI. Stations with shorter silence durations get prior position in the ISD-V and are added to the head of the PL accordingly. Getting a position at the head of the PL does not seem to serve any purpose with regards to the priority/fairness to more talkative voice stations when we look with the perspective of cyclic polling (such as in round robin polling). In the cyclic polling, it does not matter which station is polled first as each station is polled in the cyclic order. However, with another perspective, getting prior position in the PL seems very much useful in terms of efficient bandwidth utilization. The H-CFI has some maximum limit depending on the delay bounds of the admitted services. Therefore, all the admitted stations may not be polled during one H-CFI (see in Fig. 6.1, Chapter 6). In such a situation, a station with frequent traffic arrival pattern as determined by the D-PLM function, that always get prior position in the PL will definitely avail more active H-CFIs where it is polled rather than being not polled due to the maximum limit of the H-CFI. The ISD-V determines the traffic arrival behaviors at individual voice stations. A station with shorter silence duration shows more frequent traffic arrival pattern than that with the longer silence periods. The D-PLM function decides whether a voice station has more frequent traffic arrival pattern on the basis of its last IS-T values. Intuitively, a station having more frequent arrival patterns of busrty traffic will need priority over the others. Therefore, the D-PLM algorithm smartly provides fair share of the bandwidth according to the requirements

of stations.

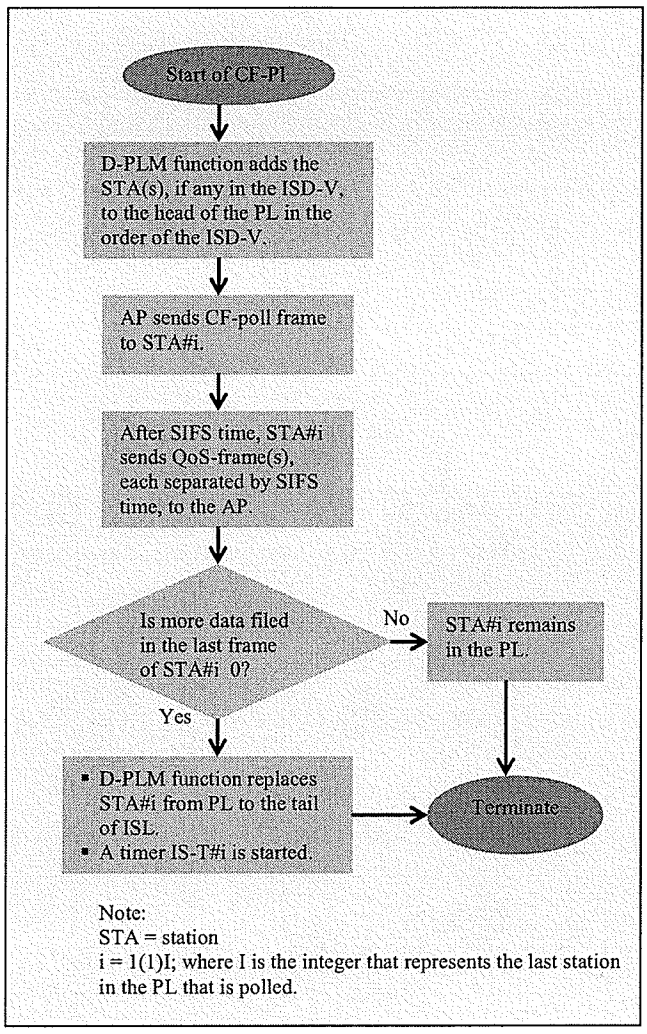

Figure 5.4. *Sequence of actions during contention-free polling interval in H-CFA.*

Active to idle state change of a station can be addressed by the D-PLM algorithm in two ways. A simple (but less efficient) way is one used in the legacy IEEE 802.11 standard where the station keeps quiet in case of transition from active to idle state. The AP does not know this status change in the current H-CFI until the next one when it polls this station and waits for PIFS time to know that the user has become silent. Another way is through in-band signaling, as in [28], [35]. This is a proactive approach but either needs modification of the MAC frame format by adding a "more data field" with a logic value 0 or 1 bit (with 0 indicating the last frame, and with

1 indicating more frames to follow) or needs help from the upper layers. In the H-CFA protocol, we propose the second option. With a minor change in the MAC frame format, this scheme can save considerable bandwidth wastage that is caused in polling idle stations and then waiting for PIFS time period or for a null frame. The sequences of actions during CF-PI are shown in Fig. 5.4.

5.3 Performance Evaluation

In this section, we compare the performance of H-CFA protocol with other related research works through simulations. We do not consider the schemes that address the fairness or priority-QoS issues or use (fully or partially) contention-based medium access approach for uplink status notification, because these works do not conform to our objectives. In the perspective of large scale commercial deployment of such schemes, the increased traffic load causes exponential growth in non-deterministic medium access delays. Since the H-CFA protocol is designed by combining the the two contention-free wireless medium access approaches, i.e., polling and TDMA-like time slots assignments, we compare H-CFA with such schemes which are also purely contention-free. These schemes are adaptive polling MAC scheme in [35]; CSSR polling scheme in [34]; and round-robin polling scheme of IEEE 802.11a/b/g/e MAC standards [3],[4].

5.3.1 Performance Metrics and Simulation Parameters

We evaluate two very important performance metrics. One is the system capacity that relates to the large scale commercial deployment of the scheme. Capacity of the system is the maximum number of voice stations that the AP can serve (poll) during the allowable length of a contention-free repetition interval. The other one is the wireless medium access waiting time delays to the first talk-spurt frame(s) of the bursty time-sensitive voice traffic when the periodic idle to active state change occurs. The waiting time delay is time in seconds that the first talk-spurt frame(s), arriving at the event of idle to active state change, has to wait due to the imposition of the threshold time intervals of the talk-spurt detection algorithm.

For bursty voice traffic, we consider a two-state Markov model of uplink voice

Table 5.1. *Simulation parameters.*

Parameters	Value
Average idle period (sec)	1.35
Average active period (sec)	1
Parameter p for [35] and [34]	0.08
Traffic arrival rate (Kbps)	64
Traffic transmission rate (Mbps)	2
Hybrid Contention-free Interval (msec)	20
PIFS length (μsec)	30
SIFS length (μsec)	10
CF-poll (bytes)	20
Data frame length (bytes)	160
Beacon frame length (bytes)	88
H-CF poll length (bytes)	30
Broadcast Response (BR) length (bytes)	14

activity. One is the idle state with mean duration of 1.35 seconds and the other one is the active state with mean duration of 1 second. Both idle and active periods have exponential distribution as used in [29],[26],[35]. During the active state, packet generation rate is assumed as 64 $Kbps$, i.e., 160 bytes per beacon/H-CFI cycle of 20 $msec$ as in IEEE 802.11 [3], [4]. The transmission rate is assumed to be 2 $Mbps$. We assume that the channel is loss-free, and the stations are always in power-awake mode, or at least in the soft power-sleep mode where they can always hear the broadcasts during the H-CFI's. We do not take into account the background contention-based data traffic as our schemes are fully contention-free. Values of the different parameters used in our simulations are shown in Table 5.1.

5.3.2 Performance Results

As we can see in Fig. 5.5, the adaptive polling MAC scheme in [35] supports 38 stations (with $p = 0.08$). This scheme saves bandwidth by not polling the stations for decreasing threshold time intervals during their idle state. The CSSR polling scheme

in [34] supports up to 40 stations for the same parameter value. The reason behind this is that it uses fixed threshold time intervals for not polling the stations during their idle state, and thus, saves more bandwidth than the scheme in [35]. The round-robin scheme of IEEE 802.11a/b/g/e supports only 25 stations due to the reason that it causes bandwidth wastage in polling the stations all the time whether they are idle or active.

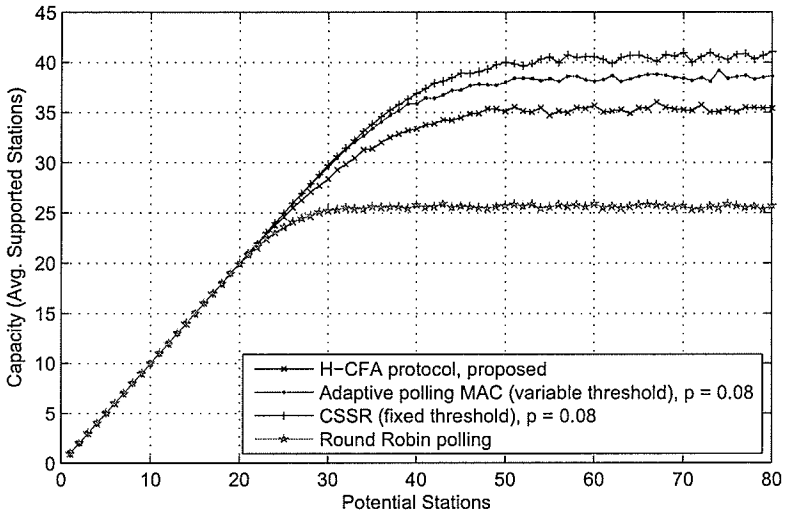

Figure 5.5. *Variations in WLAN capacities under different contention-free MAC schemes vs. round-robin scheme.*

As we can observe in Fig. 5.6, the average waiting time delay to first talk-spurt frame in case of the adaptive polling MAC scheme in [35] is 50 *msec.* This is due to the decreasing threshold time intervals during the idle state when the frame(s) has to wait for the station to be polled after the threshold time interval is over. In case of CSSR scheme in [34], the waiting time delay is even larger, i.e., 60 *msec.* This is due to the fact that the fixed threshold time intervals offer more time to arriving frames for waiting. The values of the waiting times in these two schemes indicate that the arriving frames have to wait for more than two consecutive H-CFI's which is not tolerable to the time-sensitive voice services.

As we can see in Figs. 5.5 and 5.6, the proposed H-CFA protocol outperforms all its counterparts. It can support 35 stations by not polling the stations during their silent periods as opposed to the round-robin polling. It does not enhance as much capacity as in [35], [34] due to the reason that it has to use a minimal bandwidth

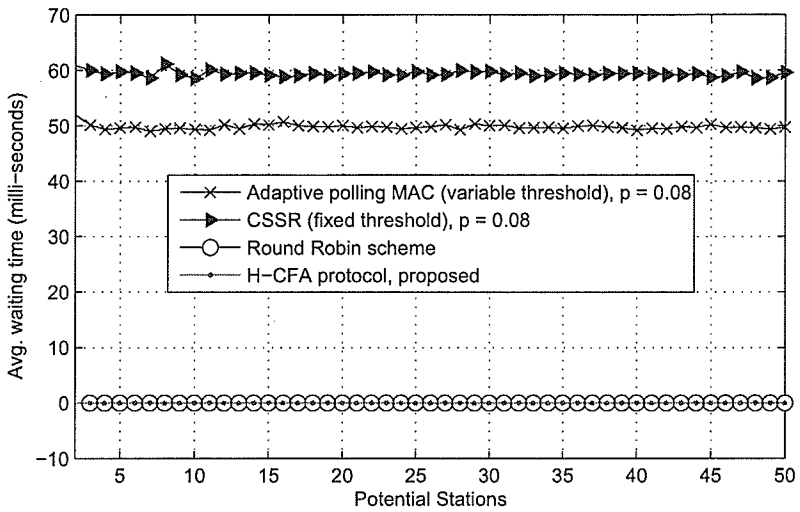

Figure 5.6. *Waiting time delay experienced by the first talk-spurt frame(s) in different schemes vs. in round-robin scheme.*

for Contention-Free Activity Detection (CF-AD) algorithm to work. However, as opposed to [35], [34], the H-CFA protocol reduces the bandwidth wastage efficiently and independently without the cost of increased waiting time delays or consecutive frame losses.

5.4 Chapter Summary

In this chapter, one of the core QoS issues, i.e., efficient silence suppression, has been addressed. A novel wireless MAC protocol called Hybrid Contention-Free Access (H-CFA) protocol has been proposed. The H-CFA protocol provides substantial multiplexing gains through silence suppression. This protocol provides purely contention-free medium access mainly through an intelligent round-robin type polling algorithm adjoining it with a TDMA-like time slot algorithm for Contention-Free Activity Detection from idle to active state change. It suppresses the idle periods of real-time voice efficiently at no cost of increased waiting time delays.

The performance analysis of H-CFA in comparison to other contention-free protocols, including the round-robin scheme (which is used in IEEE 802.11), reveals that the H-CFA protocol outperforms all its counterparts with regard to system capacity and waiting time delays. The H-CFA protocol enhances the capacity of the round-

robin scheme of IEEE 802.11a/b/g/e standard up to 40% at no cost except some minor structural changes in the PCF or CAP.

Chapter 6

QoS and Capacity Enhancement for VoIP in WiFi Networks: A Measurement-Based Call Admission Control (CAC) Scheme

6.1 Introduction

In the previous chapter, we have introduced our new Hybrid Contention-Free Access (H-CFA) MAC protocol that suppresses alternating silence periods of bursty voice traffic smartly and provides a fair share of the bandwidth according to the traffic arrival behavior at the individual voice station. Performance evaluation results of the H-CFA protocol have shown substantial QoS-aware capacity enhancement in WLANs. Since wireless medium bandwidth is limited, the delay sensitive voice traffic would need some minimum acceptable level of parametric-QoS guarantees such as maximum end-to-end delay and packet loss rate. The length of the Hybrid Contention-Free Interval (H-CFI) can not go beyond a certain maximum limit in order to keep the delays under the acceptable limits to the voice traffic. Therefore, a call admission control mechanism is required to limit the admitted voice services for guaranteeing them the negotiated delay bounds. To this end, we introduce a measurement-based call admission control (CAC) scheme for wireless VoIP, called Traffic Stream Admission Control (TS-AC) algorithm.

The TS-AC algorithm ensures efficient admission control for consistent delay-

bound guarantees and further maximizes the capacity through exploiting the voice characteristic that it can tolerate some level of non-consecutive packet loss. We apply our proposed TS-AC algorithm to the H-CFA protocol described in the previous chapter for carrying out its performance analysis through simulations. At different acceptable QoS levels of non-consecutive frame loss, the performance results show substantial increase in the system capacity while satisfying the voice users accordingly.

6.2 The Traffic Stream Admission Control (TS-AC) Mechanism

Our objective is to develop a smart wireless MAC protocol that can guarantee the requisite parametric performance metrics such as capacity enhancement, end-to-end delay bounds, consecutive packet loss control etc. None of the schemes that we have described in the previous chapters, except a few such as the IEEE 802.11e, considers the call admission control (CAC) issue which is very important for guaranteeing consistent parametric QoS to the admitted stations. While developing the measurement-based TS Admission Control (TS-AC) scheme, we put forth following targets to achieve:

- enhancement of the capacity of AP for large scale commercial deployment,
- capacity enhancement at no cost like increased waiting time delay to first MAC frame(s) that arrives when the silent to talk-spurt state change occurs,
- conforming to the maximum delay-bound limits, and
- no consecutive frame loss.

6.2.1 Protocol Overview

In order to provide a consistent level of parametric QoS, specially, the delay bounds and the acceptable level of packet loss, to already admitted traffic streams through the negotiation of TSPEC's as has been envisaged in [4], it is imperative to have an efficient control over the admission of new stations. For guaranteed delay-bound service a Call Admission Control (CAC) scheme, called Traffic Stream Admission Control (TS-AC) algorithm, is developed which maximizes the capacity gain through

exploiting the characteristic of bursty voice traffic that it can tolerate inconsecutive frame loss to some acceptable maximum level. This TS-AC algorithm keeps the number of admitted stations below some measured maximum count.

6.2.2 Protocol Description

The TS-AC function is incorporated in the AP that provides measurement based call admission control. Since the H-CFI must fulfill the delay bound requirements of all the admitted traffic streams, its length is limited (say $H\_CFI\_MAXlength$) according to the minimum negotiated delay bound of all the TSPEC's and cannot go beyond that maximum limit. The TS-AC function keeps track of the unused bandwidth (say H_CFI_Unused) during each H-CFI (i.e., $H\_CFI\_Unused = H\_CFI\_MAXlength - H\_CFI\_Timer$, where H_CFI_Timer is the actual bandwidth used during the current H-CFI). When a new call admission request comes, the TS-AC function differentiates the requested bandwidth with the measured unused bandwidth in the current H-CFI. If the differential is positive, it admits the call, otherwise, the call admission request is denied.

The proposed TS-AC algorithm maximizes the capacity by exploiting one characteristic of voice service that it can tolerate inconsecutive packet loss to some acceptable level. This means that all of the admitted traffic streams may not be served during one H-CFI. Therefore, in the worst scenario, the maximum number of admissible traffic streams depends on the fact that not a single user in the PL should remain un-served during two consecutive H-CFIs, so that the un-served station must not undergo consecutive frame loss. Therefore, the AP polls active stations in the PL in a round-robin fashion but stops where the H-CFI maximum limit comes. In the next H-CFI, it starts polling from the station where it had left in the last H-CFI. The TS-AC algorithm quantifies the increased capacity from the measured maximum number of TS's that will definitely be polled during the $H\_CFI\_MAXlength$ period. This quantification mechanism is based on the maximum negotiated QoS level in terms of percentage of times an admitted voice user may undergo inconsecutive packet loss. According to this algorithm, the greater the maximum negotiated QoS level, the larger will be the capacity, and vice versa. The D-PLM and the TS-AC functions are demonstrated in Fig. 6.1.

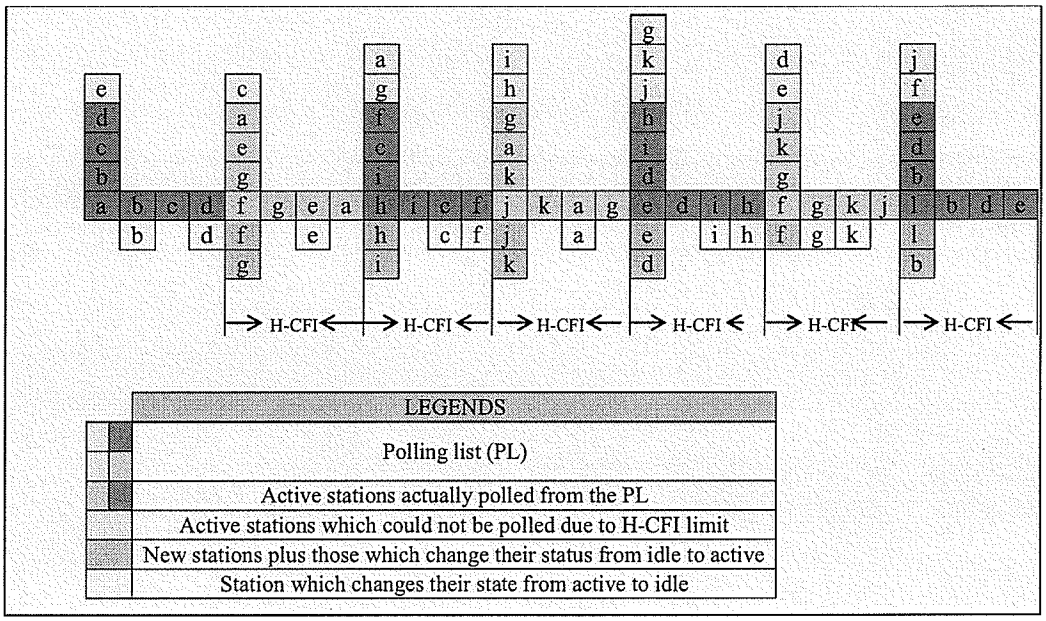

Figure 6.1. *Dynamic polling list management (D-PLM) and traffic stream addmission control (TS-AC) Algorithms.*

Let N_p, QoS_max, and N_a represent, respectively, the maximum number of TS's which are polled during a H-CFI; the maximum negotiated QoS level in terms of percentage of times an admitted voice station may undergo inconsecutive packet loss during its life-cycle of voice session; the resulting increased number of TS's (i.e., the sum of the number of TS's which are polled during the H-CFI and those which can not be polled due to $H\_CFI\_MAXlength$). Then

$$QoS\_max = \left[\frac{N_a - N_p}{N_p}\right] \times 100 \Rightarrow \; N_a = round\_off\left(\frac{N_p}{\left(1 - \left(\frac{QoS\_max}{100}\right)\right)}\right) \qquad (6.1)$$

6.3 Performance Evaluation

In this section we evaluate the proposed TS-AC algorithm through simulations. For this purpose, we apply our proposed CAC scheme to our proposed Hybrid Contention-Free Access (H-CFA) protocol to evaluate the variance in the system capacity at different levels of maximum negotiated QoS.

6.3.1 Performance Metrics and Simulation Parameters

We evaluate a very important performance measure. This is the system capacity, i.e., the maximum number of voice users that the AP can poll during each service interval or that can be satisfied according to the TSPEC parameters under the proposed call admission control mechanism. In case of TS-AC algorithm, we determine the number of satisfied voice stations which are among the potential stations served with their guaranteed delay bound requirements and do not suffer a single consecutive or inconsecutive packet loss below the agreed QoS level.

We consider a two-state Markov model of uplink voice activity. One of the states is the idle state with mean duration of 1.35 seconds and the other one is the active state with mean duration of 1 second (both idle and active periods have exponential distribution). We assume a packets generation rate of 64 $Kbps$, i.e., 160 bytes per beacon/H-CFI cycle of 20 $msec$. The transmission rate is assumed to be 2 $Mbps$. We do not take into account the background contention-based data traffic as our schemes are fully contention-free. We assume that the channel is loss-free, and the users are always in power-awake mode, or at least in the soft power-sleep mode where they can always hear the broadcasts during the H-CFI's. Various parameters used in our simulations are same as in Table 2 (chapter 5).

6.3.2 Performance Results

We apply the TS-AC algorithm along with the H-CFA protocol to evaluate the maximum number of voice stations that can be satisfied for different levels of maximum negotiated QoS. The results are shown in Fig. 6.2. At $QoS\_max = 0\%$ percent, which implies that the inconsecutive packet loss is acceptable 0 percent of times (i.e., each admitted station is polled in each H-CFI), the average number of satisfied stations (N_a) is about 35. At $QoS\_max = 10\%$, $N_a = 39$. At $QoS\_max = 25\%$, $N_a = 47$. At $QoS_{max} = 50\%$, $N_a = 70$. Without the application of the TS-AC scheme at $QoS\_max = 0\%$ and at $QoS_{max} = 50\%$, the results show that the number of satisfied users drops to zero after the respective maximum limit is reached.

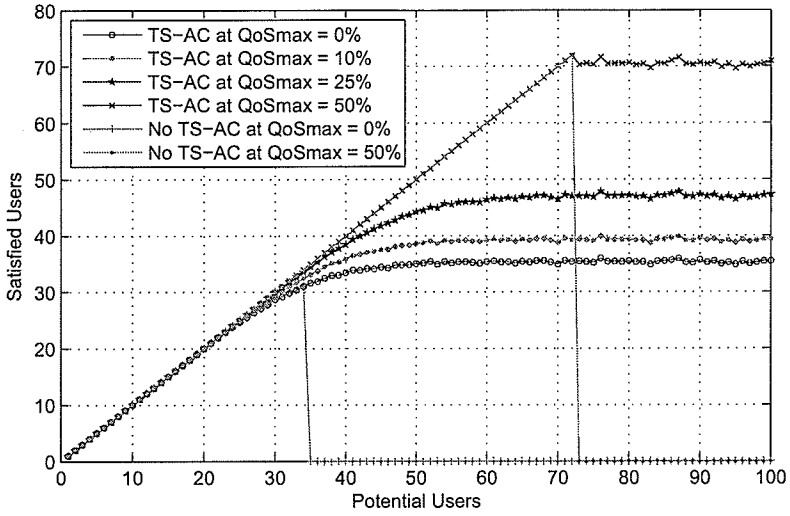

Figure 6.2. *Variations in the number of satisfied voice users in the proposed H-CFA protocol at various QoS levels with and without TS-AC scheme.*

6.4 Chapter Summary

In this chapter, one of the core QoS issues, i.e., call admission control, has been addressed. In order to provide a consistent level of parametric QoS, specially, the delay bounds and the acceptable level of packet loss to already admitted traffic streams through the negotiation of TSPEC's as has been envisaged in [4], we have introduced a Traffic Stream Admission Control (TS-AC) algorithm. This algorithm keeps the number of admitted stations below some measured maximum count. It maximizes the capacity by exploiting one characteristic of voice service that it can tolerate inconsecutive packet loss to some acceptable level.

The TS-AC algorithm quantifies the increased capacity from the measured maximum number of TS's that will definitely be polled during the $H\_CFI\_MAXlength$ period. This quantification mechanism is based on the maximum negotiated QoS level in terms of percentage of times an admitted voice user may undergo inconsecutive packet loss. In the TS-AC algorithm, we determine the number of satisfied voice stations, which are among the potential stations served with their guaranteed delay bound requirements and do not suffer a single consecutive or inconsecutive packet loss below the agreed QoS level.

We have applied the TS-AC algorithm along with the H-CFA protocol to evaluate

the maximum number of voice stations that can be satisfied for different levels of maximum negotiated QoS. The performance results of the TS-AC algorithm have shown that it does not only provide consistent delay satisfaction but also increases the capacity gain significantly at the cost of different accepted QoS levels.

Chapter 7

Conclusion

In this chapter we provide the summary of the research that has been discussed and presented in this thesis.

7.1 Summary

The chapter-wise summary provided bellow concludes our research.

- ***QoS issues in the IEEE 802.11 wireless LANs:*** we have presented a comprehensive study of IEEE 802.11 Wireless LAN MAC standard. In this context, we have explored and critically examined different important aspects like: IEEE 802.11 MAC architectural aspects; its limitations on QoS provisioning to real-time services; QoS requirements of real-time services; basic survey of QoS enhancement aspects; and efforts done so far within the standard and out of the standard. The enhanced version of the standard, i.e., IEEE 802.11e, defines two sub-functions, EDCA and HCCA, in its new Hybrid Coordination Function (HCF) for prioritized-QoS and parameterized-QoS enhancement respectively. The EDCA, which provides prioritized-QoS, is not suitable for delay sensitive services as it uses the non-deterministic wireless medium access approach, i.e., CSMA/CA, however, it is suited to enhance QoS for the best-effort data services.

  The HCCA that provides parameterized-QoS is suited to enhance QoS requirements of the delay sensitive services as it uses the deterministic channel access approach, i.e., round-robin polling. Besides, the HCF also outlines call admission control (CAC) strategies for both EDCA and HCCA for guaranteeing

service requirements of admitted services. The direct link protocol (DLP) that enables the peer QSTA's to communicate directly without QAP, and the block acknowledgement scheme for efficient utilization of the bandwidth are also discussed and explored.

Different error correction approaches, such as the reverse error correction and the forward error correction along with a joint-venture of these two approaches called hybrid FEC-ARQ error correction, are discussed with reference to its implication on the high speed error prone wireless channel. In view of the critical study of the WLAN standard, following research issues/aspects are explored. In the EDCA mode, the IEEE 802.11e MAC standard allows to dynamically change the parameters that affect prioritized-QoS provisioning, depending on the network conditions but does not define how to change these dynamically. The multi-rate PHYs capabilities defined in the IEEE 802.11 standard supports for dynamic rate switching according to the medium conditions to improve the throughput. However, the standard does not define the mechanism or the algorithm that can decide the best rate channel momentarily and then can efficiently or opportunistically switch to that channel momentarily.

In case of delay-sensitive services supported in the HCCA or in the PCF, the same round-robin type polling scheduler is proposed in all versions of the IEEE 802.11 MAC standard. As the round-robin scheduler does not suppress silence periods of bursty voice traffic, the standard allows to implement any optimized algorithm or scheme to do this job but does not specify any. The average values of the parameters that the round-robin scheduler uses for service reservation information are only suitable for constant bit rate (CBR) traffic. For variable bit rate (VBR) type of traffic, the standard does not define specific scheme. An other important research issue is the optimization of the trade-off between the channel efficiency, priority and fairness.

- ***QoS provisioning in the IEEE 802.11 WLANs:*** We have discussed and elaborated the core QoS-related design issue as our research problem that is common in all versions of IEEE 802.11 MAC standard (weather 802.11a/b/g or 802.11e). That is the inefficiency of the WiFi MAC standard in exploiting a very important characteristic of packet-switched networks, i.e., multiplexing

gains and quality of service provisioning in case of bursty type of traffic through suppressing the alternating silence periods. We have also explained the motivation and gravity for solving this QoS and capacity-related design problem that causes in-efficient utilization of the scarce wireless bandwidth.

We have presented a consolidated survey of the major research works on QoS-aware wireless MAC protocols in the recent literature and have provided their qualitative comparison categorically. We conclude that the fairness and priority based QoS schemes and the fully or partially contention-based schemes, due to their non-deterministic medium access nature, are not suited to fulfill the parametric-QoS requirements of delay sensitive services. However, the fully contention-free schemes, and call admission control schemes, are important in dealing with the capacity enhancement and parametric-QoS provisioning to real-time services. Although the fully contention-free schemes provide significant improvement in reducing the polling overhead, they are not efficient. Because they incur an increased waiting time delay to stations at the time when they change their state from idle to active, and thereby, increase the end-to-end delay substantially.

- ***Performance analysis of contention-free approaches for silence suppression in voice calls:*** We carried out and presented the performance evaluation of two well-known fully contention-free schemes which suppress the silence periods in voice calls. Mathematical models are developed to evaluate the two important performance metrics, i.e., the average polling overhead time that the scheduler saves for each voice station during its one uplink voice activity cycle, and the unnecessary average waiting time delay that the first talk-spurt frame at each voice station suffers.

  Numerical results show that both of the schemes reduce the polling overhead time considerably during the silence periods of voice calls. But these schemes cause waiting time delays for medium access on the order of tens of mile-seconds to the first talk-spurt frames. Keeping in view the maximum end-to-end delay that the voice call can tolerate and other contributing delays to it such as queuing, processing, transmission, and propagation delays, within the core of the IP network, this unnecessary channel access waiting time delay may not

make these schemes viable for voice services.

- ***A novel QoS-aware MAC protocol for voice services over IEEE 802.11-based WLANs:*** One of the core QoS issues, i.e., efficient silence suppression, is addressed. For efficient silence suppression and capacity enhancement, we have introduced a novel quality of service (QoS)-aware wireless medium access control (MAC) protocol, called Hybrid Contention-Free Access (H-CFA) protocol. This protocol provides purely contention-free medium access mainly through an intelligent round-robin type polling algorithm adjoining it with a TDMA-like time slot algorithm for contention-free activity detection from idle to active state change. It suppresses the idle periods of real-time voice efficiently at no cost of increased waiting time delays.

  The performance analysis of H-CFA protocol through simulations in comparison to other contention-free protocols, including the round-robin scheme, reveals that the H-CFA protocol outperforms all its counterparts with regard to system capacity and waiting time delays. The H-CFA protocol enhances the capacity of the round-robin polling scheme up to 40% at no cost except some minor structural changes in the PCF or CAP.

- ***QoS and capacity enhancement through a measurement based call admission control (CAC) scheme:*** One of the core QoS issues, i.e., call admission control, is addressed. In order to provide a consistent level of parametric QoS and the acceptable level of packet loss, to already admitted traffic streams, we have introduced a Traffic Stream Admission Control (TS-AC) algorithm. This TS-AC algorithm keeps the number of admitted stations below some measured maximum count. It maximizes the capacity by exploiting one characteristic of voice service that it can tolerate inconsecutive packet loss to some acceptable level. This capacity enhancement is based on the maximum negotiated QoS level in terms of the percentage of times an admitted voice user may undergo inconsecutive packet loss. The TS-AC algorithm determines the number of satisfied voice stations which are among the potential stations served with their guaranteed delay bound requirements and do not suffer a single consecutive or inconsecutive packet loss below the agreed QoS level.

  We have applied the TS-AC algorithm along with the H-CFA protocol to carry

out its performance analysis through simulations. As an important performance measure, we evaluate the maximum number of voice stations that can be satisfied for different levels of maximum negotiated QoS. The performance results of the TS-AC algorithm shows that it does not only provide consistent delay satisfaction but also increases the capacity gain significantly at the cost of different accepted QoS levels.

7.2 Future Work

The following outlines a few directions for future research in the area of QoS and capacity enhancement and bandwidth adaptation in the next generation broadband wireless IP networks (e.g., WiFi and WiMAX):

- Our ongoing research is focused on improving the H-CFA MAC protocol and the Traffic Stream Admission Control (TS-AC) algorithm and to carry out its analytical performance evaluation.
- In wireless networks, multi-path fading of the mobile radio channels is its fundamental trait induced by the changing strength of each path and the changing interference between these paths. Traditionally, channel fading is viewed as a source of unreliability and has to be mitigated. Recent and current research suggests another view of transmitting information opportunistically when and where the channel is strong by exploiting channel fluctuations, and it can enhance the throughput gains. In order for that, IEEE 802.11 standard supports multi-rate PHYs capabilities for dynamic rate switching that can improve the throughput. However, it does not define the mechanism or algorithm that can decide the best rate channel momentarily and then can efficiently or opportunistically switch to that channel momentarily. An opportunistic channel switching scheme or algorithm that takes the switching decisions based on the recently measured conditions on different channels can be developed in this context.
- The deployment of multi-radio PHY-layer technologies, such as MIMO (multiple-input multiple-output), in the emerging broadband IP technologies (i.e., IEEE 802.11n and IEEE 802.16) promises huge capacity enhancements and, at the same time, put forth new research dimension in how to optimize network per-

formance by intelligently and efficiently exploiting the multiple radios in the time and space dependent wireless medium. Even with these provisions, the IEEE standards have not matured enough to efficiently exploit these multi-radio capabilities in efficient manner. Therefore, one direction is to develop QoS-aware adaptive resource allocation MAC protocols for efficient bandwidth utilization in order for optimum performance in terms of throughput, QoS, and capacity gains.

- Another important issue is analytical modeling to evaluate the efficiency and performance of EDCF packet bursting and Contention-Free Burst (CFB) in WLANs.

Bibliography

[1] P. Slaby, "Front Lines: The Dangers of VoIP," December 1, 2005, availabl at: http://www.globetechnology.com/servlet/story/RIGAM.2005101.gtslabyoct12 /BNStory/Technology.

[2] B. H. Ashai, "Press Release: Enterprise VoIP over Wi-Fi Equipment Market to Reach $15bn by 2012, led by Cisco Systems," 03-2007, available at: http://www.juniperresearch.com/shop/viewpressrelease.php?pr=48.

[3] IEEE std. 802.11-1999, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications. 1999.

[4] IEEE Std. 802.11e/D13.0, "Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Amendment: Medium Access Control (MAC) Quality of Service (QoS) Enhancements," Jan. 2005.

[5] Q. Ni, L. Romdhani, and T. Turletti "A Survey of QoS Enhancements for IEEE 802.11 Wireless LAN," *Journal of Wireless Communications and Mobile Computing, Wiley*, Volume 4, Issue 5: pp.547-566, 2004.

[6] ITU-T, "General Characteristics of International Telephone Connections and International Telephone Circuits One- Way Transmission Time," G.114, Feb. 1996.

[7] R. Braden, D. Clark, and S. Shenker. "Integrated services in the internet architecture: an overview," IETF, RFC 1633, June 1994.

[8] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An architecture for differentiated services," IETF, RFC 2475, 1998.

[9] "Integrated Services over Specific Link Layers (issll)", Charter, ref. http://www.ietf.org/html.charters/issll-charter.html, Access date: June 2002.

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

[11] J. L. Sobrinho and A. S. Krishnakumar AS, "Real-time traffic over the IEEE 802.11 medium access control layer," *Bell Labs Technical Journal*, 172-187, 1996.

[12] A. Veres, A. T. Campbell, M. Barry, and L. H. Sun, "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, 19(10):2094-2104, 2001.

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

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

[15] A. Ganz, A. Phonphoem, and Z. Ganz, "Robust SuperPoll with Chaining Protocol for IEEE 802.11 Wireless LANs in Support of Multimedia Applications," *Wireless Networks Journal*, Springer Netherlands, 7: 65-73, 2001.

[16] I. Aad, and C. Castelluccia, "Remarks on Per-Flow Differentiation in IEEE 802.11," in *Proc. of European Wireless (EW2002)*, Florence, Italy, February 2002.

[17] L. Romdhani, Q. Ni, and T. Turletti, "Adaptive EDCF: Enhanced Service Differentiation for IEEE 802.11 Wireless Ad Hoc Networks," *IEEE WCNC'03 (Wireless Communications and Networking Conference)*, INRIA, France, 2:1373-1378, March 16-20, 2003.

[18] N. Ramos, D. Panigrahi, and S. Dey, "Quality of Service Provisioning in 802.11e Networks: Challenges, Approaches, and Future Directions," *IEEE Network* 0890-8044/05, July/August 2005

[19] A. Grilo, M. Macedo, and M. Nunes, "A Scheduling Algorithm for QoS Support in IEEE 802.11E Networks," *IEEE Wireless Communnication*, pp. 36-43, June 2003.

[20] P. Ansel, Q. Ni, and T. Turletti, "An Efficient Scheduling Scheme for IEEE 802.11e," in *Proc. Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks*, 2004.

[21] N. Ramos, D. Panigrahi, and S. Dey, "Dynamic Adaptation Policies to Improve Quality of Service of Multimedia Applications in WLAN Networks," in *Proc. International Workshop on Broadband Wireless Multimedia*, San Jose, CA, 2004.

[22] P. Brady, "A Model for Generating On-Off Speech Patterns in Two-Way Conversation," *Bell Syst. Tech. Journal*, vol. 48, no. 7, pp. 2245-2272, Sept. 1969.

[23] J. Charzinski, "Activity Polling and Activity Contention in Media Access Control Protocols," *IEEE Journal on Selected Area in Communications*, 18(9), 1562-1571, 2000.

[24] J. Y. Yeh and C.Chen, "Support of Multimedia Services with the IEEE 802.11 MAC Protocol," in *Proc. Of IEEE International Conference On Communications (ICC'02)*, pp. 600-604, 2002.

[25] T. Suzuki and S. Tasaka,"Performance Evaluation of Priority-based Multimedia Transmission with the PCF in an IEEE 802.11 Standard Wireless LAN," in *Proc.*

of IEEE Symposium on Personal, Indoor, and Mobile Radio Communications, pp. G-70- G-77, 2001.

[26] R. S. Ranasinghe, L. L. H. Andrew, and D. Everitt, "Impact of Polling Strategy on Capacity of 802.11 Based Wireless Multimedia LANs," in *Proc. IEEE International Conference on Networks (ICON)*, Brisbane, Australia, 1999.

[27] A. Kopsel and A. Wolisz, "Voice transmission in an IEEE 802.11 WLAN based access network," in *Proc. 4th ACM International Workshop on Wireless Mobile Multimedia (WoWMoM)*, pp. 24-33, Rome, Italy, 21 July, 2001.

[28] O. Sharon and E. Altman, "An Efficient Polling MAC for Wireless LANs," *IEEE/ACM Transactions on Networking*, vol. 9, no. 4, August 2001.

[29] H.-H. Liu and J.-L. C. Wu, "A Scheme for Supporting Voice over IEEE 802.11 Wireless Local Area Network," in *Proc. National Science Council ROC(A)*, vol. 25, no. 4, pp. 259-268, 2001, availabl at: http://nr.stpi.org.tw/ejournal/proceedingA/v25n4/259-268.pdf.

[30] QoS Baseline Ad Hoc Group, "QoS Baseline Proposal - Revision 2," IEEE doc. 802.11-00/360r2, Nov. 2000.

[31] P. Wang, H. Jiang, and W. Zhuang, "IEEE 802.11e Enhancement for Voice Service," *IEEE Wireless Communications*, vol. 13, no. 1, pp. 30-35, Feb. 2006.

[32] R. Y. W. Lam, V. C. M. Leung, and H. C. B. Chan, "Polling-based Protocols for Packet Voice Transport over IEEE 802.11 Wireless Local Area Networks," *IEEE Wireless Communications*, vol. 13, no. 1, pp. 22-29, Feb. 2006.

[33] N. Movahhedinia, G. Stamatelos, and H.M. Hafez, "Polling-Based Multiple Access for Indoor Broadband Wireless Systems," in *Proc. PIMRC 95*, Vol. 3, pp. 1052-6, 1995.

[34] E. Ziouva and T. Antonakopoulos, "Efficient Voice Communications Over IEEE 802.11 WLANs Using Improved PCF Procedures," in *Proc. Third International Network Conference (INC'02)*, University of Plymouth, UK, 16-18 July, 2002.

[35] Y.- J. Kim and Y.-J. Suh, "Adaptive Polling MAC Schemes for IEEE 802.11 Wireless LANs Supporting Voice-over-IP (VoIP) Services," *Wireless Commun. Mob. Comp.*, vol. 4, pp. 903-916, 2004.

[36] M. Veeraraghavan, N. Cocker, and T. Moors, "Support of Voice Services in IEEE 802.11 Wireless LANs," in *Proc. of IEEE INFOCOM'01*, 488-497 vol. 1, 22-26 April 2001.

[37] A. Bazzi, M. Diolaiti, and G. Pasolini, "Measurement Based Call Admission Control Strategies in Infrastructured IEEE8O2.11 WLANs," *IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 2093-2098, 2005.

[38] A. Bazzi, M. Diolaiti, C. Gambetti, and G. Pasolini, "WLAN Call Admission Control Strategies for Voice Traffic Over Integrated 3G/WLAN Networks," in *Consumer Communications and Networking Conference (CCNC) 2006*, vol. 2, pp.1234-1238, Jan. 2006.

PUBLICATIONS

The research and development work presented in this thesis is in the process of following publications.

1. **Irshad Qaimkhani**, Ekram Hossain, "Efficient Silence Suppression and Call Admission Control through Contention-Free Medium Access for VoIP in Wi-Fi Networks," *IEEE Communication Magazine*, to appear in Jan 2008.
2. **Irshad Qaimkhani**, Ekram Hossain, "A Novel QoS-Aware MAC Protocol for Voice Services over IEEE 802.11-Based WLANs," submitted to *Wireless Communications and Mobile Computing Journal*, Aug 2007.
3. **Irshad Qaimkhani**, Ekram Hossain, "Contention-Free Approaches in Wi-Fi MAC Design for VoIP Services: Performance Analysis and Comparison," submitted to *Wireless Communications and Mobile Computing Journal*, Nov 2007.