UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Departament de Teoria
del Senyal i Comunicacions

CTTC
Centre
Tecnològic
de Telecomunicacions
de Catalunya

Ph.D. Thesis

Cross-Layer Design and Optimization of Medium Access Control Protocols for WLANs

Author: Elli Kartsakli

Advisors: Luis Alonso, Ph. D.
Associate Professor
Universitat Politècnica de Catalunya (UPC)

Christos Verikoukis, Ph. D.
Senior Researcher
Telecommunications Technological Center
of Catalonia (CTTC)

Wireless Communications and Technologies Group
Department of Signal Theory and Communications
Universitat Politècnica de Catalunya

Barcelona, January 2012

Abstract

The widespread deployment of Wireless Local Area Networks (WLANs) has led to an increased need for efficient communication protocols that support high data rates and provide Quality of Service (QoS) guarantees. The Medium Access Control (MAC) layer plays a very important role in WLANs since it is mainly responsible for the regulation of channel access among the system users, the scheduling and resource allocation decisions, the encapsulation of data from upper layers into MAC frames and the selection of several transmission parameters.

This thesis provides a contribution to the field of MAC layer protocol design for WLANs by proposing and evaluating mechanisms that enhance different aspects of the network performance. These enhancements are achieved through the exchange of information between different layers of the traditional protocol stack, a concept known as Cross-Layer (CL) design. The main thesis contributions are divided into two parts that will be described next.

The first part of the thesis introduces Distributed Queuing Collision Avoidance (DQCA), a novel protocol for the MAC layer. DQCA behaves as a reservation scheme that ensures collision-free data transmissions at the majority of the time and switches automatically to an Aloha-like random access mechanism when the traffic load is low. After the detailed description of the protocol rules and operation, a link adaptation mechanism is proposed for the selection of the transmission rate according to the channel state of each user. Theoretical analysis and computer-based simulations show the performance improvement offered by DQCA with respect to the widely employed IEEE 802.11 standard.

The basic version of DQCA prescribes a First-In-First-Out (FIFO) scheduling order. However, DQCA can be enriched by more advanced scheduling algorithms based on a CL dialogue between the MAC and other protocol layers. High throughput performance can be obtained through channel-aware opportunistic scheduling schemes that employ information on the link quality provided by the Physical (PHY) layer. Furthermore, QoS provisioning, which is fundamental in the case of multimedia applications with stringent delay constraints, can be achieved through service-aware policies that determine the scheduling priority based on the requirements imposed by the application layer. The thesis proposes a number of scheduling algorithms applied over DQCA and discusses the performance enhancements and potential trade-offs of each scheme.

The close dependence of the channel-aware CL schemes on the availability of accurate information on the state of the wireless link has led to the design of a mechanism for the acquisition of periodic channel state updates. The obtained results show that the benefit of having valid channel state feedback overshadows the cost of the additional control overhead.

The second part of the thesis explores a different challenge in MAC layer design, related to the ability of multiple antenna systems to offer point-to-multipoint communications. Some modifications to the recently approved IEEE 802.11n standard are proposed in order to handle simultaneous multiuser downlink transmissions. A beamforming technique, which is based on the generation of random orthogonal beams and is compatible with the standard, has been employed at the PHY layer. At the MAC layer, a number of multiuser schemes that handle channel access and scheduling issues and provide mechanisms for the acquisition of the feedback information required by the PHY layer transmission technique are presented. The proposed schemes exploit multiuser diversity by opportunistically selecting the best set of users to minimize interference and increase the achieved throughput. The obtained performance enhancement is demonstrated with the help of both theoretical analysis and simulation obtained results.

Για το Νικόλα

Acknowledgements

The elaboration a PhD thesis is a long process with many ups and downs! I would like to express my gratitude to all the people that helped me during this phase of my life, both to my professional colleagues for keeping me on track in my research and to my friends for keeping me (almost) sane on a personal level.

First of all, I would like to thank my advisors, Dr. Luis Alonso and Dr. Christos Verikoukis for their continuous support, motivation and guidance through all the stages of the doctorate program. I would also like to thank my DQ accomplice, Dr. Jesús Alonso-Zárate, for his contributions in the world of distributed queuing! I deeply appreciate the help of Dr. Nizar Zorba for his help and collaboration during the latest stages of my work.

I would like to acknowledge the help of the GCRM group in the initial steps of my PhD, as well as the ongoing support from all the members of the WiComTec group that made possible the completion of this thesis.

Special thanks go to all my friends, to the old ones that followed me faithfully through this journey and to the new ones that did not hesitate to jump into this adventure. My gratitude is also extended to my colleagues in EETAC (mis compañeros de despacho y de la planta 2 – y algunos de la planta 1).

Finally, I could not thank enough my family, my parents for their love and support during these years, my sister for her calls that broke the monotony of my long days in the office and Nikolas for his enormous patience with me!

Contents

List of Tables

List of Figures

Acronyms

AC	Access Category
ACK	Acknowledgment
AMC	Adaptive Modulation and Coding
AP	Access Point
ARS	Access Request Sequence
CAA	Channel Access Algorithm
CDF	Cumulative Distribution Function
CL	Cross-Layer
CRA	Collision Resolution Algorithm
CRC	Cyclic Redundancy Check
CRQ	Collision Resolution Queue
CSI	Channel State Information
CSMA/CA	Carrier Sensing Multiple Access with Collision Avoidance
CTS	Clear to Send
CW	Contention Window
DCF	Distributed Coordination Function
DIFS	DCF Inter Frame Space
DQCA	Distributed Queuing Collision Avoidance
DTQ	Data Transmission Queue
EDCA	Enhanced Distributed Channel Access
ETI	Enable Transmission Interval
FBP	FeedBack Packet
FCS	Frame Check Sequence
FIFO	First-In First-Out
HCF	Hybrid Coordination Function
IEEE	Institute of Electrical and Electronics Engineers

ISO	International Organization for Standardization
MAC	Medium Access Control Layer
MCS	Modulation and Coding Schemes
MIMO	Multiple-Input Multiple-Output
MISO	Multiple-Input Single-Output
MOB	Multibeam Opportunistic Beamforming
MPDU	MAC Protocol Data Unit
MSDU	MAC Service Data unit
NACK	Negative Acknowledgment
NACK	Negative Acknowledgment
NAV	Network Allocation Vector
OFDM	Orthogonal Frequency Division Multiplexing
OSI	Open Systems Interconnection
PCF	Point Coordination Function
PHY	Physical Layer
PPDU	PHY Protocol Data Unit
QoS	Quality of Service
RSSI	Received Signal Strength Indicator
RTS	Request to Send
SIFS	Short Inter Frame Space
SIMO	Single-Input Multiple-Output
SISO	Single-Input Single-Output
SNIR	Signal-to-Noise-and-Interference Ratio
SNR	Signal-to-Noise Ratio
TXOP	Transmission Opportunity
WLAN	Wireless Local Area Network

Chapter 1

Introduction

1.1 Motivation

In the last decade, the Wireless Local Area Network (WLAN) market has been experiencing an impressive growth that began with the broad acceptance of the IEEE 802.11 standard [1]. The time when Internet access was a privilege reserved for few scientific or military applications seems long past: nowadays online wireless connectivity has become a part of everyday life, resulting to the expansion of the WLAN industry and the emergence of new technological challenges.

Technological advancement is closely interconnected to the available application scenarios and user requirements. As WLAN technology progresses, devices become cheaper and easier to deploy and maintain, while offering the same or more capabilities. New applications are designed to exploit the available technology and attract new users. In turn, the growth in the market and user demands serves as a driving force for further technological innovation.

The rapid development of WLAN systems has triggered several changes. The surge of new WLAN devices, such as smartphones, tablets and netbooks, has increased the need for mobility, flexibility and ubiquitous connectivity. As a result, WLANs are being deployed not only at home and office environments, but also at airports, hotels, hospitals, universities and other public hot spots. In addition, new demanding services for both personal and business applications have emerged, including Voice over IP (VoIP), web services, multimedia streaming, online gaming and video conferencing.

As the popularity of WLANs grows, more challenges need to be met. To begin with, networks must be able to support the increasing number of wireless users that contend for a limited amount of resources. Then, apart from connectivity, there is an insatiable need for faster wireless access with higher transmission rates and Quality of Service (QoS) guarantees, especially to enhance user satisfaction for time-sensitive multimedia applications. However, since wireless networks are limited by the scarcity of bandwidth and the time-varying and error prone nature of the wireless channel, the need for innovative technologies and mechanisms that provide increased spectral efficiency and robustness is imperative.

Improving the performance of WLANs is a multifaceted problem that engages different areas of the research community. Advances in the Physical layer (PHY) lead to sophisticated transmission techniques and advanced Modulation and Coding Schemes (MCS) that eventually enable faster and more reliable transmissions. A big step forward to PHY layer design has been made with the introduction of multiple antenna systems with advanced signal processing capabilities. To exploit the available PHY layer resources, there is a need to implement efficient channel access, scheduling and resource allocation algorithms at the Medium Access Control (MAC) layer. Finally, WLANs can also benefit from research on the upper layers, from advanced routing and security functions at the network layer to application layer management.

This thesis provides a contribution to the field of MAC layer protocol design for WLANs by proposing and evaluating mechanisms that enhance different aspects of the network performance. The main motivation for this work has stemmed from the following two factors:

- the *Cross-Layer (CL) design* principle. The separation of the network functionalities and responsibilities to different layers has been the traditional approach to network design. However, in an effort to meet the challenges of modern wireless networks and the increasing user demands, a new trend known as CL design has emerged. CL design is a wide term that encompasses all schemes that violate the principle of the layered architecture, from the exchange of information between layers to the joint layer design and optimization. The MAC layer, in particular, offers fertile ground for CL design since it constitutes the natural connection point between the PHY layer that deals with all the characteristics of wireless transmissions and the upper layers that impose QoS constraints.

- the support of *multiuser transmissions* by the PHY layer. The use of multiple antennas at the transmitter side (and optionally at the receiver side) in combination with advanced signal processing has offered the possibility of achieving simultaneous point-to-multipoint transmissions so that multiple users can be served at the same time, through the same frequency and code. This new capability opens many challenges at the MAC layer that must be able to handle multiuser channel access and scheduling and provide mechanisms for the acquisition of feedback information on the state of the wireless link that is typically required for the implementation of the PHY transmission techniques.

The main contributions of this work and the structure of the thesis will be discussed in detail in the following section.

1.2 Structure of the Thesis and Contributions

There are effectively two conceptual roads to MAC layer enhancement. The first lies in the modification of existing standards, with the aim to improve their performance or to provide them with new capabilities. In the context of WLANs, the prevalent standard is the IEEE 802.11 specification and its amendments which, besides their popularity, have several identified weaknesses and leave many issues open for research. The second approach is to propose more innovative solutions by designing novel MAC layer protocols outside the specifications. Naturally, advantages and disadvantages can be found with respect to both options, but there are also many valuable lessons to be learned. This thesis has followed both paths in order to study two different aspects of WLANs: CL design and multiuser transmission schemes.

The remaining part of the thesis consists of six chapters. Chapter 2 provides some necessary background information concerning the MAC layer functionalities in WLANs, the main features of the IEEE 802.11 specification, a description of the CL design principle and the most representative related works in the literature and, finally, the state of the art on multiuser MAC layer protocols. The innovative contributions of the thesis are organized into two parts. The first part consists of Chapters 3, 4 and 5 and is dedicated to a novel MAC layer protocol named Distributed Queuing Collision Avoidance (DQCA) and its enhancement through CL design. The second part of the thesis is formed by Chapter 6 and investigates possible modifications to the IEEE 802.11 standard to support multiuser communications in multiple antenna systems. Finally, Chapter 7 discusses the conclusions of the presented work and identifies potential lines for future investigation. In continuation, the main contributions of the thesis will be outlined in more detail.

The DQCA MAC protocol is the heart of the first part of the thesis, developed in Chapters 3 to 5. DQCA is a stable near-optimum MAC protocol that behaves as a random access mechanism under light traffic load and switches automatically to a reservation scheme as the traffic load grows. The inherent architecture of DQCA has several features that facilitate the incorporation of CL concepts. These include the distributed nature of scheduling that is based on two distributed queues and the structure of the DQCA frame sequence that enables the frequent exchange of feedback information within the network, usually formed by an Access Point (AP) and the associated users. Chapter 3 provides a detailed presentation of the DQCA protocol, including a set of algorithmic rules, the thorough description of the DQCA frame formats and operation examples.

DQCA provides an efficient channel access mechanism to handle contentions among users and guarantee the collision-free data transmission of data for the majority of the time. However, the basic DQCA protocol definition does not prescribe any methods for the selection of the transmission rate, in the case that multiple rates are available at the PHY layer. Chapter 4 presents a link adaptation scheme, seamlessly incorporated into the DQCA operation, that acquires information on the channel condition of each user and adapts accordingly the transmission rate to provide the desired bit error rate performance.

The most important contribution of this chapter is a mathematical model for the throughput and mean delay analysis of the DQCA protocol with link adaptation. This model is a solid tool for the evaluation of the DQCA performance under a time variant multi-rate channel for which the long term probabilities of supporting each available transmission rate exist and are known. Chapter 4 closes with the performance evaluation of DQCA under single-rate and multi-rate PHY layers (i.e., basic DQCA operation and DQCA with link adaptation, respectively).

Chapter 5 continues with one of the key objectives of this thesis, the incorporation of more advanced scheduling schemes based on a CL dialogue between the MAC and other protocol layers. High throughput and QoS provisioning are the main performance goals of the proposed CL schemes, with further considerations on other performance metrics such as delay and fairness. Knowledge of the condition of the wireless channel link at the MAC, provided by the CL interaction with the PHY layer, has led to the design of channel-aware opportunistic schemes. The main idea is that system throughput can be generally increased by encouraging transmissions when the channel condition supports the use of higher data rates. On the other hand, QoS provisioning can be achieved through service-aware policies that assign scheduling priorities on each traffic flow depending on the application service type and QoS requirements. With these concepts in mind, four CL-based scheduling algorithms are proposed, described in detail in Chapter 5. In order to investigate the enhancements offered by each scheme and identify the different performance trade-offs, three study cases will be considered for homogeneous (data only) and heterogeneous traffic conditions.

The last part of Chapter 5 addresses an issue that often affects channel-aware scheduling policies: the presence of outdated Channel State Information (CSI). This problem occurs when the CSI obtained through a link estimation mechanism does not reflect the link condition at the time of transmission, due to the time-varying nature of the wireless channel. To alleviate this issue, a mechanism for the acquisition of periodic CSI updates is proposed. It will be shown that, despite the additional overhead, a performance enhancement is achieved when accurate CSI is available.

The second part of the thesis is devoted to the investigation of MAC layer mechanisms for point-to-multipoint communications in multiple antenna systems. The recently approved IEEE 802.11n specification [2] supports advanced transmission techniques such as beamforming and spatial multiplexing but does not contemplate the possibility of simultaneous transmission to multiple destinations. This open challenge is addressed in Chapter 6 where a number of multiuser MAC schemes compatible with the IEEE 802.11n standard are presented.

The proposed schemes refer to the downlink communication direction and are based on a low-complexity beamforming PHY layer technique that serves users on random orthogonal beams. Their objectives are twofold. On the one hand they aim to increase the system capacity by transmitting to multiple users at the same time. On the other hand, they exploit multiuser diversity by opportunistically selecting the set of users with the best channel conditions and the lowest interference among them, thus supporting transmissions at higher rates. A mathematical model for the theoretical calculation of the throughout performance is also provided and is employed, along with simulation results, to explore the potential enhancements and trade-offs of the proposed MAC schemes.

1.3 Research Contributions

The novel proposals discussed in this thesis have been published in several research contributions. The work presented in the first part of this thesis, concerning the DQCA MAC protocol and its enhancement through link adaptation and CL-based scheduling algorithms, has been published in one book chapter, five journals and seven international conferences, cited next:

[BC1] **E. Kartsakli**, J. Alonso-Zárate, L. Alonso, and C. Verikoukis, “Cross-Layer Scheduling with QoS Support over a near-optimum distributed queuing protocol for wireless LAN”. *Wireless Network Traffic and Quality of Service Support: Trends and Standards*, IGI Global Publishing, USA, Mar. 2010.

[J1] A. Antonopoulos, J. Alonso-Zárate, **E. Kartsakli**, L. Alonso, and C. Verikoukis, “Cross Layer Access Point Selection Mechanisms for a Distributed Queuing MAC Protocol,” accepted for publication in the *Special Issue on Mobility Management in Future Internet of the Springer Telecommunications Systems Journal*, vol. 84, Feb. 2011.

[J2] **E. Kartsakli**, J. Alonso-Zárate, L. Alonso, and C. Verikoukis, “Cross-Layer Scheduling with QoS Support over a Distributed Queuing MAC for Wireless LANs,” *ACM/Springer Mobile Networks and Applications (MONET), Special Issue on Recent Advances in IEEE 802.11 WLANs: Protocols, Solutions and Future Directions*, vol. 14, pp. 709–724, Dec. 2009.

[J3] **E. Kartsakli**, C. Verikoukis, and L. Alonso, “Performance Analysis of the Distributed Queuing Collision Avoidance (DQCA) Protocol with Link Adaptation,” *IEEE Transactions on Wireless Communications*, vol. 8, pp. 644–647, Feb. 2009.

[J4] **E. Kartsakli**, A. Cateura, J. Alonso-Zárate, C. Verikoukis, and L. Alonso, “Cross-Layer Enhancement for WLAN Systems with Heterogeneous Traffic based on DQCA,” *IEEE Communications Magazine*, vol. 46, pp. 60–66, June 2008.

[J5] J. Alonso-Zárate, C. Verikoukis, **E. Kartsakli**, A. Cateura, and L. Alonso, “A near-optimum cross-layered distributed queuing protocol for wireless LAN,” *IEEE Wireless Communications Magazine [medium access control protocols for wireless LANs]*, vol. 15, pp. 48–55, Feb. 2008.

[C1] J. Alonso-Zárate, C. Verikoukis, **E. Kartsakli**, and L. Alonso, “Coexistence of a Novel Medium Access Control Protocol for Wireless Ad Hoc Networks and the IEEE 802.11,” in *Proc. of IEEE International Conference on Communications (ICC 2010)*, pp. 1–5, May 2010.

[C2] J. Alonso-Zárate, **E. Kartsakli**, A. Antonopoulos, L. Alonso, and C. Verikoukis, "Handoff Functions for a Distributed Queuing Collision Avoidance Medium Access Control Protocol for Wireless LANs," in *Proc. of International Conference on Ultra Modern Telecommunications Workshops (ICUMT 2009)*, Oct. 2009.

[C3] **E. Kartsakli**, J. Alonso-Zárate, L. Alonso, and C. Verikoukis, "QoS Guarantee for Wireless LAN with Heterogeneous Traffic," in *Proc. of ICT Mobile and Wireless Communications Summit (ICT-MobileSummit 2009)*, June 2009.

[C4] **E. Kartsakli**, A. Cateura, J. Alonso-Zárate, C. Verikoukis, and L. Alonso, "Cross-Layer Enhancement for WLAN Systems with Heterogeneous Traffic Based on DQCA," in *Proc. of IEEE International Conference on Communications (ICC 2007)*, pp. 5708–5713, June 2007.

[C5] **E. Kartsakli**, A. Cateura, J. Alonso-Zárate, C. Verikoukis, and L. Alonso, "Opportunistic Scheduling using an Enhanced Channel State Information Update Scheme for WLAN Systems with DQCA," in *Proc. of IEEE 65th Vehicular Technology Conference (VTC 2007 Spring)*, pp. 1021–1025, Apr. 2007.

[C6] C. Verikoukis, J. Alonso-Zárate, **E. Kartsakli**, A. Cateura, and L. Alonso, "Cross-Layer Enhancement for WLAN Systems based on a Distributed Queuing MAC protocol," in *Proc. of IEEE 63rd Vehicular Technology Conference (VTC 2006 Spring)*, pp. 1293–1297, May 2006.

[C7] **E. Kartsakli**, A. Cateura, C. Verikoukis, and L. Alonso, "A CL Scheduling Algorithm for DQCA-based WLAN Systems with Heterogeneous Voice-Data Traffic," in *Proc. of 4th IEEE Workshop on Local and Metropolitan Area Networks (LANMAN 2005)*, Sept. 2005.

The DQCA protocol description has also been included in a book on collision resolution algorithms and a book chapter on CL design:

[B1] **E. Kartsakli**, J. Alonso-Zárate, A. Cateura, C. Verikoukis, and L. Alonso, *"Contention-Based Collision-Resolution Medium Access Control Algorithms"*. Nova Science Publishers Inc., Apr. 2009.

[BC2] **E. Kartsakli**, J. Alonso-Zárate, A. Antonopoulos, and L. Alonso, "MAC Protocols with Cross-Layer Design". *Cross Layer Designs in WLAN Systems*, Troubador Publishing Ltd., Sept. 2011.

The multiuser MAC schemes, discussed in the second part of this thesis, have been presented in one journal and two international conferences and a patent (filed in 2009):

[J6] **E. Kartsakli**, N.Zorba, L. Alonso, and C. Verikoukis, "A Threshold-Selective Multiuser Downlink MAC scheme for 802.11n Wireless Networks," *IEEE Transactions on Wireless Communications*, vol. 10, pp. 857–867, Mar. 2011.

[C8] **E. Kartsakli**, N. Zorba, C. Verikoukis, and L. Alonso, "A Threshold-Selective Multiuser Downlink MAC Scheme for 802.11n Wireless Networks," in *Proc. of IEEE International Conference on Communications (ICC 2010)*, May 2010.

[C9] **E. Kartsakli**, N. Zorba, C. Verikoukis, and L. Alonso, "Multiuser MAC Protocols for 802.11n Wireless Networks," in *Proc. of IEEE International Conference on Communications (ICC 2009)*, June 2009.

[P1] C. Verikoukis, N. Zorba, **E. Kartsakli** and L. Alonso, "Method and Apparatus for Medium Access Control in a Wireless Broadband System with Multiple-Input Multiple-Output or Multiple-Input Single-Output Technology with Multiuser Capabilities," *Patent number PCT/EP2009/057276*, filed June 2009.

Apart from publications directly related to the thesis contributions, a number of other research works have been carried out during the elaboration of this thesis. In particular, two book chapters have been produced:

[BC3] A. Antonopoulos, **E. Kartsakli**, L. Alonso, and C. Verikoukis, "Dealing with VoIP Calls During Busy Hour in LTE, Recent Advances in Wireless Communications and Networks". *LTE, Recent Advances in Wireless Communications and Networks*, Jia-Chin Lin (Ed.), 2011.

[BC4] J. Alonso-Zárate, **E. Kartsakli**, L. Alonso, and C. Verikoukis, "Cooperative ARQ: A Medium Access Control (MAC) Layer Perspective". *Radio Communications*, Alessandro Bazzi (Ed.), INTECH, Apr. 2010.

Another line of investigation has been the adaptation of the DQCA MAC protocol for operation over distributed ad hoc networks. The work in this field has been published in four journals that will be cited next and several international conferences that will not be explicitly mentioned here for brevity.

[J7] J. Alonso-Zárate, **E. Kartsakli**, M. Katz, L. Alonso, and C. Verikoukis, "Multi- Radio Cooperative ARQ in wireless cellular networks: a MAC layer perspective," *Telecommunication Systems (Springer), Special Issue on Challenges in Next-Generation and Resource-Constrained Networks*, 2011.

[J8] J. Alonso-Zárate, **E. Kartsakli**, L. Alonso, and C. Verikoukis, "Performance Analysis of a Cluster-Based MAC Protocol for Wireless Ad Hoc Networks," *EURASIP Journal on Wireless Communications and Networking*, 2010.

[J9] J. Alonso-Zárate, **E. Kartsakli**, C. Skianis, C. Verikoukis, and L. Alonso, "Saturation Throughput Analysis of a Cluster-based Medium Access Control Protocol for Single-hop Ad Hoc Wireless Networks," *SIMULATION Transactions of The Society for Modeling and Simulation International*, vol. 84, pp. 619–633, Dec. 2008.

[J10] J. Alonso-Zárate, **E. Kartsakli**, C. Verikoukis, and L. Alonso, "Persistent RC- SMA: A MAC Protocol for a Distributed Cooperative ARQ Scheme in Wireless Networks," *EURASIP Journal on Advanced Signal Processing, Special Issue on Wireless Cooperative Networks*, May 2008.

Chapter 2

Background

2.1 Introduction

The main objective of this thesis is to propose and evaluate medium access, scheduling and resource allocation algorithms implemented at the MAC layer and enhanced through CL interactions. This chapter will provide some background information on MAC protocols and the relevant state of the art that will facilitate the understanding of the contributions of this thesis.

This chapter begins with an overview of the MAC layer functions and the most representative families of MAC layer protocols, presented in Section 2.2. A significant part of the existing work on MAC layer enhancement for WLANs is based on the widespread IEEE 802.11 specification and its amendments. Hence, the main features of the standard will be discussed in Section 2.3, focusing on the legacy IEEE 802.11 version of the standard, the IEEE 802.11e for QoS provisioning and IEEE 802.11n for Multiple-Input Multiple-Output (MIMO) systems.

Section 2.4 opens with a general description of the CL design principles. After presenting the most common CL classification methods found in the literature, a MAC-centric taxonomy is adopted that organizes CL-based schemes into four categories depending on their major objective: link adaptation and scheduling, power control, application adaptation and parameter tuning. More emphasis will be given in the link adaptation and scheduling and parameter tuning schemes that have served as a motivation of this thesis.

Finally, the use of MIMO technology opens the road to the design of multiuser schemes where point-to-multipoint communication is possible. Section 2.5 will discuss the state of the art on multiuser MAC layer protocols and the additional challenges that arise from simultaneous transmissions.

2.2 The MAC layer

2.2.1 Functions of the MAC layer

The rapid development of computer and communications technology towards the end of the 1970s boosted the popularity of networking and accentuated its commercial potential. In order to prevent the deployment of multiple and incompatible network architectures by different vendors, the International Organization for Standardization (ISO) developed the Open Systems Interconnection (OSI) reference model [3].

The OSI reference model provides an abstract framework for network design that enables the interconnection of heterogeneous networks. It can be visualized as a vertical stack of seven independent layers, with the upper layers dedicated to application-related issues and the lower layers to data transmission. Each layer is an entity that addresses a specific functionality, even though the actual protocols that implement the layer functions may vary depending on the system. Communication is possible only between adjacent layers and is limited to the exchange of a set of primitives through well-defined interfaces. In other words, a layer provides services to the adjacent higher layer and, in turn, receives services from the layer below.

The MAC is the lower of the two sublayers of the Data Link Control layer, i.e., the second layer of the OSI reference model. It plays a very important role in WLANs since it is responsible for the regulation of channel access among the system nodes: it defines the rules by which the nodes compete for access to the shared medium and provides mechanisms for the resolution of collisions. The term nodes or users will be employed interchangeably in this thesis to denote wireless devices that form part of the WLAN. Depending on the network topology, the nodes can talk directly to each other in an ad hoc mode (peer-to-peer), or communicate through an AP if an infrastructure scenario is considered.

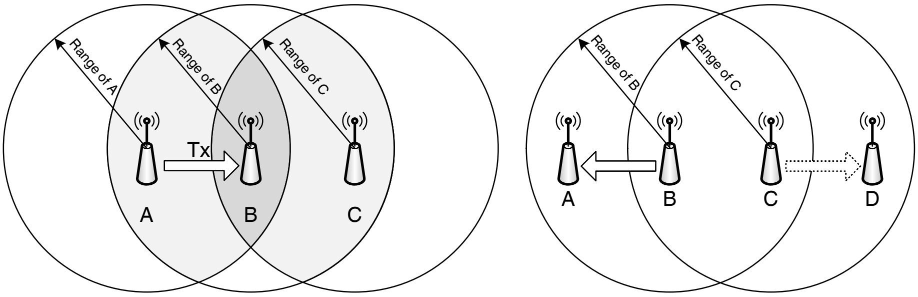
Another essential function of the MAC is the scheduling of transmissions by or towards a particular node. Scheduling has a strong impact on the network performance and can be selected to serve specific performance goals, such as throughput maximization or fairness among users. With the proliferation of multimedia applications, QoS-aware MAC scheduling aiming to satisfy specific application requirements, usually concerning time-delay constraints and tolerated error margins, has become a critical task.

The MAC layer is also responsible for the data encapsulation that includes the assembly of frames before transmission and the frame parsing and error detection upon reception. In the frame assembly process, the data from upper layers of the protocol stack are encapsulated into MAC layer frames. The specific details of the frame formation depend on the MAC protocol, but typically a header with the required control information and address fields and a Cyclic Redundancy Check (CRC) for error detection are appended to each frame. After the reception of a frame, the CRC is checked to determine whether the frame has been received correctly or with errors. In the case of errors, the frame is discarded and retransmissions of the frame may be requested, depending on the MAC protocol rules.

Wireless communications introduce additional challenges that should be handled by the MAC layer. First, the detection of collisions that occur when nodes access the medium simultaneously is difficult, especially by the transmitter. Hence, wireless MAC protocols are expected to minimize collisions through collision avoidance or reservation schemes and provide efficient mechanisms for the collision resolution.

Second, the MAC should take into consideration two well-known issues often encountered in WLANs regarding the presence of hidden and exposed terminals. Consider the case illustrated in Figure 2.1 (a), where node B lies within the range of A and C but A and C cannot listen to each other. Assume that there is an ongoing transmission from A to B. Node C, being out of the range of node A (i.e., node A is hidden from C), will falsely sense the channel idle and may initiate a transmission, thus causing a collision at B. This situation is known as the hidden terminal problem.

Now assume that there is a fourth node D within the range of node C but out of the range of node B, as shown in Figure 2.1 (b). In this particular setup, two simultaneous transmissions could take place, from B to A and from C to D (since A and D are sufficiently far from C and B, respectively). However, if B initiates a transmission to A, node C will sense the medium busy and will defer from transmitting to D. This situation is referred to as the exposed terminal problem.



(a) The hidden terminal problem: Node C cannot hear the ongoing transmission from A to B.

(b) The exposed terminal problem

Figure 2.1: The hidden and exposed terminal problems

Finally, other issues that should be handled by the MAC include user mobility and energy efficiency. In practice, it is very difficult to design a MAC protocol that efficiently tackles all these challenges, especially given the wide variety of application scenarios, deployment conditions and constraints in WLANs. Hence, all MAC protocols have strengths and weaknesses and their efficiency depends on the obtained performance trade-offs.

2.2.2 MAC Protocols for WLANs

Different strategies can be adopted to achieve the aforementioned MAC layer functionalities in the context of WLANs, leading to a vast number of MAC protocols proposed in the literature. Without attempting to provide an exhaustive overview of MAC protocols, this section will discuss some representative examples of MAC strategies which are more relevant to the contributions of this thesis.

The main objective of MAC protocols is to share the available resources among the system users. There are four basic multiple access strategies:

- *Time Division Multiple Access (TDMA).* In TDMA, users are served in different portions of time (usually named time slots), using all the available frequency bandwidth of the system.
- *Frequency Division Multiple Access (FDMA).* In FDMA, users are assigned different portions of the available frequency bandwidth. In other words, users transmit at different frequencies at the same time.
- *Code Division Multiple Access (CDMA).* In CDMA, users are assigned different pseudo-random spreading codes that enable them to transmit at the same time and frequency.
- *Space Division Multiple Access (SDMA).* When multiple antennas are available, the spatial multiplexing can be employed to allow multiple simultaneous transmissions at the same frequency.

These access methods provide collision-free access to the channel, assuming there is centralized coordination and some prior reservation phase during which the allocation of resources takes place. When this is not the case, users must compete for channel access and collisions often occur. This situation is handled by contention-based MAC schemes that must perform two functions: the Channel-Access Algorithm (CAA) that handles the channel access attempt of a user upon the arrival of a new packet and the Collision Resolution Algorithm (CRA) that defines the retransmission scheme of the collided packets in such a way so as to reduce the probability of further collisions.

Without being exhaustive, the most typical CAA policies are the following:

- *Free Access* algorithms in which users can attempt to access the channel as soon as they have packets to transmit. Random access schemes where users attempt transmission at a given time with a certain probability can be considered as a variation of free access algorithms.
- *Blocked Access* algorithms in which users are not allowed to attempt the transmission of new packets as long as there are pending collisions to be resolved.
- *Carrier Sensing Multiple Access (CSMA)* algorithms in which users must sense the channel idle before attempting a transmission.

- *Polling* schemes must receive a polling request by a central controller (usually the AP) in order to attempt transmission.

With respect to the CRA policies, the following categories can be considered:

- *Random* algorithms in which users wait a random interval of time before attempting a retransmission, in case of collision. The random defer time can be selected based on a given probability distribution, or more elaborate backoff mechanisms can be employed.
- *Tree* or *Splitting* algorithms in which the users involved in a collision are divided into groups. Then, users within a given group compete again for channel access with a reduced probability of collision.

There is an extensive number of MAC protocols in the literature and some comprehensive overviews can be found in [4] and [5]. Clearly, each scheme has unique features and capabilities and may be better adapted to a specific network scenario. Nevertheless, combinations of the aforementioned CAA and CRA policies can be found in the core of most existing MAC protocols. In continuation, some specific MAC protocol examples will be given.

The simplest example of contention-based MAC is the Aloha family of protocols. Aloha was the first implemented MAC for wireless packet data networks, invented in the 1970s [6]. It is a free access protocol in which users with new packets attempt to access the channel immediately. In the case of collisions, they defer access for a random amount of time before attempting a retransmission. Slotted Aloha is another well known protocol of the Aloha family that divides time into slots and restricts the initiation of any transmission attempt in the beginning of a time slot. This method doubles the throughput with respect to pure Aloha at the cost of time synchronization.

Another large family of protocols is known under the name of tree or splitting algorithms. Splitting algorithms usually operate in a time slotted manner, they support either free or blocked channel-access method and and are based on some channel feedback about the state of the channel. The first splitting algorithm is the Binary Tree algorithm, proposed independently and almost concurrently by Capetanakis [7], [8] and Tsybakov and Mikailov [9] in 1978. The idea is simple: once a collision occurs, the involved users are split into two subsets by flipping an unbiased coin (i.e. both sides have the same probability). The subset of users that flipped one of the sides transmits in the next slot whereas the other subset defers transmission until all the users of the first subset have transmitted successfully. This procedure is applied recursively to resolve any collisions among the users of the same subset. Several enhancements and alternatives of the basic scheme have been proposed in the literature and overviews can be found in [10] and [11].

Originally, splitting algorithms were designed for slotted Aloha-type channels with applications in wired Ethernet, satellite communications, and mobile wireless systems [12]. In the context of WLANs, splitting CRAs are often employed in reservation schemes, to resolve collisions among channel request attempts that take place within a reservation phase. The idea is to restrict collisions in the reservation phase where small control packets are employed and reduce or eliminate collisions during the data transmission. This type of protocols has served as a basis for the distributed protocol DQCA that will be presented in Chapter 3.

The most popular channel access method in WLANs is based on carrier sensing. CSMA-based algorithms adopt the process of listening to the channel and attempt a transmission only after the medium is sensed idle. Several variations of this scheme exist depending on how the users act upon finding the channel busy:

- In *1-persistent CSMA*, users sense the channel continuously and attempt a transmission as soon as the channel becomes idle again.
- In *non-persistent CSMA*, users defer sensing for a random amount of time. After the random backoff time elapses, they sense the channel again and the process is repeated until the channel is found idle.
- In *p-persistent CSMA*, users sense the channel continuously. Once the channel is sensed idle, they attempt a transmission with a random probability p.

The mandatory access mode of the widely used IEEE 802.11 is based on CSMA with Collision Avoidance (CSMA/CA), a variation of p-persistent CSMA that employs a random backoff mechanism to avoid collisions. The following section will describe some of the most important features included in the standard.

2.3 The IEEE 802.11 Specification for WLANs

The IEEE 802.11 specification for the PHY and the MAC layers is the most popular technology adopted in WLANs. The first version of the standard, often called the legacy IEEE 802.11, was issued in 1997. Since then, several amendments have been approved by the IEEE, providing enhancements such as higher data rates, QoS provisioning and increased security. A revised and corrected version of the standard that includes all major amendments (mainly IEEE 802.11 a/b/e/g) was issued in 2007 [1]. The newest addition in the 802.11 family is the IEEE 802.11n standard for higher throughput that introduces new functionalities for both MAC and PHY [2]. The IEEE 802.11 protocol architecture is illustrated in Figure 2.2.

The PHY layer defines radio transmission related parameters such as the frequency band, the data rate, the transmission technique and the number of antennas. The IEEE 802.11 specification offers multirate transmission in the 2.4 GHz and 5 GHz spectrum bands. The maximum available data rate varies depending on the version of the standard. Initially, the legacy version only supported rates up to 2 Mbps, but with the IEEE 802.11b amendment capacity increased to 11 Mbps. The adoption of Orthogonal Frequency Division Multiplexing (OFDM) in IEEE 802.11a/g resulted to significantly higher data rates of up to 54 Mbps, with a cost on the coverage range since higher rates require high modulation schemes that are

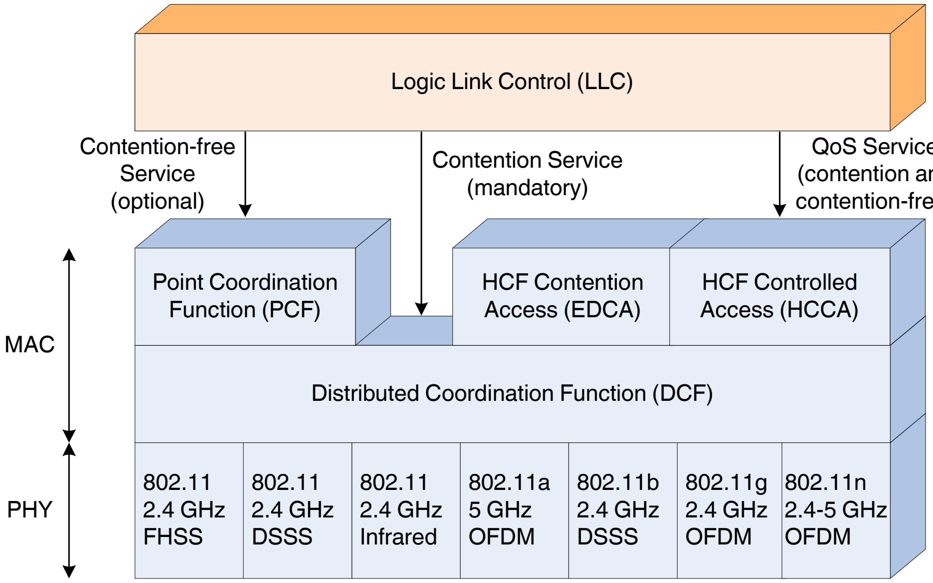

Logic Link Control (LLC)
Contention-free Service (optional)
Contention Service (mandatory)
QoS Service (contention and contention-free)
MAC
PHY
Point Coordination Function (PCF)
HCF Contention Access (EDCA)
HCF Controlled Access (HCCA)
Distributed Coordination Function (DCF)
802.11 2.4 GHz FHSS
802.11 2.4 GHz DSSS
802.11 2.4 GHz Infrared
802.11a 5 GHz OFDM
802.11b 2.4 GHz DSSS
802.11g 2.4 GHz OFDM
802.11n 2.4-5 GHz OFDM

Figure 2.2: The IEEE 802.11 protocol architecture

more susceptible to errors. Finally, the recently approved IEEE 802.11n standard supports MIMO technology with rates exceeding 100 Mbps (and up to 600 Mbps, depending on the number of antennas). A summary of the main features of the IEEE 802.11 and its amendments is given in Table 2.1.

The remaining of this section is divided into five parts. Section 2.3.1 briefly describes the IEEE 802.11 frame encapsulation. With respect to the MAC layer functions, the legacy IEEE 802.11 defined two access methods: the mandatory Distributed Coordination Function (DCF) and the optional Point Coordination Function (PCF). The DCF, described in Section 2.3.2 is the fundamental access method, used both in infrastructure and ad hoc configurations, whereas the PCF, described in Section 2.3.3, is a contention-free centralized polling scheme.

The IEEE 802.11e amendment added the Hybrid Coordination Function (HCF) that combines two access methods: a distributed scheme called Enhanced Distributed Channel Access (EDCA), and a centralized scheme called HCF Controlled Channel Access (HCCA). These access schemes extend the functionality of the DCF and the PCF, respectively, to provide service differentiation and QoS support. An overview of the EDCA and the service differentiation mechanism of IEEE 802.11e is given in Section 2.3.4.

Finally, Section 2.3.5 provides a brief description of the main PHY and MAC layer characteristics included in the IEEE 802.11n amendment for MIMO systems.

Table 2.1: Summary of IEEE 802.11 PHY layer specifications

Standard	*Date*	*Frequency (GHz)*	*Maximum Rates (Mbps)*	*Technology*
802.11	1997	2.4	2	DSSS [a], FHSS [b], Infrared (IR)
802.11a	1999	5	54	OFDM
802.11b	1999	2.4	11	High-Rate (HR) DSSS
802.11g	2003	2.4, 5	11 (DSSS), 54 (OFDM)	Extended Rate (ERP) DSSS and OFDM
802.11n	2010	2.4, 5	from 65 to 600 [c]	High-Throughput PHY (MIMO-OFDM, multiple antennas)

[a] Direct Sequence Spread Spectrum

[b] Frequency-Hopping Spread Spectrum

[c] Depending on configuration (number of spatial streams, bandwidth, etc.)

2.3.1 Frame Encapsulation in IEEE 802.11

As mentioned in Section 2.2.1, the MAC layer is responsible for the encapsulation of data from upper layers of the protocol stack into MAC layer frames. This process is illustrated in Figure 2.3. According to the IEEE 802.11 terminology, the upper protocol message is called MAC Service Data unit (MSDU) and the MAC frames are known as MAC Protocol Data Units (MPDUs).

The MSDU is the information payload that arrives to the MAC from the upper layers (e.g., a TCP/IP or UDP frame). Depending on its size, the MSDU is either encapsulated within a single MPDU (as in the example of Figure 2.3) or it is fragmented into multiple MPDUs that are transmitted sequentially by the MAC. The MPDU is formed by three parts: the header with the necessary MAC-dependent control information, the frame body that contains the MSDU (or a part of it, if fragmentation has taken place) and the Frame Check Sequence (FCS), which is a CRC for error detection, calculated over the MAC header and the frame body. Finally, the MPDUs are passed down to the PHY layer, where a header and a preamble are added, thus forming the PHY Protocol Data Unit (PPDU). The header contains PHY related control information, whereas the preamble is required for detection, synchronization and channel estimation by the receiver. The reverse procedure takes place upon data reception, where data units move upwards the protocol stack.

This terminology is also adopted in this thesis. However, for the sake of readability, the MSDU is often referred to as the data or the application message. Similarly,

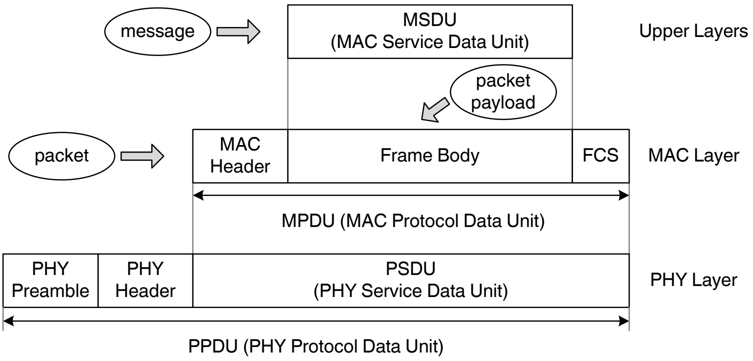


Figure 2.3: MAC layer frame encapsulation

the MPDUs are often referred to as packets and the frame body of each packet is denoted as the packet payload or simply payload. A maximum payload size can be contained within each packet that depends on the MAC protocol. As a result, messages are often fragmented to multiple payload segments that are encapsulated into MAC layer packets.

2.3.2 The Distributed Coordination Function (DCF)

The DCF is an asynchronous transmission mode based on Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) and is the mandatory channel access scheme of the IEEE 802.11 standard. To understand this mechanism, consider a scenario where a source station (or transmitter) wants to transmit a data packet to the destination station (or receiver). Before transmitting, the source station must listen to the channel for a predefined time interval called DCF Inter Frame Space (DIFS). If the channel is sensed idle during the DIFS period, the station seizes the channel and initiates the data packet (DATA) transmission. Otherwise, if the channel is sensed busy, the Binary Exponential Backoff (BEB) algorithm is executed.

According to the BEB algorithm, the station maintains a backoff counter that is set to a random value uniformly distributed within the interval $[0, CW]$. CW is referred to as the Contention Window, and it is initialized at a predefined minimum value CW_{min}. The source station keeps listening to the channel and for every idle slot time (i.e., a time unit defined at the PHY layer) it decreases the backoff counter by one unit. When the counter expires, the station attempts the data transmission. If the transmission is successful, the CW size is reset to the minimum value. Otherwise, in case of failure, the CW is doubled, according to the expression:

$$CW_i = 2^i \cdot CW_{min} \tag{2.1}$$

and the a new value for the backoff counter is selected from the interval $[0, CW_i]$ for the ith transmission attempt. The parameter i can have any integer value from 0 up to a maximum value known as the maximum backoff stage, defined in the PHY layer specification. By obliging the stations to defer transmission for a random amount of time when the channel is busy or a transmission failure occurs, the backoff algorithm reduces the probability of collision among stations that are competing for access to the wireless medium.

If the number of failed transmission attempts exceeds the retransmission limit (which is a MAC-dependent parameter), the station discards the data packet and resets the *CW* to the minimum value, in order to proceed with the transmission of the next buffered packet. Otherwise, upon the correct reception of a data packet, the destination station sends back an acknowledgment frame (ACK) after waiting for a Short Inter Frame Space (SIFS). The SIFS is introduced to compensate for propagation delays and radio transceivers turnaround times required to switch from receiving to transmitting mode. Note that the SIFS is shorter than the DIFS, to ensure that acknowledgments are given greater priority than regular data traffic. An example of the DCF operation is illustrated in Figure 2.4.

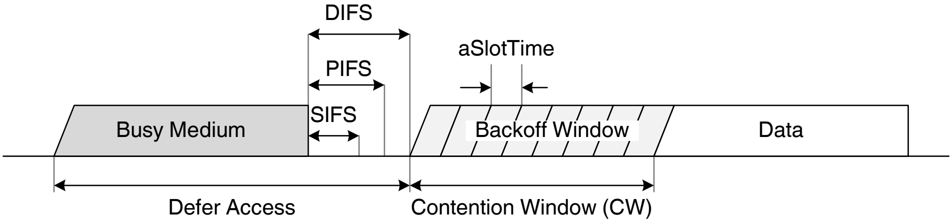


Figure 2.4: The DCF access method

The DCF also defines an optional RTS/CTS (Request to Send / Clear to Send) handshake mechanism, aimed to alleviate the hidden terminal problem. This handshake takes place as follows. When a station gains access to the channel according to the DCF rules (i.e., after sensing an idle medium for a DIFS time or after the expiration of the backoff counter), it transmits an RTS frame, instead of directly transmitting the data packet. If the destination receives the RTS, it replies with a CTS. After a successful RTS/CTS exchange, the source is enabled to transmit its data packet and waits for the reception of the ACK. This procedure is shown in Figure 2.5

The RTS/CTS exchange is in fact a means to announce the impeding use of the medium and reduce the probability of collisions among DATA packets. The source station estimates the time required for the complete transmission of its data packet (counting from the completion of the RTS transmission until the ACK reception) and includes this value in the duration field of the RTS frame. This information is copied to the subsequent CTS and DATA frames, accordingly modified to reflect the remaining time of the transmission sequence. When a station not directly involved in the ongoing transmission sequence receives any of these frames, it reads the duration field and updates the Network Allocation Vector (NAV). The NAV can be thought of as a counter that contains the time that the medium is expected to be occupied by an ongoing transmission and is decreased at a uniform rate. A station must defer transmission until the NAV counter has expired, even if it detects no

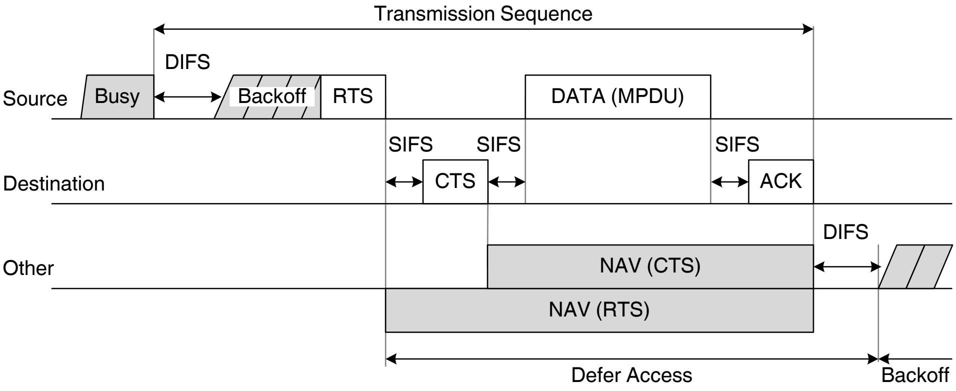


Figure 2.5: The RTS/CTS mechanism

transmission activity in the wireless medium (which is a situation that may occur if the station is outside the range of the transmitter). This process is known as the Virtual Carrier Sensing mechanism.

2.3.3 The Point Coordination Function (PCF)

The PCF is an optional access mechanism for infrastructure-based networks, in which the communication between the station is coordinated by the AP. The PCF is a hybrid scheme in which time is divided into two intervals:

- the Contention-Free Periods (CFP) during which the PCF polling scheme is implemented
- the Contention Periods (CP) during which the basic access mode, DCF, is employed

The interchange between these two phases is controlled by the AP, as shown in Figure 2.6. To indicate the beginning of the CFP, the AP transmits a beacon (B) to all users. However, before transmitting the beacon, the AP must gain access to the channel according to the DCF rules (i.e., sense the channel idle and execute the backoff algorithm if necessary). However, a different sensing time is defined for the PCF access. Unlike the DCF where a user must sense the channel idle during a DIFS, a shorter PCF Inter Frame Space (PIFS) is employed by the AP. The duration of a PIFS is shorter than a DIFS but longer than a SIFS, thus providing the initiation of a CFP with less priority than the transmission of control packets (CTS or ACK), but with higher priority than the transmission of data packets.

Once the AP gains access to the channel, it initiates a polling scheme and sends poll messages that offer transmission opportunities to the stations. Stations are allowed to transmit data packets after being directly polled by the AP. If a polled station has no data to transmit, it responds with a NULL packet. During the CFP,

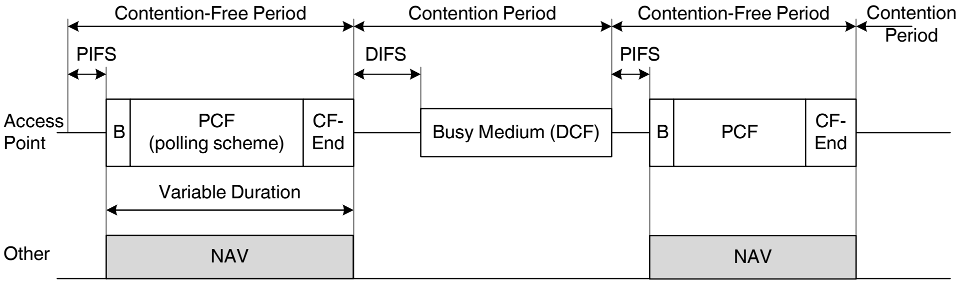


Figure 2.6: The PCF access method

the AP periodically transmits beacons that contain information regarding the duration of both the CFP and the CP and allow a new station to be associated to the network. The CFP is completed whenever the AP transmits a CF-End control packet.

2.3.4 The Hybrid Coordination Function (HCF)

The HCF, defined in the IEEE 802.11e specification, combines two access methods that support QoS provisioning: the contention-based EDCA, which will be further described in this section, and the centralized HCCA, which bears some similarities to the legacy PCF and can be further consulted in the standard specification [1].

The IEEE 802.11e introduced several new features to the legacy IEEE 802.11 mechanism. The major innovation has been the definition of eight priority levels for data traffic, mapped into four Access Categories (AC) for voice, video, best-effort and background data services (in order of decreasing priority). Every station with QoS support has four transmission queues, one for each AC, that implement an enhanced variant of the legacy DCF. In EDCA, the four ACs contend for the medium following the same set of rules but with different access probabilities, depending on their priority level. Service differentiation is achieved by providing each AC with a different set of values for the following key parameters:

- The minimum and maximum size of the contention window (CW_{min} and CW_{max} respectively), from which the backoff counter is calculated. ACs with higher priority are assigned smaller CW sizes, thus increasing the probability of selecting smaller backoff values.

- The time interval required to determine that the medium is idle. Instead of the DIFS employed in the legacy DCF, each AC is assigned a different interval called Arbitration Inter-Frame Space (AIFS). The AIFS is shorter for higher priority ACs, thus giving them the opportunity to seize the channel before ACs of lower priority. The exact duration of the AIFS is calculated as follows. For each AC, a integer parameter called Arbitration Inter-Frame Space Number (AIFSN) is defined. Then, the AIFS is calculated according to the following expression:

$$AIFS[AC] = AIFSN[AC] \cdot aSlotTime + SIFS \tag{2.2}$$

- The Transmission Opportunity (TXOP) limit. The TXOP is an optional feature introduced by IEEE 802.11e and is defined as the maximum time interval during which a station is permitted to hold the medium, once channel access is gained. Low priority ACs are assigned a low TXOP limit that enables them to complete the transmission of one DATA frame, including any necessary control frames (ACK, RTS/CTS). On the other hand, high priority ACs may hold the channel for a longer time and thus transmit multiple DATA packets with a single channel access.

Figure 2.7 gives an example of the EDCA access method. The default EDCA parameters defined in the standard are given in Table 2.2. The CW size is expressed as a function of the parameters $aCWmin$ and $aCWmax$ that depend on the PHY layer specification. For reference, in the ERP-OFDM PHY layer (defined in the IEEE 802.11g specification) the $aCWmin$ and $aCWmax$ values have been set to 15 and 1023, respectively.

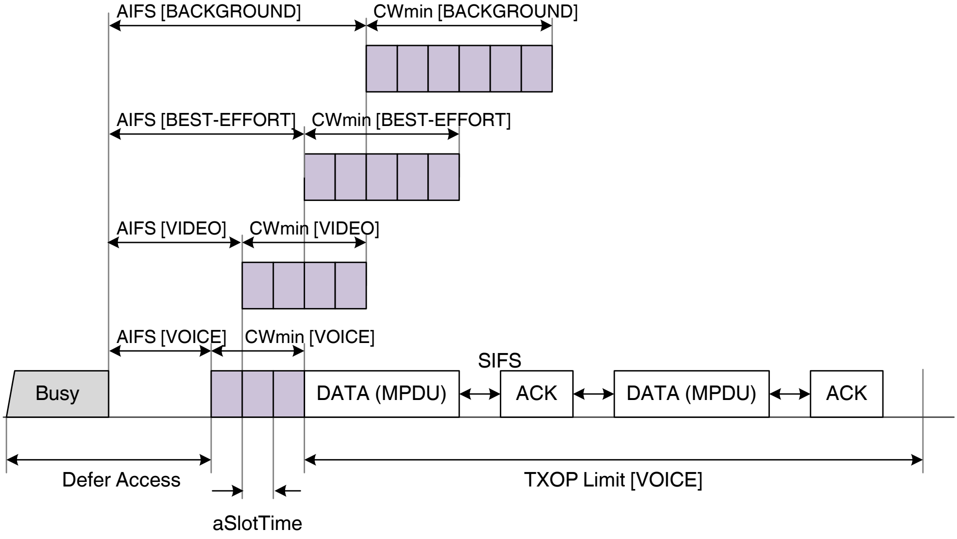


Figure 2.7: The HCF EDCA access method

Other optional modifications of the IEEE 802.11e include the use of Block Acknowledgment (BA) frames, employed to acknowledge a group of frames that are allowed to be transmitted without the need for individual ACKs, and the Direct Link Protocol (DLP) that enables the direct communication of two users in an infrastructure network without the participation of the AP.

Table 2.2: Default EDCA parameters (Table 7-37 in [1])

EDCA Parameters			
AC	CW_{min}	CW_{max}	AIFSN
Voice	$(aCWmin + 1)/4 - 1$	$(aCWmin + 1)/2 - 1$	2
Video	$(aCWmin + 1)/2 - 1$	aCWmin	2
Best-Effort	aCWmin	aCWmax	3
Background	aCWmin	aCWmax	7

2.3.5 The IEEE 802.11n Specification for Higher Throughput

The emerging IEEE 802.11n standard is an amendment to the legacy IEEE 802.11 standard that provides PHY and MAC enhancements for high throughput performance [2]. The new standard defines many new mandatory and optional features for both PHY and MAC layers, but also maintains compatibility with previous standard versions (IEEE 802.11 standard and its amendments a/b/d/e/g/h/j) [1].

The proposed PHY standard is based on MIMO/OFDM technology and can operate in either the 2.4 GHz or the 5 GHz band with a channel bandwidth of 20 MHz. The IEEE 802.11n devices must support two spatial streams and therefore must be equipped with a minimum of 2 antenna elements. The PHY defines 16 obligatory MCSs with 12 available rates (6.5, 13, 19.5, 26, 39, 52, 58.5, 65, 78, 104, 117 and 130 Mbps). A new minimum time distance of 2 μs between consecutive transmissions is defined, named RIFS (Reduced Interframe Space). New PHY sounding frames are introduced, in order to facilitate MIMO channel measurements and antenna calibration. Finally, two PHY operation modes (legacy and mixed mode) are defined, in which legacy compatible preambles are transmitted to support the coexistence of legacy devices.

The optional features of the PHY standard include the support of 40 MHz channels and up to four spatial streams. Advanced MIMO techniques such Space•Time Block Coding (STBC), transmit beamforming, and spatial multiplexing are also supported. For further enhancement, an optional operation mode, known as Greenfield, is described, in which PHY preambles do not include a legacy compatible part and therefore have reduced overhead. A shorter guard interval (GI) of 400 ns may be used, thus reducing the OFDM symbol duration to 3.6 μs (instead of 4 μs as in the case of a, g). Considering the various optional features, a set of 61 additional MCS is described and the maximum achievable throughput can reach 600 Mbps. A summary of these features is given in Table 2.3.

The MAC protocol assumes the framework defined by IEEE 802.11 standard and its later amendments (a/b/e/g). The main access mechanism is the IEEE 802.11e HCF, described in the previous section, although the DCF and the PCF are also supported to provide compatibility with legacy devices with no QoS support. The

Table 2.3: Summary of IEEE 802.11n PHY layer

Features	*Mandatory*	*Optional*
Minimum number of antennas	2 antennas (at least 2 spatial streams)	4 antennas (up to 4 spatial streams)
Channel Bandwidth	20 MHz	40 MHz
Modulation and Coding Schemes (MCS)	16 MCSs (up to 130 Mbps)	61 MCSs (up to 600 Mbps)
Operation Mode	Legacy and Mixed mode (compatible with legacy devices)	Greenfield mode (non-compatible with legacy devices)
Transmission Modes	Channel sounding capabilities	Transmit beamforming, spatial expansion, STBC
OFDM Issues	RIFS (Reduced Interframe Space)	400 ns Guard Interval

standard defines two mandatory aggregation schemes, A-MSDU and A-MPDU, to reduce control overhead. The BA feature introduced in IEEE 802.11e is enhanced with two mandatory schemes (N-immediate and Implicit BA) and the use of a compressed bit map to reduce the size of the BA frames has been included. The MAC standard also defines several protection mechanisms (long NAV, RIFS protection, PHY level spoofing, Greenfield protection) that cause legacy devices to defer transmission while MIMO transmissions take place. Mechanisms to manage the coexistence of 20 and 40 MHz channels and the operation under MIMO power save mode have been also considered. Optional MAC features include additional block acknowledgment (delayed BA) and protection mechanisms.

In order to support optional PHY mechanisms, such as antenna selection and calibration and STBC, new control frames formats are specified. The standard also defines two new optional transmission modes. The first is the reverse direction transmission that allows the bi-directional exchange of data frames between two nodes within the same session. The second is an advanced power save scheme called Power Save Multi-Poll (PSMP) which handles multi-destination (scheduled or unscheduled) uplink and downlink transmissions. Finally, the protocol supports a link adaptation scheme in which the transmission rate is adapted to the MIMO channel condition, after the exchange of information on the channel state in special control frame fields. The basic MAC features are summarized in Table 2.4.

Table 2.4: Summary of IEEE 802.11n MAC layer

Features	*Mandatory*	*Optional*
Transmission Sequences	Frame aggregation schemes (A-MSDU, A-MPDU)	Bi-directional data flow
Block Acknowledgment (BA) Scheme	N-Immediate, Implicit, compressed bit map BA	Delayed BA
Protection Mechanisms	Long NAV, PHY level spoofing, RIFS and Green Field protection	LSIG-TXOP
Power Save Mechanisms	MIMO power save	Power Save Multi-Poll (PSMP)
PHY Related	20/40MHz bandwidth coexistence	Antenna selection, calibration, STBC control frames
Other Issues		• Fast link adaptation

2.4 Cross-Layer (CL) Design

2.4.1 Breaking the OSI Layer Stack

The layering principle of the OSI model has influenced greatly the design of wired networks and has served as a reference architecture by which networks are compared. The current Internet architecture with the TCP/IP protocol suite is a successful implementation example of the OSI design paradigm. Although it does not strictly comply with it, it follows the OSI modular principle and maintains a limited amount of communication between adjacent layers [13].

The advantages of a good architectural design, as pointed out in [14], cannot be lightly overlooked. The OSI layered approach enables compatibility among vendors and different devices and makes possible to optimize each layer operation independently of the others, thus facilitating the implementation. However, the proliferation of wireless and mobile networking and the increased demand for higher performance requirements, especially in terms of QoS guarantees for multimedia applications, have posed challenges and opened new possibilities that could not be addressed with the traditional layered approach. Unlike wired links that are considerably static and predictable along time, the wireless channel changes over time and space with small and large scale variations that are often difficult to predict.

This inconvenience can be turned into an advantage with the use of sophisticated communication policies, such as the opportunistic transmission of packets when the channel conditions are favorable. In addition, the randomness in wireless propagation and the broadcast nature of the radio channel create new modalities of communication, such as multi-packet reception or user cooperation that are not feasible in wired networks [15]. All these factors have leveraged the need for more flexibility in network design, aiming to adapt the system operation to a dynamically changing channel as well as to the network characteristics and, finally, to enhance the overall communications performance.

This need has led to CL design, a concept that encompasses all schemes that violate the rigid architecture of the OSI reference model. It is a very wide term that spans from an interlayer dialog and exchange of information to the joint layer design and optimization. The number of participating layers may vary and communication may take place between any layers of the protocol stack. In less conservative approaches, it is also possible to merge layers or even define new external entities to control and coordinate CL interactions. The next section will present some classification methods for the numerous CL schemes available in the literature.

2.4.2 Classification of CL Schemes

Even though a unified taxonomy framework for CL design schemes has not been established, several classification proposals can be found in the literature. In this section, the more prevalent approaches are briefly prevented and the classification method adopted for the remaining of this section is described. Unless otherwise stated, a reference architecture model formed by five layers is considered, namely the application, transport, network, link and physical layers. The two additional layers of the OSI protocol stack (i.e., the presentation and session layers) will be omitted, following the example of the majority of publications in this topic.

General Classification Proposals for CL Design

The authors in [16] introduced the concept of interlayer coordination planes that span vertically across the protocol stack and focus on the resolution of a specific set of problems encountered in wireless mobile systems. More specifically, they have defined four planes devoted to wireless security and encryption, QoS provisioning, mobility issues and link adaptation, as illustrated in Figure 2.8. Each plane is an objective that can be achieved through CL design and thus CL algorithms can be classified in the coordinating planes depending on their targeted goal.

Another method of classification focuses on the different layers that are involved in the CL exchange of information. An example of the interactions that can take place between the layers of the OSI stack (adjacent or non-adjacent) can be seen in Figure 2.9 [17]. A detailed description of all the possible interactions between layers and the exact parameters that can be exchanged is given in [18].

One of the most common approaches uses as a taxonomy criterion the way in which the layers are coupled [15]. CL schemes are divided into four basic categories,

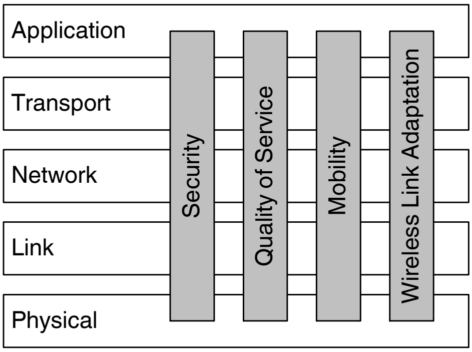


Figure 2.8: CL Coordination Planes [16]

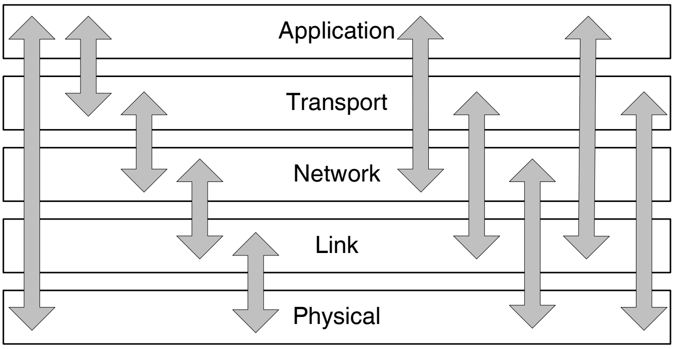


Figure 2.9: Possible interactions between layers [17]

shown in Figure 2.10, depending on the nature of the violation that has occurred in the architectural design. The four proposed categories are:

- The *exchange of information* among layers, implemented by *creating new interfaces*. The flow of information can be from lower to higher layers (upward approach, e.g., from the PHY to the MAC) or vice versa (downward approach, e.g. from the application to the MAC), whereas it is also possible to have a bidirectional iterative flow between two layers.

- The *merging of adjacent layers* into a new entity with enhanced functionalities, with the PHY and the MAC layers being the most likely candidates in this approach.

- The *design coupling without new interfaces*, where a layer is designed considering the functionalities of another layer but with no additional information exchange at runtime.

- The *vertical calibration among layers*, meaning the tuning of parameters across the layers in a static or dynamic way.

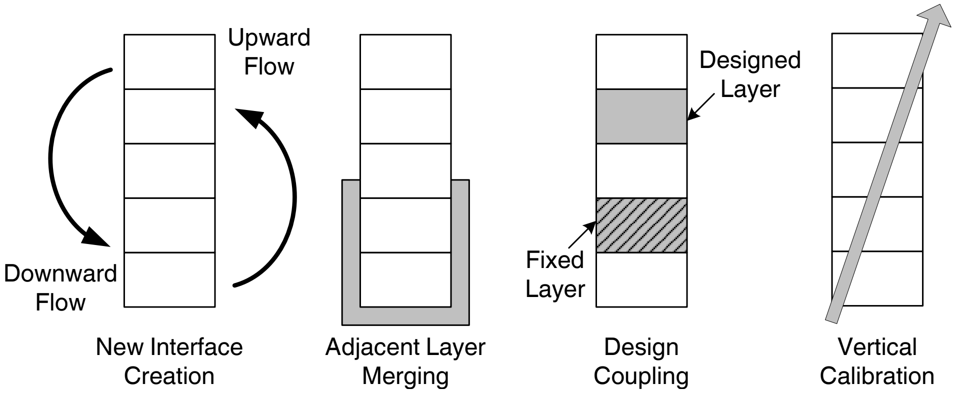


Figure 2.10: Possible coupling between layers [15]

Finally, another possible classification approach that is especially oriented towards solutions for time-sensitive multimedia applications in single hop networks is presented in [19]. In this case, the main interacting layers are the PHY, the MAC and the application layer. Five categories, illustrated in Figure 2.11, are defined depending on the hierarchical order in which the layers perform the CL interactions:

- *Top-down approach* where the higher-layer protocols select and determine the optimal parameters and strategies that concern a lower layer. This policy has been adopted in many existing systems where, for instance, the application layer dictates the MAC parameters, while the MAC selects the best PHY layer MCS.

- *Bottom-up approach* where the lower layers try to protect higher layers from losses and bandwidth variations. This approach may cause additional delays and reduce throughput, thus making it inefficient for multimedia transmissions.

- *Application-centric approach* where the application layer plays the role of a coordinator that optimizes the parameters of the other layers, adopting either a top-down or a bottom-up approach. The disadvantage of this approach is that, compared to lower layers, the application layer operates in a slower timescale and uses a coarser data granularity (i.e., the data units are multimedia files, whereas in lower layers they are packets or bits).

- *MAC-centric approach* where the MAC layer performs scheduling based on QoS information received from the application layer and also determines PHY layer parameters depending on the available channel information. The drawback of concentrating decision-making at the MAC is the inability to adapt application layer functions such as source coding to the link condition and QoS requirements.

- *Integrated approach* where joint CL optimization takes place. This is an unavoidably impractical approach, given the complexity of the optimization problem that could possibly be tackled with the use of sophisticated learning and classification techniques, such as fuzzy logic algorithms, neural networks and game theory principles.

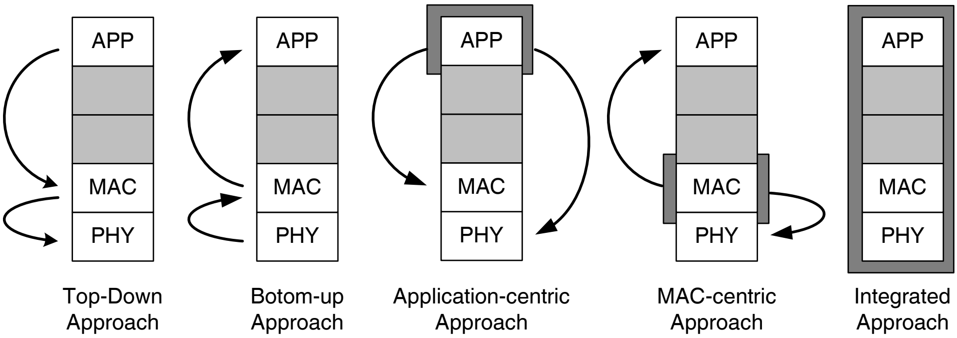


Figure 2.11: Classification for multimedia CL algorithms [19]

The classification methods described above emphasize the vast possibilities for CL design, given the numerous ways of layer coupling, the diverse CL interactions and the different objectives that can be achieved by each scheme. However, these methods are rather generic and not much oriented towards CL-based MAC mechanisms that are the main focus of this chapter. For this reason, a different classification approach will be adopted, that will be described in the next subsection.

MAC-Centric Classification Method for CL Design

Despite the diversity in the possible CL interactions between the layers, the intended goals of CL optimization are rather specific. After a thorough examination of the CL-based MAC layer schemes for WLANs that can be found in the literature, it has been deduced that the majority of the proposals can be classified into four categories that are non-exclusive, meaning that it is possible for a protocol to fit in more than one category at the same time. These categories, illustrated in Figure 2.12, along with the most common tunable parameters and functions of each OSI layer, are:

1. *Link Adaptation and Scheduling.* Although these are two different policies, they are placed within the same category since they are frequently used jointly. Link adaptation is the selection of the transmission rate that is most suitable to the channel conditions of a particular link. This mechanism usually requires a MAC-PHY CL dialog, although more layers may be involved. Scheduling is the process of allocating resources and defining the transmission order of data flows. Different policies may be selected in order to meet particular service requirements, such as fairness among users or QoS guarantees. Scheduling decisions are made by the MAC layer and can be optimized through CL interactions with any of the other layers.

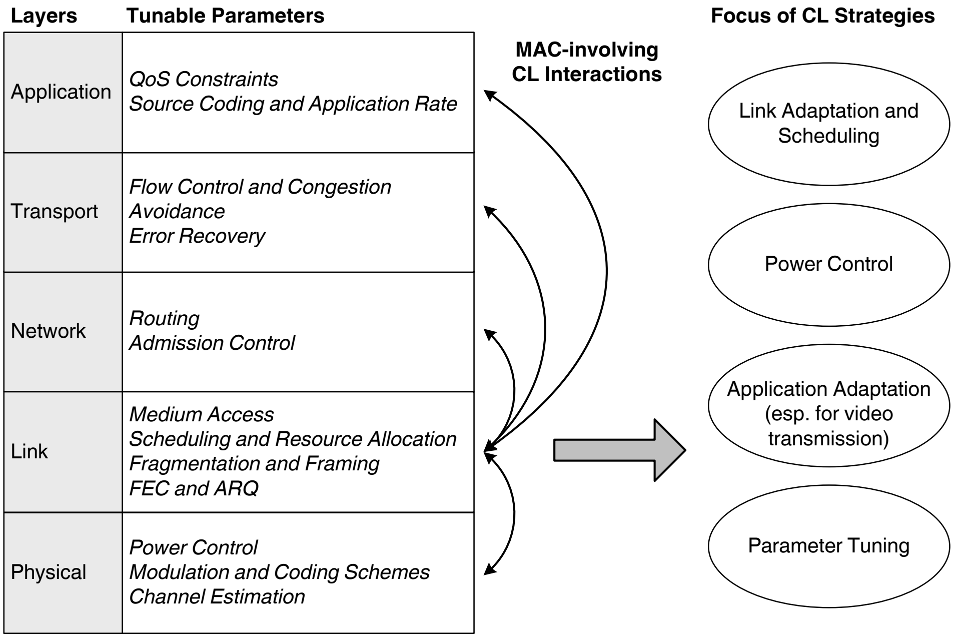

Layers
Tunable Parameters
Focus of CL Strategies
MAC-involving CL Interactions
Application
QoS Constraints
Source Coding and Application Rate
Transport
Flow Control and Congestion Avoidance
Error Recovery
Network
Routing
Admission Control
Link
Medium Access
Scheduling and Resource Allocation
Fragmentation and Framing
FEC and ARQ
Physical
Power Control
Modulation and Coding Schemes
Channel Estimation
Link Adaptation and Scheduling
Power Control
Application Adaptation (esp. for video transmission)
Parameter Tuning

Figure 2.12: CL design model

2. *Power Control.* The regulation of the transmission power is a very critical issue in WLANs. Lower power leads to the reduction of interference to other communications and the conservation of energy. This is particularly important in mobile units where an efficient power scheme will ensure the increase of battery life. On the other hand, higher transmission power may enhance connectivity under harsh channel conditions and may permit the use of higher transmission rates. Several CL-based MAC schemes consider this trade-off and offer solutions for dynamic power adaptation with the use of information passed mainly (but not exclusively) from the PHY.

3. *Application Adaptation.* This category is focused mostly on schemes that deal with the transmission of multimedia traffic. In this area, most proposals employ a CL dialog between the MAC and the application layer, where scheduling takes into consideration QoS requirements and the application layer selects the source encoding scheme (e.g. video codecs) after the exchange of information with lower layers.

4. *Parameter Tuning.* Finally, there are some proposals that aim to enhance performance (usually in terms of throughput or fairness) by tuning MAC layer parameters, such as the data packet length, the size of the contention window or the limit of retransmission attempts.

It should be mentioned that although the proposed classification encompasses the majority of the existing CL-based MAC layer schemes, it is not exhaustive.

For example, contributions that focus on the coupling of MAC and TCP layers, or routing algorithms optimized with the use of MAC metrics are not included. The reason for this is that even though those proposals involve the MAC layer, they mostly concert the transport and network layer design rather than the implementation of CL-based MAC layer protocols, which is the main objective of this thesis. Furthermore, it should be stressed that the classification of the presented schemes is not always an easy task since some algorithms may tackle issues that belong to more than one category. In continuation, the focus will be laid on the link adaptation and scheduling schemes and parameter tuning proposals that are more relevant to the contributions of this thesis.

2.4.3 Link Adaptation and Scheduling

The objective of link adaptation, also known as Adaptive Modulation and Coding (AMC), is to adjust the transmission rate to the time-varying quality of the wireless channel. Better channel conditions permit the use of higher transmission rates, thus improving the system throughput. On the other hand, when the link quality is not very good, the use of lower rates ensures connectivity, since robustness against errors is increased. The transmission rate is the result of the adjustment of the MCS implemented at the PHY, but the actual decision on which rate will be used is made by the MAC. Hence, link adaptation requires a CL dialog between the two layers.

The objective of scheduling is to determine the transmission order of data flows. These flows are groups of packets that may be defined in different ways. For instance, one flow may consist of packets either addressed to the same destination or associated to an application with specific delay constraints. A scheduling algorithm may give priority to users with better channel conditions, therefore higher available transmission rates, or to delay-sensitive applications, in order to reduce their waiting times. Scheduling policies that break the First-In-First-Out (FIFO) hierarchy are usually more efficient, especially when there is diversity in the system (e.g., in terms of available user rates, traffic QoS requirements, etc.), at the cost of undesired unfairness. However, the CL design principle has led to the implementation of more sophisticated scheduling schemes. A very popular approach is to make scheduling decisions in conjunction with link adaptation. The MAC can also incorporate CL information from higher layers, such as QoS requirements defined by the application layer, in the scheduling tasks.

The aim of this section is to review some of the more representative contributions that can be found in the literature regarding the design of MAC protocols with CL design to implement link adaptation or scheduling. The section is divided into four parts. Initially, some analytical models for the MAC layer performance that indicate the dependence between layers and can be used as a base for CL design are presented. The second part discusses common link adaptation techniques. In continuation, CL schemes developed within the context of the IEEE 802.11 are reviewed, first considering the DCF (Section 2.3.2), which constitutes the main access method, and then presenting an example of CL design based on the PCF (Section 2.3.3).

2.4.4 Analytical Formulation for PHY-MAC CL Interactions

This section reviews some contributions that develop analytical models to explore the relationship between the application, the MAC and the PHY layers. Even though these works do not propose a specific CL design scheme, they demonstrate theoretically the close dependency between the lower layers of the protocol stack, thus emphasizing the suitability and potential benefits of CL design in wireless networks. Moreover, they present a framework that enables the designers to understand how a CL dialog can be synthesized in practical MAC layer protocols.

The authors in [20] present a theoretical analysis of an IEEE 802.11a based network and show that the overall system throughput can be maximized when a rate adaptation technique that combines the selection of the modulation scheme implemented at the PHY layer and the definition of the frame fragmentation size at the MAC is employed. Frame fragmentation is one of the basic functions of the MAC layer. It consists in breaking long MSDUs, into shorter MAC frames (MPDUs) that will be passed for transmission to the PHY. The transmitting station has to contend for channel access only once: when channel access is granted, all the fragments are sent sequentially. However, each MPDU that results from the fragmentation of a larger packet is transmitted independently and has to be acknowledged by the MAC layer separately from the rest of the fragments. As a result, if an error occurs in the transmission of an MPDU, only the retransmission of the affected MPDU is required. Therefore, the selection of the fragmentation size can affect performance, since the exchange of long, unfragmented data frames may be more efficient, given that less control overhead is involved, but is less robust and could require many retransmission attempts under the presence of channel errors.

Another relevant example can be found in [21], where an analytical framework is presented to compute the end-to-end throughput of a multi-hop p-persistent CSMA/CA system with MIMO capabilities. Obtained equations depend on system parameters, such as the user density or the maximum allowable distance between source and sink, on MAC parameters, such as the transmission radius and the persistence parameter, and on PHY parameters, such as the modulation index and the transmission power. This analysis demonstrates the relationship between system, MAC and PHY parameters that can be exploited in the implementation of efficient CL designs.

2.4.5 Link Adaptation Techniques

Arguably, the most representative example of CL-based link adaptation technique can be found in a scheme named Opportunistic Auto Rate (OAR), presented in [22]. This mechanism exploits CL information to overcome an issue that occurs in IEEE 802.11 networks with multirate capability, known as the anomaly problem. This problem was comprehensively analyzed by Heusse et al. [23] and resides in the fact that when multiple rates are available, transmissions at higher rates are faster and occupy the channel less time than transmissions at lower rates. Nevertheless, the IEEE 802.11 access method provides all the users of a network with the same channel access opportunities in the long-term. As a result, users transmitting at low rates occupy the channel for longer periods of time than those users with high transmission rates. This leads to unfairness in the allocation of the resources (in terms of time share) and converts slow-transmitting stations into a network bottleneck, thus limiting the overall capacity.

OAR was proposed as a smart mechanism to improve the performance of multirate networks by exploiting PHY layer information. The key of OAR is to dynamically tune the number of transmitted packets per channel access, and consequently the achievable transmission rate, as a function of the channel quality. OAR is motivated by the fact that the channel coherence time in WLANs, which is the time interval for which the condition of a wireless channel link is considered to remain stable, is typically in the order of the transmission time of multiple packets. Hence, users who, once gaining channel access, encounter a good link condition are encouraged to transmit several packets in a row, since it is quite probable that the channel condition will not deteriorate during the transmission time. To ensure fairness, fast-transmitting users cannot occupy the channel for a period that exceeds the time required for the transmission of a single packet by a low-rate user. With this time-sharing concept, OAR maintains the long-term fairness of the IEEE 802.11 access, since otherwise fast-transmitting stations would end up dominating the channel. On the other hand, a potential problem of OAR is the fact that different flows perceive different throughputs.

In the performance evaluation presented in [22], OAR is used in conjunction with two mechanisms for the acquisition of PHY layer information on the channel state:

1. The *Auto Rate Fallback (ARF)* [24]. It constitutes the first commercial implementation that exploits the multirate capability of the IEEE 802.11 PHY, where the rate is adjusted according to the history of previous transmissions. Despite being a very simple and therefore interesting scheme, ARF has the main drawback of not being able to react in a timely manner to the dynamic changes of the wireless channel. As a result, it may overreact when the channel maintains a good condition for long periods of time by attempting to increase the transmission rate until errors occur.

2. The *Receiver Based Auto Rate (RBAR)* [25]. In RBAR, the rate is adapted to the channel characteristics with the estimation of the channel quality obtained from the reception of a RTS packet. The final decision is made by the receiver and notified to the transmitter through the CTS packet. The RBAR performs better than the ARF as it adapts faster to the radio channel variations.

The theoretical analysis and the performance evaluation of OAR described in [22] show that OAR achieves throughput gains of 40-50% over RBAR and, in addition, the throughput improvement increases with the number of nodes due to the reduced contention attained with OAR. However, OAR requires the network to operate with RTS/CTS access method in order to be able to extract PHY layer information to proceed with the link adaptation. This might limit its applicability in certain applications that use short packet lengths and do not need to perform any handshake between sender and receiver. A CL-based solution to this problem was presented in [26], where the signal strength of received frames and the number of retransmissions is used to determine the channel and receiver conditions in a relative manner. The main idea is that the transmission rate of a user is determined by the signal strength measured from previous transmissions from the AP intended to that user and by counting the number of required retransmissions that have been needed to achieve a successful transmission. Although this approach overcomes the need of RTS/CTS handshake, it shares the limited ability of ARF to adapt to the fast changes of the channel conditions since it is based on the history of previous transmissions whose information might be stale at the time of a new transmission.

2.4.6 Scheduling Algorithms Based on the IEEE 802.11 DCF

This section will discuss scheduling algorithms that enhance the performance of the IEEE 802.11 DCF. A number of these CL design schemes aim to alleviate the Head-of-Line (HOL) blocking problem. This problem is mainly due to the use of strict buffer policies such as FIFO in the context of multiuser networks. The problem arises when the link quality between the transmitter and the receiver involved in the transmission of the packet at the head of the scheduler (i.e., the HOL packet) is low. In this case, channel errors may cause retransmissions of the packet and, consequently, delays. This, in addition, holds back the transmission of other packets waiting in the queue, despite the fact that the link between the respective recipients and the transmitter may have a better quality.

One way to overcome this limitation is by exploiting multiuser diversity that stems from the fact that the links between a transmitter and different intended receivers experience independent fading and interference conditions. In other words, even though the channel may be in a bad state for a given user, it might be in a good state for another. As a result, a transmitter with packets for multiple destinations may decide to transmit to the one with the best channel conditions at a given time.

A representative MAC protocol that adopts this concept is the Opportunistic packet Scheduling and Media Access (OSMA) [27]. The framework of OSMA is represented in Figure 2.13. Each user has an independent data buffer for each candidate receiver. These buffers are dynamically reordered by means of a weighting scheme that is updated after each data transmission. Although the weighting scheme is not defined in [27], it is mentioned that the algorithm should ensure long-term fairness in terms of throughput. Whenever a node is granted transmission access, it multicasts a probing message to the first N candidate receivers, according to the priority list established by the weighting algorithm. The probing message is a multicast RTS frame, illustrated in Figure 2.14, that includes a Receiver Address (RA) and a duration field for each of the N candidate receivers.

Based on the PHY layer analysis of the received RTS, each candidate evaluates the quality of the channel with the sender (assuming channel symmetry where the channel quality of the forward link is equal to the channel quality of the reverse

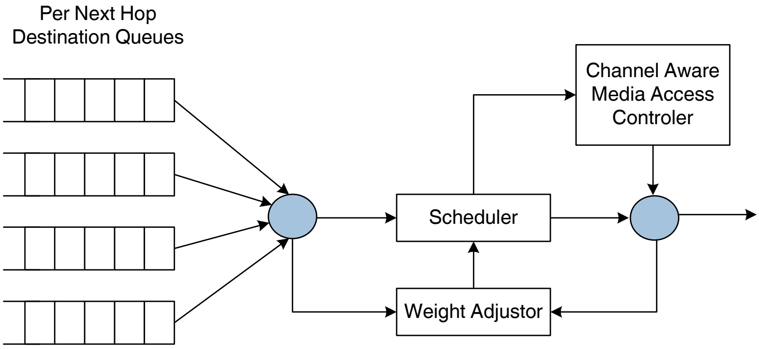


Figure 2.13: The OSMA framework [27]

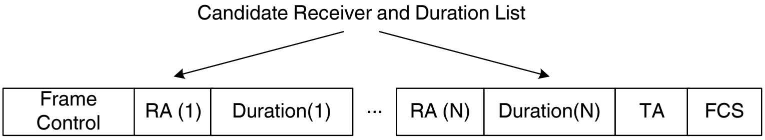


Figure 2.14: Multicast RTS of OSMA [27]

link). If the channel quality is better than a given threshold, the candidate receiver is allowed to transmit a CTS packet. In order to avoid collisions of CTS packets in case that more than one receiver fulfills this condition, the order of the candidate receivers specified in the RTS establishes a priority list. Then, each of the potential receivers waits before transmitting the CTS for a time equal to a $SIFS + (n+1) \cdot aSlotTime$, with n being the order in the priority list, and $aSlotTime$ the duration of a backoff slot defined in IEEE 802.11 specification. Therefore, the candidate receiver with highest priority will transmit first. However, this mechanism requires that the DIFS duration (i.e., the time that a station must listen to the channel before attempting to transmit) is increased to $SIFS + N \cdot aSlotTime$ to prevent a new transmission to be initiated within this waiting period. As a result, there is a trade-off between maximum diversity (high values of N) and overhead. Note that the greater the value of N the higher the probability that at least one user has good channel condition, but on the other hand, the larger the size of the multicast RTS, the longer the duration of the DIFS.

Other examples of CL channel-aware schemes for multiuser diversity exploitation in WLANs are the Weighted Fair Scheduling based on Adaptive Rate Control (WFS-ARC) protocol presented in [21], or the Channel State Dependent Packet Scheduling (CSDPS) scheme proposed in [28]. In the former, the data transmission rate is tuned with the PHY layer information available at the receiver side in combination with a scheduler that opportunistically selects the most promising user to attain both good performance and fairness. The latter uses a link status monitor to continuously track the channel quality and transmit through a “good channel” when the HOL packet is suffering “bad channel” conditions. The main drawbacks of CSDPS are that the binary model for the quality of the channel is rather simple and does not take into account the real nature of the wireless channel and the fact that the channel monitor function is not defined in the IEEE 802.11 specification.

2.4.7 Scheduling Algorithms Based on the IEEE 802.11 PCF

Although the DCF has drawn most of the attention, CL design has been also applied to the IEEE 802.11 PCF. As it was explained in Section 2.3.3, in PCF the AP polls the users following some predetermined pattern. This approach can support flow differentiation, but it does not distinguish between traffic types. A clear example of CL design as a means of attaining QoS provisioning in the context of PCF can be found in [29]. Two levels of CL dialog are implemented in this proposal.

First, a CL PHY-MAC interaction is used to determine whether a user should transmit or not. PHY layer information is attached to all exchanged frames, using the SNR (Signal-to-Noise Ratio) as the metric for the quality of the channel. Although the SNR could be mapped into delay or packet error rate conditions, the particular approach in [29] maps SNR levels into transmission rates. When a user is polled by the AP, he decides to transmit if the available rate is appropriate for his data type and QoS requirements. However, it is not specified in [29] how this SNR information is obtained.

Second, CL information is also exchanged between the PHY and upper layers for scheduling tasks. Three different traffic types are considered: voice, video and data. They are managed by using a queued system capable of interlacing between different traffic types. The management of these queues is made locally at each node with information about the PHY layer. There are three queues allocated to audio, video and best effort traffic, respectively. Once a node decides to transmit, it decides on which queue is served based on the SNR and the history of previous decisions. However, the specific mechanism to do this is not specified in [29] and remains an open issue for future work.

2.4.8 Parameter Tuning

CL techniques can be applied in different ways, modifying to some extent any of the layers of the OSI protocol stack and creating interfaces between any set of layers. Within these options, and without losing the focus on the MAC layer, one of the simplest mechanisms consists in modifying the MAC parameters, taking into account the information coming from other layers and especially from the PHY. Some algorithms that perform MAC parameter tuning will be described in this section. These examples are limited but interesting since they constitute a simple and easy to implement approach to CL design.

The authors in [30] propose a solution to cope with the fairness problem that arises when two or more WLAN stations have different channel link qualities and transmit with different rates. Their scheme consists in adjusting the message size used by the different stations to ensure that all stations occupy the channel for an equal (or as close to equal as possible) amount of time. However, there are practical problems that arise from the need for the AP to centrally control the message size of the station. The most important one is that there is no mechanism specified in the IEEE 802.11 that would allow this, so backwards compatibility cannot be achieved. The same adjustment of the packet size of the transmitting station is also adopted in [31], taking into account the presence of hidden terminals. Again, practical implementations demonstrate the benefits that can be obtained when adjusting the MAC parameters under time-varying channel conditions.

Cao et al. propose a Dynamic Binding Multi-Channel MAC (DB-MCMAC) that uses a *CW* adaption scheme to track link conditions and a dynamic binding scheme to achieve opportunistic (receiver, channel) pair selection [32]. Their scheme is backwards compatible with the IEEE 802.11 standard since the same control packets and frames are considered. This new protocol has been designed in order to mitigate the effects of fading and interference by exploiting MAC diversity. MAC diversities arise from the fact that links to different intended receivers, or over different frequency channels experience independent time-varying fading and interference conditions. These are respectively termed as multi-receiver and multi-channel diversity.

The architecture of DB-MCMAC is depicted in Figure 2.15. The protocol assumes that multiple channels can be used for transmission and is based on three key ideas:

1. The transmitter maintains multiple transmission queues, with each queue containing data packets intended for a particular destination.
2. A per-channel per-receiver adaptive CW (i.e. the DCF Contention Window, defined in Section III.A, is used to discover the fading and interference conditions of every receiver on each channel. Note that the size of the CW is representative of the history of previous transmission attempts.
3. A dynamic binding scheme is implemented to perform opportunistic selection of the best receiver and channel combination.

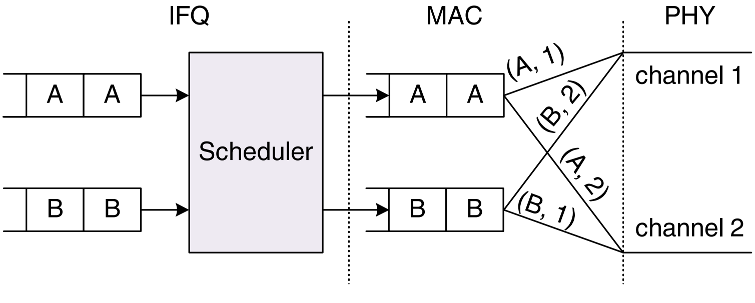


Figure 2.15: Architecture of DB-MCMAC protocol [32]

A necessary condition for exploiting diversity gains in fading/interference environments is the ability of the transmitter to track the link conditions on each channel to each receiver. In order for this to be achieved in DB-MCMAC, the successful or unsuccessful transmission of RTS packet is used to indicate the channel fading conditions. Specifically, the transmitter tries to acquire the floor on that channel by sending an RTS to the intended receiver. Depending on whether the attempt was successful or not, the contention window is decreased or increased respectively according to a multiple increase, multiple decrease (MIMD) rule [10]. This way, DB-MCMAC exploits diversity by assigning the transmission opportunity to the best link, thus enhancing MAC performance in terms of throughput.

2.5 PHY and MAC Layer Design for Multiuser Systems

The last section of this chapter will focus on multiuser schemes where point-to-multipoint communication can take place. Multiuser transmissions require the use of multiple antenna transmission techniques and advanced signal processing at the PHY layer, as well as more complex MAC layer schemes. An overview of the most representative transmission techniques for smart antennas and MIMO systems is given in Section 2.5.1, whereas the available multiuser MAC schemes are presented in Section 2.5.2.

2.5.1 PHY Layer Techniques for Multiple Antenna Systems

In recent years, the technology of smart antennas has been widely investigated in an effort to increase the capacity of wireless networks. A smart antenna system combines multiple spatially distributed antenna elements with intelligent signal processing algorithms that optimally adjust the antenna radiation pattern in order to achieve some desired objective. Smart antennas can be classified into three categories according to their level of intelligence [33]:

- The *switched beam antennas* have the lowest intelligence and can employ beamforming towards specific, predefined directions.
- The *dynamically phased antennas* can determine the direction of arrival of a received signal and steer a beam towards that direction to enhance reception.
- Finally, the *adaptive array antennas* can additionally adjust their radiation pattern to null out interference sources.

MIMO systems employ smart antenna technology with a high level of intelligence, aiming to improve transmission rates and enhance reliability and robustness. The term MIMO implies the availability of at least two antennas at each end of the communication link. When multiple antennas are employed only at the transmitter side the system is known as Multiple-Input Single-Output (MISO), whereas Single-Input Multiple-Output (SIMO) systems imply a single transmitting and multiple receiving antennas.

There are several techniques for the exploitation of multiple antennas at the PHY layer, schematically illustrated in Figure 2.16. A brief description of each technique will be given next, but more detailed explanation can be found in [34] and [35].

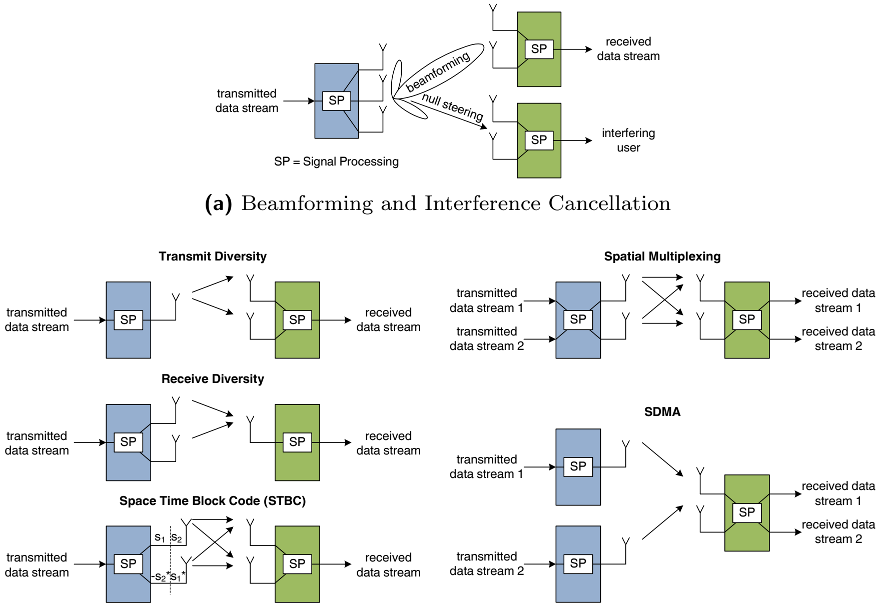
The most conventional techniques are beamforming and interference suppression (Figure 2.16 (a)). By means of beamforming, the SNR of a point-to-point communication link is increased thus resulting to higher supported data rates and extended coverage range. Interference suppression is achieved by steering the nulls of the antenna radiation pattern towards specific directions. This technique can be employed to reduce the interference produced by the transmitter but also to limit the received interference by other systems. As a result the link reliability and the spectral efficiency of the system are enhanced.

Another very efficient technique is spatial diversity that effectively mitigates multi-path fading and therefore provides increased robustness against errors (Figure 2.16 (b)). Depending on whether the multiple antenna elements are placed at the receiver (SIMO) or the transmitter (MISO), the spatial diversity schemes can be classified as receive and transmit diversity, respectively. In receive-diversity schemes, independently faded copies (due to different propagation paths) of the same signal arrive at each antenna element of the receiver and are appropriately combined or selected to enhance reception [36]. In transmit-diversity schemes the same signal is transmitted over multiple antennas after some processing has taken place to ensure that the received multiple copies of the signal will be successfully separated by the receiver [37][38][39]. Clearly, in MIMO systems, joint receive and transmit diversity schemes can be implemented.

A very powerful transmit-diversity technique that achieves both diversity and coding gain is the Space-Time Coding (STC) that involves signal coding over space (multiple antennas) and time (multiple symbol times). There are two main approaches to STC design, the Space-Time Trellis Coding (STTC) [40] and the Space-Time Block Coding (STBC) [41][42]. STTC provides considerable coding and diversity gains with the cost of high decoding complexity. On the other hand, STBC is less efficient since it mainly offers diversity gain (and minimal or zero coding gain) but has the significant property of using linear decoding at the receiver.

Another PHY layer technique is spatial multiplexing (Figure 2.16 (c)), according to which multiple independent data streams are simultaneously transmitted in the same frequency spectrum using multiple antennas. The receiver manages to extract the data streams from the received signal by employing spatial processing techniques that exploit multi-path fading. As a result, the throughput performance is increased. A very popular and spectral efficient spatial multiplexing scheme is V-BLAST (Vertical- Bell Laboratories Layered Space Time) [43]. As far as point-to-multipoint links are concerned, the spatial multiplexing of signals known as SDMA allows multiple simultaneous transmissions in the same frequency, thus multiplying the capacity of the system [44].

Summarizing, the main PHY layer techniques that are available in multiple antenna systems are beam-forming, interference cancellation, spatial diversity and spatial multiplexing. These techniques can be used separately or in combination, to obtain the desired effect. Finally, it has been demonstrated that there is a fundamental trade-off between diversity gain and spatial multiplexing gain that reflects to a design decision in favor of increased reliability or throughput, respectively [45][46].



(a) Beamforming and Interference Cancellation

(b) Spatial Diversity Techniques

(c) Spatial Multiplexing Techniques

Figure 2.16: Multiple antenna transmission techniques

2.5.2 MAC Protocols for Multiuser Transmissions

A nice overview of the most representative examples of multiuser scheduling and resource allocation can be found in [47]. The authors stress that selecting the best subset of users for each transmission is the key to achieving multiuser diversity but also point out that several practical issues arise, including the need for feedback acquisition on the link quality of the users.

A significant number of contributions has been dedicated to the development of user selection and scheduling algorithms in the context of multiple antenna systems. An early work proposes the first-fit algorithm, a sub-optimum but less complex scheduling method that selects sets of packets that can be transmitted simultaneously [48]. However, one of the basic assumptions of this work is that the channel between the base station and the users is quasi-static and is considered known by the base-station, whereas scenarios with varying channel conditions are left for future consideration. In [49] the authors propose a SDMA/TDMA scheduler that assigns packets to time slots depending on their QoS requirements. Multiple packets can be spatially multiplexed in the same slot if they satisfy a Signal-to-Noise-and-Interference Ratio (SNIR) constraint. Again, this work mainly focuses on the scheduling policy and assumes that the spatial signature and QoS requirements for each packet are acquired during an initial admission phase.

Nevertheless, in realistic scenarios the channel condition cannot be considered known and a feedback mechanism must be established. Naturally, there is a tradeoff between the feedback required to implement multiuser diversity schemes and the introduced control overhead that reduces efficiency. One way to decrease feedback is by applying a threshold to exclude users with poor channel conditions from gaining access to the channel. This idea has been extensively studied in [50]. This work offers some guidelines for the threshold selection but it does not consider a specific multiple access scheme, nor the implementation of an actual feedback acquisition mechanism. In a different approach, binary feedback (1 or 0) is used by users to express whether they satisfy threshold condition [51]. The idea is effective but assumes the presence of a dedicated low bit rate feedback channel, which is not the case in IEEE 802.11 based WLANs. Finally, another proposal combines the principle of splitting algorithms with threshold selection to determine the user with the best channel in less than three slots on average [52]. This work has been extended to provide detection of multiple users with good channel and needs on average 4.4 slots to find the best two users in the system [53].

Finally, there are some contributions that aim to include multiuser MAC schemes for IEEE 802.11 based systems. One example is the Multi-User Distributed Coordination Function (MU-DCF), presented in [54], that uses a four-way handshake that begins with a polling multiuser RTS frame. However there are several issues, mostly regarding the PHY layer implementation, that are not considered. A mathematical model for a downlink multiuser scheme for IEEE 802.11 is given in [55]. They show that performance can be improved by exploiting spatial multiplexing and conclude that there is still a need to design a modified MAC to support multiple transmissions and perform a good channel estimation mechanism.

2.6 Conclusions

This chapter has provided some background information that is relevant to the contributions of this thesis, which will be thoroughly presented in the following chapters. Initially, the functions of the MAC layer have been explained and a classification for MAC protocols has been attempted. The IEEE 802.11 standard and its main amendments for WLANs, namely b/e/g/n, have been discussed next. The largest section of this chapter has presented the principles of CL design and given an overview of CL schemes available in the literature. Finally, PHY and MAC layer schemes that exploit multiple antenna technology to achieve multiuser transmissions have been summarized.

The remaining of this thesis is organized in two parts. The first part (Chapters 3, 4 and 5) is focused on a novel MAC protocol named DQCA that combines a splitting algorithm CRA with a reservation scheme to provide almost collision-free data transmissions. A MAC-centric CL design approach has been adopted to enhance DQCA with more functionalities and better performance. Link adaptation and advanced scheduling strategies, including opportunistic transmissions and QoS provisioning, have been the main objective of the CL interactions between DQCA and the other protocol layers.

The second part of the thesis (Chapter 6) is oriented to multiuser MAC schemes. The IEEE 802.11n standard for MIMO systems has been used as a starting point and some modifications have been introduced to support point-to-multipoint transmissions. The result has been a PHY/MAC multiuser solution that advances the state of the art by jointly proposing and evaluating a low-complexity beamforming technique combined with opportunistic scheduling and a channel feedback acquisition mechanism.

Chapter 3

DQCA: A Distributed MAC Protocol for WLANs

3.1 Introduction

The continuous growth of the WLAN market in the recent years has been accompanied by increasing demands for higher transmission rates. Since, unfortunately, the available spectrum is limited, there is a strong need for efficient protocols for the lower communication layers. The focus of this thesis lays on the MAC layer and in particular on the design of channel access mechanisms and scheduling algorithms that handle efficiently the available resources.

The Distributed Queuing Collision Avoidance (DQCA) protocol is a high performance MAC scheme principally designed for WLANs. The protocol implements a reservation scheme that ensures collision-free data transmission. In addition, in order to avoid unnecessary delays when the traffic load is low, DQCA switches smoothly and automatically to an Aloha-like random access mechanism. The protocol's operation is based on two distributed queues that work in parallel and handle the resolution of collisions among channel access requests and the transmission of data, respectively.

DQCA forms part of an extended family of multiple access protocols that share the concepts of distributed queuing and network intelligence at the stations. In the early nineteen-nineties, Wenxin Xu and Graham Campbell introduced the Distributed Queuing Random Access Protocol (DQRAP), a random access scheme intended for use in a slotted broadcast channel and an infinite number of bursty stations [56], [57]. At that time, the primary application of the DQRAP was the digital data transmission on cable TV. Nevertheless, its stable behavior under all input data rates and the fact that its performance could approach that of a hypothetical perfect scheduling protocol, i.e., an M/D/1 system, have been a strong motivator to further investigate the capabilities of distributed queuing.

Many variations of the DQRAP protocol have since been developed, in an effort to adapt the efficient distributed queuing paradigm to different environments. The Extended DQRAP (XDQRAP), for example, enables stations to reserve multiple consecutive slots for data transmission with a single access request [58]. The Prioritized DQRAP (PDQRAP) introduces service differentiation by defining high and low priority transmission queues [59]. The Interleaved DQRAP, designed for satellite applications, establishes an interleaving factor to account for high propagation delays by implementing multiple DQRAP engines that operate in consecutive cycles [60]. A more recent contribution has adapted the DQRAP engine for use along with CDMA, widely used in 3G cellular networks [61]. This extensive work has been the main inspiration for the design of DQCA, which was first presented in [62].

The aim of this chapter is to provide a thorough description of the DQCA protocol for infrastructure WLANs. The main chapter body is divided into four parts. Section 3.2 introduces the DQCA protocol and describes the format of the DQCA frame sequence and the distributed queuing mechanism. Section 3.3 presents the format of the basic packets employed in DQCA, indicating the necessary control information that must be included in each frame and meant as a reference for future implementations of the protocol. A formal description of the DQCA protocol rules is given in Section 3.4, along with an example of the DQCA operation. For completeness, Section 3.5 discusses some interesting aspects of the DQCA protocol that lie outside the main focus of this thesis and provides some guidelines on how they can be tackled. Finally, the chapter closes with conclusions in Section 3.6.

3.2 Overview of the DQCA Protocol

DQCA is an efficient MAC that behaves as a random access mechanism when the traffic load of the network is low and switches smoothly and automatically to a reservation scheme as the traffic increases. The protocol is based on two logical queues, shared among all the nodes of the network, whose role is to handle two processes that take place in parallel, namely, the channel access request and the data transmission.

The operation of DQCA can be summarized as follows. Nodes may ask for channel access in a reserved slotted time interval, thus confining collisions almost exclusively to that part of the frame. Any nodes involved in a collision enter a distributed FIFO queue. The collisions are resolved in subsequent frames with the use of a blocked access m-ary tree-splitting algorithm. The nodes with a successful access request enter the second distributed queue and wait for their turn to transmit their buffered data message. DQCA implements a FIFO data transmission policy; nevertheless, priority-based scheduling schemes can be easily applied, depending on the desired system performance.

The near-optimum performance of DQCA is owed to several factors. First, the separation of the channel access and the data transmission process increases channel utilization, by eliminating idle periods that are often present in CSMA/CA schemes in which the back-off mechanism is employed, as in the case of the IEEE 802.11 DCF. Another advantage is that data transmission is practically collision-free, since collisions among data packets may only occur under very light traffic, when the system is almost empty. Finally, DQCA offers a stable throughput performance that does not deteriorate as the number of nodes increases or the traffic load grows. On the contrary, as the system approaches saturation conditions, a maximum throughput is reached and maintained even when the traffic load exceeds the channel capacity.

The DQCA operation is explained in detail in the sections that follow, starting from the description of the DQCA frame structure and the operation of the distributed queues.

3.2.1 The Structure of the DQCA Frame

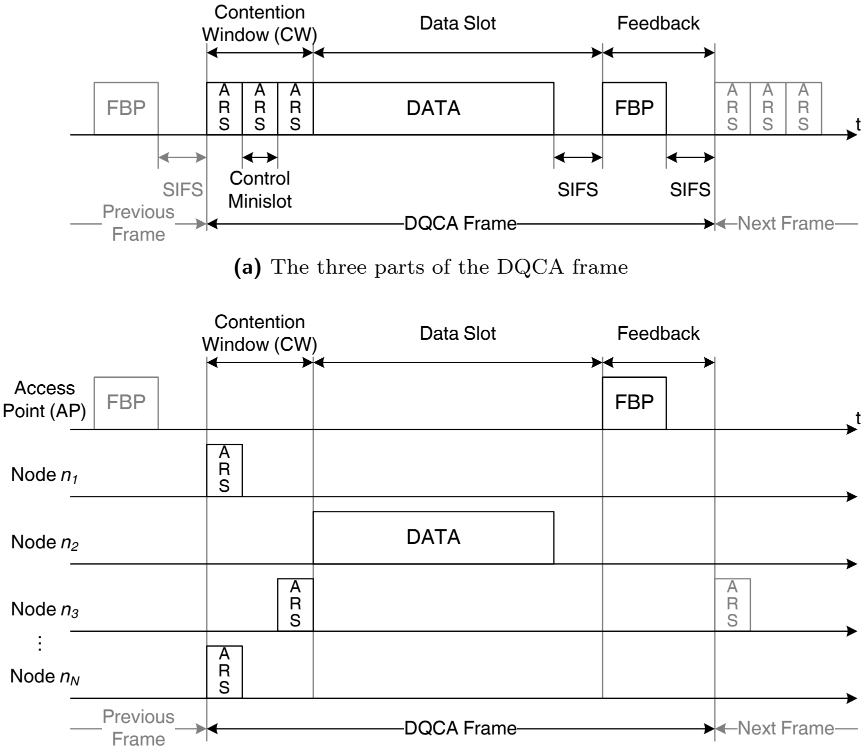
For the description of DQCA operation, a typical infrastructure WLAN scenario will be considered in which N nodes[1] share the wireless medium in order to communicate with an Access Point (AP). The structure of the DQCA frame is illustrated in Figure 3.1. As shown in plot (a), the time axis is divided into DQCA frames that consist of three parts, the Contention Window (CW), the data slot and the feedback part. An alternative view of the DQCA frame structure is given in plot (b), depicting an example of the different roles played by the AP and the users in each part of the frame (a complete example of the DQCA operation will be given in Section 3.4.2).

The first part of the DQCA frame is the CW that is further divided into m control minislots of fixed time duration (in Figure 3.1 $m = 3$ has been selected for convenience). The nodes that want to gain access to the channel can randomly select one minislot with probability $1/m$ and transmit a special reservation packet named Access Request Sequence (ARS). The ARS is a control packet defined in DQCA that has the form of a short chip (CDMA-like) sequence and contains no data information. The structure of this special frame will be discussed in detail in Section 3.3.1. Since the minislot selection by the users is a random process, there are three possible states for each minislot: empty, when no ARS transmission has taken place; success, when the minislot has been selected by a single node; and collision, when more than one ARS have been simultaneously transmitted.

The second part of the DQCA frame is reserved for the almost collision-free transmission of data packets by one node at a time. It has been considered without loss of generality, that exactly one data packet of a fixed byte length can be transmitted within a data slot, meaning that the duration of the slot is variable and depends on the actual transmission rate. Large messages are fragmented into smaller packets which are transmitted in consecutive DQCA frames.[2] In any case,

[1] The term nodes or users will be employed interchangeably to denote wireless client devices that communicate with an AP in an infrastructure scenario. They can be considered as equivalents to the IEEE 802.11 defined stations (STAs) [1], with the difference that they implement the DQCA protocol at the MAC layer.

[2] As explained in Section 2.3.1, the term message has been employed to denote the MSDUs that arrive at the MAC from the upper layers. The MSDUs are encapsulated (and fragmented, if necessary) into MPDUs, which are the data packets transmitted in each DQCA data slot.



(a) The three parts of the DQCA frame

(b) The DQCA frame from the users' perspective

Figure 3.1: Structure of the DQCA frame

data packets of variable length could be easily supported without any modifications to the DQCA frame structure.

In the last part of the frame the AP broadcasts a Feedback Packet (FBP) that contains the necessary feedback information for the execution of the DQCA protocol. The fields that are strictly necessary for the DQCA operation are the following:

- it contains ternary feedback information on the state of the each control minislot, which, as mentioned before, can be empty, success or collision.

- it includes a positive acknowledgment (ACK) if a packet has been correctly received in the data slot of the DQCA frame; otherwise, if an empty data slot has been detected or a corrupted packet has been received, it includes a negative ACK (NACK).

- it contains a final-message-bit that indicates whether the last (or the only) packet of a message has been received (final-message-bit set to '1') or more packets of the same fragmented message are expected to follow (final-message-bit set to '0').

Nevertheless, additional information can be included to implement more sophisticated scheduling policies, as it will be discussed in the following chapter. It should also be noted that the AP employs the lowest available rate for the transmission of the FBP to provide a high probability of error-free reception by all the nodes.[3]

In order to compensate for propagation delays, turnaround times required for a transceiver to switch from receive to transmit mode and for processing purposes, a Short Interframe Space (SIFS) is introduced after the transmission of the data packet and the FBP.

More information on the format of the ARS, data and FBP frames employed by the DQCA protocol will be given later, in Section 3.3. First, the core of the DQCA protocol, formed by the two distributed queues that handle the ARS collisions and the data packet transmission, will be presented in the next section.

3.2.2 The Distributed Queues

The DQCA protocol operation is based on two logical distributed queues: the Collision Resolution Queue (CRQ) and the Data Transmission Queue (DTQ). The CRQ handles the resolution of collisions among ARS that are transmitted simultaneously within the same control minislot by different nodes. The DTQ is responsible for the scheduling of data transmissions by nodes that have been granted channel access by successfully transmitting an ARS, either directly (upon their first attempt) or after the execution of the collision resolution algorithm.

It has been considered that a node manages the transmission of a single message (MSDU) at a time that, depending on its length, may be de-assembled into more than one packet (MPDUs). The node enters the DQCA queuing system as soon as it sends an ARS to request channel access for the transmission of its data message and exits when the message transmission is completed.[4] During this time, the node holds a single position in the queuing system (in either the DTQ or the CRQ) related to the aforementioned message. Therefore, it can be equivalently said that each non-empty position in the protocol queues is occupied by a message or by the corresponding node.

The two queues are logical entities, each represented by a pair of integer counters that are kept by every node in the network and updated at the end of each frame after the execution of the protocol rules. The queues are distributed in the sense that their state is known to all nodes through the integer counters and consequently the nodes can make transmission decisions on their own. The first counter indicates the queue length, or in other words, the total number of positions occupied in the queue (by nodes or messages) and is a global variable that must have the same value for all nodes. The second counter, on the other hand, has a local scope for each node and refers to its own position in the queue.

[3] The issue of errors in the FBP reception will be discussed later, in Section 3.5.2.

[4] Even if more messages are accumulated in its buffer, the node must first exit the queues and then request channel access with a new ARS transmission.

To be more specific, the counters that refer to the DTQ are denoted by TQ and pTQ, whereas the CRQ related counters are RQ and pRQ, respectively. TQ is the number of nodes waiting for transmission in the DTQ and pTQ is the position of each node in the queue. The pTQ value of a node lies in the range of $[1, TQ]$, if the holds a valid position in the DTQ, and is equal to zero if the node does not belong to the queue. This counter also indicates the relative age of each node in the DTQ: nodes with smaller pTQ values have entered the DTQ before those with higher $pTQs$. Typically, with the exception of the cases when DQCA operates as an Aloha-like scheme (a situation that cab only occur under very low traffic conditions, as it will be explained in the next section), each DTQ position is occupied by a single node, thus ensuring collision-free data transmission.

In a similar way, RQ indicates the occupied positions in the CRQ and pRQ reflects the position of each node in the queue, with values within $[1, RQ]$ for nodes waiting in the queue and a zero value otherwise. In this case, however, each CRQ position is occupied by the nodes involved in an ARS collision (i.e., nodes that have transmitted an ARS in the same minislot). Therefore, each CRQ position contains at least two nodes that will share the same pRQ value. An illustration of the distributed queues and the associated counters is given in Figure 3.2.

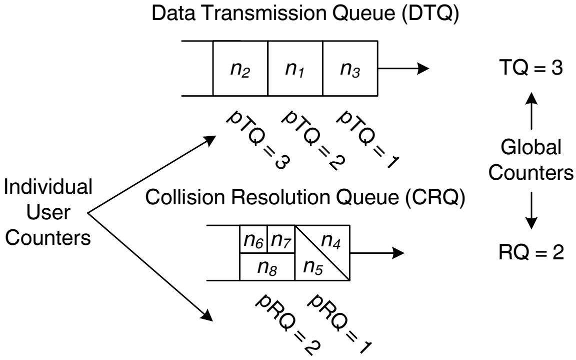


Figure 3.2: Illustration of the distributed queues and the associated counters

It should be emphasized that these counters are maintained by all the nodes in the system, even the idle nodes that have no packets to transmit and therefore do not belong to any of the queues. The counters are updated at the end of each DQCA frame (i.e., after the FBP that contains the feedback information on the outcome of the minislots and the data slot has been received), according to a set of protocol rules that will be given in Section 3.4. In the DQCA description, it has been assumed that the nodes keep track of the feedback history from the beginning of the protocol operation and that the FPB is received by all nodes with no errors. Alternatively, the AP can periodically include the global TQ and RQ values within the FBP. This approach offers some practical advantages, first in terms of energy consumption, since nodes may enter a sleep mode if idle and be notified of the state of the global counters upon waking up, and second in terms of robustness against possible counter miscalculations that may be caused by corrupted FBP packets.

3.3 The DQCA Frame Formats

Since DQCA is not yet an implemented and standardized protocol, the exact format of its frames remains an open issue. This section discusses potential formats for the basic packets employed in DQCA. Whenever possible, the frame formats defined in the IEEE 802.11 MAC specification [1] are used as a reference, in order to provide some level of compatibility between DQCA and the IEEE 802.11 standard. Modifications on the proposed frame formats could be applied to add new functionalities to DQCA. This issue will be further discussed in the following chapters.

3.3.1 The Access Request Sequence (ARS) Packet

In most MAC protocols, the frame employed to request access to the medium (the IEEE 802.11 RTS, for example) contains information such as the address of the source and the destination node, or the data packet length. The resulting frame has a size of several bytes that is further increased by the preamble added by the PHY layer before transmission. The added overhead and the fact that control frames are typically transmitted at a low rate to minimize reception errors, pose a considerable limitation to MAC layer throughput performance. In DQCA, the rationale behind ARS frames is different: their role is to indicate the request of a node for channel access within a particular minislot of the CW but, unlike other schemes, they are not required to include any information on the node identity or the amount of data to be transmitted. The sole requirement is that they must enable the AP to distinguish between the three possible events that take place in each minislot (i.e., empty slot, success and collision).

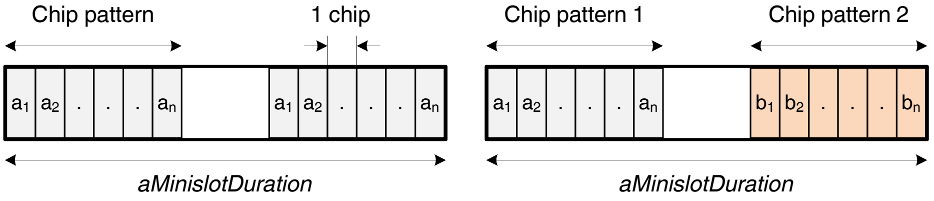
Having this condition in mind, the ARS is defined as a short chip (CDMA-like) sequence with a specific pattern that enables collision detection and is assigned to the users by the AP during an initial association phase. Then, within each control minislot, the AP can detect whether an ARS is received without errors by checking the validity of its pattern. On the other hand, when multiple ARS collide, the overlapping signal produces a corrupted pattern that does not match the original pattern of any single ARS. This technique is based on a patented method for collision detection and channel access described in [63] and [64].

One potential base for the formulation of a suitable ARS pattern is the binomial coefficient $C(n, k)$ that represents the possible ways to select a combination of k items out of a set of n items. In the present case, n would represent the number of CDMA-like chips that form the pattern and k the desired number of ‘1s’ in it. To illustrate this point, consider the coefficient $C(4, 2)$ that can generate six different sequences of four chips that contain two ‘1s’, shown in Table 3.1.

In a practical implementation of this scheme, a pattern of a sufficient length should be selected in order to support the desired number of system users. The ARS may contain multiple copies of the same sequence, resulting to a transmitted signal that will fit in exactly one minislot of duration equal to $aMinislotTime$. The proposed ARS structure is depicted in Figure 3.3 (a).

Table 3.1: Example of a pattern based on the $C(4,2)$ coefficient

User	*Pattern*
User 1	1100
User 2	1010
User 3	1001
User 4	0110
User 5	0101
User 6	0011



(a) Repetition of a single pattern **(b)** Combination of two patterns [61]

Figure 3.3: ARS frame format: two alternative implementations

The simplest way to generate such a signal is by turning the carrier on or off for a time period of t_{chip} according to the $C(n,k)$ pattern [63]. In other words, for a sequence containing k '1s', the transmitting node should turn the carrier on for a total time of $k \cdot t_{chip}$ within a control minislot. The receiver locks onto the arriving carrier and by integrating the signal over the minislot duration it can determine the presence of an empty, successful or collision slot, since:

- an empty slot corresponds to the absence of carrier
- a successful slot corresponds to the presence of carrier for a total time of $k \cdot t_{chip}$
- a collision slot corresponds to the presence of a carried for a total time exceeding $k \cdot t_{chip}$

It has been estimated that a t_{chip} of two hundred cycles of the carrier is sufficiently long for a receiver to detect and lock on the signal [63]. Consequently, for a typical WLAN system operating at the 2.4 GHz band, the minimum theoretical duration of t_{chip} could approximately 0.08 μs and the minislot duration could be of the order of a few μs.

A variation of this scheme that reduces the complexity of the receiver is proposed in [61]. In this approach, each user is assigned a pair of sequences in such a way that no other user will share the exact same sequence combination. The resulting ARS frame will contain both sequences, as illustrated in Figure 3.3 (b). The advantage of this implementation is that it reduces the number of unique sequences that must be available to the system, since with a total of F sequences up to $\binom{F}{2} = F(F-1)/2$ users can be supported. The receiver consists of a bank of F matched filters, one per sequence, that will produce a correlation peak whenever a sequence with a valid pattern is detected within the received signal. Hence, the receiver can differentiate between a successful ARS, corresponding to the detection of exactly two peaks, a collision, resulting to more than two peaks, and an empty minislot, containing no recognizable sequence.

Two basic assumptions regarding the control minislot mechanism have been considered throughput this thesis. They are discussed next:

- There is perfect minislot synchronization and all ARS within a minislot are received simultaneously by the AP. In practice, the nodes can track the beginning of the CW with relative precision, since it begins within a SIFS time from the reception of the FBP. Nevertheless, nodes located at different distances from the AP may experience different propagation delays, which may cause synchronization errors. This could be amended by extending the size of the contention minislots to include the transmission time required for an ARS plus a guard time to account for propagation delays. In any case, according to the IEEE 802.11 PHY layer specification [1], propagation delays are of the order of 1 μs for IEEE 802.11b and below 1 μs for IEEE 802.11g.

- The state of the control minislots is detected by the AP without errors. A study of the probability of slot state mis-detection can be found in [61], applied to a variation of DQCA adapted to a CDMA environment named DQRAP. This work also proposes mechanisms for system recovery in case of feedback errors, that mainly consist of algorithmic steps to update or reset the queue counters, depending on the nature of the situation. Some further discussion on this issue will take place in Section 3.5.2.

Finally, it should be noted that in order to add additional functionalities to DQCA, such as link adaptation or QoS provisioning, some modifications on the ARS structure would be required. These issues will be addressed in the following two chapters, in which CL enhancement mechanisms for DQCA will be discussed.

3.3.2 The Data Packet

The DQCA data packets follow the frame format defined in the IEEE 802.11 MAC specification [1]. The data frame, depicted in Figure 3.4, consists of three parts: the MAC header, the frame body that contains the data payload and the Frame Check Sequence (FCS). The content of each field is briefly described next and the main differences with the IEEE 802.11 MAC header are highlighted.

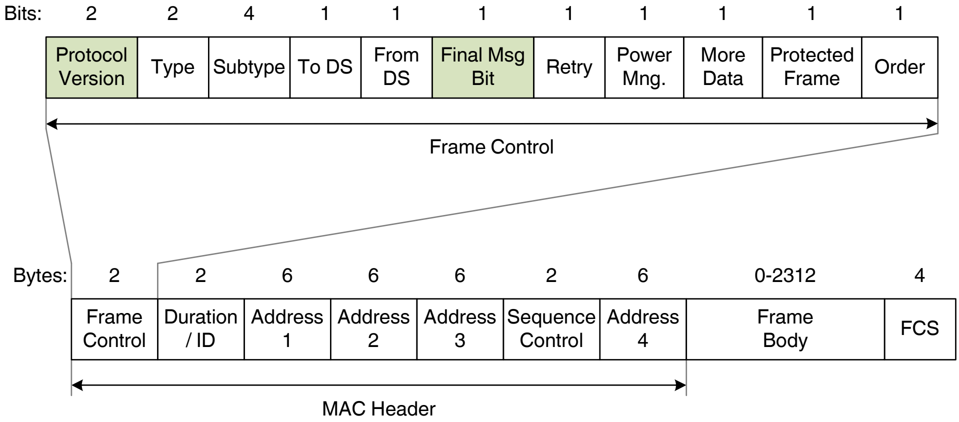


Figure 3.4: MAC header for the DQCA data frame

Frame Control Field

The Frame Control field has a total length of 16 bits and consist of eleven subfields, also depicted in Figure 3.4. Most of these subfields are not particularly necessary to the operation of DQCA but are maintained in the MAC header for compatibility with the IEEE 802.11 standard. For this reason, they are not further explained here but can be consulted in the IEEE 802.11 specification [1].

The two Frame Control subfields that have a valid meaning for the DQCA protocol are described next:

- The Protocol Version. It consists of two bits that indicate the version of the MAC protocol. For the IEEE 802.11 standard, including its amendments a/b/e/g/n, the value of the protocol version is ‘0’. Any frame with a different version number is discarded by the IEEE 802.11 MAC entity. In DQCA this field can be used in two ways. One possibility is to set the protocol version to a non zero value, thus completely separating a DQCA operating network from IEEE 802.11 stations. The other approach is to maintain the protocol version to 0 and introduce some protection mechanisms to enable the coexistence of the two MAC protocols. Some further thoughts on the coexistence scenario will be given in Section 3.5.3.

- The final-message-bit. This field is set to ‘1’ if more fragments (MPDUs) of the current data message (MSDU) are to follow in subsequent DQCA frame. Otherwise, if the current packet is the last or the only part of the data message, the final-message-bit is set to ‘0’. It should be noted that this field is called More Fragments bit in the IEEE 802.11 specification, but its function is exactly the same.

Duration/ID Field

The Duration/ID field is 16 bits in length and its usage in the IEEE 802.11 standard depends on the particular frame type it refers to. Typically, in the case of unicast data frames, the Duration/ID indicates the time (in μs) required for the transmission of the data frame and the corresponding acknowledgment. This practically means that IEEE 802.11 stations that decode the MAC header of a frame not addressed to them will consider the medium busy during the time indicated in this field.

In DQCA, the Duration/ID field is not strictly necessary, since the protocol does not employ a virtual carrier sensing mechanism. Nevertheless, an appropriate setting of this field could be the key to enable the coexistence of DQCA and IEEE 802.11, as it will be discussed in Section 3.5.3.

Address Fields

The MAC header has four 6-byte address fields in the MAC header that indicate the MAC addresses of the receiver (i.e., the immediate recipient of the data frame), the transmitter, the destination (i.e., the final data recipient) and the source node (i.e., the node that initiated the data transmission). In most cases, the fourth address is omitted from the MAC header and the possible address assignments for the first three fields can be consulted in the IEEE 802.11 standard [1].

Sequence Control

As in the IEEE 802.11 standard, the Sequence Control field is used to indicate the sequence number (12 bits) and the fragment number (4 bits) of a data frame.

Frame Check Sequence

The Frame Check Sequence (FCS) consists of a 32 bit CRC for error detection.

Frame Body

The frame body contains the data payload and its maximum size is 2304 bytes or 2312 bytes when encryption is used. Messages that exceed this length must be fragmented into smaller packets.

3.3.3 The Feedback Packet

The third frame defined in DQCA is the Feedback packet (FBP), which can be thought of as a modified version of the IEEE 802.11 ACK frame. The structure of the FBP is depicted in Figure 3.5.

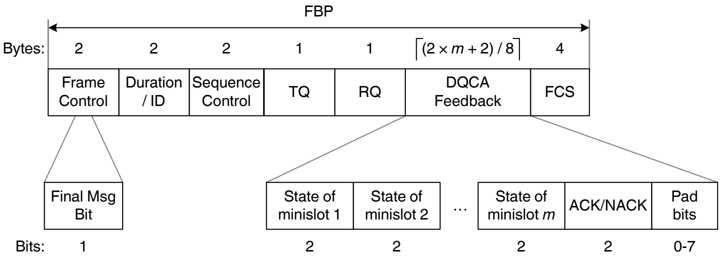


Figure 3.5: FBP frame format

It contains five fields: the Frame Control, the Duration/ID, a reserved field, the DQCA feedback and the FCS. In particular:

- The Frame Control consists of 2 bytes and contains the information described in the MAC header of the data packet (see Figure 3.4 in the previous section). The most important field is the final-message-bit that indicates whether there are more fragments of the same data message to follow. The AP copies this field from the Frame Control of the received data packet to the Frame Control of the FBP, in order to make this information available to all users, as it will be explained in more detail in Section 3.4.1.

- The Duration/ID, is an optional field but could be employed to enable the coexistence of DQCA and IEEE 802.11, as it will be discussed in Section 3.5.3.

- A sequence control field that contains the sequence number of the FBP, incremented by one for every FBP transmitted by the AP. By checking the value of this field, the nodes can deduce whether they have missed any previous FBP transmissions, a situation that can possible produce errors in the calculation of the queue counter values.

- Two 1-byte fields have been reserved for the transmission of TQ and RQ global counter values. These fields do not need to be present in all FBP, but they can be included periodically in order to allow new or idle nodes to enter the system and to increase robustness against errors caused by counter miscalculations by the nodes or by mis-detection the state of the minislots.

- The DQCA feedback contains the necessary information for the execution of the DQCA rules, namely the state of the control minislots and the data slot. This amounts to a feedback of $2 \times m$ bits for the minislots, given that two bits are sufficient for the representation of the ternary minislot state (i.e., empty, success, collision). Pad bits may be appended to the end of this field, to round up its length to the nearest multiple of a byte. The feedback part also includes an acknowledgment for the data packet transmitted in the data slot, indicating whether the packet has been correctly received by the AP or not (ACK or NACK, respectively).

In any case, the feedback field can be further extended to include additional information if required in order to implement more advanced scheduling schemes (for example, for QoS provisioning, which will be one of the main issues discussed in Chapter 5).

3.4 The DQCA Protocol Operation

Now that the main concept of DQCA and the frame formats have been fully described, the protocol operation rules will be given, along with a detailed example of the DQCA operation.

3.4.1 The DQCA Algorithm and the Protocol Rules

The DQCA protocol assumes an infrastructure topology consisting of an AP and a set of associated users. This section will provide a detailed description of the DQCA operation, by separately examining the role of the AP and the users. For the sake of simplicity, only the uplink communication direction has been considered. Hence, the nodes contend for channel access in order to transmit their messages to the AP, whereas the AP serves as a coordinator and provides the necessary feedback for the execution of the DQCA protocol rules. In any case, the presented rules can be easily modified to permit the protocol operation in the downlink, as it will be further discussed in Section 3.5.1.

Before proceeding to the detailed description of DQCA, it would be convenient to address an important implementation decision on the number of minislots m that constitute the CW. The selection of parameter m is critical to the system performance. A small number of minislots can cause congestion in the channel access process whereas a larger number may introduce unnecessary overhead. This issue has been thoroughly addressed in [56], where the conclusion was reached that, when $m \geq 3$, the collision resolution process works faster than the data transmission so near-optimum performance is ensured. As m increases, the delay experienced by the nodes is slightly reduced, with a cost on the throughput performance that deteriorates due to the additional overhead. Hence, unless otherwise stated, $m = 3$ is adopted as the most appropriate value for the CW size.

In DQCA, the AP has some additional responsibilities with respect to the associated users.[5] Hence the protocol operation will be described from the separate points of view of the AP and the users, respectively. Figure 3.6 shows a flow chart of the steps executed at the AP in each DQCA frame. The most important function is the collection of the feedback information regarding the state of the control minislots and the outcome of the packet transmission in the data slot that are included within the broadcast FBP. Initially, once the network is set up, the AP transmits a FBP that serves as a synchronization beacon.

[5] In any case, DQCA can also be applied to an ad hoc scenario in which each user takes the role of the AP for a given amount of time, assuming that no hidden nodes are present.

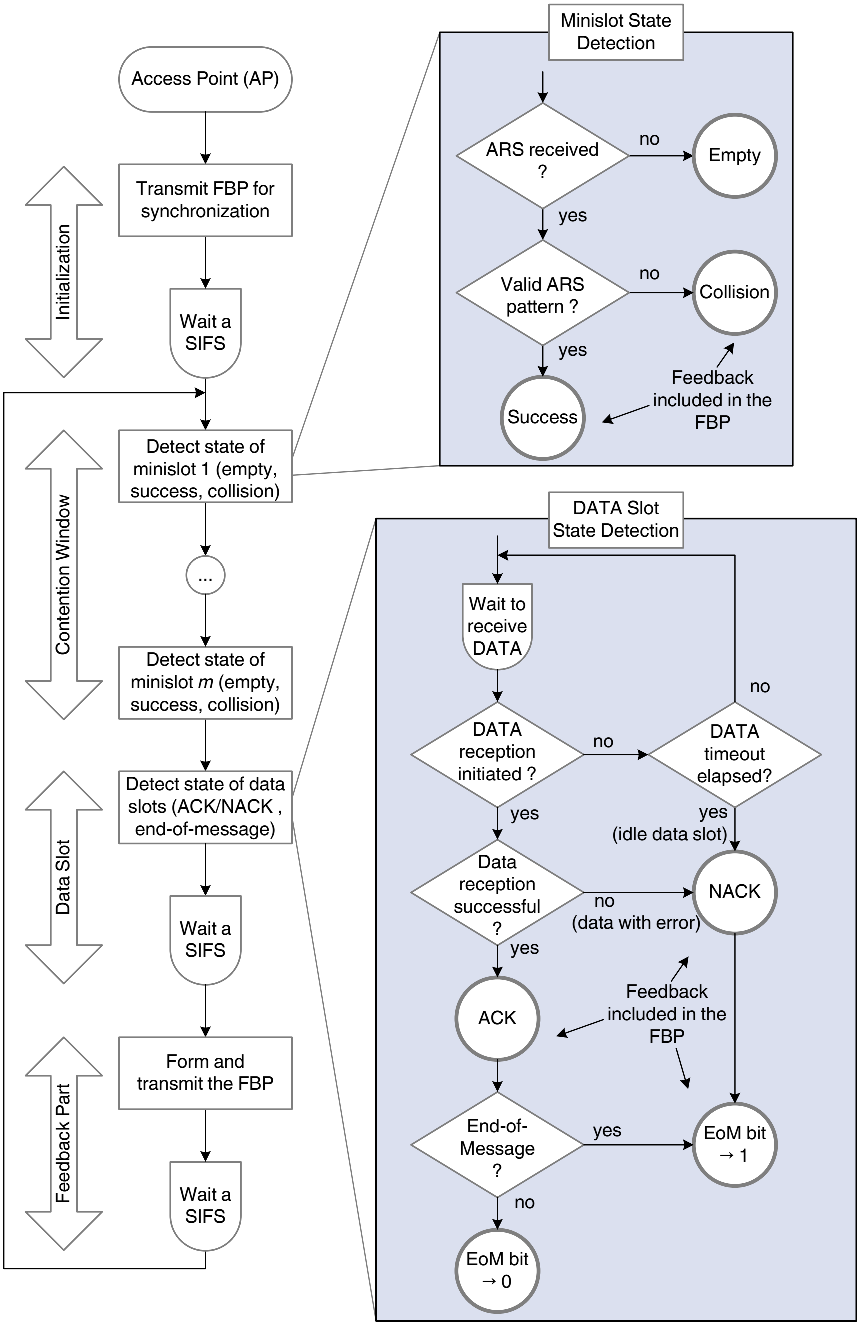

Access Point (AP)
Initialization
Transmit FBP for synchronization
Wait a SIFS
Contention Window
Detect state of minislot 1 (empty, success, collision)
...
Detect state of minislot *m* (empty, success, collision)
Data Slot
Detect state of data slots (ACK/NACK , end-of-message)
Wait a SIFS
Feedback Part
Form and transmit the FBP
Wait a SIFS
Minislot State Detection
ARS received ?
no
Empty
yes
Valid ARS pattern ?
no
Collision
yes
Success
Feedback included in the FBP
DATA Slot State Detection
Wait to receive DATA
no
DATA reception initiated ?
no
DATA timeout elapsed?
yes
yes (idle data slot)
Data reception successful ?
no (data with error)
NACK
yes
ACK
Feedback included in the FBP
End-of-Message ?
yes
EoM bit → 1
no
EoM bit → 0

Figure 3.6: Flow chart of the DQCA operation at the AP

The CW begins a SIFS time after the transmission of the FBP and is formed by m control minislots of a fixed duration. During this time, the AP listens to the channel and detects the outcome of each minislot. As mentioned before, if no ARS has been transmitted in a particular minislot, the AP detects an empty minislot state. Otherwise, depending on whether the pattern of the received signal within a minislot corresponds to the transmission of a single or multiple ARS (as explained previously in Section 3.3.1), the AP can distinguish between a successful outcome or an ARS collision. The ternary feedback on each minislot (marked within circles in the flow chart) is included in the FBP that will be transmitted by the AP at the third part of the DQCA frame.

The CW is followed by the data slot dedicated to the uplink transmission of data packets from the users to the AP. If no data transmission has been detected within a given data timeout period, the AP considers the data slot empty and a NACK is included in the FBP.[6] This scenario typically occurs under low traffic conditions, when users make sparse transmission attempts. Otherwise, the AP receives the data packet and issues an ACK, if a data packet has been received with no errors, or a NACK indicating the reception of a corrupted packet.

Apart from the state of the control minislots and the data slot, the AP must also set the final-message-bit for the FBP. If a data packet has been successfully received, the AP copies the final-message-bit of the received packet to the respective field of the FBP. As a result, all the users are notified whether the transmitting user has more pending frames to transmit in the next DQCA frame. If no packet has been received (empty data slot or collision) the AP sets the final-message-bit to ‘1’ to indicate that there is no ongoing data transmission session.

In continuation, the DQCA protocol operation is examined from the perspective of the users, with the help of the flow chart depicted in Figure 3.7. Consider that a new message arrives at an idle node with an empty buffer. The node must first check the state of the CRQ and wait until it becomes empty ($RQ = 0$). This condition is imposed by the blocked access collision resolution algorithm that does not permit access requests for newly arrived messages until all pending collisions are resolved. In continuation, the node must check the state of the DTQ. If the DTQ is empty ($TQ = 0$), a situation corresponding to very light traffic conditions, the Aloha-like transmission mechanism is initiated, as explained later in this section.

In the most typical scenario in which $TQ \geq 0$, DQCA operates as a reservation scheme. Therefore, the node randomly selects one of the m control minislots with equal probability $1/m$ and transmits an ARS packet. It then waits until the end of the frame to receive the FBP and check the state of the corresponding minislot. If the feedback indicates a collision due to the simultaneous transmission of multiple ARS in the same minislot, the node enters at the tail of the CRQ and shares the position with the other nodes involved in the same collision. When the group of collided nodes reaches the head of the queue ($pRQ = 1$) they randomly select a

[6] The possibility of falsely detecting an empty data slot despite an ongoing transmission and retransmission mechanisms for error recovery are not considered in the DQCA protocol description but will be discussed in Section 3.5.2.

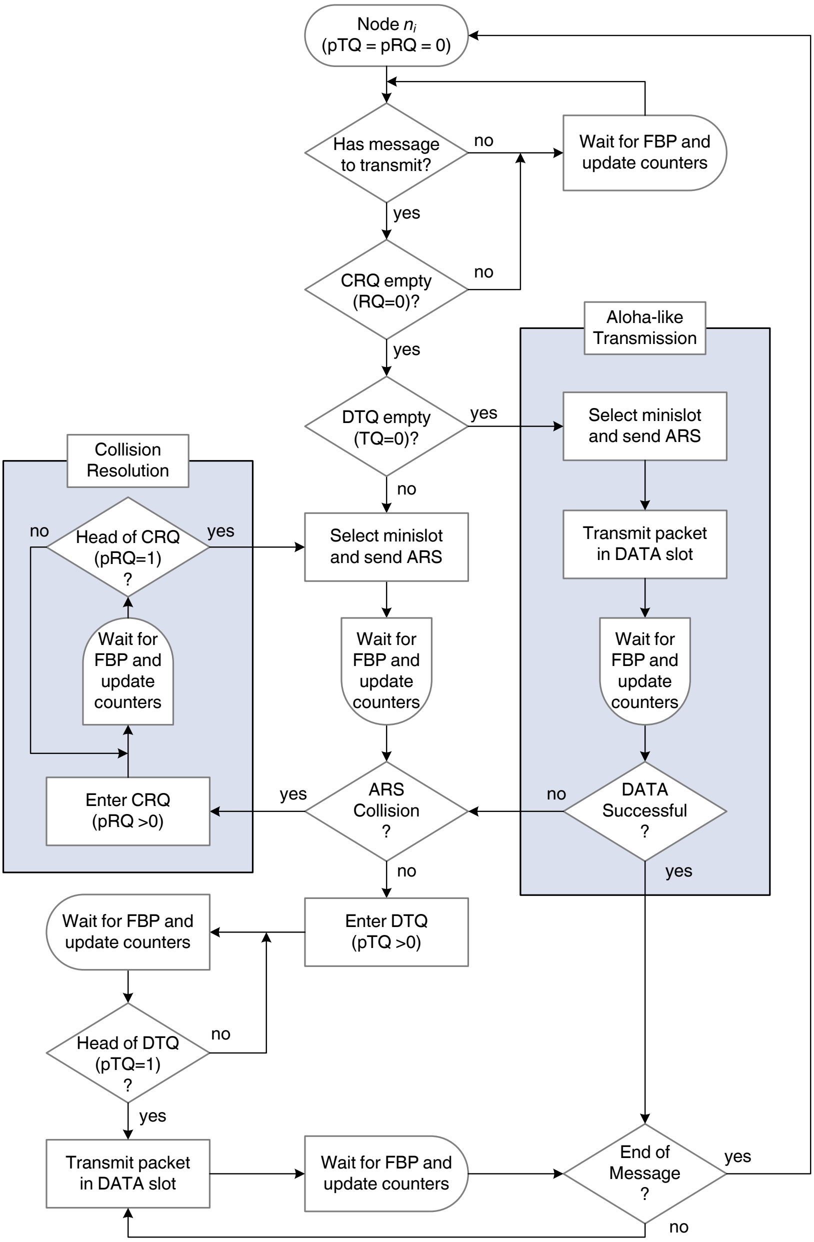

Node n_i
(pTQ = pRQ = 0)
Has message to transmit?
no
Wait for FBP and update counters
yes
CRQ empty (RQ=0)?
no
yes
Aloha-like Transmission
DTQ empty (TQ=0)?
yes
Select minislot and send ARS
Collision Resolution
no
Head of CRQ (pRQ=1) ?
yes
Select minislot and send ARS
Transmit packet in DATA slot
Wait for FBP and update counters
Wait for FBP and update counters
Wait for FBP and update counters
Enter CRQ (pRQ >0)
yes
ARS Collision ?
no
DATA Successful ?
yes
no
Enter DTQ (pTQ >0)
Wait for FBP and update counters
Head of DTQ (pTQ=1) ?
no
yes
Transmit packet in DATA slot
Wait for FBP and update counters
End of Message ?
yes
no

Figure 3.7: Flow chart of the DQCA operation at the system nodes

minislot and transmit a new ARS. This process is, in fact, an m-ary tree splitting algorithm and is repeated until the collision is resolved and the involved nodes successfully transmit an ARS.

Once a valid ARS is transmitted, the node enters the DTQ and is placed at the tail of the DTQ (by setting its pTQ counter to the corresponding value). The node advances towards the head of the DTQ by one position at a time, as the preceding nodes complete their data transmission and exit the queue. When it reaches the head of the queue ($pTQ = 1$) it begins the transmission of the data message in the data slot. If the message exceeds the predefined byte size it is fragmented into packets that are transmitted in consecutive DQCA frames. When the message transmission is completed, the node exits the queue (i.e., sets $pTQ = 0$) and has to repeat the process for every new message in its buffer.

The reservation scheme ensures collision-free data transmission, since collisions may only occur during the access request process. However, a data slot is wasted if a message arrives when the system is empty ($TQ = RQ = 0$), since the node must transmit an ARS, wait for the FBP and then transmit in the next frame. For this reason, the immediate (or free) access mode has been added: if a node with a new message finds the system empty, apart from transmitting an ARS in a randomly selected minislot, it is also allowed to transmit a data packet in the data slot of the same DQCA frame. If no other nodes employ the immediate access mode within the same frame, the data packet is successfully transmitted; otherwise, a collision in the data slot occurs. In that case, the reservation mechanism can be resumed seamlessly and the next algorithmic step for each involved node is determined by the outcome of the corresponding ARS transmission.

The DQCA algorithm is implemented as a sequence of actions executed by the nodes depending on the state of the queues and the information included in the FBP. The algorithm rules can be divided into three sets, the Data Transmission Rules (DTRs), the Request Transmission Rules (RTRs) and the Queuing Discipline Rules (QDRs). The DTRs and RTRs determine the actions of each node in the frame to follow. In particular, DTRs indicate whether a node can transmit a packet in the data slot of the subsequent frame and RTRs handle the transmission of ARS packets in the control minislots. The QDRs update the state of the queues by calculating the value of the four counters depending on the events that took place in the control minislots and the data slots, as reported in the FBP at the end of each frame. The execution of these rules takes place during the SIFS time that follows the transmission of the FBP and precedes the beginning of a new DQCA frame.

The three sets of rules are described next and are executed serially, in the presented order. If a user does not satisfy the condition of a particular rule, it simply advances to the next rule.

Data Transmission Rules (DTRs)

Each node must determine whether it is enabled to initiate the transmission of a message in the data slot of the subsequent DQCA frame. This can occur, when either of the following two conditions is met:

- If the system is empty ($TQ = RQ = 0$), every node that has data in its transmission buffer is enabled to transmit a packet in the data slot of the following frame (immediate or free access mode).
- If a node is at the head of the DTQ ($pTQ = 1$), it is enabled to transmit a packet in the data slot of the following frame. The value of the final-message-bit in the MAC header must be set to '1' when the last packet of the message is being transmitted, otherwise it must be set to '0'.

Request Transmission Rules (RTRs)

Each node must determine whether it is enabled to transmit an ARS in a randomly selected control minislot of the CW, according to the following two rules:

- If there are no collisions pending for resolution in the CRQ ($RQ = 0$), every node that has a message to transmit and does not hold a position in neither queue ($pTQ = pRQ = 0$) is enabled to transmit an ARS in a randomly selected minislot of the following frame. Note that this rule also applies to the immediate access mode.
- If a node is in the head of the CRQ ($pRQ = 1$), it is enabled to transmit an ARS in a randomly selected minislot of the following frame.

Queuing Discipline Rules (QDRs)

All nodes must update their counter values depending on the events that occurred in the preceding DQCA frame and are described through the information included in the FBP. Each node must perform the following actions in the presented order:

- Increase the value of TQ by one unit for each control minislot with a success state in the CW of the previous frame.
- Decrease the value of TQ by one unit if the last packet of a message (with the final-message-bit set to 1) has been transmitted successfully in the data slot of the previous frame.
- If there have been collisions pending for resolution ($RQ > 0$), reduce RQ by one unit to account for the collision resolution attempt of the nodes at the head of the CRQ.
- Increase the value of RQ by one unit for each control minislot with a collision state in the CW of the previous frame.
- Calculate its position in the DTQ and update the pTQ counter in the following way:

  - If the node is already waiting in the DTQ ($pTQ > 0$), it must decrease its pTQ value by one unit if the last packet of a message (with the final-message-bit set to 1) has been transmitted successfully in the data slot of the previous frame.
  - If the node has transmitted an ARS in the kth minislot of the previous frame and the state of this minislot has been marked as success, the node sets its pTQ value to point at the end of the DTQ. If more that one successful ARS have been transmitted in the CW of the previous frame, the order in which the corresponding nodes enter the DTQ follows a time arrival criterion. Therefore, the node that selected the kth minislot will enter the DTQ after the nodes that successfully transmitted an ARS in the first $(k-1)$th minislots and before the nodes that selected any of the remaining $(m-k)$th minislots.

- Calculate its position in the CRQ and update the pRQ counter as follows:
  - If the node is already waiting in the CRQ ($pRQ > 0$), it must first decrease its pRQ value by one unit and then increase it by one unit for each control minislot with a collision state in the CW of the previous frame.
  - If the node has transmitted an ARS in the kth minislot of the previous frame and the state of this minislot has been marked as collision, the node sets its pRQ value to point at the end of the CRQ. As before, if there are collisions in multiple slots, the involved nodes enter the CRQ in the time order in which the collisions have occurred.

3.4.2 An Example of the DQCA Operation

The DQCA mechanism and the execution of the algorithmic rules can be better understood with the help of an example that focuses on three consecutive DQCA frames, illustrated in Figure 3.8. The Figure can be divided into three parts: the upper part shows the frame exchange between the nodes and the AP; the middle part displays the arrival of data messages, represented by arrows on the time axis; the lower part contains a snapshot of the two distributed queues at the end of each frame, after the execution of the protocol rules has been completed.

Consider that initially the system is empty ($TQ = RQ = 0$), as well as the buffers of the nodes. Then, shortly before the ith DQCA frame messages arrive at nodes n_1 and n_2. The immediate access mode is applied, according to which the two nodes transmit an ARS in a randomly selected control minislot and also transmit a data packet in the data slot. The data packets unavoidably collide but the ARS are transmitted with success since different minislots have been selected. The FBP broadcasted by the AP reports the state of the minislots (success-empty-success) and the collision in the data slot. In the meantime, new messages arrive at nodes n_3, n_4 and n_5.

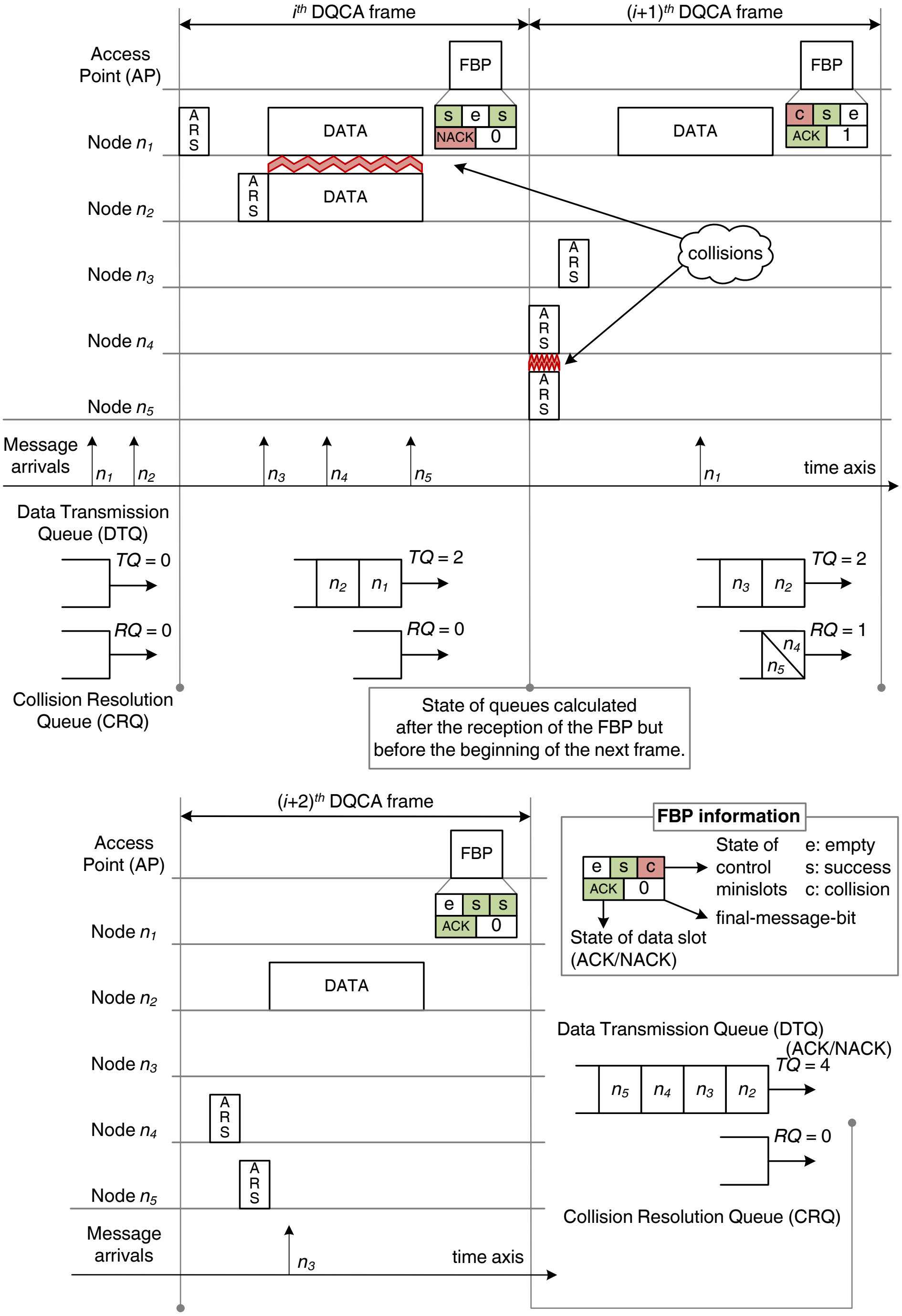


Figure 3.8: Example of the DQCA operation

Once the FBP is received, the nodes execute the QDRs to update the queue counters. As a result, n_1 who has successfully transmitted in the first minislot enters at the head of the DTQ ($pTQ = 1$) and, as the DTRs dictate, can perform a packet transmission in the data slot of the following frame. Node n_2 also enters the DTQ in the second position ($pTQ = 2$). In parallel, according to the RTRs, nodes n_3, n_4 and n_5 can attempt an ARS transmission in the CW.

In the second DQCA frame, nodes n_4 and n_5 randomly select the first minislot, resulting to a collision among the two ARS. Node n_3 achieves a successful ARS transmission by selecting the second minislot. The minislot collision does not affect the data transmission process, hence, a packet is transmitted by node n_1. In this example, it has been assumed that this message consists of a single packet, so the final-message-bit must be set to '1' to indicate that no more fragments are expected.

The FBP announces that the state of the three minislots was empty-success-collision, respectively, and the the data message has been completed with success. The state of the queues after the end of the second frame is as follows. Node n_1 exits the queuing system by setting its pTQ counter to 0 and n_2 advances to the head of the DTQ. Node n_3 enters the DTQ at the second position ($pTQ = 2$), after successfully transmitting an ARS. Nodes n_4 and n_5 enter at the head of the CRQ ($pRQ = 1$) in order to resolve their ARS collision. Note that the nodes occupy the same queue position since they were involved in the same collision. In the meanwhile, a new message arrives for n_1.

In the third frame, according to RTRs, n_4 and n_5 retransmit an ARS in the CW, this time with success as they select different minislots. On the other hand, node n_1 is not allowed to compete for access for its newly arrived message until the pending collision is resolved. In the data slot, a packet is transmitted by n_2. This time, it has been assumed that the data message consists of two packets, so the final-message-bit is set to '0', indicating that there are more fragments to follow in the next frame.

Therefore, n_2 remains at the head of the DTQ and is followed, as before, by n_3. Nodes n_4 and n_5 leave the CRQ after their successful ARS transmissions and enter the DTQ at positions 3 and 4, respectively. Now that the CRQ is empty, n_1 is allowed to attempt an ARS transmission in the following DQCA frame. Finally, another message arrives at the buffer of n_3, however the node will not deal with it until the transmission of a previously arrived message is completed.

3.5 Further Discussion and Open Issues

DQCA has been conceived as a MAC protocol for infrastructure, single hop scenarios and this is the context in which, throughout this thesis, DQCA performance has been evaluated and enhanced with the application of CL scheduling. Nevertheless, these conditions are not limiting factors for the application of DQCA. On the contrary, they constitute a starting point that opens the road for the adaptation of the protocol to different environments. For example, the DQCA concept has served as a base for a master-slave clustering mechanism that makes its application possible in ad hoc scenarios in which every node may assume the role of the AP for a bounded amount of time [65].

The main objective of this thesis has been to extend the scheduling policies supported by the DQCA MAC protocol through a CL interaction with other layers of the OSI stack. Unavoidably, as in most cases when the focus is laid on a particular aspect of a subject, some issues are left out of the scope of the study, for the sake of simplicity and to avoid deviation from the main point of interest. This section will tackle some interesting aspects of DQCA operation that have not been thoroughly considered in this thesis. In particular, the following issues will be discussed:

- The DQCA operation in both uplink and downlink communication directions.
- The impact of channel errors on the DQCA operation and some mechanisms for system recovery.
- The impact of interference on DQCA and mechanisms for coexistence with IEEE 802.11 WLANs.
- Handoff functions for DQCA operation in cellular environments.

This section does not pretend to thoroughly resolve these pending issues, but to present the problem statement and provide some guidelines for the design of possible resolution schemes in future work.

3.5.1 DQCA Operation in the Downlink

The DQCA description and examples provided so far in this chapter have considered an uplink communication scenario in which a number of nodes compete for channel access in order to transmit data to the AP. Nevertheless, in a communication system it is expected to encounter uplink and downlink transmissions and a MAC protocol should be expected to handle both communication directions.

There are two reasons why the focus has been mainly laid on the uplink. In the first place, one of the main strengths of DQCA is the channel access mechanism that allows multiple nodes to access the medium in an efficient way, without additional delays due to backoff or data collisions. This feature is more relevant to the uplink, especially when the number of contending users is high, whereas the downlink is rather related to user selection and scheduling issues managed by the AP. In the second place, the DQCA downlink operation can be considered as a special case of the uplink scenario, given that the AP is, in fact, another node but with some additional responsibilities, summarized in the flow chart presented in Figure 3.6.

The DQCA downlink operation can be achieved by introducing some minor modifications to DQCA that do not alter the fundamental characteristics of the protocol and will be described by means of an example, illustrated in Figure 3.9. In the first depicted DQCA frame, it has been assumed that no ARS transmissions have

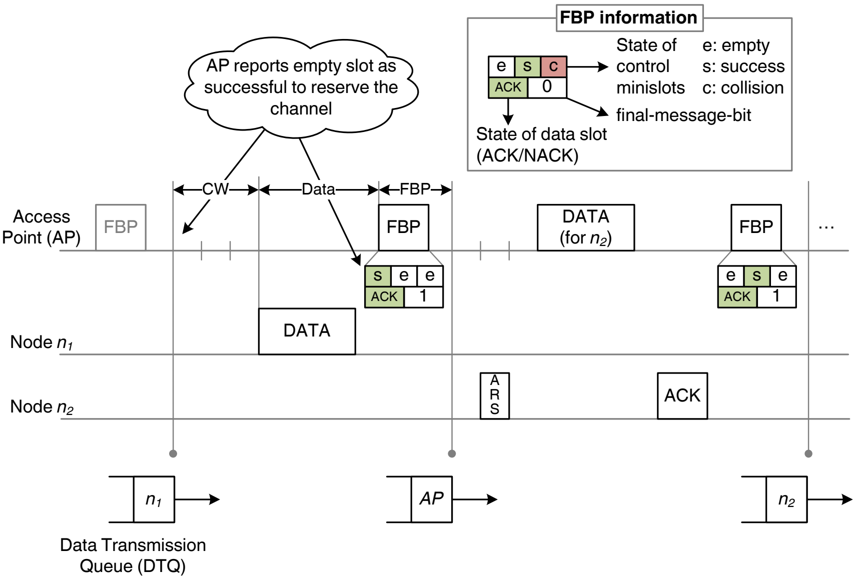


Figure 3.9: DQCA operation in the downlink

taken place in the CW and that node n_1 has completed an uplink data message transmission (hence the final-message-bit has been set to '1'). In an uplink-only scenario, the FBP would report the empty state of the control minislots and the successful data transmission. In the downlink scenario, assuming that the AP has data to transmit and needs to gain channel access, it can simply report one of the empty minislots as successful. This will be transparent to the nodes who will assume that another node has entered the DTQ and update the queue counters according to the protocol rules. Thus, the AP will hold a position in the DTQ and will transmit when the head of the queue is reached, which, in the given example, occurs in the next DQCA frame.

In the CW of the second depicted frame, node n_2 sends an ARS in the second minislot. In the data slot, the AP transmits a packet, which is coincidentally (in this example) addressed to n_2. Once the data transmission is completed, n_2 replies with an ACK frame after a SIFS time, otherwise it remains silent. Finally, the AP transmits the FBP a SIFS time after the reception of the ACK, or, in case of data failure, once after a specified ACK timeout has elapsed.

The proposed downlink implementation can be seamlessly incorporated in the DQCA operation, as presented in Section 3.2. The resource allocation between the uplink and the downlink is an open issue and smart algorithms can be implemented at the AP to determine the amount of time dedicated to each communication direction according to the system needs.

3.5.2 DQCA Operation under Channel Errors

An important assumption considered in the DQCA description is that, given the protocol operation in an infrastructure scenario, all the nodes can hear the AP and receive the FBP with no errors. In order to minimize the probability of feedback errors, even though link adaptation may be applied to the data packet transmission depending on the rate set defined by the PHY layer, the FBP is always transmitted at the lowest available rate. Nevertheless, errors are an unavoidable part of wireless communications so it is important to design mechanisms for the recovery of the system whenever they occur. In continuation, three types of errors will be discussed that may affect the system in different ways, depending on the involved part of the frame, namely the data slot, the control minislots or the FBP.

Errors in the Data Slot

The first and most straightforward type of error affects the transmission of the data packet that may be received with errors by the AP or, in a more extreme case, not be detected at all. Hence, the AP will detect either an erroneous data transmission or an idle data slot. The key in both scenarios is that, assuming that the system is not empty (i.e., there are users waiting in the queues) the AP has been expecting the reception of the data packet by the node at the head of the DTQ.

The simplest way to deal with this type of error is to define a number K of packet retransmission attempts. If the transmitting node does not receive a valid acknowledgment through the FBP after K retransmissions in consecutive DQCA frames, the message is discarded and the node leaves the DTQ. A simple example is given in Figure 3.10 where node n_1 transmits a data packet which is received with errors. The FBP indicates that no valid data has been received, so n_1 is prompted to attempt a retransmission in the following frame. To indicate that retransmissions are pending, the AP sets the final-message-bit to '0'. After K unsuccessful attempts, n_1 discards the packet and exits the DTQ. The final-message-bit is now set to '1', so that the remaining nodes in the DTQ (e.g., nodes n_2 and n_3 in the example) execute the DQCA algorithm rules and advance in the queue.

Of course, this example assumes that the FBP is correctly received by n_1, which is the case in the majority of the time since the FBP is always transmitted at the lowest transmission rate. Errors in the FBP will be considered later in this section. Furthermore, the AP can maintain some statistics on the errors that occur in the data slot and exploit them to perform more sophisticated system decisions.

Consider now the case where the AP does not detect at all the transmitted data packet and senses an empty data slot, even though the DTQ is not empty (i.e., a data transmission was being expected). If the AP proceeds with the transmission of the FBP there is a possibility of collision between the FBP and the data packet, as shown in Figure 3.11 (a). To avoid this situation, a data timeout period is employed which, in order to cover the worst case scenario, should be equal to the time required for the transmission of a packet at the lowest transmission rate. The AP should

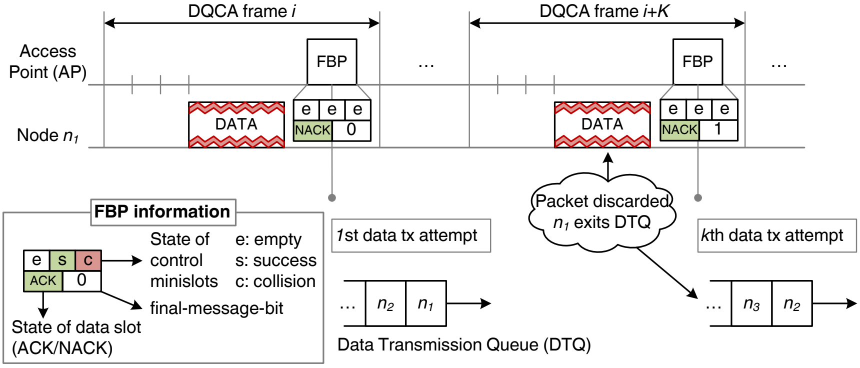
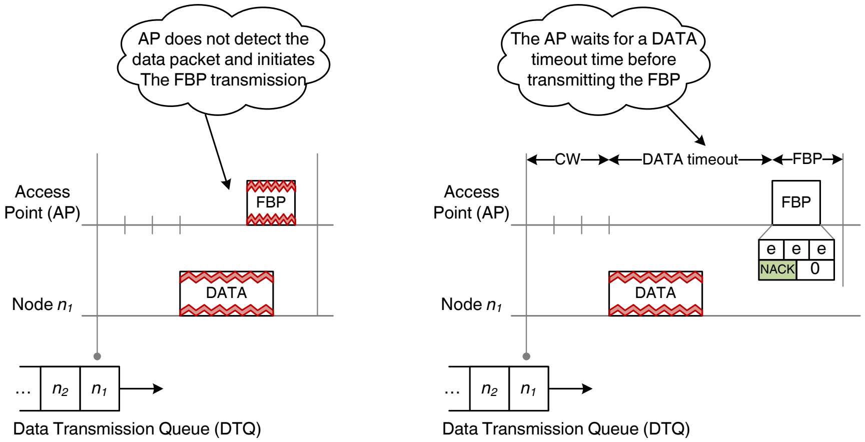


Figure 3.10: Data packet retransmission due to channel errors

transmit the FBP after this timeout has elapsed to avoid a collision, as shown in Figure 3.11 (b).


AP does not detect the data packet and initiates The FBP transmission
The AP waits for a DATA timeout time before transmitting the FBP
CW
DATA timeout
FBP
Access Point (AP)
Node n_1
FBP
DATA
e e e
NACK 0
... n_2 n_1
Data Transmission Queue (DTQ)

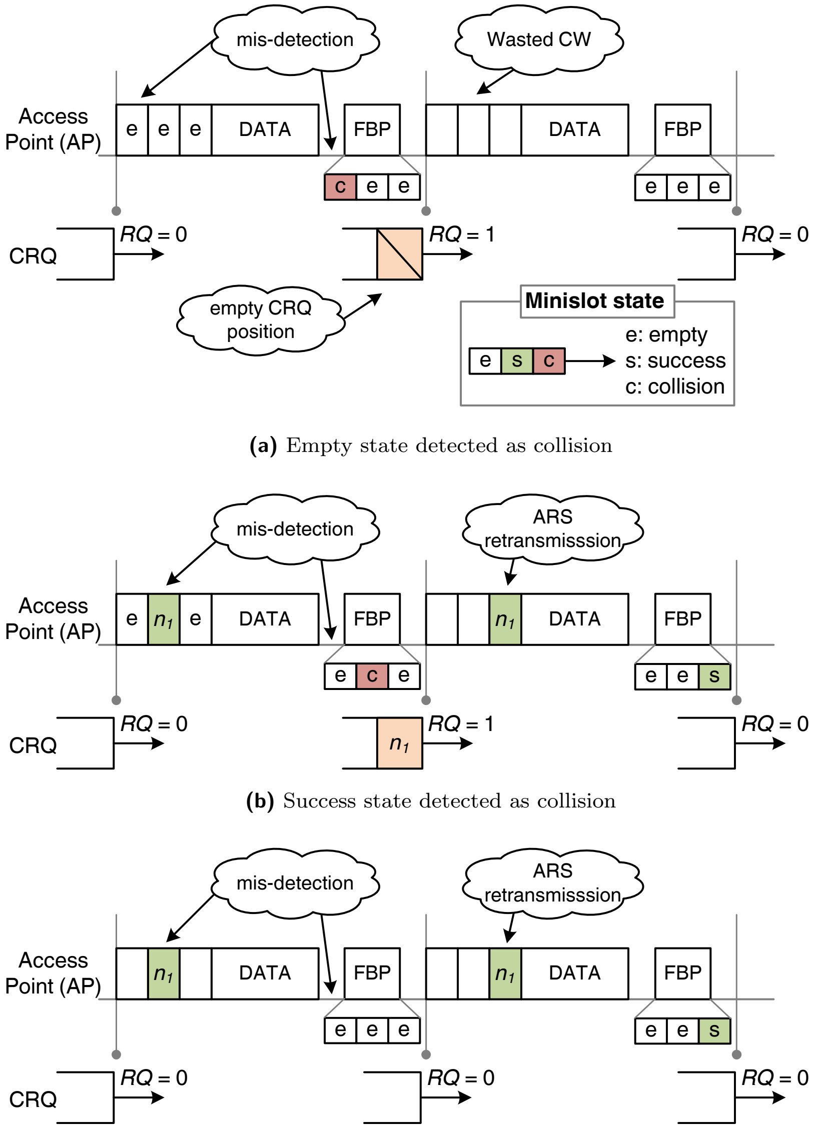
(a) Collision between the data packet and the FBP

(b) Collision avoidance with the use of the data timeout policy

Figure 3.11: Example of the data timeout policy

Mis-detection of the State of the Control Minislots

Another type of errors, briefly mentioned in Section 3.3.1, concerns the mis-detection of the state of the control minislots by the AP. As a result, the execution of the DQCA algorithmic rules can produce problematic situations since it is based on inaccurate feedback provided within the FBP. The possible mis-detection errors and their eventual resolution, more thoroughly addressed in [61], are summarized in Table 3.2. Brief examples are illustrated in Figure 3.12, plots (a) to (f).

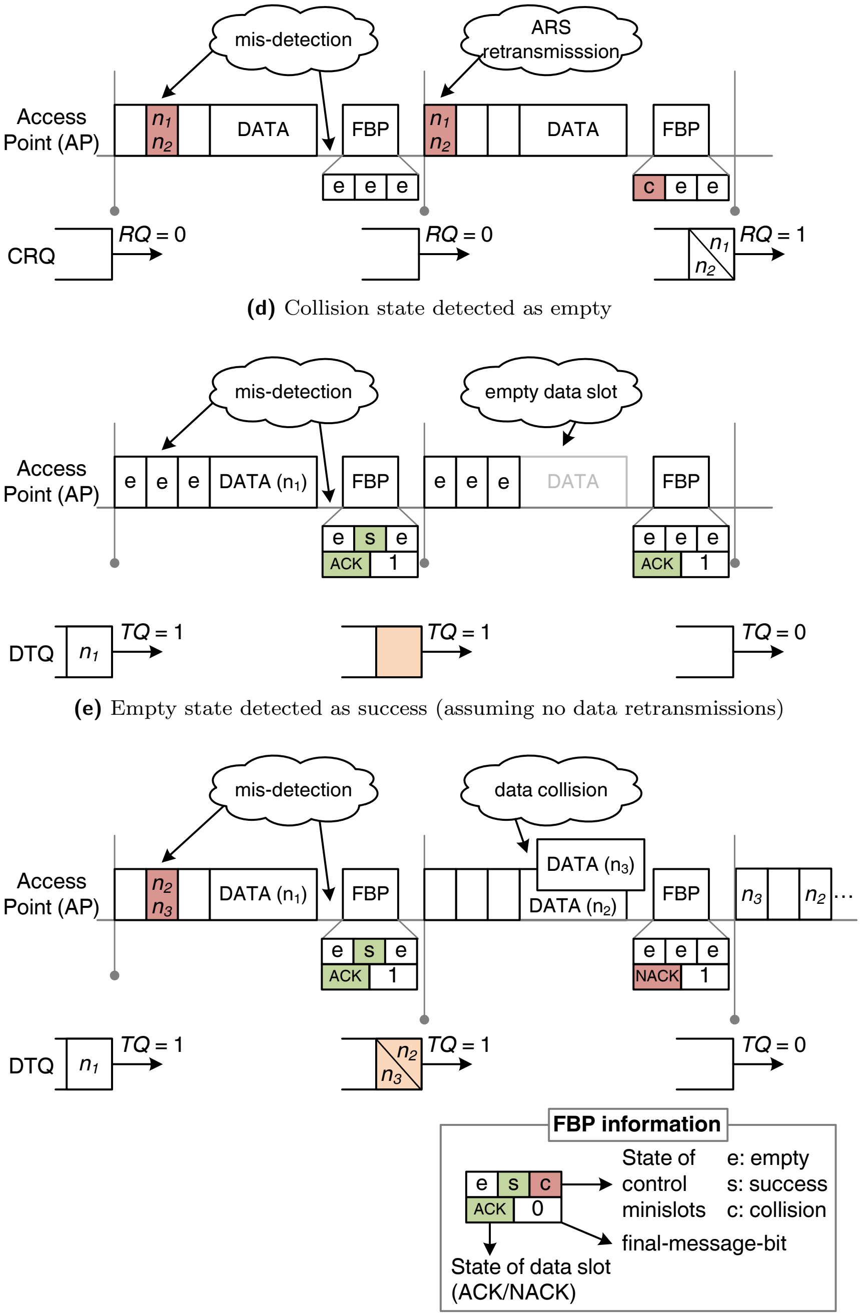


(a) Empty state detected as collision

(b) Success state detected as collision

(c) Success state detected as empty

Figure 3.12: Examples of state mis-detection of the control minislots



(d) Collision state detected as empty

(e) Empty state detected as success (assuming no data retransmissions)

(f) Collision state detected as success (assuming no data retransmissions)

Figure 3.12: Examples of state mis-detection of the control minislots (cont.)

Table 3.2: Possible outcomes of the control minislot state mis-detection

Detected State	*True State*	*Resolution*
Collision	Empty	Empty slot detected as a collision: an empty CRQ position will be falsely considered occupied by the system. No nodes are affected and the CRQ process will eventually resume after an empty CW. [Figure 3.12 (a)].
Collision	Success	Successful slot detected as a collision: the involved node will enter the CRQ and will eventually retransmit an ARS. [Figure 3.12 (b)].
Empty	Success	Successful slot detected as empty: the involved node will realize that its ARS was not received and will attempt a retransmission in a subsequent CW. [Figure 3.12 (c)].
Empty	Collision	Collision slot detected as empty: the involved nodes will realize that their ARS were not received and will attempt a retransmission in a subsequent CW. [Figure 3.12 (d)].
Success	Empty	Empty slot detected as success: an empty DTQ position will be falsely considered occupied by the system, leading to an empty data slot. In this case the system will execute the policies described in Section 3.5.2 and the DTQ process will eventually resume after K data slots[a]. [Figure 3.12 (e)].
Success	Collision	Collision slot detected as success: a DTQ position will be occupied by multiple nodes leading to collision. The AP will notify the nodes in the FBP[a] and the collision resolution algorithm will be executed. [Figure 3.12 (f)].

[a] If data retransmissions are not considered, correct protocol operation will resume after only $K = 1$ lost data slot.

The first four error scenarios have a lesser impact on the system since they mainly affect the collision resolution process and their only consequence is that some nodes suffer some additional slight delay before entering the DTQ. The last two mis-detection cases are more serious because they involve data packet collisions and require more time for their resolution. In any case, it is important to emphasize that even though some efficiency may be lost due to errors, the system is able to recover without entering in a deadlock state, ensuring the robustness of the protocol.

Errors in the FBP Reception

The FBP plays a very important role in the execution of DQCA since it contains the necessary feedback for the correct execution of the protocol rules by the nodes. Hence, in order to minimize errors, it is always transmitted at the lowest available rate. In any case, FBP errors cannot be completely avoided so it is necessary to contemplate mechanisms for the system recovery whenever they occur.

Consider the example illustrated in Figure 3.13. Node n_3 transmits a data packet which is received correctly by the AP, which, in turn, replies with the FPB acknowledging the reception and indicating the end of the transmitted message. Normally, n_3 should exit the DTQ and the node in the following DTQ position, i.e., n_1 in the example, should proceed with the data transmission in the next DQCA frame. Nodes n_1 and n_2 receive the FBP without errors and update their queue counters. Node n_3 however misses the FBP so it is not aware of the outcome of its data transmission and, in addition, has a wrong perception of the state of the DTQ.

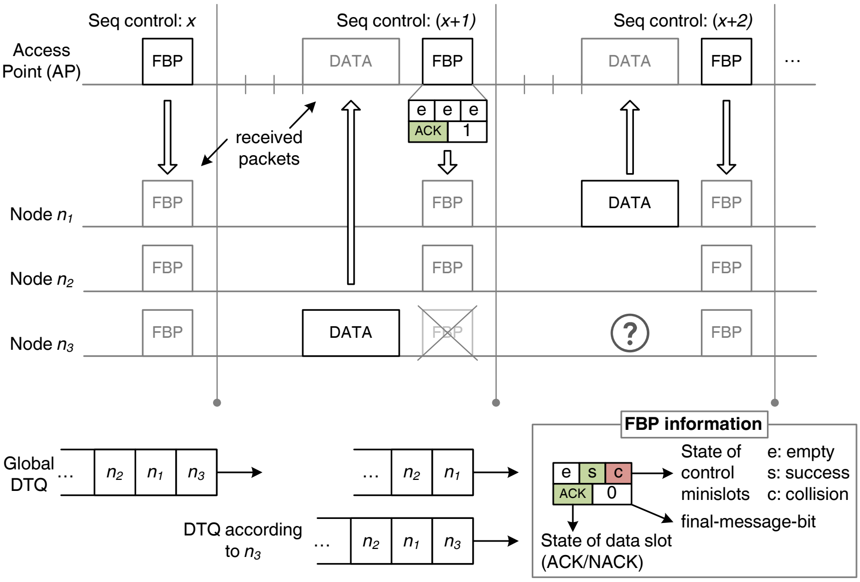


Figure 3.13: Error in FBP reception

Several problems may arise from this situation. First, n_3 does not know whether the data packet has been received and may attempt to retransmit it in the next DQCA frame. This would cause collision with the data transmission from node n_1, which could be repeated in subsequent DQCA frames if retransmissions are considered. Eventually, when the maximum number of retransmission is reached, the nodes will discard the packets and normal DQCA operation will resume. A worst case scenario could emerge if node n_3 lost synchronization[7] and attempted

[7] The data slot begins exactly a $(SIFS + m\times$ ARS) time after the end of the FBP transmission, so if a node misses completely the reception of the FBP it could lose synchronization.

the data retransmissions outside the data slot.

Some simple mechanisms could be employed to alleviate the impact of FBP errors in the network. One solution would be to employ a FBP timeout limit, defined as the maximum time interval that a node should wait for the reception of a FBP after the end of the data packet transmission. In an error-free scenario, this interval would be equal to a SIFS time. Nevertheless, if a longer timeout were appropriately set, a node that missed a FBP would refrain from retransmissions long enough to receive the FBP of the next DQCA frame, so any synchronization issues would be resolved. In addition, by keeping track of the sequence control number of the FBP that increments by one unit for consecutive frames (see Section 3.3.3), a node can determine whether any previous FBP transmission has been missed. In that case, to avoid collisions, the node should drop any retransmission attempts and reset its counters (local reset). Bear in mind that the TQ and RQ global counters are included in the FBP, so that a node can easily resume operation after a local reset, following the DQCA rules. Finally, a global reset of the counters of all nodes could be forced by the AP as an ultimate solution if multiple consecutive errors occurred in the data transmissions.

3.5.3 Interference and Coexistence with Other Systems

DQCA is a MAC layer protocol that can be applied over various PHY layers and operate in different licensed or unlicensed frequency bands. Depending on the actual protocol implementation, interference and coexistence issues with other systems sharing the same spectrum should be considered.

Careful frequency planning and deployment are keys to minimizing interference among systems. Additionally, since DQCA requires the presence of a coordinating node, it could easily include an adaptive algorithm for the selection of the operation frequency channel, if the PHY layer permits it. For example, multiple errors from different nodes during an extensive time interval could be attributed to interference by the AP who could, in turn, decide to shift the operation frequency to another available channel.

Throughout this thesis and without loss of generality, DQCA operation has been evaluated over the IEEE 802.11 b and g PHY layer specifications. An interesting issue would be to examine interference and coexistence scenarios between DQCA and the IEEE 802.11 DCF. The operation of the two protocols in the same or adjacent frequency bands would have a stronger impact on IEEE 802.11 DCF performance. Due to the IEEE 802.11 CSMA/CA mechanism and given that, with few exceptions, the maximum idle interval within a DQCA frame does not exceed a SIFS time, legacy stations would never sense the medium idle and would be forced to defer from transmission for long periods.

Nevertheless, it is possible to implement a coexistence mechanism on DQCA nodes in order to enable the joint operation of both protocols. This idea has been further developed in [66], where a mechanism for the coexistence of IEEE 802.11 with a DQCA variation for ad hoc networks has been proposed. The main idea is

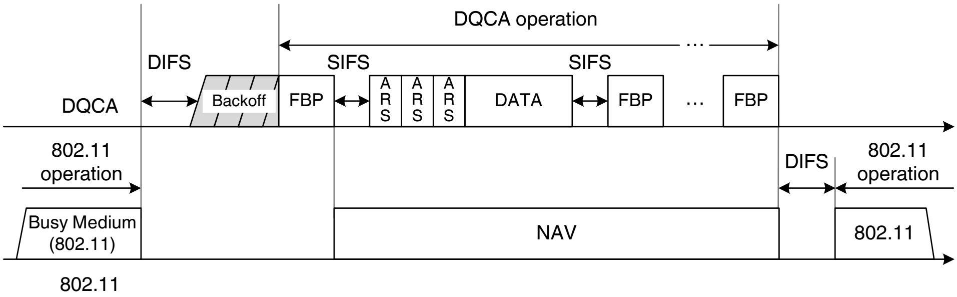


Figure 3.14: DQCA with IEEE 802.11 DCF coexistence

presented here though an example depicted in Figure 3.14. In this example, time is shared between DQCA and legacy IEEE 802.11 stations, in way similar to the alternation between contention and contention-free periods defined in the IEEE 802.11 standard. The DQCA AP seizes the channel following the DCF rules, i.e., after sensing the channel idle for a DIFS time and after executing the backoff mechanism (not depicted in the figure), and maintains it for a number of consecutive DQCA frames. By appropriately setting the Duration/ID field of the FBP (and of the DQCA data packets), the IEEE 802.11 stations can update the NAV and defer transmission until the DQCA protocol releases the channel.

3.5.4 Handoff Functions for Cellular Deployment

The presented description of DQCA considers a single-cell scenario with a centrally located AP. Some work has been conducted in [67] and [68] to extend DQCA operation to a multi-cell environment where nodes may move between adjacent cells controlled by APs that operate in different frequencies. In order to enable seamless roaming, it is necessary to define a handoff mechanism to implement a number of functions at the mobile node. These functions should handle the link status monitorization to determine if a handoff is required, the AP discovery and selection based on some predefined criteria and the reassociation process with the new AP. A flow chart that illustrates the basic steps of the DQCA handoff process is given in Figure 3.15. A brief description of these steps and how they can be incorporated in the DQCA mechanism will be provided in this section.

Handoff Initiation Process

Typically, a node initiates the handoff process after noticing a deterioration of the link quality with the connected AP. A link status monitoring function is, hence, required, in order to obtain frequent updates on the channel state. The periodic broadcast of the FBP at the end of each DQCA frame facilitates this process, since the nodes can estimate the link status on a frame-by-frame basis. If the link quality is below a predefined threshold, the node should initiate the discovery process to

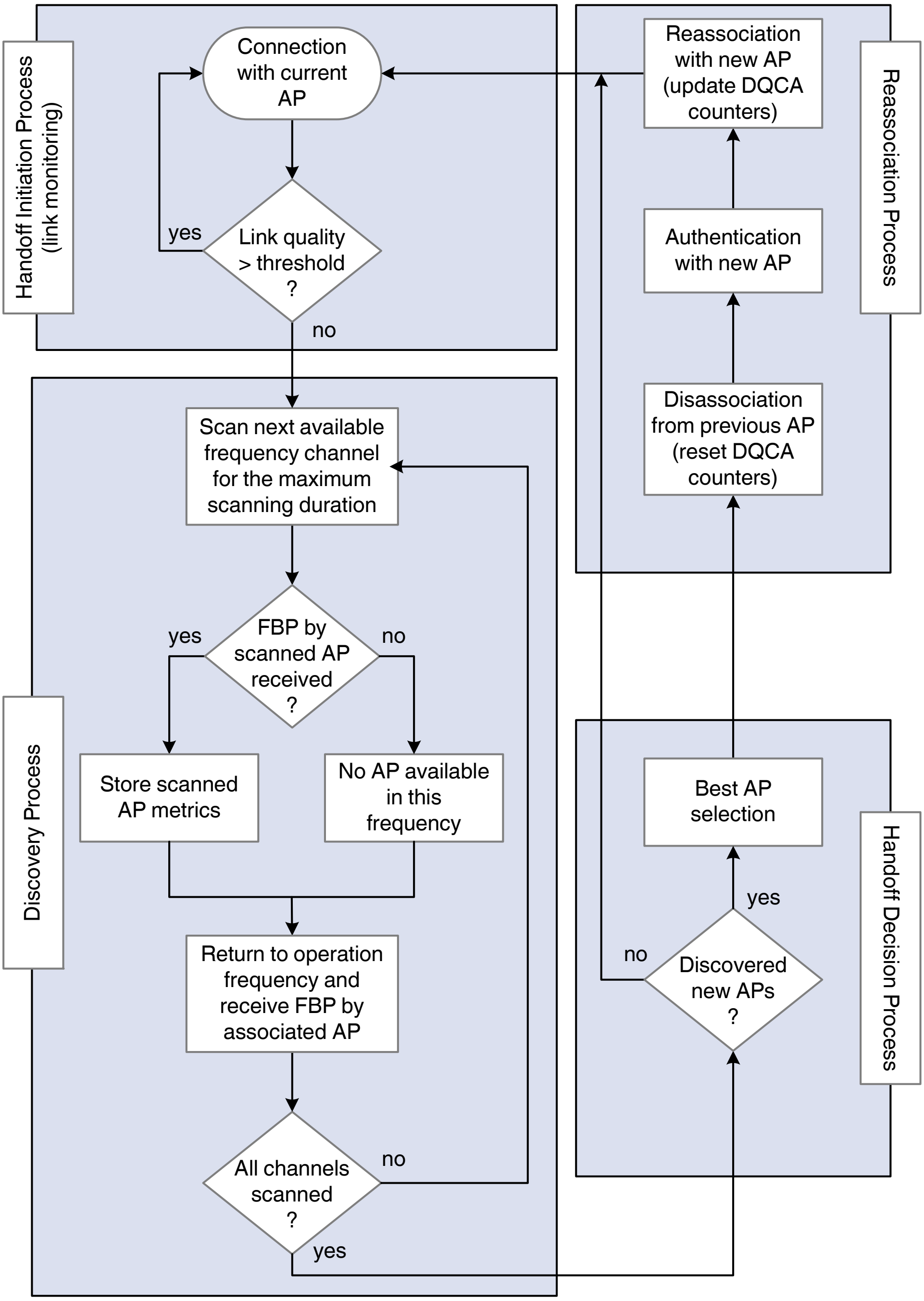

Handoff Initiation Process
(link monitoring)
Connection with current AP
yes
Link quality > threshold ?
no
Reassociation with new AP (update DQCA counters)
Authentication with new AP
Disassociation from previous AP (reset DQCA counters)
Reassociation Process
Scan next available frequency channel for the maximum scanning duration
yes
FBP by scanned AP received ?
no
Store scanned AP metrics
No AP available in this frequency
Discovery Process
Return to operation frequency and receive FBP by associated AP
no
All channels scanned ?
yes
Best AP selection
yes
no
Discovered new APs ?
Handoff Decision Process

Figure 3.15: Flow chart of the handoff process in DQCA

obtain information on the availability of neighboring APs with potentially better connection. The selection of the threshold is an important system decision that should take into consideration the following conditions:

- The threshold should not be very low, since, to ensure a seamless handoff, the discovery phase should be initiated before the connection with the current AP is lost.
- The threshold should not be very high, since frequent scans have a cost performance and may lead to unnecessary handoffs.

Apart from the link quality, another metric that can lead to a handoff initiation is the congestion state of the AP. In DQCA, the state of the distributed queues, expressed by the counters TQ and RQ, provide an indication of the traffic load within the cell and can be employed to achieve load balancing among a set of adjacent APs in a multi-cell DQCA network.

Discovery Process

Once a node has determined that a handoff is required, the discovery function is initiated. During this phase, the node must passively scan other frequency channels in a search for available APs with a better link quality. In this case, the FBP that is broadcasted by each AP at the end of every DQCA frame acts as a beacon. If a scanning node detects a FBP from a new AP, it can measure the link quality in order to decide whether to proceed with the handoff.

The challenge presented in the discovery phase is that while a node remains associated to an AP, it must receive the FBP in order to execute the DQCA algorithmic rules. In other words, the scanning process should take place during the other parts of the DQCA frame (i.e., the CW and the data slot) and the node must switch back to the operating frequency in time for the FBP reception. For this reason, the scanning time should be carefully selected, according to the following criteria:

- The scanning time must be long enough so that a node in the discovery phase may be able to receive and demodulate a complete FBP packet from the scanned AP. Hence, the scanning duration has to be at least equal to the fixed transmission time of a FBP packet plus the time needed to perform the channel switch.
- The scanning time must be short enough to allow a station to reconnect to the operating channel in time for the FBP reception. In other words, the scanning time should not exceed the joint duration of the CW and the data slot. Since the data slot is generally variable in size, the most conservative approach should be employed, by considering the shortest data slot duration that corresponds to the transmission of the smallest data packet at the highest available transmission rate.

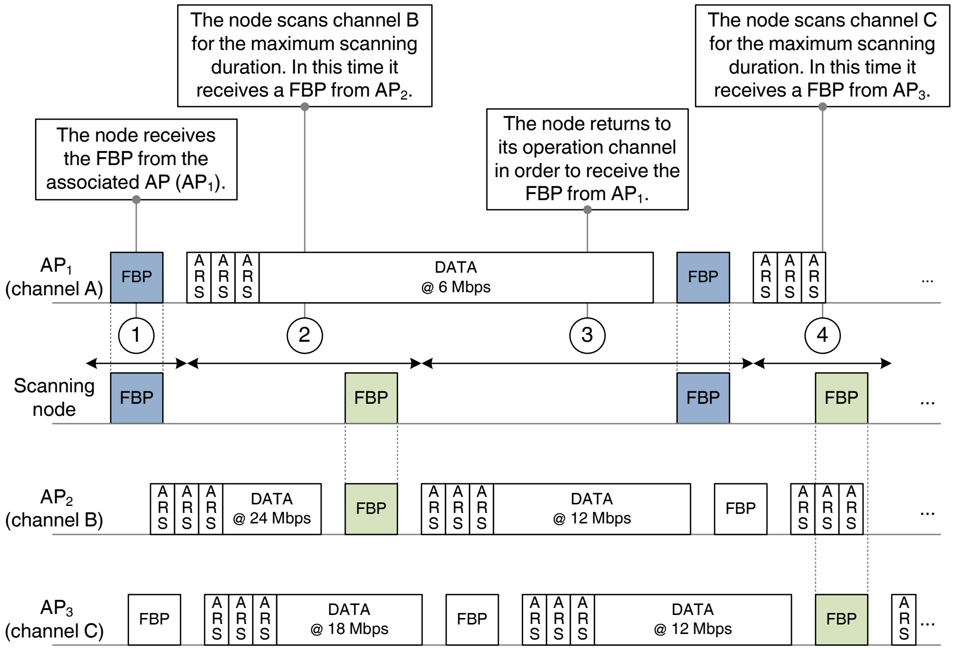

The node scans channel B for the maximum scanning duration. In this time it receives a FBP from AP2.
The node scans channel C for the maximum scanning duration. In this time it receives a FBP from AP3.
The node receives the FBP from the associated AP (AP1).
The node returns to its operation channel in order to receive the FBP from AP1.
AP1 (channel A)
Scanning node
AP2 (channel B)
AP3 (channel C)
FBP
ARS
DATA @ 6 Mbps
DATA @ 24 Mbps
DATA @ 12 Mbps
DATA @ 18 Mbps
1
2
3
4
...

Figure 3.16: Example of the discovery process in DQCA

An example of the discovery process is shown in Figure 3.16. Consider a scenario with three APs operating in non-overlapping channels A, B and C (as, for example, the channels 1, 6 and 11 of IEEE 802.11 [1]). The example focuses on a node associated to AP_1, operating on channel A, that initiates the discovery process. The node listens to the neighboring channels B and C for some time but must always return to the operating channel A to receive the FBP from AP_1. Generally, there is no intra-cell timing synchronization between the APs. In the example, once the FBP of AP_1 is received (marked on the figure as event 1) the node switches to channel B for the maximum defined scanning duration. In this time interval, it receives the FBP transmitted by AP_2. Then, the node returns to its operation channel and remains there until the reception of the FBP, sent by AP_1. Thus, the node may execute the DQCA rules and update the state of its queues. In continuation, the next available channel (i.e., channel C) is scanned and a FBP by AP_3 is received. In general, the node may not always detect the operation of an AP in a scanned channel due to the fact that the scanning duration is bounded and the FBP transmission may not take place within this time limit.

Handoff Decision Process

Once the discovery process is completed, the node must decide whether to proceed to the execution of a handoff and select the best AP if multiple choices are available. Some algorithms for the AP selection have been proposed and evaluated in [68]. The metrics identified as critical factors for the AP selection are summarized next, along with some guidelines for the implementation of the handoff decision process:

- The quality of the link between the node and the AP, measured in terms of the SNR or the Received Signal Strength Indicator (RSSI). Better channel quality implies more reliable communication and can support the use of higher transmission rates. In a simple but effective handoff decision mechanism, the node could select the AP with the highest measured link quality, after all the available frequency bands have been scanned. An alternative scheme could be the selection of the first discovered AP that yields a higher SNR with respect to the current AP. The latter option reduces the delay of the handoff process since the scanning of the whole spectrum is no longer a prerequisite, but, on the other hand, may not lead to the discovery of the best neighboring AP. Both schemes should include a hysteresis margin to ensure that the channel quality of the new AP is sufficiently better than the current link, to avoid unnecessary handoffs and reduce the ping-pong effect.

- The congestion conditions of the AP, expressed by the TQ and RQ counters. The traffic load served by an AP is directly related to the QoS that the network can provide. In an overloaded cell, the node will probably suffer long waiting times before getting a chance to transmit. Hence, load balancing algorithms can be designed to improve both the system performance and the user experience. In a simple load-balancing handoff mechanism, the node can select the AP with the smaller TQ value (i.e., the less populated waiting queue).

- An estimation of the expected queuing delay to be encountered in the new cell. The length of the DTQ, which is the metric employed in the previous point, is not the only factor that determines the waiting time of a new node in the system. This delay also depends on the service time required by the nodes in the DTQ. This is a function of various parameters such as the length of the data messages and the transmission rate of each node. By selecting the AP with the lowest expected queuing delay a more efficient load balancing handoff mechanism can be achieved, at the cost of some additional control information that should be included in the FBP.

Association Process

Once the discovery process has been completed and a new AP has been selected, the node can proceed with the handoff process. In order to establish a connection with the new AP, the node must leave the previous AP (disassociation process) and initiate the authentication and the reassociation process with the new AP.

To disassociate from the current AP, the node has to reset its four DQCA queue counters (i.e., set $TQ = RQ = pTQ = pRQ = 0$). Since this reset takes place locally at the node, the counters of the AP and the other nodes are not updated and this may slightly affect the protocol operation. Roughly, there are three possible situations, depending on the state of the node before the reset:

- The node does not have a pending message for transmission within the DQCA queuing system (i.e., its counters hold a zero value). In this case, the reset will have absolutely no effect on the system.
- The node has a message in the CRQ (i.e, $pRQ > 0$). It should be reminded that each CRQ position is occupied by the nodes involved in a previous ARS collision. Hence, if the node resets its counters and abandons its CRQ position, it actually facilitates the collision resolution of the remaining nodes that share this position.
- The node has a message in the DTQ (i.e, $pTQ > 0$). In this case, if the node leaves the system, an empty DTQ position will be created that will eventually lead to an empty data slot. The AP can resolve this situation by setting the final-message-bit of the FBP to '1' to indicate the end of the message.[8]

After the disassociation, the node must connect to the new AP through the authentication and the reassociation process. The main steps followed in the IEEE 802.11 specification [1] are adopted. During the authentication, the node must establish its identity in order to be allowed access to the network. Then, during the reassociation phase, the node and the AP exchange information on their capabilities, such as the supported transmission rates, etc.

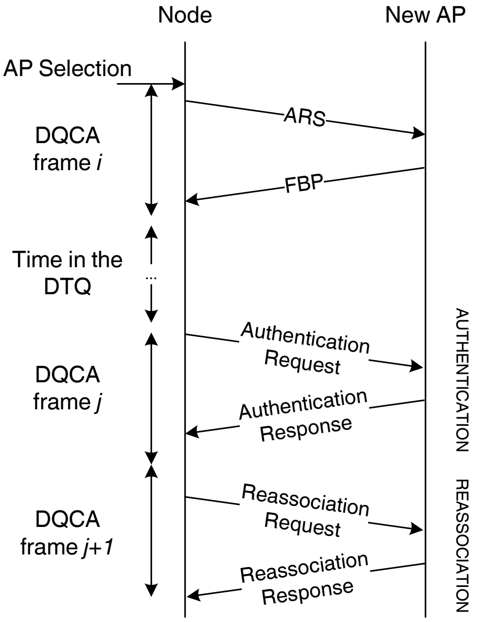


Figure 3.17: Authentication and reassociation process

In order to initiate the exchange of the authentication and the reassociation frames, the node follows the same DQCA access rules employed for the transmission

[8] This policy has been also employed in Section 3.5.2 to resolve errors in the data slot of the DQCA frame.

of a data packet (Section 3.4.1). In other words, the node transmits an ARS, waits for the FBP, resolves any collisions if necessary and eventually enters in the DTQ. The main difference is that when the node reaches the head of the DTQ, instead of a data packet, it transmits the authentication request. The AP includes the authentication reply in the FBP and the same process is repeated for the association frames. This frame exchange is schematically depicted in Figure 3.17.

3.6 Conclusions

DQCA is a novel distributed MAC protocol that serves as an a alternative solution to the widespread DCF defined in the IEEE 802.11 specification. The novelty of DQCA lies in the management of the resolution of collisions among channel request and the transmission of data as two parallel processes, handled by a pair of distributed logical queues. Several advantages are gained by this design, summarized as follows:

- The backoff window procedure which is a necessary part of CSMA/CA-based protocols and can cause long idle periods is no longer required. In DQCA, users are given the opportunity to compete for channel access in a limited section of every frame, whereas the biggest part of the frame is efficiently devoted to data transmission.
- Instead of the typical RTS frames, DQCA employs a very short chip sequence for the channel access request process. This patented technology requires some minor modifications at the physical layer but reduces significantly the control overhead of the MAC protocol. Nevertheless, even if a legacy PHY layer was employed, thus introducing PHY headers and preambles, the resulting frame would still be much smaller than the legacy RTS due to the fact that DQCA does not require any MAC layer information (e.g., the destination address) during the access request process.
- The maximum throughput can be achieved with a small number of control minislots, even for a large number of competing users. This is mainly due to the fact that, in general, the data transmission is a slower process compared to the collision resolution, in the sense that the transmission of a message, depending on its size, may take place in more than one DQCA frame. So, as long as there are users scheduled for transmission in the data queue, a channel access collision will not directly affect the system throughput, even if it requires several consecutive contention windows for its resolution. Hence, in most scenarios, a small number of control minislots is sufficient, without excluding the possibility of a dynamic adaptation of the CW size by the AP.
- DQCA maintains a stable near-optimum (TDMA-like) throughput performance under saturation. As long as there are users waiting in the data queue (which, in long term, will be the case under saturation), the data slot of the frame will be occupied by transmissions, thus yielding the maximum throughput performance that can be achieved by DQCA.

- The transmission of data packets is practically collision-free when there are no channel errors. Collisions only take place on rare occasions when the system is empty and DQCA enters in an Aloha-like transmission mode. In any case, these collisions do not affect the system performance since they are very rapidly resolved and only occur under very low traffic conditions.

- Despite the unavoidable cost that channel errors have on the system performance, DQCA maintains stability and does not enter in deadlock situations. The impact of errors occurring in different parts of the DQCA frame may vary in severity, but in general the system recovers smoothly with small delays.

This chapter has presented a novel MAC protocol for WLANs that exploits efficiently the available network resources and allocates them fairly among the users. Advanced allocation strategies can be designed at the MAC layer if additional information is available through a CL dialogue between different layers of the protocol stack. The next chapter will discuss the incorporation of CL techniques into the DQCA mechanism and will propose several channel-aware scheduling algorithms that aim to improve performance and provide QoS guarantees for multimedia applications.

Chapter 4

DQCA with Link Adaptation

4.1 Introduction

The previous chapter has provided a detailed description of the DQCA protocol operation, with all the rules that handle channel access and data transmission. This chapter will focus on the incorporation of a link adaptation mechanism in DQCA to exploit the multirate capabilities of the PHY layer by adapting the transmission rate to the channel condition of each user.

In general, the set of transmission rates employed at the MAC is determined by the MCSs defined at the PHY. High-order MCSs offer higher rates but are more prone to channel errors, whereas low-order MCSs are less efficient but more robust. The link quality between two wireless nodes is usually measured in terms of the SNR that expresses the strength of the received signal relative to the background noise. Link adaptation is the process of dynamically adjusting the MCS to match the time-varying channel condition in order to enhance spectral efficiency while adhering to a target error performance.

The inherent feedback mechanism of DQCA facilitates the practically seamless incorporation of link adaptation capabilities to the protocol operation. In addition, the proposed scheme will serve as a parting point for the design of channel-aware CL policies that include the rate capabilities and the channel conditions of the users in the scheduling decisions, which will be discussed in the following chapter.

The main body of this chapter is divided into three sections. Section 4.2 presents the link adaptation scheme and describes a feedback mechanism for the acquisition of SNR measurements. The most important contribution, given in Section 4.3, is a mathematical model for the calculation of the throughput and mean delay of DQCA with link adaptation, based on queuing theory tools. The evaluation of the DQCA performance is given in Section 4.4. The presented results have been obtained through both simulations, based on a custom made C++ simulation tool, and the application of the theoretical model. Finally, Section 4.5 is devoted to conclusions.

4.2 The Link Adaptation Mechanism

The frame structure of DQCA, described in detail in the previous chapter, facilitates the incorporation of a link adaptation scheme with few modifications to the protocol. The proposed scheme is implemented in the following, straightforward way. For each ARS that has been correctly received within a control minislot of the CW, the AP measures the corresponding signal strength and determines the maximum rate that can be employed for the transmission of data transmission on this particular link. The rate selection is achieved with the help of a lookup table that contains the minimum SNR level required by each available transmission rate defined at the underlying PHY layer, in order to ensure performance within a target error rate. The estimated rate is included in the FBP broadcasted by the AP at the end of the frame and will be employed by the node for the transmission of the data packet.

An example of this process is shown in Figure 4.1. In the first depicted DQCA frame (frame i), node n_1 transmits an ARS in the second minislot of the CW. Since the ARS reception is successful (i.e., no collision occurs), the AP is able to estimate the most appropriate transmission rate that can be supported by the link quality of n_1. The estimated rate value, denoted by r, is included in the FBP transmitted at the end of the frame. Hence, at the end of the frame, after executing the DQCA protocol rules, given in Section 3.4.1, n_1 enters the DTQ in the second position. Additionally, in the data slot of the same frame, n_2 transmits a data packet with final-message-bit set to '0', meaning that there are more packet to follow in the subsequent DQCA frames. Eventually, n_2 completes the transmission of its message at the $(j-1)$th frame and the exits the DTQ (the intermediate DQCA frames are not depicted in the figure, since they do not affect the link adaptation mechanism and due to space limitations). Finally, in the jth frame, n_1 advances to the head of the DTQ and employs the estimated rate r for the transmission of its data packet.

The proposed link adaptation scheme considers the following assumptions:

- The AP is able to measure the SNR and estimate the quality of the link from a correctly received ARS. In a more conservative approach, the ARS frame format could be extended to include a PHY layer preamble with training symbols. The exact length of the training field would depend on the PHY layer specification, but it would most probably be of the order of a few μs.[1]

- It has been assumed that the SNR value estimated upon the transmission of an ARS by a given node is still representative of the link quality by the time the node initiates the data transmission. The validity of this assumption depends mainly on two factors: the specific characteristics of the time-varying channel and the amount of time that a user has to wait in the DQCA data queue (i.e., the DTQ), which is equivalent to the elapsed time between the SNR measurement and the beginning of the data transmission. Without excluding the possibility of occasional errors, the proposed link adaptation scheme is

[1] For example, the training fields employed in the IEEE 802.11g OFDM PHY layer specification have a duration of 16 μs.

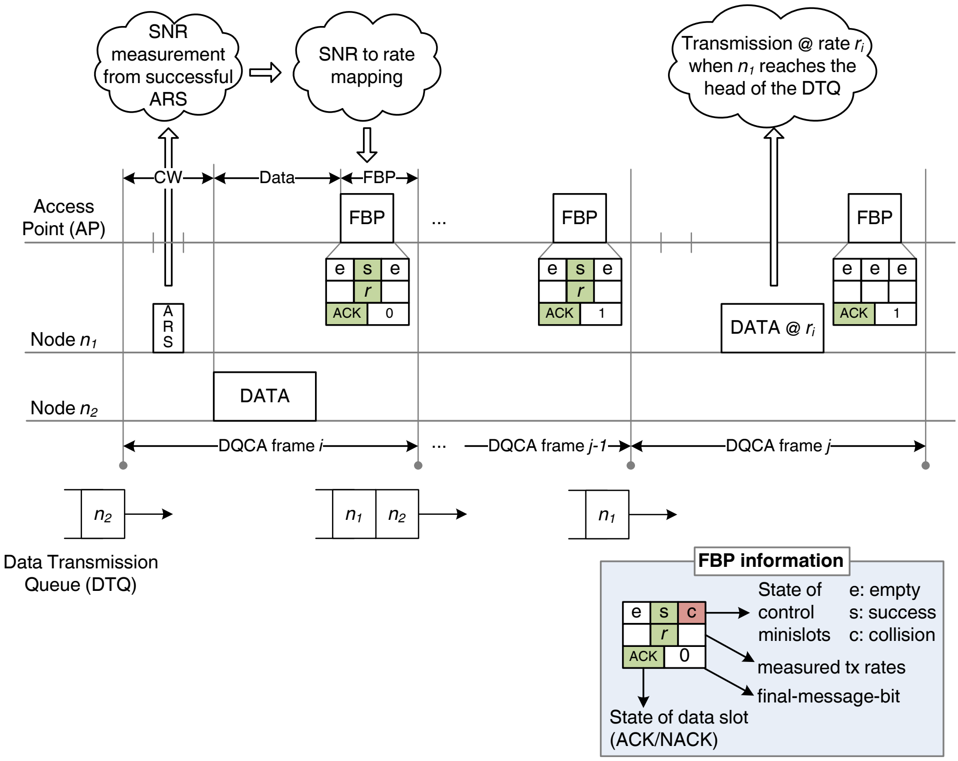


Figure 4.1: DQCA with link adaptation

suitable for scenarios in which the DQCA queuing delays do not exceed the coherence time τ_c of the wireless channel, defined as the time during which the channel condition does not experience significant variations. This is a valid assumption provided that indoor quasi-static scenarios are considered and that the DQCA protocol is not saturated, meaning that the incoming traffic does not systematically exceed the DQCA capacity and therefore queuing delays are not very long. This issue will be further discussed in Section 5.4 of the following chapter, where a mechanism to acquire updated information on the state of the channel is proposed.

If for any reason the selected bit rate does not reflect the link quality at the time of transmission, the following outcomes are considered:

- If the employed rate is lower than the rate supported by the channel condition, the packet is successfully received. If the node has more packets of the same message to transmit in subsequent DQCA frames, it can either employ the same sub-optimal rate (simplest case) or select a higher rate based on new estimations of the link quality obtained from the exchange of the previous data packet and the FBP.

- If the employed rate is higher than the rate supported by the channel, the packet is received with errors and is considered lost. In that case, the node selects the immediately lower rate of the available rate set and retransmits the packet in the next DQCA frame. This process is repeated until the packet transmission succeeds or the number of permitted retransmissions is reached.

For the implementation of the link adaptation scheme it is necessary to include an additional rate vector in the FBP, to indicate the estimated transmission rate for every successfully received ARS. Given a total of m control minislots in the CW, the additional overhead would include at most m transmission rate values. The number of bits required for the representation of a given rate value depends on the number of available rates defined in the PHY rate set. As a reference, the IEEE 802.11b rate set defines four rates (1, 2, 5.5 and 11 Mbps) that can be expressed by 2 bits, whereas 3 bits would be sufficient to represent the set of eight rates (6, 9, 12, 18, 24, 36, 48 and 54 Mbps) defined in the IEEE 802.11a/g specifications [1]. Thus, in a typical scenario where $m = 3$ minislots are employed in DQCA, the additional overhead in each FBP would be in the order of 6 to 9 bits. The structure of the modified FBP is shown in Figure 4.2.

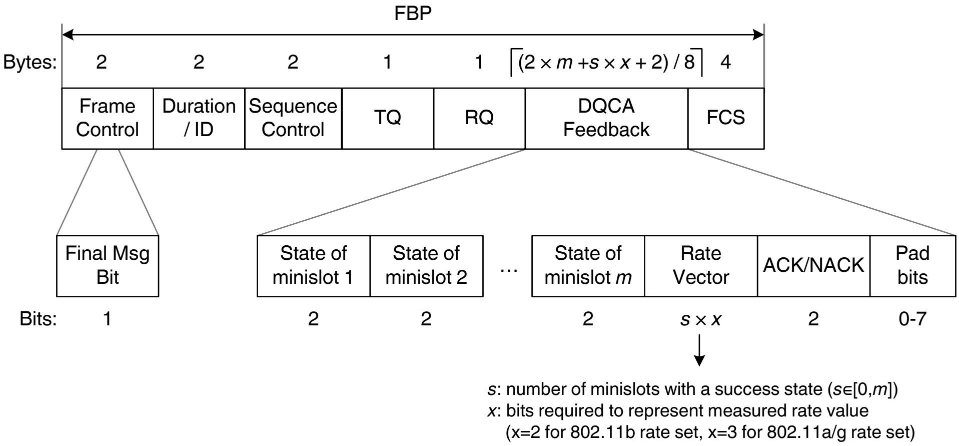


Figure 4.2: The modified FBP frame for DQCA with link adaptation

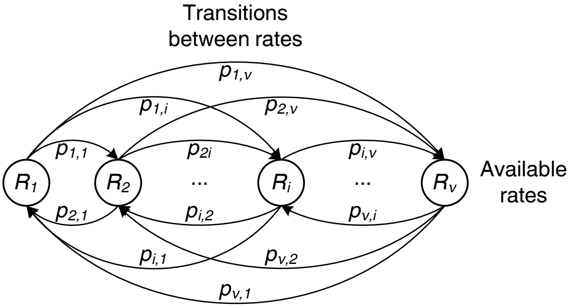
4.3 Analytical Model for the Performance Evaluation of DQCA with Link Adaptation

This section will present the proposed analytical model for the calculation of the throughput and mean delay performance of DQCA with link adaptation. The first two parts discuss the main assumptions required for the mathematical analysis and the queuing model employed to describe the DQCA operation. The throughput and delay analysis will, then, follow in the last two parts of the section.

4.3.1 Modeling Considerations

A Poisson distributed message (MSDU) arrival process has been considered for the traffic generation, with a known mean rate of λ messages per second. Each message has an exponentially distributed length with a known mean value of κ packets (MPDUs) per message.[2] As a result, the message transmission time can be approximated by the exponential distribution.

Depending on their size, messages can be fragmented into packets, each one of them with fixed length of L bytes. Exactly one packet can be transmitted within the data slot of each DQCA frame and all the packets of a message are transmitted at the same rate, estimated through the link adaptation mechanism. Each user manages the transmission of a single message, consisting of one or multiple packets, at a time, transmitted in the data slots of consecutive DQCA frames. The user exits the DQCA system once the message transmission is completed, but can reenter in subsequent DQCA frames if a new message arrives.



(a) The Markov chain

$$
\begin{array}{cc}
 & \overbrace{\hphantom{R_1 \quad R_2 \quad \cdots \quad R_\nu}}^{\text{future state}} \\
 & \begin{array}{cccc} R_1 & R_2 & \cdots & R_\nu \end{array} \\
T = & \left[\begin{array}{cccc} p_{1,1} & p_{1,2} & \cdots & p_{1,\nu} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,\nu} \\ \vdots & \vdots & \ddots & \vdots \\ p_{\nu,1} & p_{\nu,2} & \cdots & p_{\nu,\nu} \end{array}\right]
\begin{array}{c} R_1 \\ R_2 \\ \vdots \\ R_\nu \end{array}
\left.\vphantom{\begin{array}{c} R_1 \\ R_2 \\ \vdots \\ R_\nu \end{array}}\right\} \text{current state}
\end{array}
$$

(b) The transition matrix

Figure 4.3: The channel model

A multi-rate PHY layer has been considered with a set of ν available transmission rates $R_i \in [R_1, R_\nu]$, sorted from lowest to highest. Link adaptation applies only to data packets, whereas the FBP is sent with the lowest rate R_1. The link condition of the wireless nodes has been modeled in the following way. It has been assumed that

[2] The notation explained in Sections 2.3.1 and 3.2.1 has been adopted.

although wireless channels are characterized by fast-fading, some correlation is very likely to exist between the current and the immediately next state of the link for a given user (slow varying channel). In addition, the channel is characterized by a coherence time τ_c during which the time-domain signal may be considered correlated on average. In other words, it can be assumed that the link state of a given user does not experiment significant variations during an average interval equal to the channel τ_c. Finally, a coherence time τ_c longer than the average frame duration has been considered, to ensure that the link condition remains unchanged from the moment of the SNR measurement until the beginning of the data transmission.

With these considerations in mind, the wireless channel conditions of each user have been modeled as a ν-state Markov chain (Figure 4.3 (a)) where each state corresponds to a particular transmission rate. The Markov chain is represented by a transition matrix T that contains the set of probabilities $p_{i,j}$ with which a user with a current available rate R_i will select a future rate R_j once the coherence time τ_c has elapsed (Figure 4.3 (b)). It has also been assumed that the steady-state probability π_i of a user being at the ith state exists and is known for all states. Each channel state corresponds to a transmission rate and therefore it can be said that π_i is the probability of selecting R_i for transmission.

4.3.2 System Model

The DQCA system can be divided into the collision resolution (CR) subsystem and the data transmission (DT) subsystem, as shown in Figure 4.4. The Enable Transmission Interval (ETI) is the time from the arrival of a new message at a node until the beginning of the next DQCA frame. The node, then, enters the CRQ and is served (i.e. sends an ARS) when it reaches the head of the queue. After that, the node either exits the CR subsystem with probability $P(\lambda)$, which is the probability of selecting an empty minislot, or reenters in the CRQ with probability $1 - P(\lambda)$, which is the probability of ARS collision. Eventually, the node enters the DT subsystem, waits in the DTQ and is served by the ith server (i.e., employs rate R_i for the data transmission) with probability π_i. Since a single rate is used for the transmission of a data packet, only one of the ν servers may be active at a time.

The analysis of the CR subsystem is not trivial but it will be shown that it can be approximated by an $M/M/1$ queuing system [61, 69]. Based on this assumption it can be claimed that the output process of the CR subsystem (and subsequently the input of the DT subsystem) has the same statistics as the input process, i.e., it is Poisson distributed with mean rate λ. As far as the DT subsystem is concerned, the service rate of each of the ν servers is exponentially distributed with a different mean μ_{TQi} (with $1/\mu_{TQi}$ being the mean message service time). The total service time of the subsystem is a random variable denoted by x that follows an Hyperexponential distribution with ν stages and therefore the subsystem can be modeled as a $M/H_\nu/1$ [70].

Finally, it should be noted that the proposed model can also be employed for the throughput and delay analysis of the DQCA protocol without link adaptation,

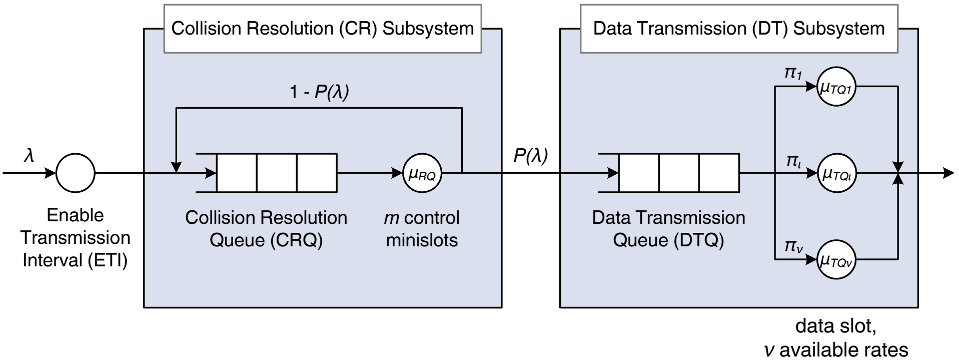

Collision Resolution (CR) Subsystem
Data Transmission (DT) Subsystem
1 - P(λ)
λ
P(λ)
μRQ
Enable Transmission Interval (ETI)
Collision Resolution Queue (CRQ)
m control minislots
Data Transmission Queue (DTQ)
π1
πi
πv
μTQ1
μTQi
μTQv
data slot, v available rates

Figure 4.4: DQCA with link adaptation system model

for the particular case of $\nu = 1$, when, in other words, there is only one transmission rate available at the PHY layer. A summary of the main parameters employed in the DQCA model analysis is given in Table 4.1.

4.3.3 Throughput Analysis

Throughput S is defined as the number of transmitted information data bits per second, or equally, as the average data transmission rate of the system. By accepting that the same amount of traffic flows into and out of the CR subsystem, meaning that all collisions are resolved sooner or later, it can be deduced that the average throughput is only affected by the DT subsystem and can be expressed as

$$S = \sum_{i=1}^{\nu} \gamma_i \rho_i R_i \quad \text{(bps)} \tag{4.1}$$

where γ_i is the relative MAC layer throughput achieved when rate R_i is employed for data transmission while ρ_i is the percentage of time during which rate R_i is used. These terms will be described in more detail next.

The relative throughput γ_i expresses the portion of the DQCA frame that is dedicated to the transmission of useful data bits, when rate R_i is employed. Ideally, a relative throughput of 1 could be achieved if no control information were included in the DQCA frame. In practice, however, control information is unavoidable and the relative throughput is less than one. For a given rate R_i, γ_i can be calculated as the ratio of the time required for the transmission of a data packet $T_{data,i}$ to the

Table 4.1: Main parameters for the DQCA model analysis

Symbol	*Description*
γ_i	Relative MAC throughput for rate R_i
κ	Mean message(MSDU) size (Exponential) (packets/msg)
λ	Mean message arrival rate (Poisson) (msg/s)
μ_{RQ}	Mean message service rate of the CR subsystem (msg/s)
μ_{TQi}	Mean message service rate for rate R_i (msg/s)
π_i	Steady-state probability of rate R_i being available to a user
ρ_i	Percentage of time that R_i is used
ρ	System utilization factor
L	Packet (MPDU) size (bytes)
$P(\lambda)$	Probability of successful ARS
R_i	Data transmission rate (bps), $R_i \in [R_1, R_\nu]$
S	Throughput (bps)
$T_{data,i}$	Transmission time of a data packet at rate R_i (s)
$T_{overhead}$	Transmission time of PHY and MAC overhead (s)
$T_{frame,i}$	DQCA frame time at rate R_i ($T_{data,i} + T_{overhead}$) (s)
$\overline{T}_f$	Mean DQCA frame duration (s)
t_{ETI}	Enable Transmission Interval (ETI) (s)
t_{CR}	Message time in the Collision Resolution subsystem (s)
t_{DT}	Message time in the Data Transmission subsystem (s)
t_c	Message delay caused by data collision (s)
t_T	DQCA message delay (s)
w	Waiting (queuing) time of a message in the DT (s)
x	Service time of a message in the DT (s)

total DQCA frame duration $T_{frame,i}$

$$\gamma_i = \frac{T_{data,i}}{T_{frame,i}} = \frac{T_{data,i}}{T_{data,i} + T_{overhead}} = \frac{\dfrac{8L}{R_i}}{\dfrac{8L}{R_i} + T_{overhead}} = \frac{1}{1 + \left(\dfrac{R_i}{8L} T_{overhead}\right)}. \tag{4.2}$$

$T_{overhead}$ is a known parameter that includes the duration of the CW that consists of m control minislots of fixed length, the PHY and MAC layer headers added to the data packet, the transmission time of the FBP at the lowest transmission rate R_1 and two SIFS (according to the DQCA transmission sequence described in

Section 3.2.1. The relative throughput γ_i also expresses the maximum normalized throughput [3] achieved for rate R_i under saturation traffic conditions, when the data slot of every DQCA frame is occupied by a packet transmission.

In non-saturation regime, depending on the incoming traffic, some data slots remain empty. The term ρ_i is defined to express the percentage of time during which data transmissions take place at rate R_i. With respect to the system model depicted in Figure 4.4, ρ_i is equivalent to the percentage of time that the ith server is busy (utilization of server i) and can be derived from the analysis of the $M/H_\nu/1$ system [70]

$$\rho_i = \lambda \frac{\pi_i}{\mu_{TQi}} \tag{4.3}$$

where λ is the mean message arrival rate (in messages per second), π_i the steady-state probability of using rate R_i and μ_{TQi} is the mean message service rate for rate R_i (also in messages per second). The probabilities π_i depend on the channel and are considered known. Given that each message consists of κ packets on average, according to the considered traffic distribution, and each packet is served within exactly one DQCA frame the term μ_{TQi} can be calculated as

$$\mu_{TQi} = \frac{\mu_{TQi,packet}}{\kappa} = \frac{1}{\kappa}\frac{1}{T_{frame,i}} = \frac{1}{\kappa\left(\dfrac{8L}{R_i} + T_{overhead}\right)}. \tag{4.4}$$

By making use of (4.2), (4.3) and (4.4) in (4.1), throughput can be calculated as a function of λ:

$$S = \sum_{i=1}^{\nu} \frac{\lambda\kappa\left(\dfrac{8L}{R_i} + T_{overhead}\right)\pi_i R_i}{1 + \left(\dfrac{R_i}{8L}T_{overhead}\right)} \quad \text{(bps)}. \tag{4.5}$$

Finally, the system utilization factor ρ is defined as the percentage of time during which the DT subsystem is busy (i.e., packets are being transmitted within the data slot of the DQCA frame). For the $M/H_\nu/1$ system ρ is given by

$$\rho = \sum_{i=1}^{\nu} \rho_i = \lambda \sum_{i=1}^{\nu} \frac{\pi_i}{\mu_{TQi}} = \lambda E[x]. \tag{4.6}$$

The above equation can also be used to calculate the mean message service time $E[x]$.

The system is stable when the system utilization factor ρ satisfies the condition $\rho \leq 1$, which holds when the system is not saturated (i.e., the message arrival rate does not exceed the message service rate). The maximum incoming traffic rate λ_{max}

[3] To obtain the throughput in bps, the normalized value γ_i should be multiplied by the respective rate R_i.

for which the system remains stable can be calculated from (4.6) for $\rho = 1$

$$\lambda_{max} = \frac{1}{\sum_{i=1}^{\nu} \frac{\pi_i}{\mu_{TQi}}}. \tag{4.7}$$

The maximum stable throughput S_{max} can hence be calculated as

$$S_{max} = \sum_{i=1}^{\nu} \gamma_i \lambda_{max} \frac{\pi_i}{\mu_{TQi}} R_i \quad \text{(bps)}. \tag{4.8}$$

4.3.4 Mean Delay Analysis

The mean delay $E[t_T]$ of the system can be expressed as the sum of four components

$$E[t_T] = E[t_{ETI}] + E[t_{CR}] + E[t_{DT}] + E[t_c] \tag{4.9}$$

where $E[t_{ETI}]$ corresponds to the mean ETI, $E[t_{CR}]$ is the mean delay in the CR subsystem , $E[t_{DT}]$ the mean time spent in the DT subsystem and $E[t_c]$ the mean delay induced by collisions of data packets. The proposed delay analysis follows the framework provided in [61] and [69]. Although several modifications with respect to the previous work have been required, the main innovation lies in the calculation of $E[t_{DT}]$, which is directly affected by the link adaptation scheme.

Since a message can arrive at any time within a frame, the t_{ETI} is a uniformly distributed random variable in the interval $(0, \overline{T}_f)$ where $\overline{T}_f$ is the mean frame duration. Hence

$$E[t_{ETI}] = \frac{1}{2}\overline{T}_f = \frac{1}{2}\sum_{i=1}^{\nu} \pi_i T_{frame,i} \tag{4.10}$$

with $T_{frame,i}$ the frame duration when rate R_i is used, given by (4.2).

The service time of the CR subsystem follows a geometrical distribution with parameter $P(\lambda)$, since this is the probability that a node will exit the CR subsystem after transmitting a successful ARS. Therefore, the CR subsystem is an $M/G/1$ queue and its analysis is fairly complicated. However, if the continuous time equivalent is considered, as indicated in [61] and [69], the service time could be approximated by the exponential distribution normalized by the mean duration of the DQCA frame $\overline{T}_f$, with mean rate

$$\mu_{RQ} = \ln\left(\frac{1}{1 - P(\lambda)}\right)\frac{1}{\overline{T}_f}. \tag{4.11}$$

With this approximation the model becomes an $M/M/1$ and the total delay $E[t_{RQ}]$, which includes the waiting time in the CRQ and the service time, is given by

$$E[t_{RQ}] = \frac{1}{\mu_{RQ} - \lambda} = \frac{1}{\ln\left(\dfrac{1}{1-P(\lambda)}\right)\dfrac{1}{\overline{T}_f} - \lambda}. \tag{4.12}$$

Finally, for a system with m minislots, $P(\lambda)$ can be computed as the probability that a node selects a free minislot given that there are n nodes contending for access in the same frame, multiplied by the probability of actually having the n arrivals in that frame

$$\begin{aligned}
P(\lambda) &= \sum_{n=0}^{\infty} P\left(\text{free minislot}|k=n\right) P\left(k=n\right)\\
&= P(k=0) + \sum_{n=1}^{\infty}\left[m\left(\frac{1}{m}\right)\left(1-\frac{1}{m}\right)^n P\left(k=n\right)\right]\\
&= e^{-\lambda\,\overline{T}_f} + \sum_{n=1}^{\infty}\left[\left(1-\frac{1}{m}\right)^n e^{-\lambda\,\overline{T}_f}\frac{\left[\lambda\overline{T}_f\right]^n}{n!}\right] \Rightarrow\\
P(\lambda) &= e^{-(\lambda/m)\,\overline{T}_f}.
\end{aligned} \tag{4.13}$$

Once the CR subsystem has been approximated by an $M/M/1$ queue, it is known, according to Burke's theorem, that the output flow has the same statistics as the input flow. Hence, the input traffic of the DT subsystem can be considered Poisson distributed with a mean rate of λ. Thus, the DT subsystem can be modeled as an $M/H_\nu/1$ queue with Poisson arrivals and hyperexponential service time. The mean waiting time $E[w]$ of a message in the DTQ is [70]

$$E[w] = \frac{\lambda}{(1-\rho)}\sum_{i=1}^{\nu}\frac{\pi_i}{\mu_{TQi}^2} \tag{4.14}$$

where ρ is given by (4.6). The mean service time $E[x]$ can also be deduced from (4.6). Hence the mean total message delay $E[t_{DT}]$ in the DT subsystem is

$$E[t_{DT}] = E[w] + E[x] = \frac{\lambda}{(1-\rho)}\sum_{i=1}^{\nu}\frac{\pi_i}{\mu_{TQi}^2} + \sum_{i=1}^{\nu}\frac{\pi_i}{\mu_{TQi}}. \tag{4.15}$$

Finally, data collisions may only occur when a new message finds the system empty, according to the protocol rules described in [61]. The mean delay $E[t_c]$ caused by this event is equal to the probability of data collision multiplied by the mean frame duration, since if a data collision occurs, the message will enter any of the two subsystems of the model (depending on whether its ARS had succeeded or collided) in the next frame. Hence

$$\begin{aligned} E[t_c] =& P\,(\text{system is empty}) \\ & P\,(\text{more than 1 msg arrive}|\text{system is empty})\,\overline{T}_f \\ =& \rho_0 \sum_{n=2}^{\infty} \left[e^{-\lambda \overline{T}_f} \frac{\left[\lambda\,\overline{T}_f\right]^n}{n!} \right] \overline{T}_f \Rightarrow \\ E[t_c] =& \rho_0 \left[1 - e^{-\lambda\,\overline{T}_f} \left(1 + \lambda \overline{T}_f\right)\right] \overline{T}_f \end{aligned} \tag{4.16}$$

where ρ_0 is the probability that the system is empty and is equal to

$$\rho_0 = 1 - \rho, \quad \rho \geq 1 \tag{4.17}$$

since ρ, given by (4.6), is the time during which the system is busy.

This section has presented a mathematical model calculation of the throughput and mean delay performance of DQCA with link adaptation. In continuation, the model will be employed for the performance evaluation of DQCA and its validity will be demonstrated with the help of simulations (especially in Section 4.4.4 where the non-saturation case is examined).

4.4 DQCA Performance Evaluation

This section is dedicated to the performance evaluation of DQCA and the discussion of the obtained results. The mathematical analysis presented previously has been employed to calculate the benchmark DQCA performance. The numerical evaluation of the theoretical DQCA performance has been based on MATLAB$^{\text{TM}}$. For the validation of the theoretical model, simulations have also been conducted by employing a custom-made simulation software based on the C++ programming language. The decision to develop a custom simulation tool was based on the need for flexibility and total control over the simulation environment, the possibility to incorporate traffic and channel models and the reutilization of existing related work by our research group. The object-oriented software tool executed a frame-by-frame simulation of the DQCA operation. In each simulation iteration, every user updated its traffic and channel conditions according to the considered models and executed the DQCA protocol rules to update its counters and determine its course of action (whether to transmit an ARS or a data frame, or whether to remain idle).

Section 4.4.1 provides information on the simulation setup that includes the employed parameters of the PHY and the MAC layers, the wireless channel model, the traffic generation and the evaluated metrics. The performance evaluation follows next, divided in three study cases:

- Section 4.4.2 compares the DQCA and IEEE 802.11 DCF performance under saturated traffic conditions for a single-rate channel. This scenario aims to evaluate the efficiency of the DQCA protocol for a given transmission rate.

- Section 4.4.3 compares the DQCA and IEEE 802.11 DCF performance under saturated traffic conditions for a multi-rate channel. In this scenario, the link adaptation mechanism is employed. The selection of some basic design parameters of DQCA, such as the number of control minislots and the ARS duration, is also discussed.

- Section 4.4.4 continues with the performance evaluation of DQCA in the non-saturation regime. Results obtained by the use mathematical model for the throughput and delay performance of DQCA with link adaptation are contrasted with simulations to demonstrate the validity of the model.

4.4.1 Simulation Setup

This section discusses the most relevant aspects of the simulation setup which are essential for the comprehension of the obtained results. First the underlying PHY layer specifications that are based on the IEEE 802.11 standard amendments b and g are given. Then, the MAC layer parameters for both DQCA and the IEEE 802.11 DCF are provided. The channel model is discussed next, based on a simple but effective transition matrix that represents a discrete Markov chain model. In the last part of this section, the traffic generation model and the employed performance metrics are described.

The Physical layer

The DQCA protocol is a flexible MAC scheme that can be applied over different PHY layer protocol stacks. Without losing generality, the underlying PHY layer for the presented results in this chapter is based on the widely deployed IEEE 802.11 standard. Two particular PHY layer specifications have been considered:

- The High Rate/Direct Sequence Spread Spectrum Physical Layer (HR/DSSS PHY) defined in the IEEE 802.11b specification. This standard supports four rates up to 11 Mbps and has been adopted in the initial stages of this work.

- The Extended Rate PHY (ERP-OFDM) defined in the IEEE 802.11g standard, supporting a rate set of eight rates up to 54 Mbps. The majority of the presented results is based on this specification, which is the most widely deployed WLAN protocol up to date, with the emerging IEEE 802.11n gradually taking its place.

Table 4.2 summarizes the PHY layer simulation parameters. Note that only the first three parameters are relevant to the DQCA protocol whereas the backoff slot time and the CW size are only meaningful for the IEEE 802.11 simulation.

Table 4.2: Summary of PHY layer simulation parameters

Parameters applicable to DQCA and IEEE 802.11 DCF operation		
Parameter	*IEEE 802.11b PHY*	*IEEE 802.11g PHY*
Rate set	1, 2, 5.5 and 11 Mbps	6, 9, 12, 18, 24, 36, 48 and 54 Mbps
PLCP preamble and PHY header	96 μs	20 μs
aSifsTime	10 μs	10 μs
Parameters applicable only to IEEE 802.11 DCF operation		
Parameter	*IEEE 802.11b PHY*	*IEEE 802.11g PHY*
aSlotTime	20 μs	9 μs
aCWmin	31	15
aCWmax	1023	1023

The MAC layer

The MAC layer parameters employed in the simulations of the IEEE 802.11 DCF and DQCA protocols are summarized in Table 4.3.

Table 4.3: Summary of MAC layer simulation parameters

IEEE 802.11 b/g		DQCA	
Parameter	*Value*	*Parameter*	*Value*
RTS	20 bytes	**ARS**	10 μs
CTS	14 bytes	**minislots** m	3
ACK	14 bytes	**FBP**	13 bytes
MAC Header	34 bytes	**MAC Header**	34 bytes

The Channel Model

The PHY layer specification determines the set of rates that are available for transmission. However, the actual rate to be employed by each user depends on the condition of the wireless link between the transmitter and the receiver at the time of data transmission. Higher rates can be supported under a good channel condition, usually measured in terms of the SNR (or SNIR, if interference is taken into account), whereas lower rates are required to reduce transmission errors when the channel deteriorates. Simulating a detailed wireless channel model that encompasses realistic propagation loss and user mobility models can be a challenging task, requiring increased computational power and simulation time.

The decision on whether such complexity is necessary in order to obtain meaningful results and conclusions depends on the considered scenario. For example, a realistic channel representation may be decisive in the evaluation of beamforming techniques or in the application of resource allocation algorithms in multi-cell environments where handover and interference issues must be considered. On the other hand, for the MAC layer performance of DQCA in single-cell infrastructure WLAN networks a less complex channel model can be employed without compromising the validity of the obtained results.

Hence, the wireless channel has been modeled as a discrete Markov chain of R states, with R being the size of the available rate set (i.e., 4 for IEEE 802.11b and 8 for IEEE 802.11g PHY layer specifications) [71]. As described in Section 4.3.1, the Markov chain is represented by a transition matrix T that contains the set of probabilities with which a user with a given present rate will select any of the R available rates once the coherence time τ_c has elapsed.

The transition matrices are not unique; on the contrary, their entries can be selected in various ways in order to represent different simulation environments. The following matrix has been adopted for the representation of the IEEE 802.11b channel. The steady state probabilities of transmitting at each rate of the defined rate set $R_b = [1, 2, 5.5, 11]$ are $\pi_b = [0.1764, 0.2941, 0.2941, 0.2353]$, respectively. In other words, the majority of transmissions are likely to take place at the rates of 2 or 5.5 Mbps.

$$\begin{array}{cc} & \overbrace{\begin{array}{cccc} 1 & 2 & 5.5 & 11 \end{array}}^{\text{future state}} \\ T_b = & \left.\begin{bmatrix} 0.5 & 0.4 & 0.1 & 0 \\ 0.2 & 0.5 & 0.2 & 0.1 \\ 0.1 & 0.1 & 0.5 & 0.3 \\ 0 & 0.2 & 0.3 & 0.5 \end{bmatrix} \begin{array}{c} 1 \\ 2 \\ 5.5 \\ 11 \end{array} \right\} \text{current state} \end{array}$$

In a similar manner, the IEEE 802.11g channel has been modeled with the help of an $8x8$ transition matrix. In this case, the rate set consists of eight rates, namely $R_g = [6, 9, 12, 18, 24, 36, 48, 54]$, and the steady state probabilities of transmitting at each rate are $\pi_g = [0.0004, 0.0022, 0.0114, 0.0571, 0.2967, 0.3467, 0.2039, 0.0816]$, respectively, with 24 and 36 Mbps being the most likely rates to be employed.

$$\begin{array}{cl}
 & \overbrace{\begin{array}{cccccccc} 6 & 9 & 12 & 18 & 24 & 36 & 48 & 54 \end{array}}^{\text{future state}} \\
T_g = & \left.\begin{bmatrix}
0.4 & 0.5 & 0.1 & 0 & 0 & 0 & 0 & 0 \\
0.1 & 0.4 & 0.5 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.1 & 0.4 & 0.1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0.1 & 0.4 & 0.4 & 0.1 & 0 & 0 \\
0 & 0 & 0 & 0.1 & 0.5 & 0.4 & 0 & 0 \\
0 & 0 & 0 & 0 & 0.3 & 0.5 & 0.2 & 0 \\
0 & 0 & 0 & 0 & 0.1 & 0.2 & 0.5 & 0.2 \\
0 & 0 & 0 & 0 & 0 & 0.1 & 0.4 & 0.5
\end{bmatrix}
\begin{array}{c} 6 \\ 9 \\ 12 \\ 18 \\ 24 \\ 36 \\ 48 \\ 54 \end{array}
\right\} \text{current state}
\end{array}$$

A coherence time of $\tau_c = 150$ ms has been considered for all users. In the performed simulations, each user maintains a counter for the coherence time. Once this time elapses, the user calculates its new available transmission rate according to its previous channel condition and the transition probabilities given before. In general, the considered coherence time is higher than the average frame duration[4], so it has been assumed that the channel condition does not change drastically during the transmission of a frame. It has also been assumed that, regardless of the channel conditions, the appropriate rate selection guarantees that no errors are introduced during transmissions.

Most results presented in this chapter will employ these channel models that will be, hence, referred to as IEEE 802.11 b and g channel models, respectively. However, in some special cases, a static channel model will be considered according to which only a single, fixed rate will be available for transmission.

Traffic Generation Model

In this chapter, only Poisson generated data traffic has been considered in order to match the mathematical model assumptions described in Section 4.3.1. Hence, the data traffic sources have been assumed to follow a Poisson distributed generation process with an arrival rate of λ messages per second. Each message consists of an exponentially distributed number of packets, with κ packets per message on average. The packets have a fixed length of L bytes. The average offered load generated with this model is $C_{data} = \kappa\lambda L(8 \cdot 10^{-6})$ Mbps.

In the presented study cases, the average message size has been set to $\kappa = 10$, whereas packet sizes L of 100, 512, 1000, 1500 and 2312 bytes have been considered. Since data messages are relatively large, their transmission takes place in consecutive DQCA frames, with L being the maximum number of bytes transmitted per frame. A summary of these parameters is given in Table 4.4.

[4] For reference, the longest DQCA frame duration, considering a packet size of $L = 2312$ bytes

Table 4.4: Summary of the data traffic generation model

Parameter	*Value*
Message Arrival	Poisson with rate λ (msg/s)
Message Size	Exponentially distributed with mean $k = 10$ packets/msg
Packet Size	Fixed packet size, $L = [100, 512, 1000, 1500$ and $2312]$ bytes
Offered Load	$C_{data} = \kappa\lambda L(8 \cdot 10^{-6})$ Mbps

More traffic models for multimedia applications will be considered in the following chapter.

Definition of Performance Metrics

Finally, before proceeding to the simulation results, some brief definitions of the performance metrics will be given in this section.

- *Throughput* is defined as the rate of transmitted bits per second and is calculated as the ratio of the total number of successfully transmitted data bits to the duration of the simulation experiment (average throughput). Throughput is evaluated at the MAC layer, meaning that the data bits include the application payload and any headers added by higher layers, whereas the MAC and PHY layer headers are considered overhead. Unless otherwise stated, the throughput values presented in the next section refer to the total system throughput, calculated as the sum of the throughput performance of all system users.

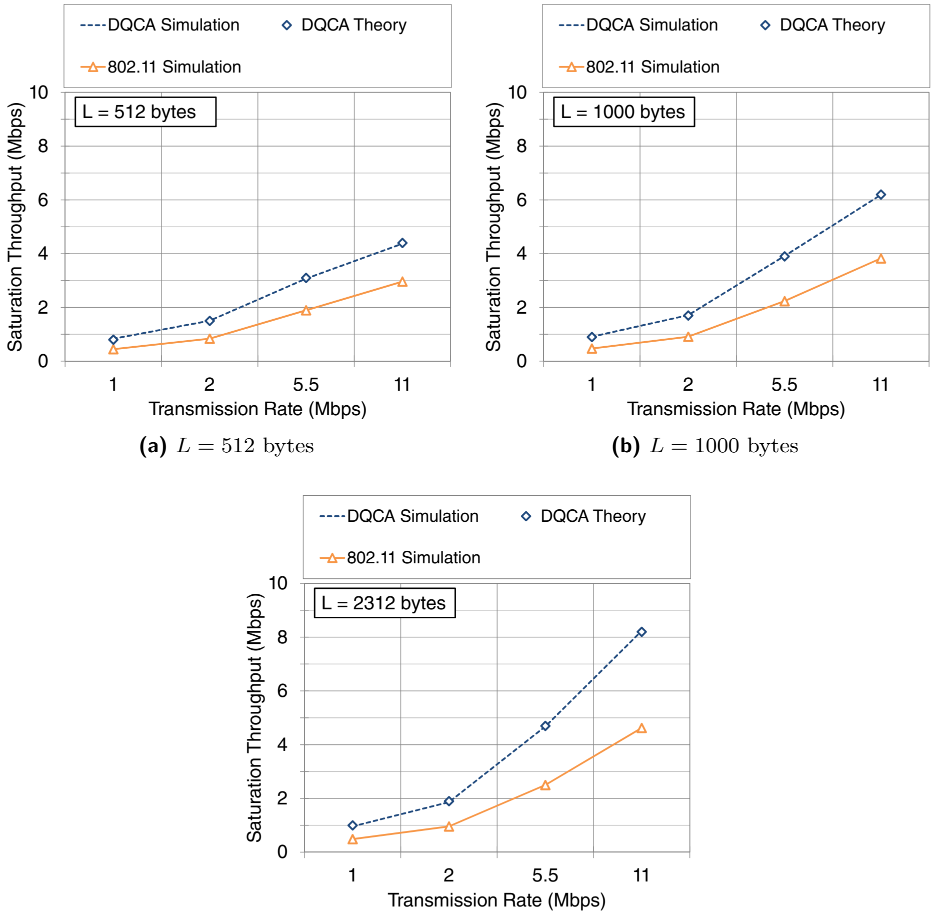
- *Mean message delay* is defined as the average time from the generation of a data message until its complete transmission. This time consists of the waiting time of the message in the MAC layer buffer (queuing time), the time required for the respective user to gain access to the medium (access time) and the actual transmission time of the message (including the transmission of ACK or FBP, in the case of IEEE 802.11 and DQCA, respectively). The standard deviation of the message delay is also considered, as a metric of the dispersion of the message delay experienced by users with respect to the average value.

4.4.2 Saturation Analysis for a Single-Rate Channel

The first results compare the saturation throughput performance of DQCA with respect to IEEE 802.11 DCF MAC with RTS/CTS, described in Section 2.3.2 (denoted for brevity as 802.11 in the figures). In this initial approach, a single-rate

and the lowest transmission rate of $R_1 = 1$ Mbps, is approximately 19 ms.

channel has been considered, meaning that for each simulation run all users share the same rate R, which in turn represents the channel capacity. In order to measure the maximum achievable throughput, saturated traffic has been considered, employing the traffic generation model parameters presented in Table 4.4. To bring the system under saturation, a high value for the Poisson message arrival rate λ has been selected to ensure that there are always messages waiting to be transmitted in the buffer of each user. The number of users has been set to $N = 20$.



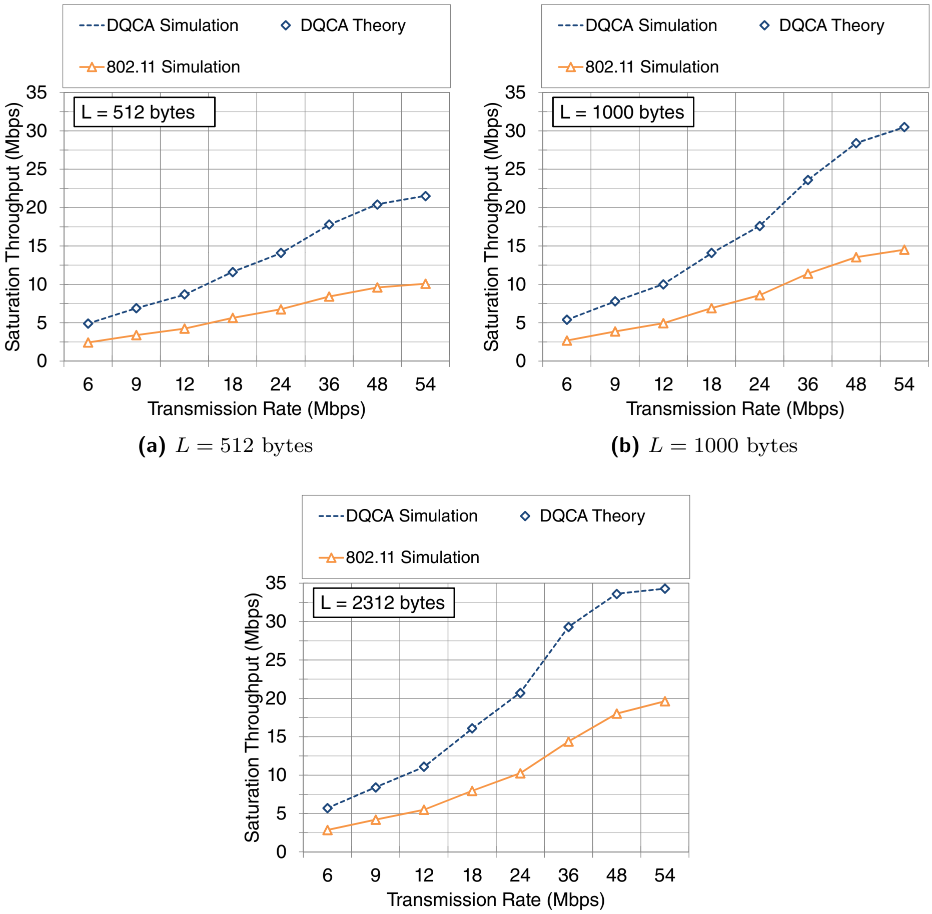
(a) $L = 512$ bytes

(b) $L = 1000$ bytes

(c) $L = 2312$ bytes

Figure 4.5: Saturation throughput comparison between DQCA (theory and simulations) and IEEE 802.11b DCF (four available transmission rates)

Figures 4.5 and 4.6 present the saturation throughput performance obtained for each transmission rate available at the IEEE 802.11b and g PHY layer specifications, respectively. The plotted values refer to the total system throughput, which is the aggregate throughput of all $N = 20$ users. Three different packet sizes have been considered, L=512, 1000 and 2312 bytes, producing the in plots (a) to (c), respectively.



(a) $L = 512$ bytes

(b) $L = 1000$ bytes

(c) $L = 2312$ bytes

Figure 4.6: Saturation throughput comparison between DQCA (theory and simulations) and IEEE 802.11g DCF (eight available transmission rates)

Each plot contains three set of results:

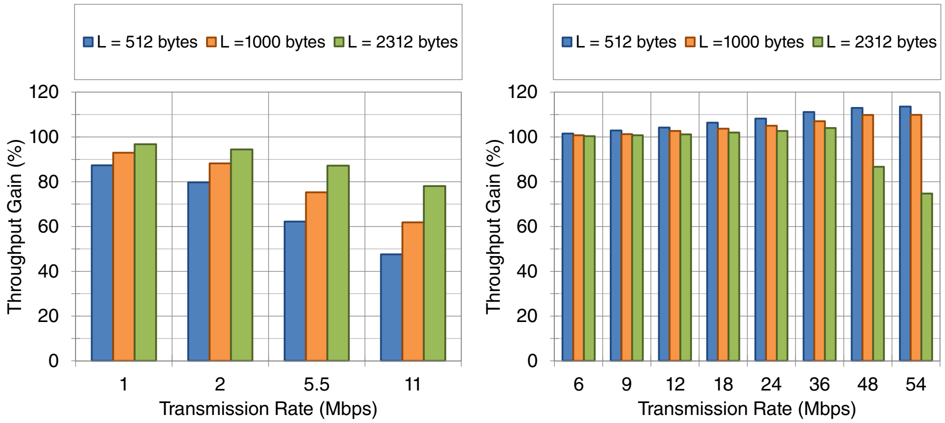
- the dashed line represents the saturation throughput of the DQCA performance obtained by simulations carried out with the custom made C++ simulation tool.
- the markers indicate the theoretical DQCA saturation throughput obtained through a MATLAB^{TM} implementation of the theoretical model presented in Section 4.3.3. The saturation throughput values have been obtained by considering the maximum message arrival rate λ_{max} for which the system

is stable (given by equation 4.7). It should be noted that, in this case, the analytical model has been evaluated for the specific case of $\nu = 1$ rate, since a single-rate channel has been considered.

- the solid line with the triangle markers represents the saturation throughput obtained by the IEEE 802.11 DCF with the RTS/CTS handshake.

First of all, a close match between the theoretical and the simulated values can be observed in all figures. Of course, these results apply to the specific case of all users having a single available transmission rate, so the link adaptation mechanism is not employed (results on multi-rate channels will be given in the next sections).

The most important observation on the presented results is the clear performance improvement achieved by DQCA with respect to the IEEE 802.11 DCF. The throughput gain obtained by DQCA has been plotted in Figure 4.7, with plots (a) and (b) corresponding to the IEEE 802.11 b and g PHY layers, respectively. For the IEEE 802.11b channel, DQCA provides a performance gain that varies from 47% to 96%, but is generally above 75% for most combinations of rate and packet size. The gain is increased in the case of IEEE 802.11g rates where the DQCA practically doubles the throughput with respect to the IEEE 802.11 DCF.



(a) IEEE 802.11b PHY rates

(b) IEEE 802.11g PHY rates

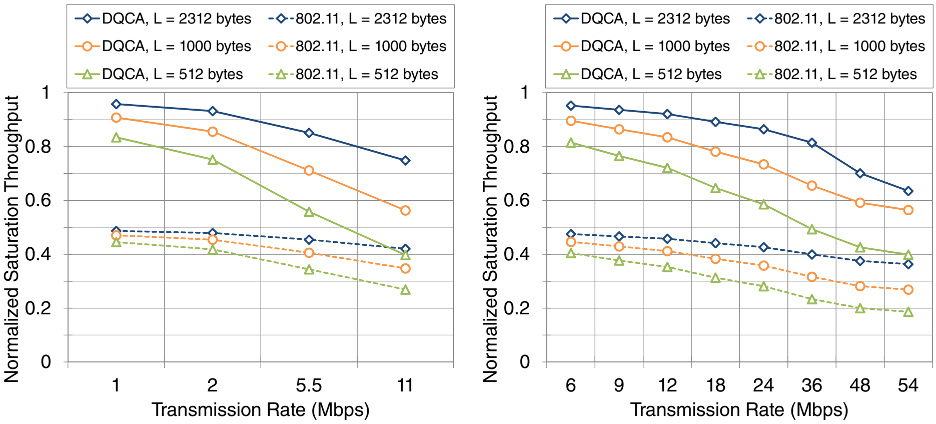
Figure 4.7: Throughput gain percentage of DQCA over the IEEE 802.11 b/g DCF

This improvement is the combined result of the following factors:

- the reduced DQCA overhead. The DQCA control overhead consists of the $m = 3$ control minislots for the transmission of the 10 μs ARS frames and the 13 byte FBP transmitted at the lowest transmission rate (which is either 1 Mbps or 6 Mbps, for the IEEE 802.11 b and g channel, respectively). On the other hand, the control information of the IEEE 802.11 DCF consists of the RTS/CTS handshake (20 and 14 bytes transmitted at the lowest transmission rate) and the ACK frame (14 bytes, transmitted at the data rate).

- the elimination of backoff period. As explained in Section 2.3.2, the DCF employs the binary exponential backoff mechanism to avoid and resolve access collisions among users. The initial backoff period depends on the minimum size of the backoff window aCW_{min} which is set to 31 slots for the IEEE 802.11b and 15 slots for the IEEE 802.11g specification (with the slot size equal to 20 μs and 9 μs for each case, respectively). In case of collision, the CW is doubled, until the maximum defined value (equal to 1023 slots) is reached. Under saturation traffic conditions and for $N = 20$ users, the occurrence of access collisions are high and a significant amount of time is bound to be consumed by the backoff mechanism. On the other hand, DQCA does not employ backoff since ARS collisions are confined within a short CW and are resolved according to a m-ary splitting algorithm.

- the parallel operation of the collision resolution and data transmission processes in DQCA. As a result, collision-free packet transmissions can take place in the data slots of each DQCA frame, even if ARS collisions take place and are being resolved in the CW.



(a) IEEE 802.11b PHY rates

(b) IEEE 802.11g PHY rates

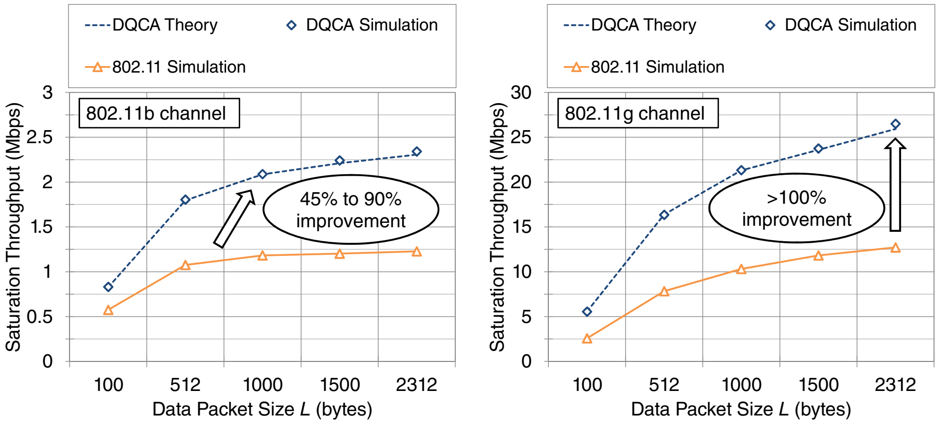
Figure 4.8: Throughput efficiency comparison of DQCA and IEEE 802.11 b/g DCF

Finally, the same results are presented in a different way in order to illustrate the MAC layer efficiency of the two MAC protocols. Figure 4.8 depicts the normalized saturation throughput achieved for each transmission rate. The normalized values are obtained by dividing the throughput by the channel capacity (i.e., the employed rate). Hence, a normalized throughput of 1 would correspond to the performance of an ideal MAC scheme with no control overhead, in which all resources are devoted to the transmission of useful data bits. The presented results show that DQCA is closer to the ideal performance with respect to the IEEE 802.11 DCF. Higher MAC efficiency is achieved when larger packet sizes are employed. In addition, MAC efficiency drops as the rate is increased. This occurs due to the fact that higher rates result to faster data transmissions and consequently the amount of time consumed by the PHY and MAC control overhead becomes more significant with respect to the time dedicated for the transmission of useful data. Efficiency could be increased if a higher rate were employed for the transmission of the control frames, but this would decrease the robustness of the system against channel errors.

4.4.3 Saturation Analysis with Link Adaptation

The next step has been to evaluate the saturation throughput performance of DQCA with link adaptation. To this end, the channel models described in Section 4.4.1 have been considered for the IEEE 802.11b/g PHY rate sets. As mentioned before, the available rate of each user is evaluated with the help of the transmission matrices, based on the previous link condition of the user, and is maintained during a period equal to the channel coherence time τ_c, set to 150 ms in the simulated scenario.

A scenario consisting of $N = 20$ users with saturated data traffic conditions has been initially assumed. Figure 4.9 compares the saturation system throughput of DQCA with link adaptation versus the IEEE 802.11 DCF MAC with RTS/CTS for the IEEE 802.11 b and g channel models (plots (a) and (b), respectively). Different data packet sizes L have been considered, varying from 100 to 2312 bytes. Simulated and theoretical values have been obtained for the performance of DQCA and show a close match. The theoretical saturation throughput has been calculated from equation (4.8) by employing the steady state probabilities given in Section 4.4.1 for the two channel models.



(a) IEEE 802.11b PHY rates

(b) IEEE 802.11g PHY rates

Figure 4.9: Saturation throughput for DQCA with link adaptation

According to the selected channel matrices, the employed transition matrices favor the rates of 24 and 36 Mbps for the 802.11g channel and the rates of 2 and 5.5 Mbps for the IEEE 802.11b channel. As the size of the data packet increases, it can be observed that the DQCA saturation throughput lies between the values of these rates. The improvement gained with respect to the IEEE 802.11 DCF mechanism is significant and, as in the first case study scenario, it exceeds 100% for the IEEE 802.11g channel and varies from 45 to 90% for the IEEE 802.11b channel.

Figure 4.10 plots the total saturation throughput versus the number of users N that varies from 1 to 80. In this case, the IEEE 802.11 DCF MAC has been evaluated through simulations with and without the RTS/CTS mechanism. The DQCA performance has also been obtained through simulations, however the throughput value obtained through the theoretical model has also been plotted as a reference (dotted line).

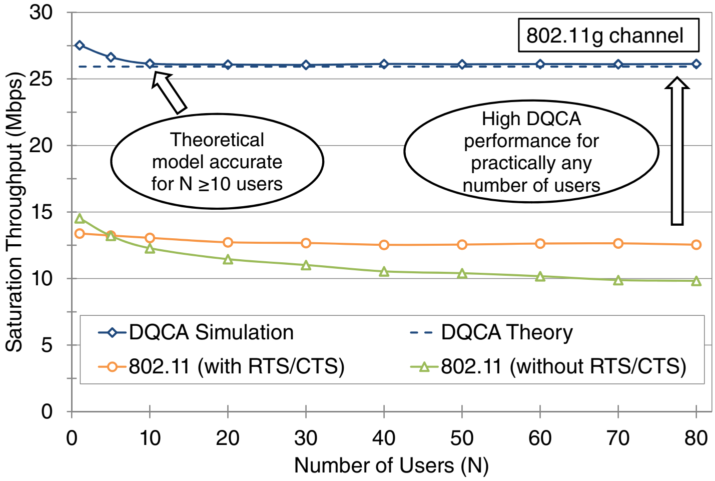


Figure 4.10: Saturation throughput comparison between DQCA and IEEE 802.11g DCF versus the number of users N

At this point, it should be noted that the throughput formulation of the DQCA analytical model, as given by equation (4.1), is not given as a function of the number of users. This is due to the fact that the queuing analysis of the model is based on a Poisson message arrival process, thus implying an infinite number of traffic sources (i.e., users). In practice, as shown in Figure 4.10, the theoretical model matches accurately the simulation results for $N \geq 10$ users.

Furthermore, Figure 4.10 shows that DQCA outperforms both IEEE 802.11 DCF schemes (with and without the RTS/CTS mechanism) for all values of N. The total system throughput is practically doubled when DQCA is used, with a gain up to 165% with respect to the IEEE 802.11 basic access (without the RTS/CTS) for $N = 80$ users. This improvement stems from the fact that in DQCA the contention among ARS frames is limited within the $m = 3$ minislots of the CW, whereas the remaining part of the frame is dedicated to the data transmission by users waiting in the DTQ. Hence, under high traffic conditions, DQCA practically operates as a TDMA scheme in which data packets are transmitted in a collision-free manner within the data slot of each DQCA frame. On the other hand, the IEEE 802.11 DCF suffers from long backoff periods since the probability of finding the channel busy when attempting to initiate transmission increases as the number of users grows. When the RTS/CTS mechanism is not employed, the DCF performance deteriorates further, since collisions among users involve data packets that require a much longer transmission time compared to the RTS/CTS frames or the DQCA ARS.

Figure 4.11 displays the saturation DQCA throughput versus the number of control minislots m that ranges from 2 to 10 and for $N = 10$, 20, 40 and 80 users. To facilitate the representation of the results, the throughput has been normalized with respect to the maximum value obtained for each user number N. These values are given in Table 4.5.

Table 4.5: Maximum saturation throughput used for normalization in Figure 4.10

	Number of Users			
	$N = 10$	$N = 20$	$N = 40$	$N = 80$
optimum minislots m	2	2	2	3
max. throughput (Mbps)	26.80	26.17	26.18	26.15

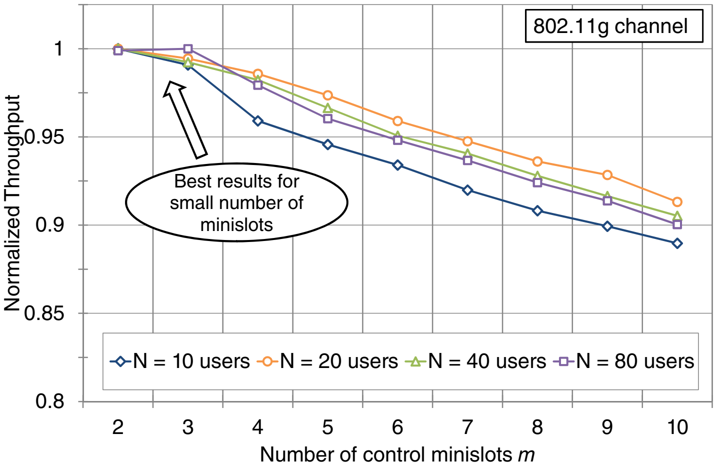


Figure 4.11: Normalized saturation throughput of DQCA versus the number of control minislots m

It can be observed that for $N \geq 20$ the obtained results do not vary much, a fact that was also illustrated in the previous figure (Figure 4.10). The most interesting remark, however, is that the maximum throughput is obtained for either 2 or 3 control minislots, even when there are 80 users contending for channel access in the system. This occurs because the collision resolution process works faster than the data transmission process in the sense that once permission to transmit is granted (by reaching the head of the transmission queue), a user typically holds the channel for a number of consecutive DQCA frames until the compete transmission of its message.[5] In the meantime, the collision resolution algorithm is executed in the beginning of each frame so that at least one user is likely to resolve the collision and enter the DTQ by the time the transmitting user exits the system. Furthermore, each collision resolution process involves only a fraction of the total number of system users since, once an ARS collision takes place, non-involved users are blocked from the contention process until the collision is fully resolved.

The last plot of this section, Figure 4.12, studies displays the DQCA saturation throughput as a function of the ARS duration. As explained in Section 3.3.1, the ARS is a short chip sequence that is transmitted within a control minislot by any user who wants to gain access to the channel. Even though it is very short, the ARS contains a specific pattern that permits the distinction between an idle minislot, a successful ARS transmission and a collision [64]. Throughout this thesis, the length of the ARS has been set to 10 μs, assuming that this duration is sufficiently long for an ARS to be detected. The small duration of the ARS is one of the key factors for the low control overhead introduced by DQCA. For instance, for $m = 3$, the CW of the DQCA frame will last 30 μs, whereas the RTS/CTS exchange in the IEEE 802.11 DCF exceeds 100 μs.[6]

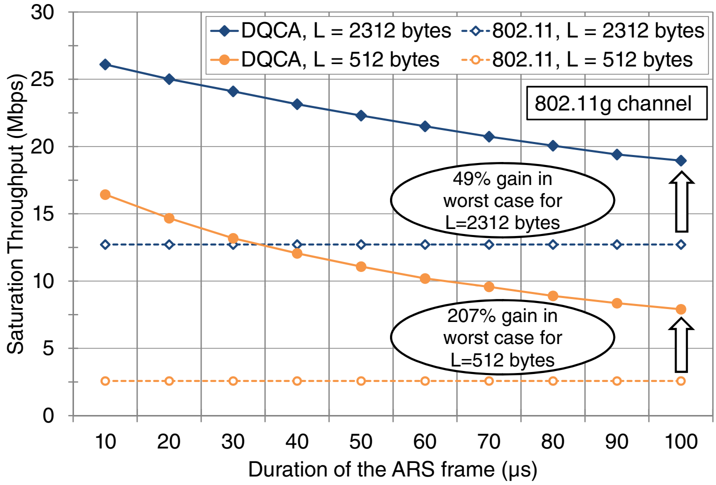


Figure 4.12: DQCA saturation throughput versus the ARS length

In any case, since a testbed implementation of DQCA is not yet available, the exact ARS duration cannot be specified with accuracy. For this reason, Figure 4.12 investigates how the use of longer ARS frames may affect the system performance. It can be observed that when longer ARS are employed (with $m = 3$ control minislots), the DQCA throughput unavoidably drops due to the additional control overhead.

[5] According to the adopted traffic generation model, each data message consists of an exponentially distributed number of packets of length L. A single packet is transmitted within a DQCA frame and the transmitting user holds the channel for the number of frames required for the completion of the message.

[6] The calculation has been made for an RTS of 20 bytes, a CTS of 14 bytes and employing the lowest rate of 6 Mbps prescribed in the IEEE 802.11g PHY specification.

However, there still remains a considerable improvement with respect to the IEEE 802.11 DCF, marked by dashed lines. For the worst represented case where the ARS duration is equal to 100 μs, this gain is in the scale of 49% when packets of 2312 bytes are employed and up to 207% for packets of 512 bytes. Anyway, bear in mind that in the IEEE 802.11g specification, the PHY layer header and preamble have a joint a duration of 20 μs which is sufficient for detection and synchronization purposes. Thus, it would be reasonable to assume that an ARS duration close to this value should be sufficient, given that this frame does not contain any MAC layer information.

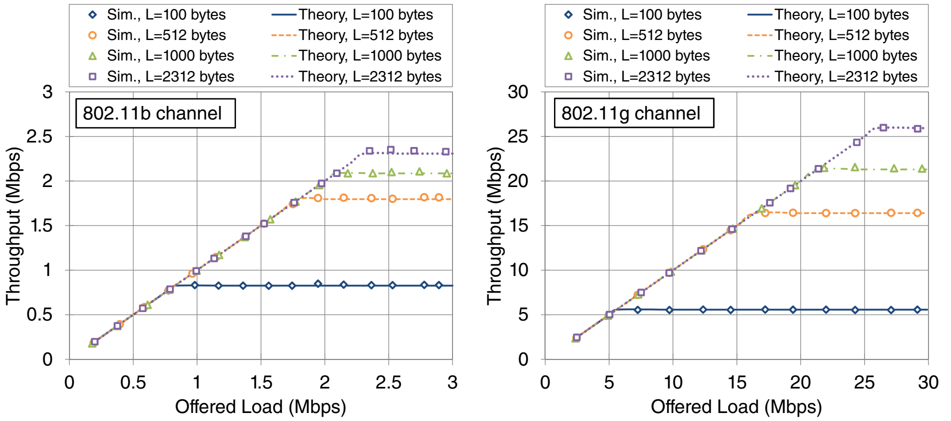
4.4.4 Non-Saturation Analysis with Link Adaptation

Finally, the throughput and delay performance under non-saturated traffic conditions for DQCA with link adaptation has been evaluated. The number of users has been set to $N = 20$ and a channel coherence time of $\tau_c = 150$ ms has been selected. The data traffic generation model explained in Section 4.4.1 has been employed, with Poisson message arrivals at different average rates λ (in msg/s). As a result, performance is evaluated as a function of the offered traffic load, calculated as $C_{data} = \kappa\lambda L(8 \cdot 10^{-6})$ Mbps, with $\kappa = 10$ the mean number of packets per message and L the packet size. The theoretical results have been based on the analytical model given in Section 4.3, using the steady state probabilities of the channel models described in Section 4.4.1.

The DQCA throughput as a function of the offered load has been plotted in Figure 4.13 for the IEEE 802.11b/g rate sets (plots (a) and (b), respectively). Four packet sizes have been considered, varying from smaller ($L = 100$ bytes) to longer ($L = 2312$ bytes) packets. The close match between the theoretical values, corresponding to the markers, and the simulated values, represented by lines, is clear, thus confirming the validity of the model. As far as the DQCA performance is concerned, it can be observed that the throughput curves increase linearly with the offered load until a maximum value is reached. This maximum value corresponds to the saturation throughput and remains stable regardless of the amount of the offered traffic load.

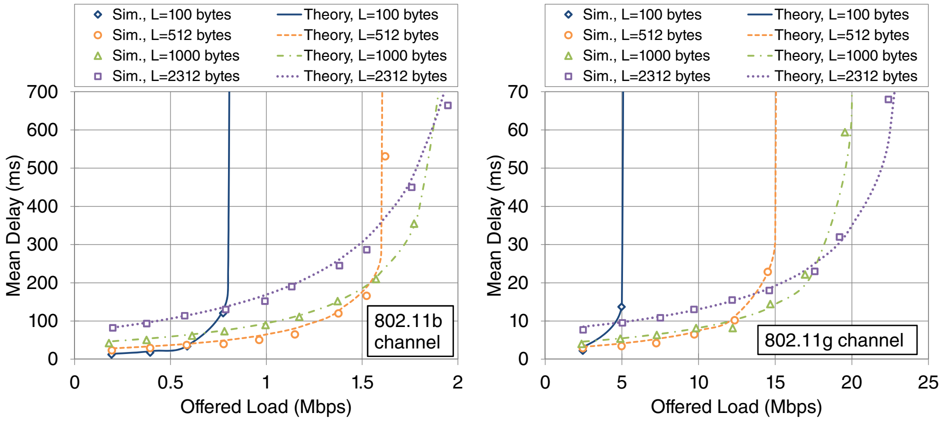
The mean message delay for the two channel models has been plotted in Figures 4.14 (a) and (b). Again, the theoretical model provides a good approximation to the simulated system performance, although some slight differences appear due to the simplifications that were considered in the queuing analysis. In general, the mean message delay remains low as long as the incoming traffic load does not lead the system to saturation. As expected, delay is lower for smaller data packet sizes, due to the reduced transmission time that is required, and when higher transmission rates are employed, as in the case of IEEE 802.11g rate set. On the other hand, when the system becomes saturated, the generated traffic exceeds the system capacity and messages begin to accumulate in the user buffers, a fact reflected by a steep increase in the mean delay plots.

In continuation, a comparison between the performance of DQCA and IEEE



(a) IEEE 802.11b PHY rates **(b)** IEEE 802.11g PHY rates

Figure 4.13: DQCA throughput as a function of the offered load, theoretical model versus simulations



(a) IEEE 802.11b PHY rates **(b)** IEEE 802.11g PHY rates

Figure 4.14: Mean DQCA message delay as a function of the offered load, theoretical model versus simulations

802.11 DCF in terms of throughput and delay is presented. The total number of users has been set to $N = 20$ and packets of $L = 2312$ bytes have been considered. As shown in Figure 4.15, throughput for both systems increases linearly with the offered load when the traffic is low. However, the improvement introduced by DQCA becomes apparent as the offered load grows. The capacity of the IEEE 802.11 system under the current conditions is approximately 15 Mbps, whereas DQCA provides a 78% gain with the maximum achieved throughput exceeding 25 Mbps.

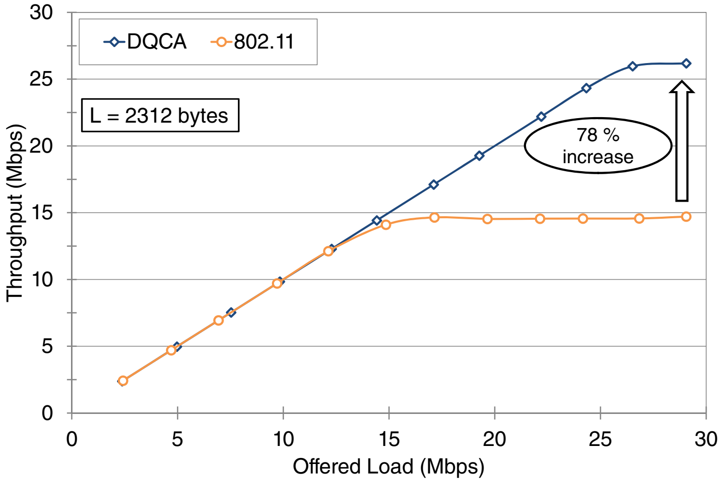
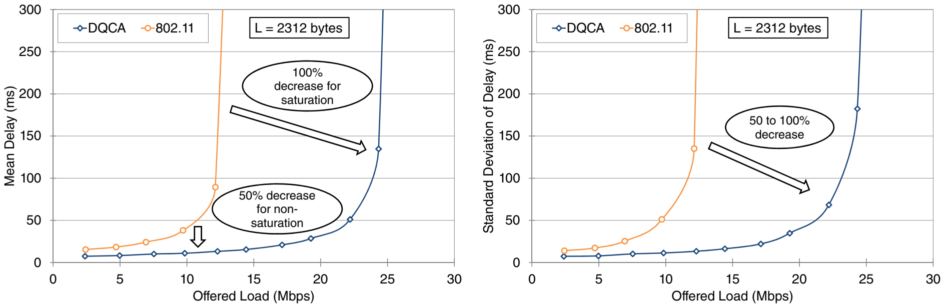


Figure 4.15: DQCA versus IEEE 802.11g DCF throughput performance

The delay performance of the two systems has been depicted in Figure 4.16. When both systems operate in the non saturated region, DQCA ensures a 50% decrease in the mean message delay, with values below 15 ms (subplot (a)). As the IEEE 802.11 approaches saturation, for an offered load of approximately 15 Mbps, the performance difference becomes more pronounced and a 100% decrease in the delay of DQCA is observed. The standard deviation (std.) of the message delay, plotted in Figure 4.16 (b), exhibits a similar behavior. A lower std. value indicates that all transmitted messages suffer from delays relatively close to the mean value and this, in turn, is an indicator of fairness among the users of the network.



(a) Mean message delay **(b)** Standard deviation of message delay

Figure 4.16: DQCA versus IEEE 802.11g DCF delay performance

4.5 Conclusions

This chapter has presented a link adaptation mechanism that provides the DQCA MAC protocol with the additional capability of adapting the transmission rate to the time-varying wireless channel conditions. The main novelty of this proposal has been the development of an analytical model for the calculation of the throughput and the mean message delay performance of DQCA. To account for the link adaptation mechanism, the DQCA queuing system has been modeled as an $M/H_{\nu}/1$ system with Poisson message arrivals and Hyperexponential service time of ν stages that correspond to the ν available transmission rates. The model can also be applied to single-rate channels for the special case of $\nu = 1$.

Some interesting conclusions can be drawn from the results presented in this chapter and are summarized next:

- A close match has been obtained between the theoretical DQCA performance calculated according to the mathematical model and the simulation results. In particular, the theoretical model provides accurate throughput and delay performance estimations when at least $N = 10$ users are present in the system under the non-saturation regime. The model can also be employed to estimate the maximum achieved throughput under saturation (delay metrics in a saturated system are not meaningful).

- The comparison between the DQCA and IEEE 802.11 DCF performance under saturated traffic conditions has revealed the increased efficiency of DQCA as MAC protocol. DQCA guarantees the practically collision-free transmission of data while maintaining low control overhead and eliminating backoff periods. This leads to an enhanced the MAC layer performance that is much closer to the system capacity. The presented results have shown that the throughput gain of DQCA over the IEEE 802.11 is above 50%, for all the transmission rates available to the IEEE 802.11 b and g PHY specifications (HR/DSSS and ERP-OFDM PHY layers, respectively) and in many cases the gain overcomes 100%.

- The saturation performance of DQCA with link adaptation shows a 45% to 90% gain with respect to the IEEE 802.11b DCF and over 100% for the IEEE 802.11g DCF, under the considered channel models. Better results are obtained as the packet size grows.

- The high throughput performance of DQCA is maintained even when the number of users grows. The key behind this behavior is that collisions among users are limited to the CW and the data transmissions take place in a TDMA-like manner within the data slot. On the contrary, the IEEE 802.11 basic access DCF performance deteriorates as the user number increases, due to multiple collisions and extended backoff periods. This is slightly improved with the use of the RTS/CTS mechanism in the DCF, however, performance is still much lower compared to DQCA.

- It has been shown that the m-ary splitting algorithm employed for the resolution of collisions among ARS frames can efficiently handle collisions for a small number of control minislots. In addition, the impact of the ARS frames size (and hence the duration of the control minislots) on performance has been studied. Clearly, throughput is higher for smaller ARS, however in the case that longer ARS were necessary (to account for PHY related synchronization issues or to reduce the probability of detection errors) DQCA still achieves a gain of more than 50% with respect to the legacy IEEE 802.11 DCF.

- Finally, studies on the non-saturation regime have shown improvements in both the throughput and delay performance of the users.

Chapter 5

Cross-Layer Enhancements for the DQCA Protocol

5.1 Introduction

DQCA is a fair protocol that treats all users equally, without taking into consideration any differences they may have in terms of capabilities and service requirements. However, as technology advances and the popularity of WLANs steadily grows, networks tend to become heterogeneous and new challenges arise. CL design, based on the exchange of information between different layers of the protocol stack, opens the road for the implementation of more sophisticated MAC protocols. By collecting parameters from other layers, the MAC layer obtains a more complete view of the system composition and user requirements and can adapt accordingly the scheduling decisions to achieve a better performance or provide QoS provisioning.

CL design is not a straightforward process, several implementation approaches are possible that depend on the protocols involved in the CL dialogue, the way in which the exchange of information is accomplished, as well as the desired scheduling objectives. In order to incorporate CL scheduling algorithms in DQCA, the following steps should be considered:

- determine the desired performance goals, including throughput performance and QoS requirements, and identify the possible trade-offs that are associated with them.

- identify which CL parameters should be available at the MAC layer and establish mechanisms for their retrieval through a communication between different protocol layers.

- modify the basic DQCA design to incorporate the CL scheduling schemes and provide the necessary feedback mechanisms.

Even though there is no single correct approach to these questions, this thesis will set a framework for CL design in DQCA and propose some mechanisms for the CL dialogue between different protocol layers.

In the context of this thesis, two main CL strategies have been investigated. The first is a channel-aware strategy that involves communication between the MAC and the PHY layers, to obtain the Channel State Information (CSI) and the available physical resources. The link adaptation mechanism presented in the previous chapter is employed in the core of the channel-aware schemes that exploit the knowledge on the link quality to make smart scheduling decisions. The second strategy aims to provide QoS-oriented scheduling by including information on the service type and the QoS requirements of the traffic flows, obtained through a MAC layer interaction with the application layer. In the considered CL design, all the information is gathered by the MAC layer, which in this case is based on the DQCA protocol, and may result in modifications of the MAC functions, namely the channel access, the scheduling and the frame encapsulation processes. This concept has been illustrated in Figure 5.1.

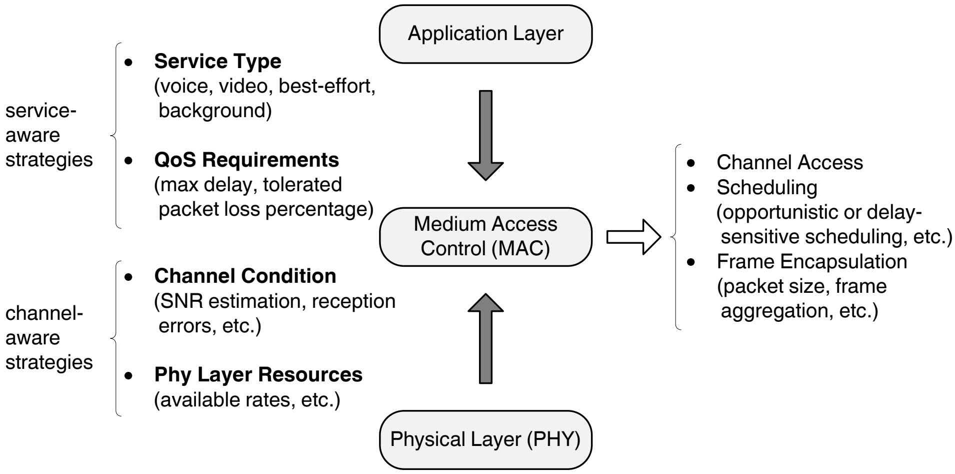

Application Layer
service-aware strategies
Service Type
(voice, video, best-effort, background)
QoS Requirements
(max delay, tolerated packet loss percentage)
Medium Access Control (MAC)
Channel Access
Scheduling
(opportunistic or delay-sensitive scheduling, etc.)
Frame Encapsulation
(packet size, frame aggregation, etc.)
channel-aware strategies
Channel Condition
(SNR estimation, reception errors, etc.)
Phy Layer Resources
(available rates, etc.)
Physical Layer (PHY)

Figure 5.1: CL design strategies for DQCA

The main body of the chapter is divided into three sections. Section 5.2 is dedicated to the description of the proposed CL-based scheduling algorithms that aim to incorporate channel and service-aware capabilities to the DQCA protocol. The enhancements and the performance trade-offs obtained by the application of these CL schemes are discussed in Section 5.3 where simulation results on three study cases are presented. The efficiency of channel-aware schemes is closely linked to the validity of the available information on the state of the wireless channel. Section 5.4 discusses the impact of outdated CSI on the system performance and presents a feedback mechanism that periodically collects updated information on the link quality of the users. Finally, the chapter closes with conclusions in Section 5.5.

5.2 Description of the CL Scheduling Algorithms

In the basic design of DQCA, users are served on a FIFO basis, in the order with which they enter the DTQ. The MAC layer scheduling process can be enhanced and oriented to achieve specific performance goals and QoS requirements by taking into account information available to different protocol layers through a CL interaction.

In this section, four CL scheduling algorithms will be presented, stressing the different goals and implementation details of each scheme. The differences between the CL policies will be further emphasized with the help of scheduling examples, provided in the last part of this section. It should be mentioned that the basic DQCA protocol used as a reference and as a base for CL enhancement in this chapter includes the link adaptation mechanism presented in Chapter 4.

5.2.1 CL-alg: A Strict Opportunistic Scheme

The first proposed scheduling algorithm, named *CL-alg*, constitutes a channel-aware scheme that exploits the time-varying nature of the wireless channel. As mentioned in Section 2.4.5, different users in a WLAN may experience independent fading and interference conditions that result to diverse link quality, an effect known as multiuser diversity. The link adaptation scheme proposed in the previous chapter provides a mechanism to select the higher rate that a user may employ for transmission, based on information of the channel condition and the desired bit error rate. However, link adaptation per se does not alter scheduling decisions; it mainly aims to improve the performance of individual users once channel access has been granted to them through the FIFO discipline of basic DQCA.

CL-alg implements an opportunistic policy that assigns priorities based solely on the available transmission rate of each user. In other words, at any given time, the user with the highest available rate is scheduled for transmission. This algorithm does not distinguish between traffic flows of different service types and, therefore, it does not aim to provide QoS guarantees to delay-sensitive applications.

By exploiting multiuser diversity and encouraging transmissions at high data rates the total system throughput is maximized. On the other hand, the performance perceived by individual users depends on their particular channel statistics. Ideally, if all users experience the same fading on average, then, even though some users may defer transmission due to bad link quality at a particular moment, they will be allowed to transmit when their channel conditions eventually improve. As a result, all users receive an equal share of the system resources on the long term. Fairness is not ensured, however, in scenarios where the fading statistics are not the same for everyone. In this case, the less fortunate users that have worse average channel conditions will be given less opportunities to transmit and thus attain low throughput and experience long delays.

Some modifications must be made to the DQCA protocol in order to implement the CL-alg:

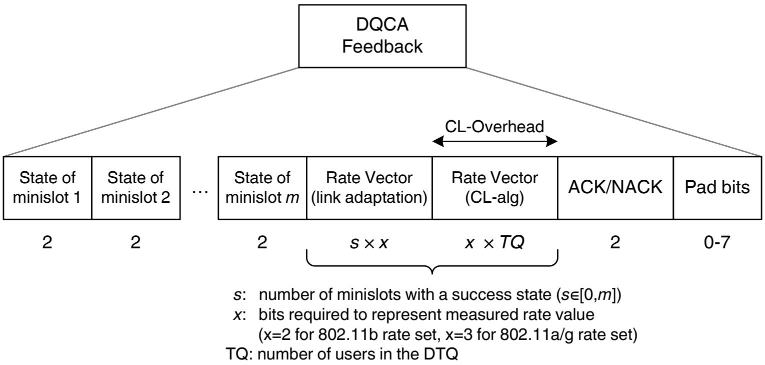

DQCA Feedback
CL-Overhead
State of minislot 1
State of minislot 2
…
State of minislot m
Rate Vector (link adaptation)
Rate Vector (CL-alg)
ACK/NACK
Pad bits
2
2
2
s × x
x × TQ
2
0-7
s: number of minislots with a success state (s∈[0,m])
x: bits required to represent measured rate value
(x=2 for 802.11b rate set, x=3 for 802.11a/g rate set)
TQ: number of users in the DTQ

Figure 5.2: DQCA feedback information field in the FBP for CL-alg

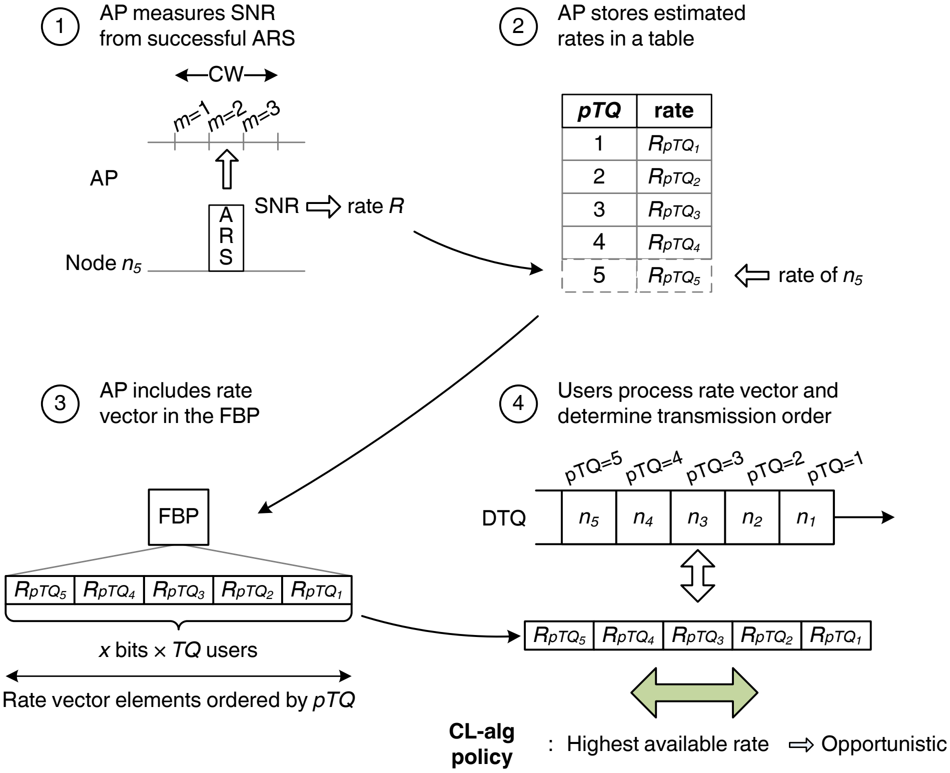

1 AP measures SNR from successful ARS
CW
m=1
m=2
m=3
AP
A R S
SNR ⇨ rate R
Node n5
2 AP stores estimated rates in a table
pTQ
rate
1 RpTQ1
2 RpTQ2
3 RpTQ3
4 RpTQ4
5 RpTQ5
⇦ rate of n5
3 AP includes rate vector in the FBP
FBP
RpTQ5 RpTQ4 RpTQ3 RpTQ2 RpTQ1
x bits × TQ users
Rate vector elements ordered by pTQ
4 Users process rate vector and determine transmission order
pTQ=5 pTQ=4 pTQ=3 pTQ=2 pTQ=1
DTQ
n5 n4 n3 n2 n1
RpTQ5 RpTQ4 RpTQ3 RpTQ2 RpTQ1
CL-alg policy
: Highest available rate ⇨ Opportunistic

Figure 5.3: Schematic representation of the CL-alg

- A link adaptation scheme must be incorporated for the estimation of the channel condition between the AP and each of the users waiting in the DTQ. The solution proposed in Section 4.2 has been adopted, according to which the SNR of the link between the AP and a user is measured whenever an ARS is successfully received. A lookup table is employed to match the SNR level to the most appropriate bit rate, in order to adhere to a targeted error performance.

- The AP maintains a vector with the estimated transmission rates of all the users in the DTQ, acquired through the link adaptation scheme, sorted by the position of the users in the queue (expressed by the counter pTQ). A predefined encoding scheme is usually employed so that each rate can be represented by a few bits per node (e.g., $x = 2$ bits for the four transmission rates of IEEE 802.11b and $x = 3$ bits for the eight rates of IEEE 802.11g).

- The estimated transmission rates must be fed back to the users. Hence, the rate vector is included in the FBP transmitted at the end of the DQCA frame. The FBP has the same structure defined in Section 4.2 (Figure 4.2), with an additional rate vector introduced in the DQCA feedback field, as shown in Figure 5.2. The exact amount of the CL overhead that must be added in the FBP depends on the rate codification scheme (x bits/rate) and the number of users in the DTQ (expressed by the counter TQ), but can be generally expressed as:

$$\text{CL-Overhead}_{\;CL-alg} = x \times TQ \quad \text{(bits per frame)}. \tag{5.1}$$

- Finally, upon the reception of the FBP the users extract the rate vector and can therefore determine their assigned transmission rate, as well as the available rates of the other users in the DTQ. By rearranging the vector entries in descending order, with the higher rates placed in the first positions, the users can decide in a distributed way whether they are enabled to transmit.

This process is schematically shown in Figure 5.3. In this example, node n_5 successfully transmits an ARS and the AP is able to measure the SNR of the link and estimate the appropriate transmission rate. This value is stored in the AP and transmitted in the form of a vector in the FBP. According to the CL-alg opportunistic policy, the user with the highest transmission rate and the smallest pTQ value (i.e., the first user with the highest transmission rate to be found closer to the head of the DTQ) is scheduled to transmit first.

5.2.2 SP-alg: A Strict Service Differentiation Scheme

The second CL scheme, *SP-alg*, implements a strict service priority policy that aims to enhance QoS for delay sensitive applications. Before explaining in detail the proposed scheduling policy, some implementation issues will be discussed, which are necessary for the better understanding of the SP-alg. In particular, the following modifications should be applied to the DQCA protocol:

- First, traffic flows are mapped onto a set of service classes according to their QoS demands, a concept known as service differentiation. In the general case, P service classes are considered ($P > 1$), with class $p = P$ having the most demanding requirements, usually expressed as most stringent delay and jitter constraints or tolerable percentage of lost packets. A practical example of this concept, which will be adopted in this thesis, is the establishment of four service classes ($P = 4$) for voice, video, best-effort and background data traffic applications, defined in the IEEE 802.11e specification [1]. In this case, voice is the highest priority service class ($p = 4$) followed by the other three classes in descending order of priority. It has been assumed that the service differentiation takes place at an upper layer and therefore when a message arrives at the MAC layer buffers its service class is known.

- Second, the users must notify the AP of the type of traffic they are planning to transmit. In other words, the service type of each message must be contained in the ARS. To this end, P distinguishable groups (patterns) of ARS are formed, one for each service class. When a user wants to request access for a particular message, it transmits an ARS from the group that corresponds to the service class it belongs to. For simplicity, it has been considered that a user handles the transmission of a single message of a given application at a time. The user exits the DQCA queuing system when the transmission of this message is completed and can enter again with a new ARS for a subsequent message (that may belong to the same or a different service class). Hence, it is said that a user belongs to a particular service class when its currently transmitted message corresponds to that service class. In any case, as far as the system is concerned, a user with two active traffic flows that belong to different service classes can be equivalently modeled as two independent users with a single traffic flow each.

- Finally, all the users must be informed of the service class corresponding to a successful ARS in order to execute the DQCA protocol rules at the end of the frame and determine the transmission order according to the scheduling algorithm. This information is included in the FBP, together with the state of the control minislots. For reference, considering $m = 3$ control minislots in the CW and $P = 4$ service classes (represented with $y = 2$ bits), the additional overhead would include at most 6 bits in the FBP. In general, the maximum additional overhead can be calculated as:

  $$\text{CL-Overhead}_{\ SP-alg} = y \times s \quad \text{(bits per frame).} \tag{5.2}$$

  with s being the number of successful ARS frames received in the control minislots ($s \in [0, m]$). The modified DQCA feedback field of the FBP is shown in Figure 5.4.

The scheduling objective of the SP-alg is to provide absolute priority to the users that belong to the highest priority service class P. As a result, the applications with stringent delay constrains have a better chance to satisfy their QoS requirements. On the other hand, low priority applications may have to wait longer times in order

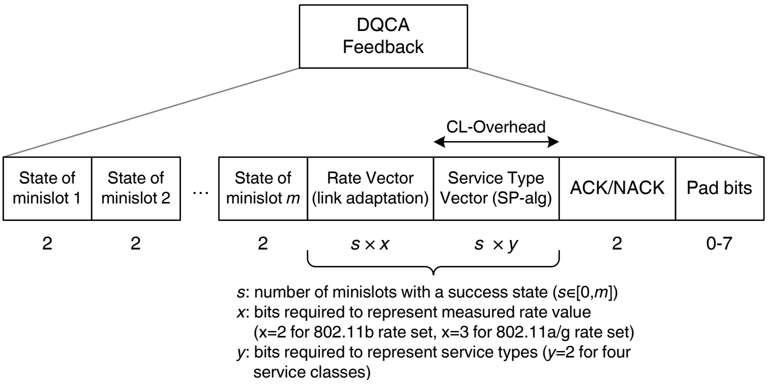

DQCA Feedback
CL-Overhead
State of minislot 1
State of minislot 2
...
State of minislot m
Rate Vector (link adaptation)
Service Type Vector (SP-alg)
ACK/NACK
Pad bits
2
2
2
s × x
s × y
2
0-7
s: number of minislots with a success state (s∈[0,m])
x: bits required to represent measured rate value (x=2 for 802.11b rate set, x=3 for 802.11a/g rate set)
y: bits required to represent service types (y=2 for four service classes)

Figure 5.4: DQCA feedback information field in the FBP for SP-alg

to transmit, however this is an acceptable trade-off given that these applications can tolerate longer delays.

This mechanism can be better understood by visualizing a system of P Data Transmission Queues (denoted by DTQ_p with $1 \geq p \geq P$), instead of a single DTQ, as was the case in DQCA. Each queue handles the users of the respective service class. Hence, P counters are defined, denoted by TQ_p ($1 \geq p \geq P$), that represent the number of elements (i.e. nodes) in each DTQ_i. Another counter (pTQ) is required to indicate the age of the particular node in the queue to which it belongs, given that a node can only belong to one queue at a time. Therefore, the position of a node in the data transmission subsystem can be fully described by the pair of integers (p, pTQ_p), where p denotes the service class and consequently the queue in which the node belongs. Summarizing, $(P + 1)$ counters must be maintained at every user for the data transmission subsystem and two counters (RQ and pRQ) for the collision resolution subsystem, as in the basic DQCA operation.

A node belonging to the pth queue (DTQ_p) can initiate transmission only if the following two conditions are met:

- All the queues with a higher priority level q are empty (i.e., $TQ_q = 0$ for all $q > p$).
- The node has reached the head of the pth (i.e, has $pTQ = 1$). In other words, the queues follow a FIFO discipline and nodes that belong to the same service are served in order of arrival.

Once a node is granted access according to the aforementioned rules, it maintains control of the medium until the completion of its message. In other words, a non-preemptive policy is adopted and subsequent arrivals of higher priority traffic do not affect ongoing transmissions.

A schematic representation of the SP-alg operation is depicted in Figure 5.5. The example shows how two nodes, n_4 and n_5, select ARS of different types to indicate their intention to transmit background and voice traffic flows, respectively. The FBP contains the state of the control minislots and the service type of the two successful ARS, thus enabling the nodes to enter in corresponding background and voice data queues. According to the SP-alg, the voice users will be the first to transmit in order of their *pTQ* values (i.e., FIFO with respect to the voice data queue) and will be followed by the users of the remaining three queues.

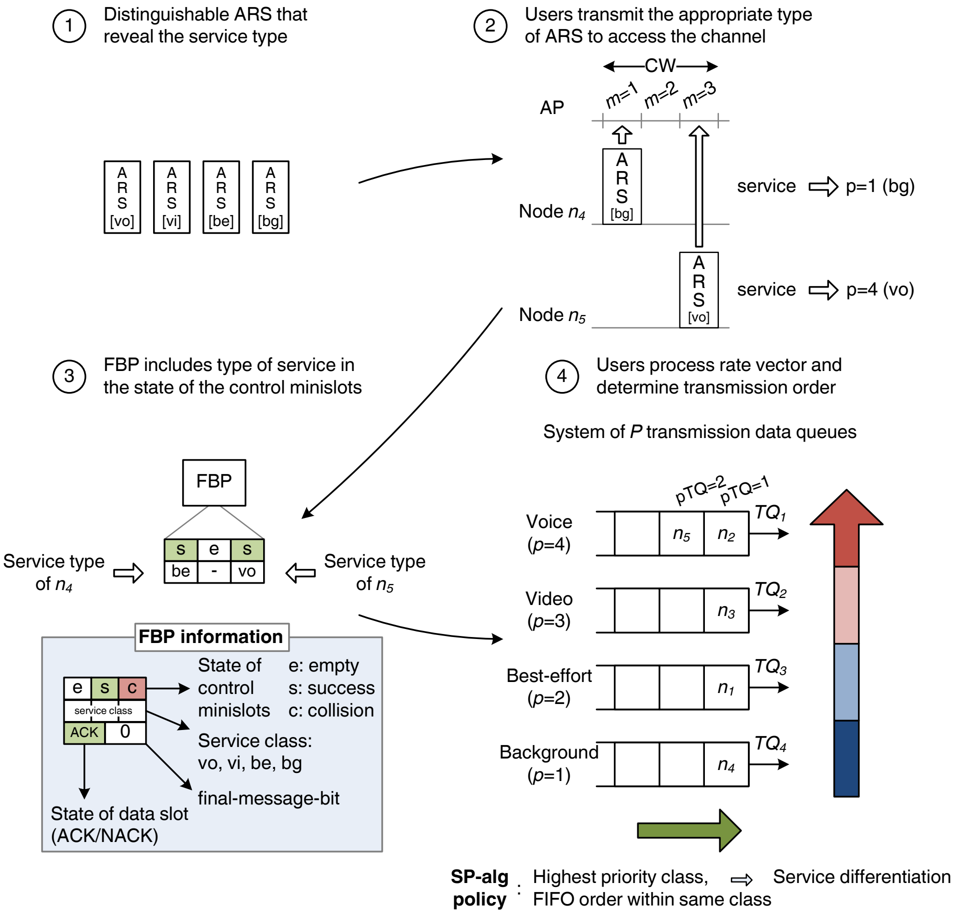
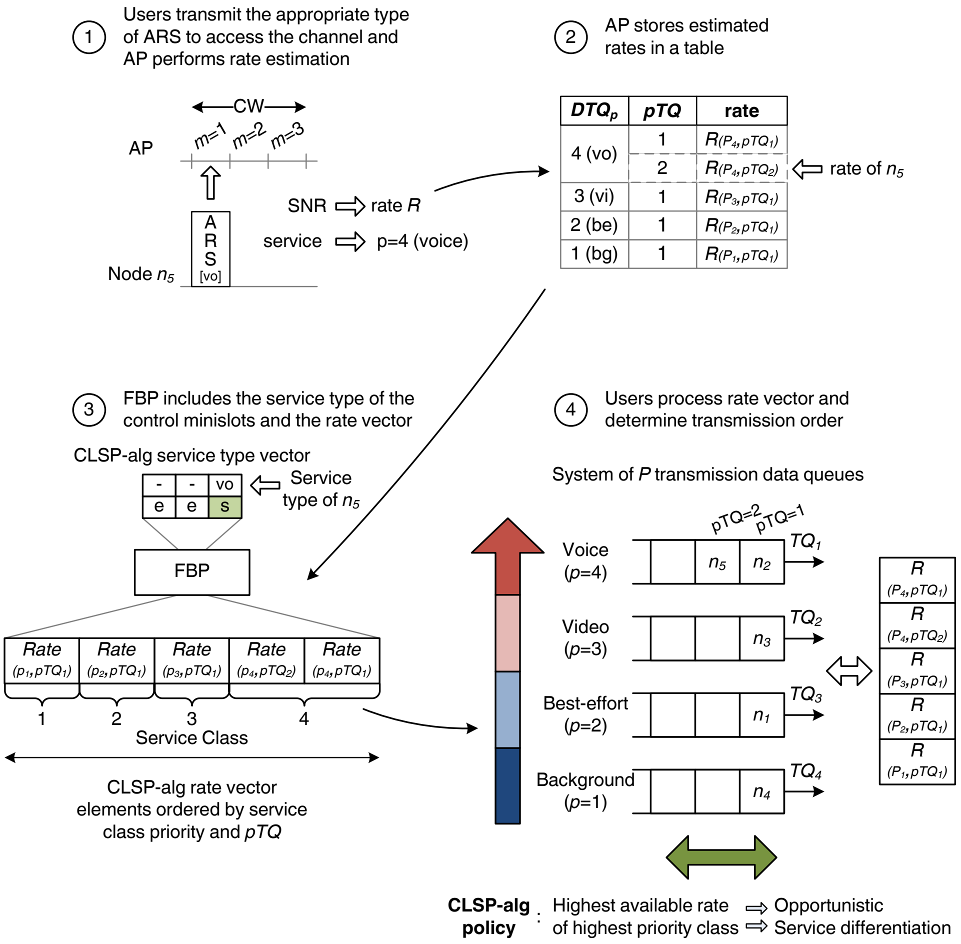

1 Distinguishable ARS that reveal the service type
ARS [vo]
ARS [vi]
ARS [be]
ARS [bg]
2 Users transmit the appropriate type of ARS to access the channel
CW
AP
m=1
m=2
m=3
Node n_4
ARS [bg]
service ⇨ p=1 (bg)
Node n_5
ARS [vo]
service ⇨ p=4 (vo)
3 FBP includes type of service in the state of the control minislots
FBP
s e s
be - vo
Service type of n_4
Service type of n_5
FBP information
e s c
service class
ACK 0
State of control minislots
e: empty
s: success
c: collision
Service class: vo, vi, be, bg
final-message-bit
State of data slot (ACK/NACK)
4 Users process rate vector and determine transmission order
System of *P* transmission data queues
pTQ=2
pTQ=1
Voice (p=4)
n_5
n_2
TQ_1
Video (p=3)
n_3
TQ_2
Best-effort (p=2)
n_1
TQ_3
Background (p=1)
n_4
TQ_4
SP-alg policy : Highest priority class, FIFO order within same class ⇨ Service differentiation

Figure 5.5: Schematic representation of the SP-alg

5.2.3 CLSP-alg: A Strict Opportunistic Scheme with Service Differentiation

So far, two different policies have been presented, the first implementing an opportunistic transmission scheme and the second a QoS-oriented service differentiation strategy. These two concepts are combined to form the third proposed scheme, *CLSP-alg*. CLSP-alg adopts the strict priority scheme defined in the SP-alg, but also employs the opportunistic scheduling of CL-alg among users that belong to the same service class. A schematic representation of the CLSP-alg is given in Figure 5.6.


1 Users transmit the appropriate type of ARS to access the channel and AP performs rate estimation
CW
AP
m=1
m=2
m=3
A R S [vo]
Node n_5
SNR ⇨ rate R
service ⇨ p=4 (voice)
2 AP stores estimated rates in a table
⇦ rate of n_5
3 FBP includes the service type of the control minislots and the rate vector
CLSP-alg service type vector
- - vo
e e s
⇦ Service type of n_5
FBP
Rate (p_1,pTQ_1)
Rate (p_2,pTQ_1)
Rate (p_3,pTQ_1)
Rate (p_4,pTQ_2)
Rate (p_4,pTQ_1)
1
2
3
4
Service Class
CLSP-alg rate vector elements ordered by service class priority and pTQ
4 Users process rate vector and determine transmission order
System of P transmission data queues
pTQ=2
pTQ=1
Voice (p=4)
Video (p=3)
Best-effort (p=2)
Background (p=1)
n_5
n_2
n_3
n_1
n_4
TQ_1
TQ_2
TQ_3
TQ_4
R (P_4,pTQ_1)
R (P_4,pTQ_2)
R (P_3,pTQ_1)
R (P_2,pTQ_1)
R (P_1,pTQ_1)
CLSP-alg policy : Highest available rate ⇨ Opportunistic
of highest priority class ⇨ Service differentiation

DTQ_p	pTQ	rate
4 (vo)	1	$R(P_4,pTQ_1)$
	2	$R(P_4,pTQ_2)$
3 (vi)	1	$R(P_3,pTQ_1)$
2 (be)	1	$R(P_2,pTQ_1)$
1 (bg)	1	$R(P_1,pTQ_1)$

Figure 5.6: Schematic representation of the CLSP-alg

The implementation of this algorithm requires the following steps:

- As in the case of the SP-alg, the data transmission subsystem consists of P DTQs. Service-aware ARS are employed for the channel access request and the service type of successful ARS is included in the FBP in order to enable users to update the TQ_p counter values of the corresponding pth data queue.

- Upon the reception of an ARS, the AP measures the SNR and estimates the quality of the link. A rate vector is then formed, as in the case of the CL-alg, containing the estimated transmission rates of all the users waiting in the DTQs. In this case, however, the vector entries are first sorted by order of the service class. Then, the entries that belong to the same service class are ordered according to the position of the nodes in each DTQ.

- The FBP contains the additional CL overhead due to the rate vector and the service type of the successful ARS. In general, the added CL overhead can be calculated as

$$\text{CL-Overhead}\ _{CLSP-alg} = x \times \sum_{p=1}^{P} TQ_p + s \times y \quad \text{(bits per frame)}, \qquad (5.3)$$

considering that x bits (typically x=2 or 3) are employed to represent the available rate values and y bits are required to represent the P service classes for the s successful control minislots (typically $P = 4$ and therefore $y = 2$). The parameter TQ_p expresses the number of users of the pth service class that are waiting for transmission. The modified DQCA feedback field of the FBP is shown in Figure 5.7.

- Finally, the users extract the rate values from the FBP and determine the transmission order. According to the CLSP-alg, channel access is granted to the user that belongs to the highest priority service class and has the highest available transmission rate.

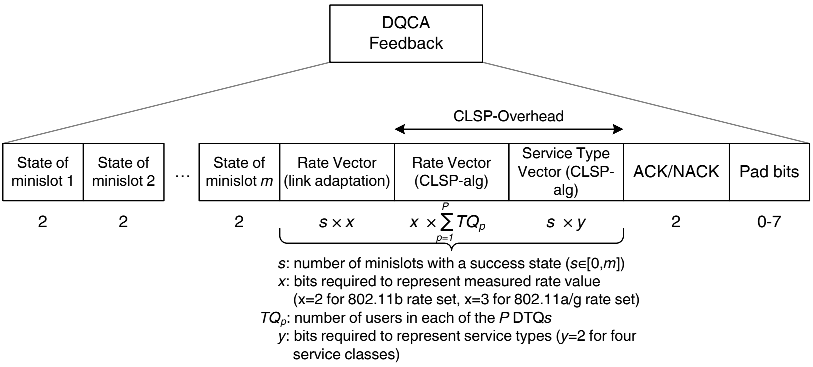

DQCA Feedback
CLSP-Overhead
State of minislot 1
State of minislot 2
...
State of minislot m
Rate Vector (link adaptation)
Rate Vector (CLSP-alg)
Service Type Vector (CLSP-alg)
ACK/NACK
Pad bits
2
2
2
s×x
$x \times \sum_{p=1}^{P} TQ_p$
s ×y
2
0-7
s: number of minislots with a success state (s∈[0,m])
x: bits required to represent measured rate value
(x=2 for 802.11b rate set, x=3 for 802.11a/g rate set)
TQp: number of users in each of the P DTQs
y: bits required to represent service types (y=2 for four service classes)

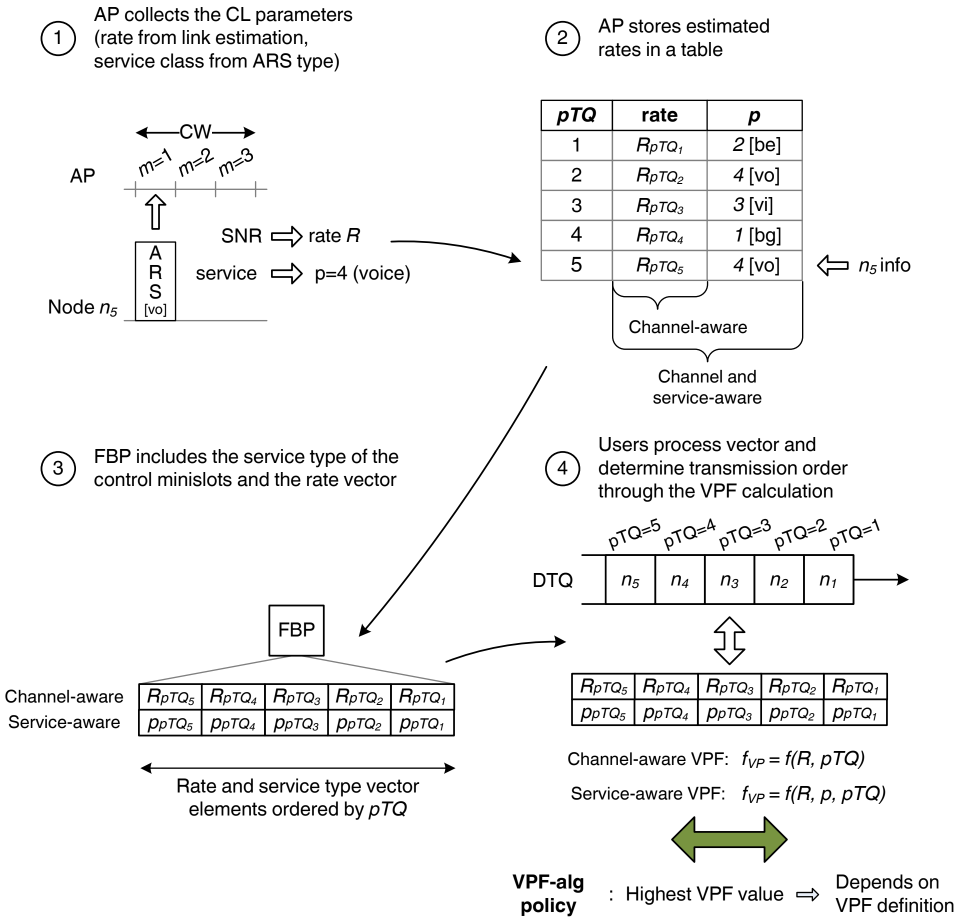
Figure 5.7: DQCA feedback information field in the FBP for CLSP-alg

5.2.4 VPF-alg: The Virtual Priority Function Concept

Finally, a more generic and flexible approach for the incorporation CL scheduling within the DQCA protocol has been considered. The fourth technique, named *VPF-alg*, defines a Virtual Priority Function (*VPF*) that determines the transmission order of the users in the DTQ. Based on a CL dialog aimed to provide channel and service aware scheduling strategies, the VPF can be generally defined as a function of PHY, MAC and APP layer parameters:

$$\text{VPF-alg} : f_{VP} = f(\mathit{PHY\ parameters,\ MAC\ parameters,\ APP\ parameters}). \qquad (5.4)$$

In accordance with the distributed character of DQCA, the VPF definition must be known to all users. The AP is responsible for collecting the CL parameters required for the calculation of the VPF values of every user in the DTQ. These parameters are included in the FBP in the form of a vector, ordered by the position of the users in the DTQ as expressed by the pTQ counter. A predefined encoding scheme is usually employed so that these parameters can be represented by a few bits per node. This mechanism is similar to the rate vector employed for channel-aware scheduling in CL-alg and CLSP-alg, although in this case multiple CL parameters may be included per user instead of only the available transmission rate. At the end of each frame every user calculates its VPF value, as well as the VPF values of the other users waiting in the DTQ and the one with the highest VPF value is scheduled to transmit first. In the case where multiple users share the same VPF value, priority is given to the one with the longest waiting time in the DTQ (i.e. the user with the smallest pTQ value). A schematic representation of the VPF-alg is depicted in Figure 5.8.


1
AP collects the CL parameters (rate from link estimation, service class from ARS type)
CW
AP
m=1
m=2
m=3
A R S [vo]
Node n_5
SNR ⇨ rate R
service ⇨ p=4 (voice)
2
AP stores estimated rates in a table
⇦ n_5 info
Channel-aware
Channel and service-aware
3
FBP includes the service type of the control minislots and the rate vector
FBP
Channel-aware
Service-aware
Rate and service type vector elements ordered by pTQ
4
Users process vector and determine transmission order through the VPF calculation
pTQ=5
pTQ=4
pTQ=3
pTQ=2
pTQ=1
DTQ
n_5
n_4
n_3
n_2
n_1
Channel-aware VPF: $f_{VP} = f(R, pTQ)$
Service-aware VPF: $f_{VP} = f(R, p, pTQ)$
VPF-alg policy : Highest VPF value ⇨ Depends on VPF definition

pTQ	**rate**	***p***
1	R_{pTQ_1}	*2* [be]
2	R_{pTQ_2}	*4* [vo]
3	R_{pTQ_3}	*3* [vi]
4	R_{pTQ_4}	*1* [bg]
5	R_{pTQ_5}	*4* [vo]

R_{pTQ_5}	R_{pTQ_4}	R_{pTQ_3}	R_{pTQ_2}	R_{pTQ_1}
p_{pTQ_5}	p_{pTQ_4}	p_{pTQ_3}	p_{pTQ_2}	p_{pTQ_1}

Figure 5.8: Schematic representation of the VPF-alg

The priority function can be selected in various ways, according to the available CL parameters and the scheduling objective which is usually a trade-off between throughput maximization and fairness. An interesting observation is that by appropriately selecting the VPF definition, some basic scheduling schemes can be implemented. For example, for $f_{VP} = 1/pTQ$, the FIFO transmission order of DQCA can be accomplished. A number of different VPF definitions will be considered in the remaining of this section. They are divided into two groups: the first group implements channel-aware scheduling whereas the second group combines opportunistic scheduling with QoS provisioning.

Rate-aware VPF Definitions

The first group of the proposed VPF definitions implements a mild opportunistic transmission scheme that takes into account the channel condition of the users but also aims to provide a level of fairness among them. The considered approach to attain this goal is by including the age of the users in the DTQ, as expressed by the pTQ counter, in the scheduling decisions. Hence, the general VPF expression has the following form:

$$\text{VPF-alg} : f_{VP} = f(R, pTQ) \tag{5.5}$$

where R is the available user rate (in Mbps) and $pTQ \in [1, TQ]$ for each user within the DTQ.

A simple example of this group is the following VPF definition, denoted by VPF-alg$_1$:

$$\text{VPF-alg}_1 : f_{VP} = \frac{R}{R_\nu} \cdot \frac{1}{pTQ} \tag{5.6}$$

where R_ν is the maximum transmission rate (in Mbps) defined in the rate set. For example, R_ν=11 Mbps for the IEEE 802.11b PHY and R_ν=54 Mbps for IEEE 802.11g.

According to this definition, the VPF value of each user is directly proportional to its available transmission rate, normalized by the maximum rate value. As a result, users with better channel conditions and, therefore, higher rates, are opportunistically assigned transmission priorities. On the other hand, by placing the pTQ value at the denominator, the VPF value increases as a user approaches the head of the DTQ. Hence, users with longer times in the system may be given a chance to transmit, even if their channel conditions are not the best.

Another VPF-alg function that achieves a similar objective by employing a different function definition is given next:

$$\text{VPF-alg}_2 : f_{VP} = \frac{\alpha^r}{pTQ}, \alpha \in \mathbb{N}. \tag{5.7}$$

In this case, α is a tunable integer parameter with values greater than one and r is an integer within $[1, \nu]$, with ν being the number of available transmission rates in the rate set. Concretely, $\nu = 4$ for the four rates defined in IEEE 802.11b and $\nu = 8$

for the eight rates defined in IEEE 802.11g.

Again, the aim of this function is to prioritize high rate users while taking into account the age of the users in the DTQ. The difference with the first VPF definition (VPF-alg$_1$) is that, in this case, the rate index is employed as an exponent over a variable base α. As a result, the rate has a stronger impact on the scheduling decisions. A more insightful comparison between the two functions will be provided in the next chapter, with the help of simulations.

For the implementation of channel-aware VPF-alg, the FBP must contain a vector with the rates of the users in the DTQ, sorted by their pTQ value. Similarly to the CL-alg, the required overhead can be calculated as:

$$\text{CL-Overhead}_{\ VPF-alg_{1,2}} = x \times TQ \quad \text{(bits per frame).} \tag{5.8}$$

with x bits being employed to represent the rate set and TQ being the number of users in the DTQ.

Channel and Service-aware VPF Definitions

The second group of the proposed VPF definitions combines opportunistic scheduling and QoS provisioning through service differentiation. In this case, the adopted general VPF expression is of the form:

$$\text{VPF-alg} : f_{VP} = f(p, R, pTQ). \tag{5.9}$$

where R is the transmission rate, pTQ the DTQ position counter for each node and p is the service type identifier with values in $[1, P]$, with P being the highest priority class (typically $P = 4$ for background, best-effort, video and voice traffic flows).

Two definitions have been considered for the evaluation of the VPF-alg. The first definition is:

$$\text{VPF-alg}_3 : f_{VP} = \alpha \cdot 2^p + (1-\alpha) \cdot \frac{2^r}{pTQ}, 0 \leq \alpha \leq 1 \tag{5.10}$$

where $r \in [1, \nu]$ is an integer index of the available transmission rates. In particular, $r = 1$ corresponds to the minimum and $r = \nu$ to the maximum rate of the rate set.

The second considered definition is:

$$\text{VPF-alg}_4 : f_{VP} = \alpha \cdot \frac{p}{P} + (1-\alpha) \cdot \frac{R}{R_\nu} \cdot \frac{1}{pTQ}, 0 \leq \alpha \leq 1 \tag{5.11}$$

where R_ν is the maximum rate defined in the rate set (specifically, R_ν=11 Mbps in IEEE 802.11b and R_ν=54 Mbps in IEEE 802.11g).

In both examples, the VPF definition has two parts. The first is a function of the service type identifier p, with higher values corresponding to traffic flows of increased priority. The second part of the definition depends, on the one hand, on the transmission rate, either expressed as the integer rate index r in the function VPF-alg$_3$, or as the rate R normalized by the maximum rate value, in function VPF-alg$_4$. On the other hand, the second part of the definition is divided by the pTQ value, thus giving the opportunity of users with lower rates but longer waiting times in the data transmission subsystem to gain access to the channel.

A tunable parameter, denoted by α with $0 \leq \alpha \leq 1$, is employed to weight the two parts of the VPF-alg definitions. Higher values of α place more weight on the service class of the traffic flow, thus making QoS provisioning the principal objective of the VPF-alg. Similarly, the mild opportunistic policy (i.e., available rate versus the DTQ position of the user) becomes the prevalent scheduling factor when smaller values of α are selected.

For the implementation of channel and service-aware VPF-alg, the FBP must contain a vector with the rates of the users in the DTQ and the type of the service class, sorted by their pTQ value. The required overhead can be calculated as:

$$\text{CL-Overhead}_{\,VPF-alg_{3,4}} = (x + y) \times TQ \quad \text{(bits)}. \tag{5.12}$$

with x being the number of bits employed to represent the rate set, y the bits required for the representation of the P service classes and TQ being the number of users in the DTQ. A summary of the four proposed VPF-alg definitions, along with the employed parameters and the values adopted in this thesis, is given in Table 5.1.

5.2.5 Overview of the CL-based Algorithms

This section has presented four scheduling algorithms applied over the DQCA protocol to attain different performance objectives. These algorithms add channel and service-aware capabilities to DQCA through a CL interaction between the MAC layer, on the one side, and the PHY and the application layers on the other. The first three schemes, CL-alg, SP-alg and CLSP-alg prescribe specific performance goals whereas the VPF-alg is a more flexible scheme whose behavior depends on the definition of the priority function. A summary of the scheduling objective of each algorithm is given in Table 5.2.

The differences between the proposed schemes can be better understood with the help of an example, illustrated in Figure 5.9. The example depicts a snapshot of the DTQ with six users waiting for transmission and shows how the transmission order is affected by each CL policy. The pTQ counter indicates the order in which the users have entered the DTQ, with $pTQ = 1$ corresponding to the oldest user in the queue. The available transmission rate and the service class type are also marked for each user. The eight rate sets of the IEEE 802.11g specification and the four service classes of IEEE 802.11e have been considered in the example. Assuming that no more users enter the system in this example, the transmission order for each algorithm is formed as follows:

(a) DQCA follows a FIFO discipline with users transmitting in order of their pTQ counter.

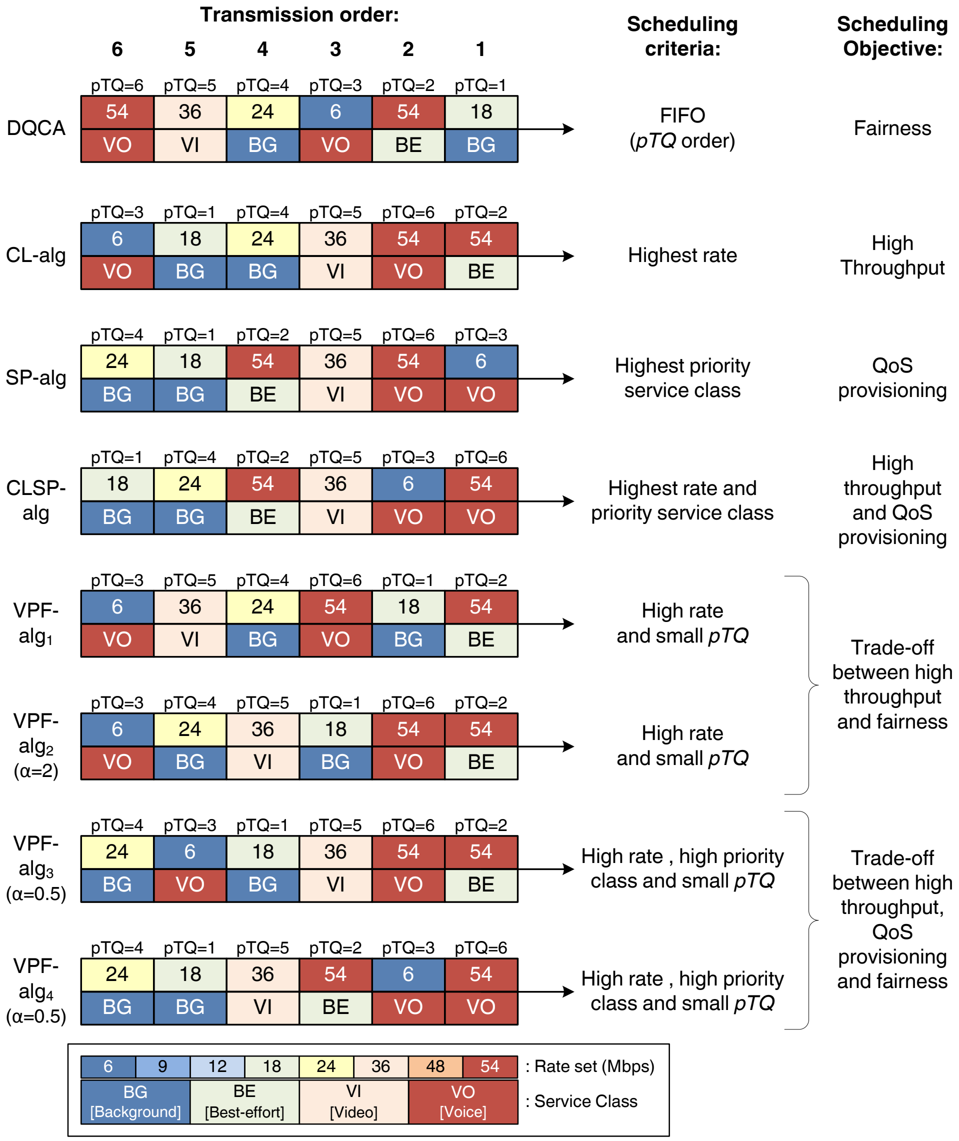

Transmission order:
6 5 4 3 2 1
Scheduling criteria:
Scheduling Objective:
DQCA
pTQ=6 pTQ=5 pTQ=4 pTQ=3 pTQ=2 pTQ=1
54 36 24 6 54 18
VO VI BG VO BE BG
FIFO (pTQ order)
Fairness
CL-alg
pTQ=3 pTQ=1 pTQ=4 pTQ=5 pTQ=6 pTQ=2
6 18 24 36 54 54
VO BG BG VI VO BE
Highest rate
High Throughput
SP-alg
pTQ=4 pTQ=1 pTQ=2 pTQ=5 pTQ=6 pTQ=3
24 18 54 36 54 6
BG BG BE VI VO VO
Highest priority service class
QoS provisioning
CLSP-alg
pTQ=1 pTQ=4 pTQ=2 pTQ=5 pTQ=3 pTQ=6
18 24 54 36 6 54
BG BG BE VI VO VO
Highest rate and priority service class
High throughput and QoS provisioning
VPF-alg1
pTQ=3 pTQ=5 pTQ=4 pTQ=6 pTQ=1 pTQ=2
6 36 24 54 18 54
VO VI BG VO BG BE
High rate and small pTQ
VPF-alg2 (α=2)
pTQ=3 pTQ=4 pTQ=5 pTQ=1 pTQ=6 pTQ=2
6 24 36 18 54 54
VO BG VI BG VO BE
High rate and small pTQ
Trade-off between high throughput and fairness
VPF-alg3 (α=0.5)
pTQ=4 pTQ=3 pTQ=1 pTQ=5 pTQ=6 pTQ=2
24 6 18 36 54 54
BG VO BG VI VO BE
High rate , high priority class and small pTQ
VPF-alg4 (α=0.5)
pTQ=4 pTQ=1 pTQ=5 pTQ=2 pTQ=3 pTQ=6
24 18 36 54 6 54
BG BG VI BE VO VO
High rate , high priority class and small pTQ
Trade-off between high throughput, QoS provisioning and fairness
6 9 12 18 24 36 48 54 : Rate set (Mbps)
BG [Background] BE [Best-effort] VI [Video] VO [Voice] : Service Class

Figure 5.9: Operation example of the CL-based algorithms

Table 5.1: Parameters employed by the VPF-alg

Proposed VPF-alg definitions		
VPF-alg$_1$	$f_{VP} = (R/R_\nu) \cdot (1/pTQ)$	
VPF-alg$_2$	$f_{VP} = \alpha^r/pTQ$	
VPF-alg$_3$	$f_{VP} = \alpha \cdot 2^p + (1-\alpha) \cdot 2^r/pTQ$	
VPF-alg$_4$	$f_{VP} = \alpha \cdot p/P + (1-\alpha) \cdot (R/R_\nu) \cdot (1/pTQ)$	
Parameter	*Description*	*Values adopted in this thesis*
ν	Size of rate set	$\nu = 4$ for 802.11b, $\nu = 8$ for 802.11g
r	Rate index	$r \in [0, \nu]$
R	Transmission rate (Mbps)	$R_{802.11b} = \{1, 2, 5.5, 11\}$, $R_{802.11g} = \{6, 9, 12, 18, 24, 36, 48, 54\}$
R_ν	Max. transmission rate	$R_\nu = 11$ (802.11b), $R_\nu = 54$ (802.11g)
P	Number of service classes	$P = 4$
p	Service class type	$p \in [1, P]$ for background, best-effort, video and voice service, respectively
pTQ	User position in the DTQ	$pTQ \in [1, TQ]$
TQ	Total users in the DTQ	integer ≥ 0
α	Tunable parameter	$\alpha = 2, 3, 4, ...$ for VPF-alg$_2$ $\alpha \in [0, 1]$ for VPF-alg$_3$, VPF-alg$_4$

(b) CL-alg implements a strict opportunistic scheme based on the available rate of each user. The goal is to maximize throughput by prioritizing users with high transmission rates. Hence, the first user to transmit is the one with the highest available rate ($R = 54$ Mbps). Since two users fulfill this condition in the example, the one with the smallest pTQ ($pTQ = 2$) is scheduled first.

(c) The SP-alg is focused solely on the service class of each node, aiming to reduce the waiting times of delay-sensitive applications. Thus, the oldest user with the highest priority class (i.e., the voice user with $pTQ = 3$) transmits first.

(d) CLSP-alg combines the two previous scheduling policies. Users are sorted based on their service class, as in the case of the SP-alg, but among those of the same class, priority is given to the one with the highest rate. Hence, the first to transmit is the voice user with $R = 54$ Mbps.

(e) VPF-alg$_1$ combines the rate and the pTQ value of each user to determine the transmission order. The aim is to encourage transmissions at higher rates but with some consideration for the arrival order of users (expressed by the pTQ), thus increasing fairness. In the example, based on the evaluation of the VFP values, the first and the third positions are assigned to the users with the highest

Table 5.2: Summary of the CL-based algorithms

CL-based algorithm	*Channel-aware (Opportunistic)*	*Service-aware (QoS provisioning)*
CL-alg	X	-
SP-alg	-	X
CLSP-alg	X	X
VPF-alg$_1$	X	-
VPF-alg$_2$	X	-
VPF-alg$_3$	X	X
VPF-alg$_4$	X	X

transmission rate of 54 Mbps. Nevertheless, the oldest user in the queue ($pTQ = 1$) is given the chance to transmit second, despite having a relatively low rate of 18 Mbps. Note that the service class of the users are not included in the scheduling decisions.

(f) VPF-alg$_2$ follows the same principles as VPF$_1$, but employs a different priority function that gives more importance to the transmission rate (by elevating the rate index to the power of the tunable parameter α). As a result, the first two transmission positions are given to the users with the highest rate of 54 Mbps. The oldest user ($pTQ = 1$) is now given the third position, which is still a fairer treatment compared to the CL, SP and CLSP-alg.

(g) VPF-alg$_3$ includes the rate, the service class and the pTQ value of each user in the scheduling decisions. Rate has a stronger impact than the service class and for the first three positions the algorithm behaves as the CL-alg. Nevertheless, the fourth position is given to the user with $pTQ = 1$. The result is a more flexible opportunistic policy that encourages high-rate transmissions without completely depriving low-rate users of the opportunity to transmit.

(h) Finally, the VPF-alg$_4$ considers the same parameters as VPF-alg$_3$ but places more importance to the service class. As a result, the highest priority voice users are scheduled first, whereas the video user is assigned the fourth position. The third position is given to a best-effort user with the highest rate of 54 Mbps, thus stressing the balance that VPF-alg$_4$ tries to achieve between opportunistic and service-aware scheduling. Lastly, the last two transmission position are given to the lowest priority background users, however, the user with the smallest pTQ is scheduled first in spite of its lower transmission rate, to account for the longer waiting time in the DTQ.

This example gives some insight on the scheduling decisions and trade-offs associated with the four proposed CL algorithms. Further discussion on these issues will take place in the following section where, with the help of simulations, the performance of the CL-based schemes in different scenarios will be compared.

5.3 Performance Evaluation of the CL Algorithms

The most solid method to evaluate and gain insight on the performance of a MAC protocol is through the development of a valid mathematical model. Mathematical analysis provides a very useful base for obtaining benchmark results for specific scenarios and usually under a set of assumptions. Analytical formulation, however, is not always feasible due to the difficulties that are often encountered in the process of modeling realistic scenarios. In the case of the proposed CL-based algorithms, performance depends on many different parameters such as traffic models, channel conditions, scheduling policies and other factors that are often interrelated, and as a result mathematical formulation is not straightforward and sometimes even not possible.

Hence, the results presented in this section are based on simulations, obtained through a custom-made C++ simulation tool (also employed for the simulation results presented in Chapter 4).

Section 5.3.1 provides information on the simulation setup. The performance evaluation follows next, divided in three study cases:

- Section 5.3.2 evaluates the performance of CL-based algorithms over DQCA in a data only environment in which the main scheduling objective is throughput maximization.
- Section 5.3.3 extends the CL-based scheduling for a mixed voice and data scenario, introducing the concept of service differentiation to achieve QoS provisioning.
- Section 5.3.4 considers the service differentiation paradigm of IEEE 802.11e with traffic sources mapped onto four service classes. Different traffic generation models are employed to simulate multimedia applications such as voice and video sessions along with best-effort and background non-real time data.

5.3.1 Simulation Setup

The performance evaluation of the CL-based scheduling algorithms has been based on the simulation framework described in Section 4.4.1 of the previous chapter.

PHY Layer Parameters and Channel Model

At the PHY layer, the HR/DSSS PHY, defined in IEEE 802.11b, with four transmission rates (1, 2, 5.5. and 11 Mbps), and the ERP-OFDM, defined in IEEE 802.11g, with eight transmission rates (6, 9, 12, 18, 24, 36, 48 and 54 Mbps), have been considered. A summary of the PHY related parameters is given in Table 5.3. The transition matrices described in Section 4.4.1 have been employed for the calculation of the channel conditions that determine the available transmission rate of each user. As a result of the selected values, the majority of transmissions are likely to take place at the rates of 2 or 5.5 Mbps, in the case of the IEEE 802.11b rate, set and at the rates of 24 and 36 Mbps, in the case of the IEEE 802.11g rate set.

Table 5.3: Summary of PHY layer simulation parameters (Section 4.4.1)

Parameters applicable to DQCA and IEEE 802.11 MAC protocol operation		
Parameter	*IEEE 802.11b PHY*	*IEEE 802.11g PHY*
Rate set	1, 2, 5.5 and 11 Mbps	6, 9, 12, 18, 24, 36, 48 and 54 Mbps
PLCP preamble and PHY header	96 μs	20 μs
aSifsTime	10 μs	10 μs
Parameters applicable only to IEEE 802.11 MAC protocol operation		
Parameter	*IEEE 802.11b PHY*	*IEEE 802.11g PHY*
aSlotTime	20 μs	9 μs
aCWmin	31	15
aCWmax	1023	1023

MAC Layer Parameters and Service Differentiation

The MAC layer parameters employed in the simulations are summarized in Table 5.4. In the case of DQCA, some additional overhead information must be added for the implementation of the CL-based scheduling algorithms, as explained in detail in the previous section. The exact number of overhead bits considered in the simulations will be provided separately in each presented study case.

Table 5.4: Summary of MAC layer simulation parameters (Section 4.4.1)

IEEE 802.11		DQCA	
Parameter	*Value*	*Parameter*	*Value*
RTS	20 bytes	**ARS**	10 μs
CTS	14 bytes	**minislots** m	3
ACK	14 bytes	**FBP**	13 bytes
MAC Header	34 bytes	**MAC Header**	34 bytes

Traffic Generation Models

The service differentiation paradigm defined in the IEEE 802.11e specification has been adopted, according to which traffic flows are mapped onto four service classes, named Access Categories (AC) in the standard. In order of descending priority, the voice (VO), video (VI), best-effort (BE) and background (BK) service classes are defined. Three types of data sources have been considered to simulate the traffic flows of the four service classes.

Data Traffic Sources

The Poisson traffic model, described in Section 4.4.1 of the previous chapter, has been employed for the generation of the best-effort and background traffic. As mentioned before, a Poisson arrival process with mean rate λ messages per second has been assumed. The message length is exponentially distributed and consists on average of $\kappa = 10$ packets of fixed length L. The average offered load generated with this model is $C_{data} = \kappa\lambda L(8 \cdot 10^{-6})$ Mbps.

The data traffic flows are not delay sensitive, thus forming the lower priority service classes. A maximum tolerable delay of 5 s has been assumed for the best-effort class, whereas no delay constraint has been set for the background class.

Voice Traffic Sources

The voice traffic generation has been modeled as a two-state transition between ON and OFF periods, as shown in Figure 5.10. The time spent at the ON and OFF state is exponentially distributed with mean values of 1 s and 1.35 s, respectively, following the Brady's model for voice conversations [72]. During the ON phase, packets of 160 bytes are generated every 20 ms resulting to a CBR of 64 kbps, as defined in the G.711 voice codec [73]. No packets are generated during the OFF state, resulting to an average load of $C_{voice} = 27.23$ kbps per traffic flow.

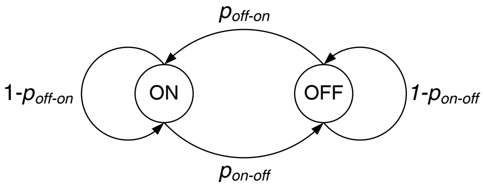


Figure 5.10: State diagram of the ON-OFF model for voice traffic generation

The packets generated during a given ON period, will be referred to as a burst of voice traffic. Since voice packets are small compared to the data packets, it has been considered that all buffered voice packets of the same burst can be transmitted within a single DQCA frame.

The voice traffic is very sensitive to delays and packets are dropped by the receiver if delivered outside a given time constraint. In the presented results, a maximum tolerated delay of 150 ms has been considered for the voice service and if exceeded the packets are dropped. To guarantee QoS, the percentage of lost voice packets should not exceed the 1% of the total transmitted voice packets.

Video Traffic Sources

A near real-time video model defined in [74] has been used for the generation of a streaming video traffic. The main concept of the model is shown in Figure 5.11. Each video frame of video arrives periodically, at a regular time interval determined by the number of frames per second (fps). Each frame is decomposed into a fixed number of slices that have a variable size that follows a truncated Pareto distribution. Each slice is transmitted as a single packet. Due to the video encoding process, there is an interarrival delay between the packets of the same frame that is also modeled by a truncated Pareto distribution.

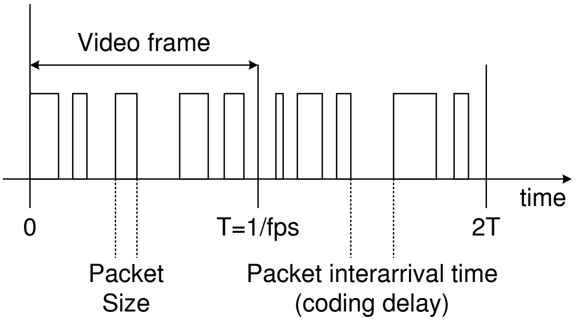


Figure 5.11: Video streaming traffic model [74]

As in the case of the voice nodes, video nodes are allowed to transmit all the buffered packets that belong to the same video frame in a single DQCA frame. The model parameters have been selected to generate a video streaming flow of 180 kbps. The traffic consists of 10 frames per second, divided into 25 packets. The packet length follows a truncated Pareto distribution with parameter $\alpha = 1.2$ and $K = 50$, resulting to packets of 50 to 200 bytes. The packet interarrival time is also Pareto distributed, with $\alpha = 1.2$ and $K = 2.5$, with values within the interval of 2.5 to 4 ms. The maximum tolerated delay for the video class has been set to 300 ms and the maximum percentage of lost video packets should remain below 1%.

Definition of Performance metrics

Finally, before proceeding to the simulation results, some brief definitions of the performance metrics will be given. The throughput and delay metrics have been also given in the previous chapter but are repeated here for convenience.

- *Throughput* is defined as the rate of transmitted bits per second and is calculated as the ratio of the total number of successfully transmitted data bits to the duration of the simulation experiment (average throughput). Throughput is evaluated at the MAC layer, meaning that the data bits include the application payload and any headers added by higher layers, whereas the MAC and PHY layer headers are considered overhead. Unless otherwise stated, the throughput values presented in the next section refer to the total system throughput, calculated as the sum of the throughput performance of all system users.

- *Relative Throughput* is defined as the ratio of the successfully transmitted useful bits to the number of generated bits. A relative throughput of 1 means that all the generated traffic is transmitted by the end of the simulation experiment, whereas smaller values mean that part of the generated traffic is not successfully transmitted, either due to network congestion or due to QoS related packet loss (i.e., packets discarded if they fail to satisfy QoS restrictions).

- *Mean message delay* is defined as the average time from the generation of a data message until its complete transmission. This time consists of the waiting time of the message in the MAC layer buffer (queuing time), the time required for the respective user to gain access to the medium (access time), and the actual transmission time of the message (including the transmission of ACK or FBP, in the case of IEEE 802.11 and DQCA, respectively). The standard deviation of the message delay is also considered, as a metric of the dispersion of the message delay experienced by users with respect to the average value.

- *Delay jitter* $J_{(i)}$ is defined as the mean deviation (smoothed absolute value) of the difference $D_{(i-1,i)}$ between the delays of two consecutive packets $i-1$ and i. Its calculation is based on the formula given in [75]:

  $$J_i = J_{(i-1)} + \left(\left|D_{(i-1,i)}\right| - J_{(i-1)}\right)/16 \ , i \geq 1 \tag{5.13}$$

  with $J_{(0)} = 0$ and $D_{(0,1)} = 0$.

- *Jain or Fairness Index F* is an indicator of the fairness for the resource allocation among users [76]. It is defined as:

  $$F = \frac{(\sum_{i=1}^{N} x_i)^2}{N \sum_{i=1}^{N} (x_i^2)} \tag{5.14}$$

  where x_i is the average throughput of the ith user and N the total number of users. It is a continuous function bounded between 0 and 1, with higher values corresponding to fairer policies. It also has an intuitive relationship with user perception. To give a straightforward example taken from [76], if 80% of the users are treated fairly and the remaining 20% are not allocated any resources, the fairness index will amount to 0.8 (or 80%).

5.3.2 CL Scheduling for Data Traffic

The first study case examines the potential performance enhancements from the application of CL-based scheduling over the main access mechanism of DQCA in the presence of data traffic. Since a single service class is considered in this scenario, the proposed service differentiation schemes are not relevant. Hence, the performance evaluation will focus on two of the algorithms, the CL-alg and the VPF-alg.

The first policy, CL-alg, schedules opportunistically users depending on their available transmission rate, estimated during the channel request process with the transmission of the ARS frame. Users with higher rates have priority and are allowed to transmit before slower users. The second policy, VPF-alg, implements a more balanced scheme that schedules users in decreasing order of their priority function value. Initially, the first proposed definition of the VPF-alg (Section 5.2.4, equation (5.6)) will be adopted, that calculates the priority function as follows:

$$\text{VPF-alg}_1 : f_{VP} = \frac{R}{R_\nu} \cdot \frac{1}{pTQ} \tag{5.15}$$

where R_ν is the maximum transmission rate (in Mbps) defined in the rate set (either R_ν=11 Mbps for the IEEE 802.11b PHY and R_ν=54 Mbps for IEEE 802.11g PHY) and pTQ the waiting position of the user in the DTQ.

Table 5.5: Summary of simulation parameters for Case Study 1

Parameter	*Value*
Number of users	$N = 20$
Number of service classes	$P = 1$ (Best-effort)
Best-effort traffic (BE)	
Traffic generation	Poisson msg arrivals, average $\kappa = 10$ packets/msg
Packet size	$L_{BE} = 2312$ bytes
Mean Offered Load	$C_{BE} = \kappa\lambda L_{BE}(8 \cdot 10^{-6})$ Mbps
PHY Layer	IEEE 802.11b and g
Coherence time	$\tau_c = 150$ ms
Evaluated Schemes	CL-alg, VPF-alg$_1$, VPF-alg$_2$
CL Overhead (per frame)	$\text{CL}_{o,CL-alg} = \text{CL}_{o,VPF-alg} = x \times TQ$ bits, $x = 2, 3$ for IEEE 802.11b and g, resp.

In order to implement the CL algorithms, the estimated transmission rates of the TQ users waiting in the DTQ must be included in the FBP, at the end of each DQCA frame. In order to map the four available rates defined in the IEEE

802.11b specification, 2 bits are required. Similarly, 3 bits are sufficient for the representation of the eight 802.11g defined rates. The required overhead, along with the main simulation parameters of this scenario are summarized in Table 5.5.

Performance over the IEEE 802.11b rate set

Initially, the IEEE 802.11b rate set has been employed for the performance comparison between DQCA and the two CL algorithms. The channel condition of the users has been modeled with the help of transmission matrices given in Section 4.4.1. Figure 5.12 shows the total throughput achieved for $N_{BE} = 20$ users as a function of the offered load. The throughput improvement gained by the use of CL scheduling is clearly marked. CL-alg achieves a maximum throughput of approximately 7.8 Mbps, providing a 224% gain with respect to the 2.3 Mbps offered by DQCA. The performance of VPF-alg is less pronounced compared to CL-alg, but still a 50% throughput gain is obtained with respect to DQCA.

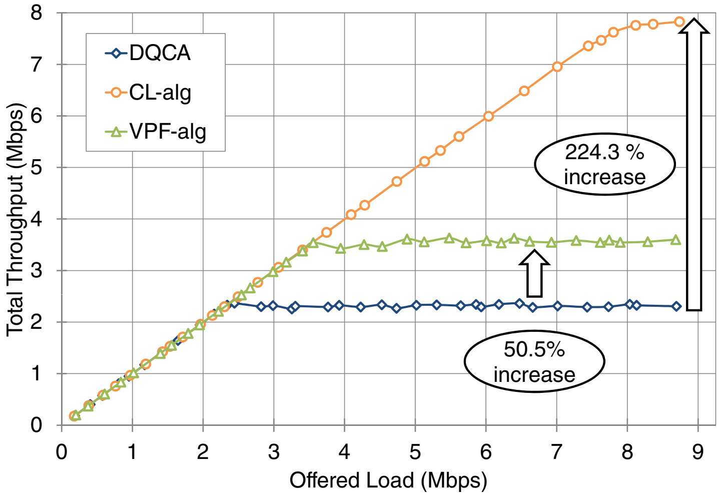
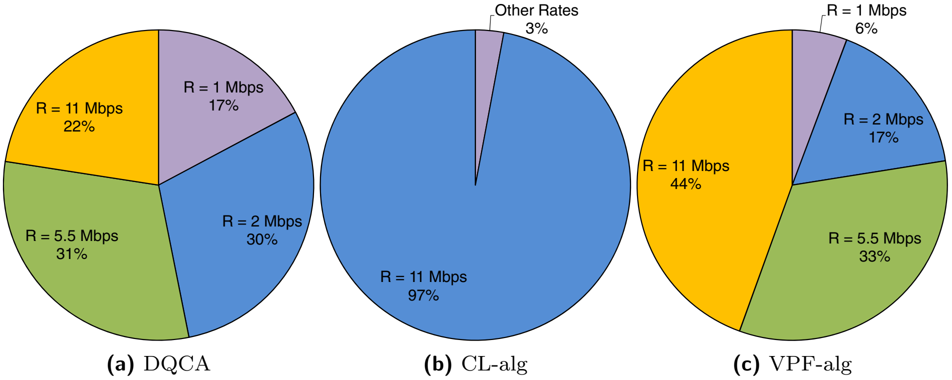


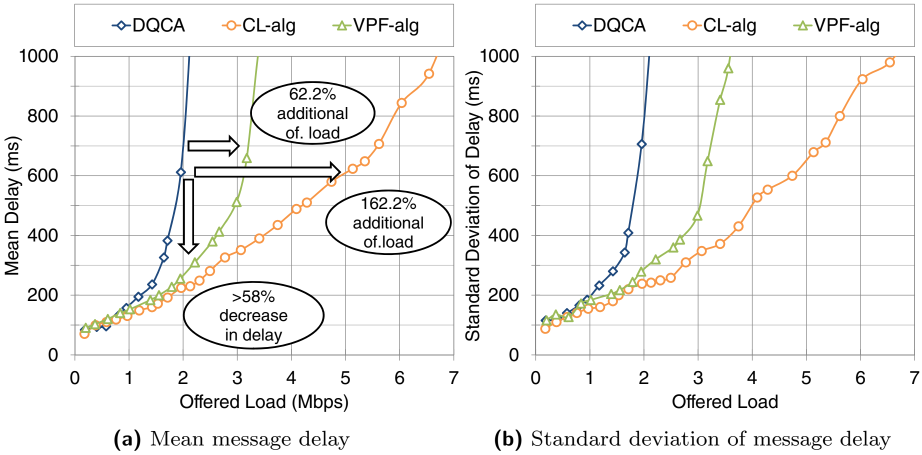
Figure 5.12: Throughput as a function of the offered load (IEEE 802.11b rates)

The throughput gain obtained with the CL algorithms is the result of the opportunistic scheduling that encourages transmissions at higher rates. To illustrate this point, the percentage of frames transmitted by the four available rates for each algorithm has been plotted in Figure 5.13. In the case of DQCA (plot (a)), where scheduling is independent of the transmission rate, the rate utilization approaches the corresponding steady state probability of the channel transmission matrix ($\alpha_b = [0.1764, 0.2941, 0.2941, 0.2353]$, as explained in Section 4.4.1). On the other hand, CL-alg strongly promotes transmissions by faster users, resulting to the 97% of transmissions being conducted at the maximum rate of 11 Mbps (b). Finally, the VPF-alg constitutes a milder opportunistic scheme that increases the percentage of high rate transmissions without completely blocking transmissions at lower rates (c).



(a) DQCA **(b)** CL-alg **(c)** VPF-alg

Figure 5.13: Percentage of frames transmitted at the IEEE 802.11b rates



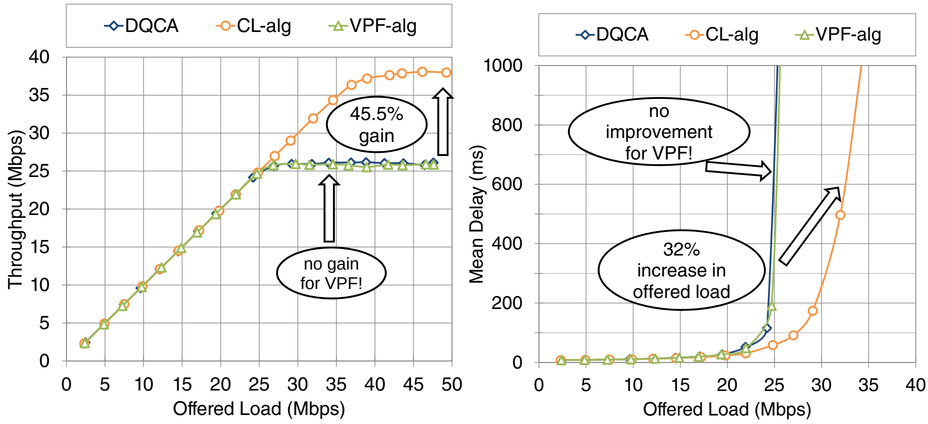
(a) Mean message delay **(b)** Standard deviation of message delay

Figure 5.14: Delay performance as a function of the offered load (IEEE 802.11b rates)

The delay performance has been plotted in Figure 5.14. Comparing the mean delay for the three schemes (plot (a)), it can be observed that even though they demonstrate similar delay under low traffic, CL schemes outperform DQCA as the offered load grows. For instance, for 2 Mbps of generated traffic, both CL scheme show a decrease in the mean delay of at least 58%, from 600 ms to less than 250 ms. A different way of examining Figure 5.14 (a) is by observing the traffic load that can be tolerated by each algorithm, in order to maintain the mean delay below a given level. For a delay threshold of 600 ms, for example, the supported offered load is approximately 2 Mbps for DQCA, 3.2 Mbps for the VPF-alg (aprox. 62% increase) and more than 5 Mbps for CL-alg (aprox. 162% increase). The standard deviation of the delay is depicted in Figure 5.14 (b). As the traffic grows, the std. also increases for all schemes, even though the CL algorithms reduce the dispersion of the message delay.

Performance over the IEEE 802.11g rate set

The respective plots have been obtained for the IEEE 802.11g rate set. The CL-alg achieves a maximum throughput of 38 Mbps, offering a 45% increase in performance with respect to the 26 Mbps obtained by DQCA, as shown in Figure 5.15 (a). This is a significant enhancement in throughput, although it appears less impressive compared to the relative gain of 224% obtained in the IEEE 802.11b scenario (Figure 5.12). In the case of the VPF, the results are more surprising, since the obtained results are very close to the performance of DQCA. Similar observations can be made regarding the mean message delay, depicted in Figure 5.15 (b). The CL-alg can support 32% more traffic load with respect to DQCA without CL, yielding an average delay below 500 ms, whereas the VPF-alg performance is indistinguishable from that of DQCA.

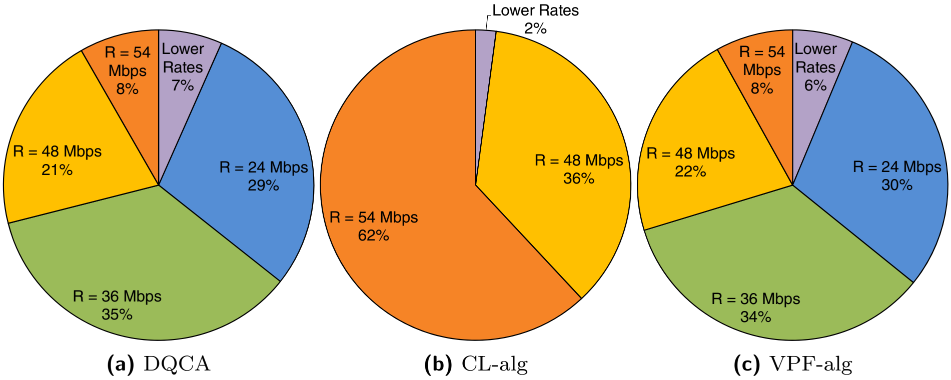


(a) Throughput **(b)** Mean message delay

Figure 5.15: Performance as a function of the offered load (IEEE 802.11g rates)

In order to better understand this behavior, the percentage of frames transmitted by the eight available rates has been plotted in Figure 5.16. In the case of DQCA (a), the 85% of transmissions are performed at either 24, 36 or 48 Mbps, revealing that with the employed channel model the link state of the users is relatively good for the majority of time. The CL-alg raises the number of transmissions at the maximum rate of 54 Mbps from 8% to 62% and practically blocks transmissions at rates below 48 Mbps (b). On the other hand, the VPF-alg slightly increases the percentage of transmissions at higher rates with respect to DQCA, however the difference is barely noticeable, resulting to the similar performance of the two schemes (c).

The main reason behind the poor performance of the VPF-alg is the selected priority function $(R/R_{min}) \cdot (1/pTQ)$. Despite being suitable for the IEEE 802.11b scenario, in this case, where there are eight available rates and the channel condition is fairly good, the VPF fails to provide the desired priority to the faster users. For example, consider a snapshot of the system where there are 3 users waiting in the



(a) DQCA **(b)** CL-alg **(c)** VPF-alg

Figure 5.16: Percentage of frames transmitted at the IEEE 802.11g rates

DTQ and the last one ($pTQ = 3$) has an available rate of 54 Mbps. According to the defined VPF, the third user would be scheduled to transmit first[1] if the user at the head of the DTQ ($pTQ = 1$) had an available rate $\leq$ 12 Mbps. This does not occur very often in the current scenario, since the available rates of the users are below 24 Mbps for only a 7% of the time (Figure 5.16 (a)).

Summarizing, the selected VPF for the IEEE 802.11g scenario does not further enhance the DQCA performance in terms of the maximum overall throughput, mainly because the system is homogeneous with all users having high available rates during the majority of the time. In any case, the VPF-alg is a scheduling concept that is not restricted to a single function definition; different functions can be selected to alter the scheduling priority. To illustrate this point, the VPF-alg performance has been evaluated by employing the second VPF definition given in Section 5.2.4 ((5.7)), which is repeated here for convenience:

$$\text{VPF-alg}_2 : f_{VP} = \frac{\alpha^r}{pTQ}, \alpha \in \mathbb{N}. \tag{5.16}$$

where r is an integer within $[1, 8]$ corresponding to the eight IEEE 802.11g rates (from smallest to highest) and pTQ the position of the users in the DTQ. The parameter α can be assigned values ≥ 1 to attain different degrees of opportunistic scheduling.

Figure 5.17 shows the maximum throughput as a function of the offered load for three different values of the VPF-alg parameter α, namely α =2,3 and 4. The VPF-alg performance is bounded between the maximum throughput attained by DQCA and the CL-alg, marked by dotted lines as lower and upper bounds, respectively. Higher values of α emphasize the role of the available transmission rate in the transmission order, whereas smaller values provide a compromise between throughput and fairness. Similar results are obtained for the delay performance, not presented

[1] The rate of the second user could also affect the scheduling order, but assume, for simplicity, that in this example it has the lowest rate of 6 Mbps.

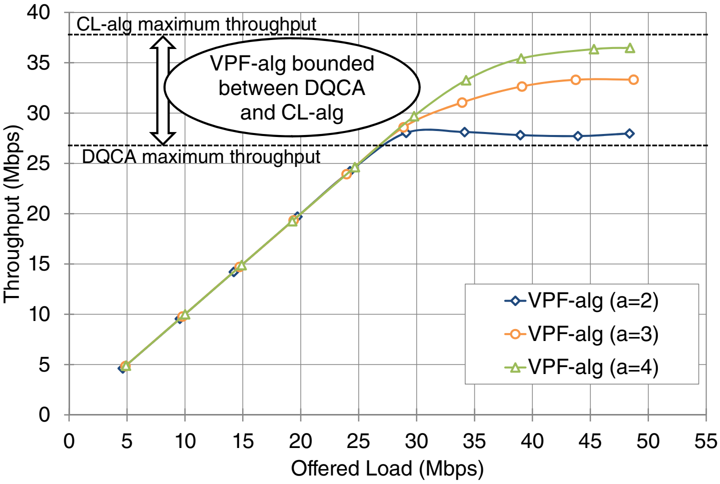
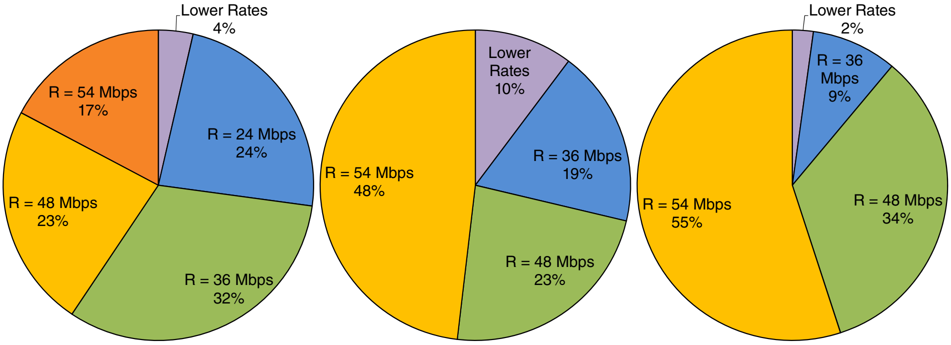


Figure 5.17: VPF-alg throughput as a function of the offered load (IEEE 802.11g rates)

here for brevity, which is again bounded within the DQCA and CL-alg performance. The percentage of the transmitted frames per rate are depicted in Figure 5.16. As expected, the percentage of transmissions at high rates has increased compared to DQCA (Figure 5.16 (a)) and grows for higher values of the parameter α.



(a) VPF-alg ($\alpha = 2$) **(b)** VPF-alg ($\alpha = 3$) **(c)** VPF-alg ($\alpha = 4$)

Figure 5.18: Percentage of frames transmitted at the IEEE 802.11g rates

Performance over a heterogeneous channel scenario

Finally, a third case has been studied to provide further insight to the performance of the CL scheduling algorithms. In this case, a heterogeneous scenario has been considered in which not all users share the same channel model. In particular, 15 of the 20 total users follow the IEEE 802.11g channel model, as before, and the other 5 users are assumed to employ a constant rate of 6 Mbps. In other words, most users have a fairly good time-varying channel whereas the remaining users suffer from a harsh link condition that does not improve during the simulation time. This scenario has been selected in order to study how the CL algorithms handle fairness when the available resources among users are not equal.

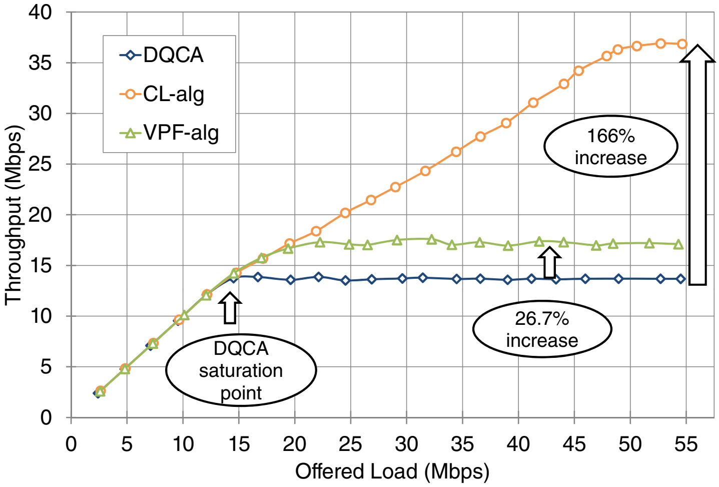


Figure 5.19: Throughput versus the offered load under a heterogeneous channel

Figure 5.19 shows the obtained throughput for DQCA, CL-alg and VPF-alg with the priority expression VPF-alg$_1$. The presence of the five low-rate users affects the performance of DQCA which achieves a maximum throughput of 13.5 Mbps, approximately 12 Mbps lower than the throughput achieved under a homogeneous channel (Figure 5.15). The CL-alg throughput performance remains practically unaffected, with the maximum throughput approaching 37 Mbps, yielding a 166% increase with respect to DQCA. The VPF-alg, on the other hand, provides a 26% increase over DQCA, which is interesting since in the homogeneous scenario it didn•t accomplish any improvement. The difference is that in this case, the the VPF-alg grants different priorities to fast and low-rate users whereas in the previous scenario all users had very similar channel conditions.

The mean message delay has been plotted in Figure 5.20. The two CL schemes manage to maintain the mean delay low for higher amounts of offered traffic. In particular, for a delay below 500 ms, 240% more traffic load compared to DQCA can be supported by the CL-alg and 40.3% by the VPF-alg. An interesting observation is that the delay curve for both CL schemes attains a local maximum for an offered load of around 15 Mbps and then decreases again until the respective saturation point of each algorithm. It is not a coincidence that this anomaly appears around the saturation point of DQCA (i.e., just below 15 Mbps of offered load), as it will be explained next.

When the system is not saturated, the throughput is equal to the offered load, meaning that all the generated traffic eventually gets transmitted (i.e., relative throughput of 1). As expected, delay grows as the traffic increases, mainly due to

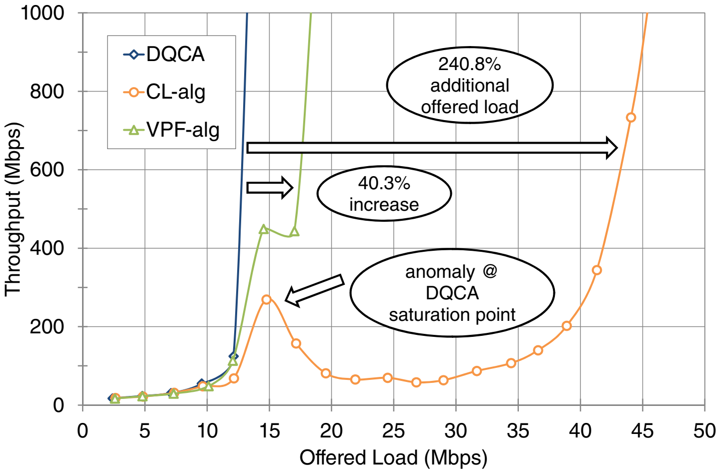


Figure 5.20: Mean message delay versus the offered load under a heterogeneous channel

the prolongation of the waiting time of the users in the DTQ and the data messages in the users' buffers. Once saturation is reached, the incoming traffic exceeds the system capacity and messages begin to accumulate in the buffers. In the case of DQCA that treats all users equally, the message delay under saturation suffers a very steep increment.

The CL schemes, however, exhibit a different behavior as a result of the opportunistic scheduling. Before saturation, all users get a chance to transmit, even though faster users are scheduled before slower users. Nevertheless, once the system becomes saturated, slower users are practically blocked from transmission (especially in the case of the CL-alg). Consequently, most transmissions take place at high data rates and the reduced transmission time produces a decrease in the average message delay with respect to the delay performance just before the saturation point (represented as the local maximum at around 15 Mbps of offered load). Eventually, messages also accumulate in the buffers of the fast users leading to a steep increase in delay (for a traffic load above 45 Mbps, in the case of the CL-alg). Bear in mind that the delay metric takes into consideration only the successfully transmitted messages. As a result, the accumulated messages in the buffers of the slower users do not have an impact on the plotted delay performance since they never get to be transmitted.[2]

Summarizing, the application of CL scheduling seems to improve the overall performance even under heterogeneous channel conditions. However, this enhancement is not experienced in an equal way by all users. In order to gain more insight in

[2] The reason why this anomaly did not appear in the previous scenarios is that in this case there is a set of slow users that have no chance of channel improvement. In the previous scenarios, the channel condition of the all users was time-varying and users with low available rates were blocked for transmission only for a limited amount of time until their channel improved.

the trade-off between the overall performance and fairness among users, plots of the Jain index and the average throughput per user are presented in continuation.

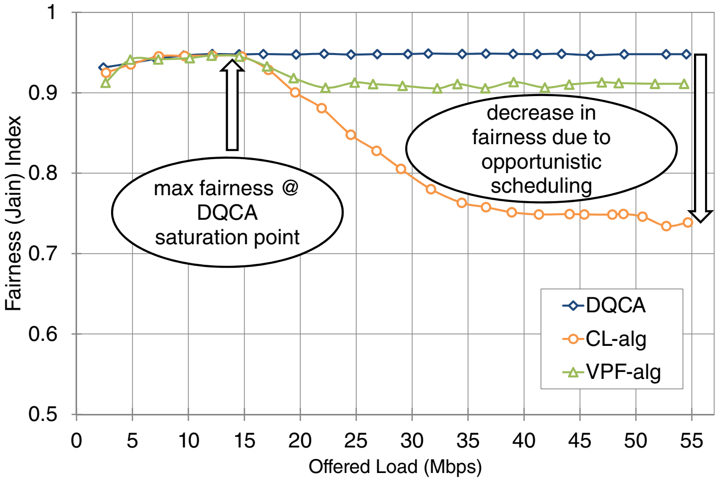
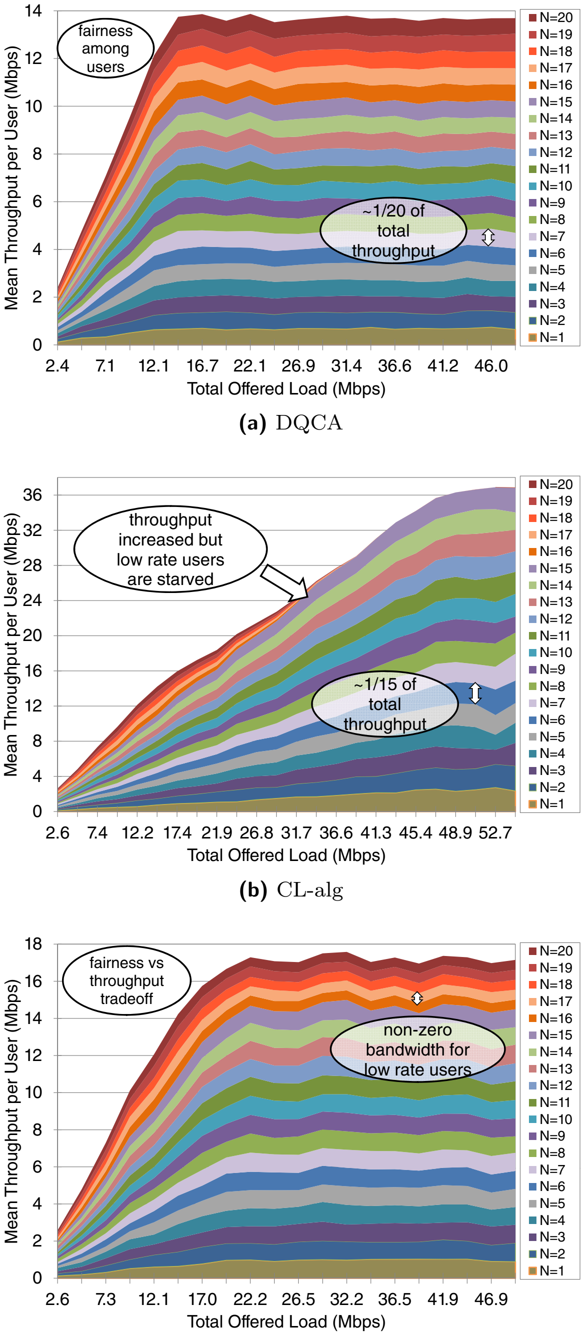


Figure 5.21: Fairness comparison under a heterogeneous channel

The Jain index for the three schemes has been plotted in Figure 5.21. For low traffic loads, the Jain index is approximately 0.95 for all schemes, meaning users are treated with considerable fairness. DQCA maintains fairness even when the traffic load grows, since it implements a FIFO policy and all users are served regardless of their available rate. On the contrary, fairness gradually decreases for the CL-alg and seems to stabilize at the value of 0.75. This result can be interpreted in an intuitive way, in line with the example provided with the definition of the Jain index in Section 5.3.1. An index of 0.75 means, in this case, that the 75% of users is treated fairly whereas a 25% of the users are starved. This is exactly what the scheduling policy of the CL-alg does: resources are shared among the 15 users with good channel conditions (with all users treated fairly in the long term due to the time-variability of the channel) and the remaining 5 users are practically denied access due to their constant rate of 6 Mbps. The fairness index for the VPF-alg is bounded by the respective values of DQCA and the CL-alg and maintains a value above 0.9.

Finally, to better visualize the allocation of resources, the average throughput per user has been plotted as a function of the total offered load in Figures 5.22 (a) to (c), for DQCA, CL-alg and VPF-alg, respectively. The five low-rate users are plotted at the upper part of each figure (N_{BE}=16 to 20). The fair resource allocation is evident in the case of DQCA, since the throughput of each user corresponds to the 1/20 to the total system throughput. CL-alg has a balanced performance for low values of offered load. However, as the traffic grows, low-rate users are assigned less resources and eventually are completely denied the opportunity to transmit, whereas the remaining 15 users get an equal share of the system throughput. Finally, the VPF-alg represents the intermediate solution in which high-rate users receive preferential treatment without completely starving the low-rate users.



(a) DQCA

(b) CL-alg

(c) VPF-alg

Figure 5.22: Average throughput per user versus the total offered load

5.3.3 CL Scheduling for Voice and Data Traffic

So far, the presented scenarios have considered only best-effort data traffic load. However, the widespread usage of VoIP applications makes it imperative to examine the system performance under the presence of heterogeneous voice and data traffic sources. There are several important differences between the two service classes that should be taken into consideration in order to achieve efficient scheduling. First, the voice traffic is typically composed by silent (OFF) periods alternated by a burst of relatively small packets, generated in constant time intervals. Then, unlike best-effort traffic, the voice service has specific QoS requirements regarding the tolerated delay and the packet loss rate.

A summary of the most significant parameters employed in this study case is given in Table 5.6. For simplicity, it has been assumed that each user generates a single flow of either best-effort data or voice traffic. The number of data users (or flows) is denoted by N_{BE} whereas N_{VO} represents the number of voice users in the system. Each data user produces a fixed load of 1 Mbps, following the Poisson model adopted in all the previous study cases. Each voice flow corresponds to an average traffic rate of 27.23 kbps, with no packet generation during OFF periods and a CBR of 64 kbps during the ON phase. For the presented results of this section, $N_{VO} = 20$ voice users have been considered, which is a reasonable number of active VoIP calls in a WLAN office scenario.

Figure 5.23 compares the performance of DQCA and IEEE 802.11e under a mixed traffic scenario composed of $N_{BE} = 10$ data and $N_{VO} = 20$ voice users. The relative throughput of the voice class (i.e., ratio of transmitted to generated bits) is depicted as a function of the size of the best-effort data packets that varies from small packets of $L_{BE} = 100$ bytes up to 2312 bytes, with a fixed generated load of 1 Mbps per user. Ideally, to ensure a high quality voice connection, the relative voice throughput should have a value that approaches 1, meaning that all generated traffic is successfully transmitted and no packets are discarded due to time constraints. This is achieved by DQCA but is not the case for IEEE 802.11e that fails to provide QoS, especially as the size of the data packets grows. The effectiveness of DQCA is mainly due to the efficient channel access mechanism that practically eliminates packet collisions even for a large number of participating users.

As far as the size of the data packets is concerned, it can be observed that mixed traffic consisting of long (i.e., above 1000 bytes) data packets and short voice packets of less than 200 bytes has a negative impact on the performance, especially for IEEE 802.11e. Even though the same amount of aggregated best-effort traffic is present and equal to 10 Mbps on average, the voice performance deteriorates when the data traffic consists of longer packets. This occurs because the additional time required for the transmission of longer data packets produces a proportional increase in the waiting time of buffered voice packets in queue for transmission, which can be detrimental for the voice QoS, given the stringent requirements for the maximum tolerated delay (150 ms) and voice packet loss and ($< 1\%$). DQCA seems unaffected by the packet size in this particular case; however, as it will be shown in continuation, longer data packets do affect the voice class performance

Table 5.6: Summary of simulation parameters for Case Study 2

Parameter	*Value*
Number of users	$N = N_{BE} + N_{VO}$
Number of service classes	$P = 2$ (Best-effort and Voice)
Best-effort traffic (BE)	
Number of data users	varying, $N_{BE} = [10, 60]$
Traffic generation	Poisson msg arrivals, average $\kappa = 10$ packets/msg
Packet size	varying, $L_{BE} = [100, 512, 1000, 1500, 2312]$ bytes
Mean offered load	1 Mbps per user
QoS demands	maximum delay of 5 s, no lost packets
Voice traffic (VO)	
Number of voice users	$N_{VO} = 20$
Traffic generation	Brady's ON-OFF model with G.711 voice codec 160 bytes/20 ms, average load of 27.23 kbps/user
Mean offered load	27.23 kbps per user
QoS demands	maximum delay of 150 ms, lost packets < 1%
PHY Layer	IEEE 802.11g
Coherence time	$\tau_c = 150$ ms
Evaluated Schemes	CL-alg, SP-alg, CLSP-alg, VPF-alg$_3$
CL Overhead (per frame)[a]	$\mathrm{CL}_{o,CL-alg} = 3 \times TQ$ bits, $\mathrm{CL}_{o,SP-alg} = 3$ bits, $\mathrm{CL}_{o,CLSP-alg} = 3 \times \sum_{p=1}^{2} TQ_p + 3$ bits, $\mathrm{CL}_{o,VPF-alg_3} = 4 \times TQ$ bits

[a] Overhead values derived from Section 5.2, for $x = 3$ bits per rate (IEEE 802.11g rate set), $m = 3$ DQCA control minislots and $y = 1$ bit per service class

when the traffic load is increased.

So far, DQCA seems to provide QoS for the voice class under heterogeneous traffic scenarios. Nevertheless, the results of Figure 5.23 correspond to a relatively low traffic scenario with an average of 10 Mbps of aggregated data traffic and 544.4 kbps of aggregated voice traffic. This traffic load is below the capacity of the DQCA system that, for the considered channel model, has been calculated

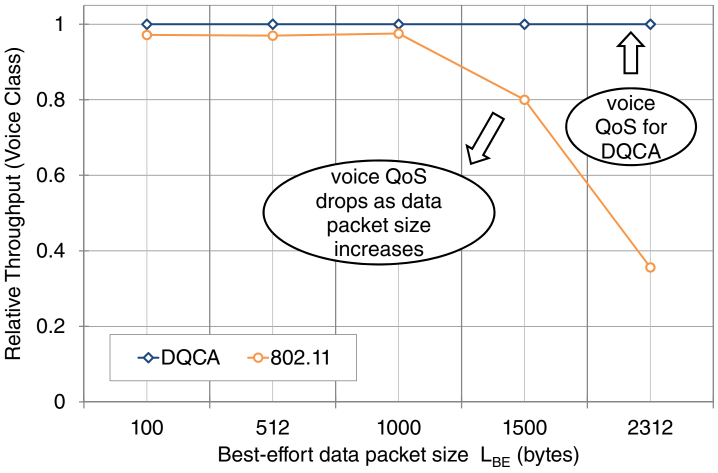
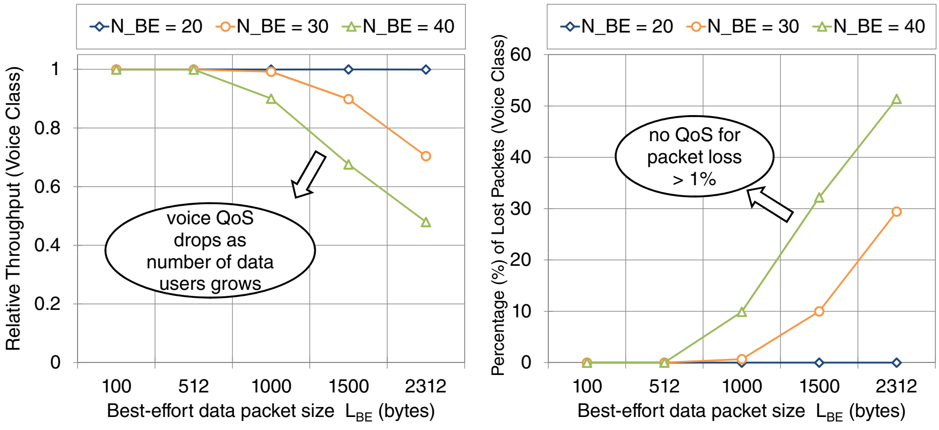


Figure 5.23: DQCA versus IEEE 802.11e relative throughput performance for the voice class ($N_{BE} = 10$ data users, $N_{VO} = 20$ voice users)

to approximately 25 Mbps (this result has been represented in Figure 4.9(b) of Chapter 4, Section 4.4.3 as a function of the packet size). The next set of plots examines the DQCA voice service performance under heavier traffic conditions, achieved by adding more best-effort traffic flows of 1 Mbps each. In particular, the case of $N_{BE} = 20$, 30 and 40 data users has been considered, whereas the number of voice flows has been fixed to $N_{VO} = 20$.



(a) Relative throughput

(b) Percentage of lost packets

Figure 5.24: DQCA versus IEEE 802.11e relative throughput performance for the voice class under heavier data traffic load ($N_{VO} = 20$ voice users)

Figures 5.24 (a) and (b) depict the relative throughput and the percentage of lost packets for the voice class, respectively, as a function of the best-effort data packet size. For $N_{BE} = 20$, DQCA manages to provide QoS to the voice class regardless of the packet size. However, when more data users are added the performance degradation is evident. The decrease in the voice throughput is the result of two factors; first, part of the generated traffic begins to accumulate in the user buffers as the system becomes more congested and second, transmitted voice packets are dropped by the receiver due to delayed delivery. It should not be forgotten that one of the QoS requirements of the voice service is that only a certain amount of packet delay can be tolerated and packets that exceed this maximum are discarded despite being successfully received. In this scenario, the maximum tolerated delay has been set to 150 ms and the acceptable percentage of lost packets should be below 1%.

Summarizing, the basic DQCA operation cannot provide QoS for the voice service under all circumstances: even though multiple voice flows can be successfully supported under mixed voice and data scenarios when the traffic load is below the system capacity, voice performance drops significantly under more congested traffic conditions. The need for QoS guarantees for voice and multimedia applications, in general, is a strong motivation factor for the introduction of CL scheduling algorithms. In the previous section, channel-aware opportunistic CL scheduling schemes applied over DQCA in order to enhance the system performance. Here, a number of CL schemes that take into account both channel condition and service class information will be considered.

The detailed description of the four CL policies can be found in Section 5.2 and are briefly explained next:

- *CL-alg*: a strict opportunistic scheme that gives priority to the user with the highest available transmission rate, without differentiating among traffic of different service classes.
- *SP-alg*: an algorithm that differentiates between service classes (voice and best-effort data in this case) and gives priority to voice users. FIFO scheduling is implementing among traffic of the same class.
- *CLSP-alg*: a combination of the above schemes that gives priority to voice users and, in addition, perform opportunistic scheduling among the users of the same class.
- VPF-alg: a more balanced scheme that employs a priority function to determine the transmission order. The priority function evaluated in this section is VPF-alg$_3$ (Section 5.2.4, equation (5.10)) defined as:

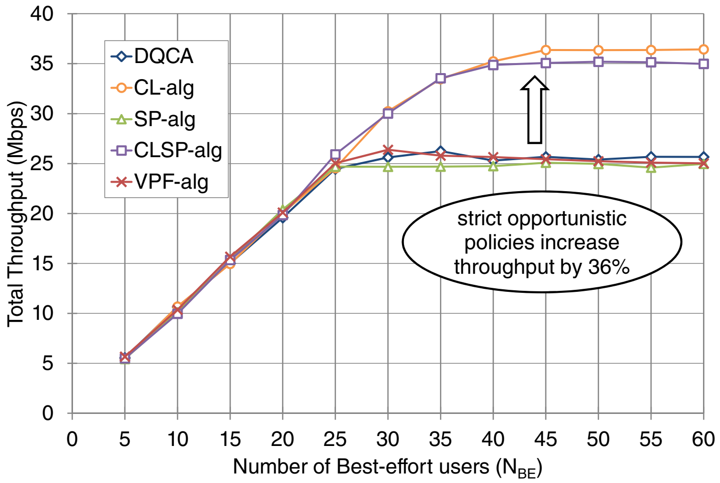
$$\text{VPF-alg}_3 : f_{VP} = \alpha \cdot 2^p + (1-\alpha) \cdot \frac{2^r}{pTQ}, 0 \leq \alpha \leq 1 \qquad (5.17)$$

  where $r \in [1, \nu]$ is an integer index of the available transmission rates, with $\nu = 8$ for the IEEE 802.1g rate set . In particular, $r = 1$ corresponds to the minimum (6 Mbps) and $r = \nu$ to the maximum (54 Mbps) rate of the rate set.

The integer p is the service class identifier with values 2 and 4 for best-effort and voice traffic, respectively. The integer pTQ indicates the position of the user in the DTQ and the tunable parameter α takes values ≤ 1 to attain different degrees of opportunistic scheduling.

The following setup has been adopted for the remaining of this section. As before, a fixed number of voice users $N_{VO} = 20$ has been considered, generating an average traffic load of approximately 545 kbps. The number of best effort data users is gradually increased, from $N_{BE} = 5$ to 60, with each user generating 1 Mbps of data in packets of $L_{BE} = 2312$ bytes. Thus, the performance of the voice service can be observed as a function of the increasing network traffic. Unless otherwise stated, the parameter α for the VPF-alg has been set to 0.6.

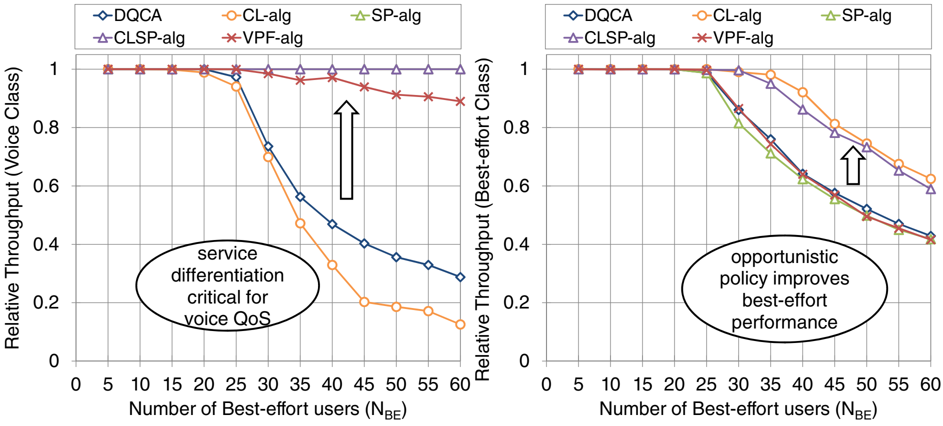
Figure 5.25 depicts the aggregated voice and data throughput as a function of the number of data users. Comparing the maximum achieved throughput obtained as the traffic load grows, the evaluated schemes can be divided into two groups: the first consists of two schemes that achieve a throughput above 35 Mbps and the second includes the remaining three algorithms that yield a throughput of approximately 25 Mbps. A closer look reveals that the more throughput efficient schemes are the opportunistic CL-alg and CLSP-alg that give priority to high rate users. DQCA and SP-alg do not give preferential treatment to faster users, whereas the mild opportunistic policy of VPF-alg does not seem to provide a clear advantage over the other schemes.



strict opportunistic policies increase throughput by 36%

Figure 5.25: Total system throughput

Nevertheless, the total throughput does not reveal much information on the performance of the QoS-demanding voice service class. For this reason, the relative throughput for voice and data services is given separately in Figures 5.26 (a) and (b), respectively. The first interesting and expected observation is that service differentiation among the voice and data traffic flows is necessary in order to provide



(a) Voice Class

(b) Best-Effort Class

Figure 5.26: Relative throughput performance per service class

QoS guarantees. As shown in plot (a), the two schemes that assign an absolute priority to voice traffic flows, SP-alg and CLSP-alg, achieve a relative voice throughput of 1, even when $N_{BE} = 60$ data users are present (corresponding to a congested scenario with 80 users in the system and a generated traffic load of approximately 60.5 Mbps). On the contrary, DQCA and CL-alg that do not distinguish between the two service classes fail to provide QoS to the voice service when the number of data users exceeds 20 and 15, respectively. The VPF-alg with $\alpha = 0.6$ implements a milder service differentiation scheme with voice traffic assigned some level of priority over best-effort data and manages to support up to 25 data users.

On the other hand, the best-effort traffic does not have any particular QoS constraints; however it constitutes the larger part of the system traffic load. Opportunistic scheduling, adopted by CL-alg and CLSP-alg, enhances significantly the best-effort performance (Figure 5.26 (b)). Through a joint observation of the voice and data throughput performance it can be deduced that the most efficient scheme is CLSP-alg, which combines service differentiation and opportunistic scheduling within each service class. The former policy ensures QoS for the voice traffic and the latter increases the average rate employed for transmissions, in such a way that voice QoS provisioning is achieved with only a slight reduction on the best-effort traffic throughput.

Figure 5.27 shows the percentage of lost voice packets as a function of the number of data users. As mentioned before, losses that exceed 1% of the generated packets cause a noticeable deterioration on the quality of a voice conversation. The presented results are consistent with the relative throughput of the voice service (Figure 5.26 (a)). The useful information provided by this plot is that the degradation of the voice throughput is mainly due to the high percentage of lost packets. In other words, the problem is not the accumulation of voice packets in the buffers due to network congestion, since voice users get access to the channel, but the fact

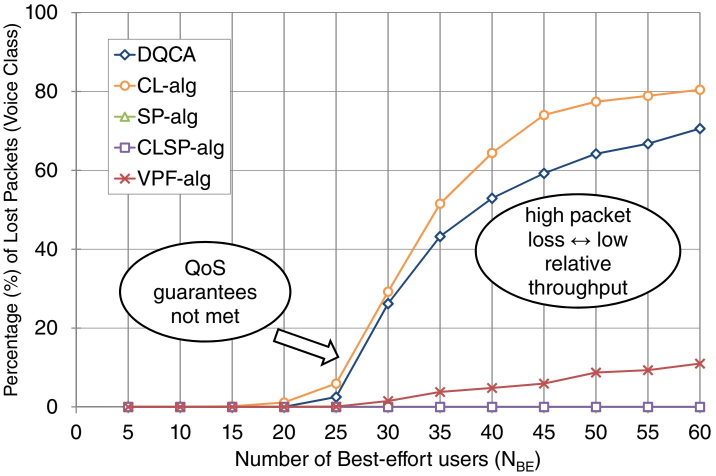


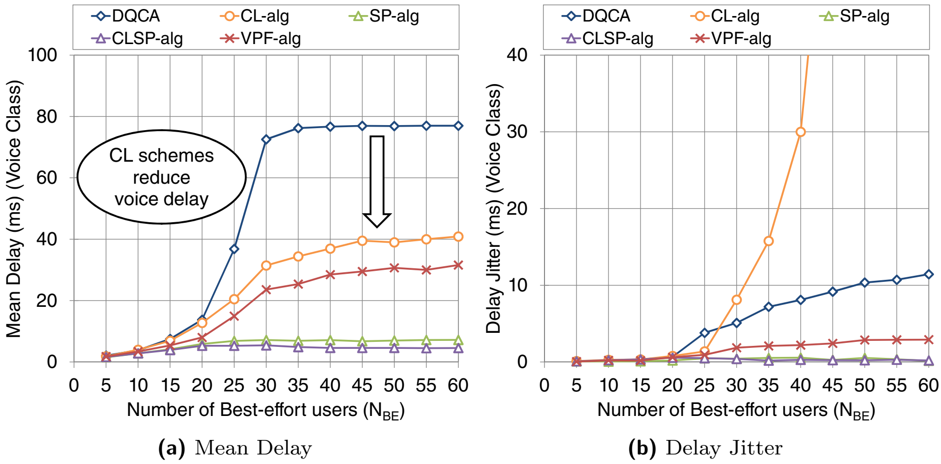
Figure 5.27: Percentage of lost packets for the voice class

that even though most voice packets are transmitted successfully, they are often discarded at the received for failing to meet the specified QoS time-constraints. This is why reducing the packet delay of the voice class is the key to providing QoS guarantees to time sensitive applications.

Figure 5.28 depicts the delay performance of the voice class in terms of mean delay and jitter (plots (a) and (b), respectively). The application of CL scheduling appears to have a positive impact on the average delay since all four CL algorithms perform better that DQCA. In any case, it should not be forgotten that the delay statistics consider only the packets received within the maximum tolerated delay restriction of 150 ms. The mean delay of DQCA, for example, that stabilizes at 77 ms as the number of data users grows, corresponds to less than half of the transmitted voice packets, given that the respective packet loss percentage is above 50% (Figure 5.27). Similarly, the mean delay of CLSP-alg and SP-alg is very low, below 7 ms, for all voice packets since no packets are lost thanks to the strict service differentiation policy.

Delay jitter is another important metric for real-time traffic application, expressing the variation in the delay between consecutive packets. High jitter values can cause significant degradation on the voice service quality. Again, CLSP-alg and SP-alg exhibit the best jitter performance below 1 ms, followed by VPF-alg and DQCA. The worst jitter performance is achieved by CLSP-alg, which is expected since the transmission of various large data packets may be interposed between two consecutive voice packets.

The last metric employed for the comparison of the four CL schemes is the Jain index, presented in Figure 5.29. The fairness is calculated from a system point of view, considering that for maximum fairness all users should be able to transmit the same percentage of the packets in their buffers, regardless of the type of service.



(a) Mean Delay **(b)** Delay Jitter

Figure 5.28: Delay performance of the voice class

In other words, a Jain index of 1 would correspond to the case where voice and data users achieve the same relative throughput. The ideal case of course would correspond to a relative throughput of 1 for all users, translated as a throughput of 1 Mbps per data user and 27.23 kbps per voice user. This occurs under low traffic (i.e., when the number of data users is low). As traffic increases, fairness drops since some users are allocated more resources than others.

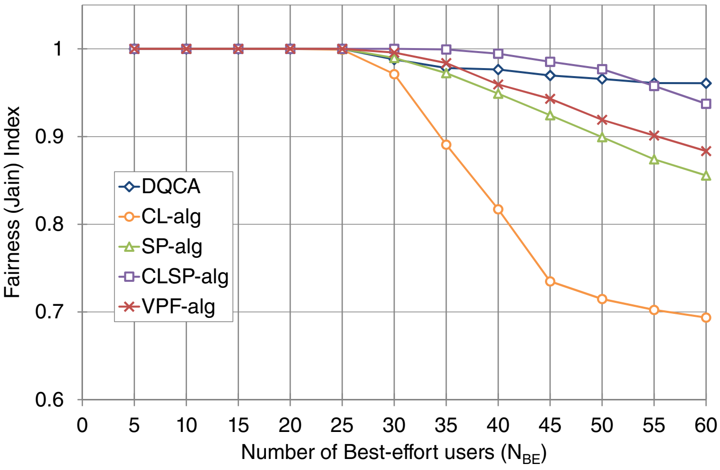
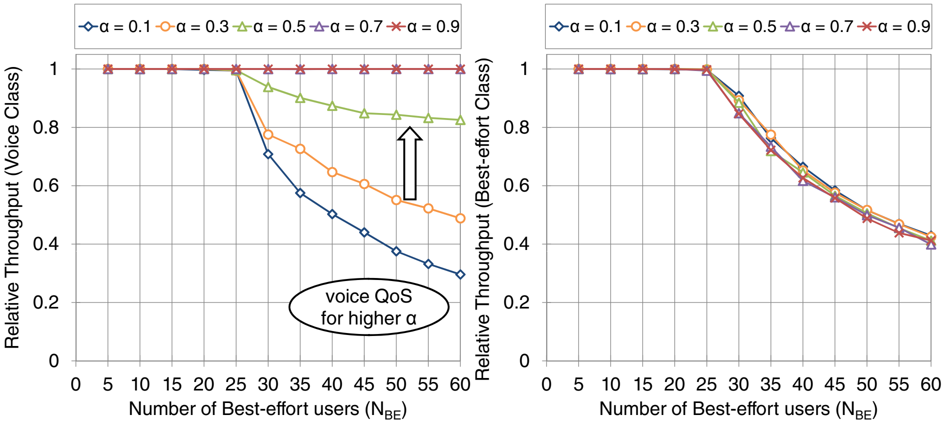


Figure 5.29: Fairness performance for the voice class

DQCA is the fairer scheme in the long term due to its FIFO scheduling policy. It is interesting to observe that CLSP-alg comes second in fairness even though it implements an unfair policy. This occurs because, thanks to the algorithm's efficient scheduling, throughput is increased for both voice and data services so that all users are relatively satisfied with the resource allocation. That is the reason why the fairness index for CLSP-alg is higher that DQCA for $N_{BE} \leq 50$ data users. Nevertheless, CLSP-alg fairness is decreased as the system approaches saturation. Then, the best-effort relative throughput drops since data users tend to be deprived of service as a result of the priority assigned to the voice class. VPF-alg with $\alpha = 0.6$ is the third more fair scheme, although its performance depends on the selection of α, as it will be discussed next. The least fair schemes are the SP-alg and the CL-alg, because they fail to provide a balanced resource allocation between the two service classes. Indeed, going back to Figure 5.26, the difference between the relative throughput performance between voice and best-effort service is more pronounced for these two schemes.



(a) Voice Class **(b)** Best-Effort Class

Figure 5.30: Relative throughput performance of VPF-alg for different values of parameter α

To complete this case study, the performance of the VPF-alg different values of the parameter α has been evaluated. According to the employed VPF definition, the parameter α weights the role of the service class of a traffic flow, giving priority to voice over data, and $(1 - \alpha)$ determines the influence of the mild opportunistic scheduling that takes into account the available transmission rate and the position of a user in the DTQ. Therefore, higher values of α enhance the relative throughput of the voice service, depicted in Figure 5.30 (a). This has the opposite effect on the best-effort relative throughput (Figure 5.30 (b)). Nevertheless, since voice traffic is significantly lower than best-effort traffic, the cost on the best-effort service is not very high.

5.3.4 CL Scheduling for Multiple Service Classes

In the previous section, the enhancement obtained with the application of four CL-based scheduling schemes over the DQCA protocol has been studied in the presence of mixed voice and data traffic. In this study case, the previous setup is extended to include traffic that can be mapped onto four service classes of different characteristics and QoS constraints. Although the main concept is the same, this scenario allows for a more profound comparison between the proposed schemes, emphasizing the trade-off between maximum throughput and QoS provisioning.

The service differentiation paradigm of IEEE 802.11e has been adopted. In particular, four classes are defined for background, best-effort, video and voice traffic, with service class ids (p) from 1 to 4, respectively. Voice traffic is assigned the highest priority due to stringent delay constraints, with a maximum tolerated delay of 150 ms per voice packet, after which the packet is dropped. For the video class, the maximum delay has been set to 300 ms per packet. Best-effort data can tolerate a delay up to 5 s and finally background data has no delay constraints. A summary of the most significant parameters employed in this study case is given in Table 5.7.

The four CL-based algorithms evaluated in this section are, again, the strict opportunistic CL-alg, the service-aware SP-alg, CLSP-alg that combines the two former policies and VPF-alg. The priority function employed in this section is VPF-alg$_3$ (Section 5.2.4, equation (5.11)) defined as:

$$\text{VPF-alg}_4 : f_{VP} = \alpha \cdot \frac{p}{P} + (1-\alpha) \cdot \frac{R}{R_\nu} \cdot \frac{1}{pTQ}, 0 \leq \alpha \leq 1 \tag{5.18}$$

where R_ν is the maximum rate defined in the rate set (R_ν=54 Mbps in IEEE 802.11g), pTQ the user position in the DTQ, p the service class identifier and α is a tunable weighting factor within $[0, 1]$. Unless otherwise stated, the value of $\alpha = 0.5$ has been employed in the results presented in this section.

As before, it has been assumed for simplicity that each user generates a single type of traffic flow. At the beginning of the simulated scenario all traffic is generated by $N_{BK} = 10$ background users, resulting to an average load of 18 Mbps. The background traffic has been maintained constant throughout the simulation and, in addition, traffic flows of the other three service classes have been gradually introduced. From this point on in this section, when referring to the number of traffic flows per service class it will be meant the voice, video and best-effort users present at the system, in addition to the 10 background traffic flows. This number, denoted by N_{tf}, is marked on the x-axis of all the figures presented in this study case (Figures 5.31 to 5.41).

In particular, for $N_{tf} = 0$, at the beginning of the x-axis, the offered load consists of the 10 background traffic flows. At $N_{tf} = 1$, three traffic flows are added, one for voice, one for video and one for best-effort traffic, thus raising the total number of users in the system to 13. At the final evaluation point, $N_{tf} = 20$, the system contains 70 users, 10 with background traffic and 20 with each of the other three service classes (i.e., $N_{BK} = 10$ and $N_{VO} = N_{VI} = N_{BE} = 20$). In other words,

Table 5.7: Summary of simulation parameters for Case Study 3

Parameter	*Value*
Number of users	$N = N_{BK} + N_{BE} + N_{VI} + N_{VO}$
Number of classes	$P = 4$ (Background, Best-effort, Video and Voice)
Background traffic	
Background users	$N_{BK} = 10$
Mean offered load	1.8 Mbps per user (Poisson, $L_{BK} = 1000$ bytes)
Best-effort traffic	
Best-effort users	varying, $N_{BE} = [0, 20]$
Mean offered load	1 Mbps per user (Poisson, $L_{BE} = 1000$ bytes)
QoS demands	maximum delay of 5 s, no lost packets
Voice traffic	
Voice users	varying, $N_{VO} = [0, 20]$
Traffic generation	Brady's ON-OFF model with G.711 voice codec
Mean offered load	27.23 kbps per user (160 bytes/20 ms)
QoS demands	maximum delay of 150 ms, lost packets $< 1\%$
Video traffic	
Voice users	varying, $N_{VI} = [0, 20]$
Traffic generation	10 frames/second, 25 packets/frame
Packet size	Truncated Pareto, min = 50 bytes, max = 200 bytes
Packet inter-arrival	Truncated Pareto,min = 2.5 ms, max = 4 ms
Mean offered load	180 kbps per user
QoS demands	maximum delay of 300 ms, lost packets $< 1\%$
PHY Layer	IEEE 802.11g
Coherence time	$\tau_c = 150$ ms
Evaluated Schemes	CL-alg, SP-alg, CLSP-alg, VPF-alg$_4$
CL Overhead (per frame)[a]	$\text{CL}_{o,CL-alg} = 3 \times TQ$ bits, $\text{CL}_{o,SP-alg} = 6$ bits, $\text{CL}_{o,CLSP-alg} = 3 \times \sum_{p=1}^{2} TQ_p + 6$ bits, $\text{CL}_{o,VPF-alg_3} = 5 \times TQ$ bits

[a] Overhead values derived from Section 5.2, for $x = 3$ bits per rate (IEEE 802.11g rate set), $m = 3$ DQCA control minislots and $y = 2$ bit per service class

at each simulation step the traffic load is incremented by 1.2 Mbps on average, decomposed into approximately 27.23 kbps of voice conversation, 180 kbps of video streaming 1 Mbps of best-effort data. The approximate value of total offered load is marked at the upper part of the figures.

The total throughput of the system is illustrated in Figure 5.31. The best performance, as expected, is achieved by the CL-alg due to the opportunistically scheduled transmissions that generally take place at a higher rate. A maximum throughput of 25 Mbps is reached and maintained even when the traffic load increases. The performance of DQCA comes next, with throughput reaching a maximum of approximately 21.5 Mbps, but later drops and stabilizes at 19 Mbps when more than 14 traffic flows per service class are present. Interestingly, the total throughput is lower for the other three proposed algorithms (SP-alg, CLSP-alg and VPF-alg) and it decreases as the traffic load grows. The CLSP-alg performs better than the SP-alg and the difference is more pronounced for fewer traffic flows. For example, for 5 traffic flows CLSP-alg yields a throughput of 22.3 Mbps, with a gain of almost 4% with respect to DQCA and 12.9% compared to SP-alg. The VPF-alg performs slightly better than the SP-alg for low traffic load. Nevertheless, it exhibits a milder decrease rate and eventually, as more traffic flows are added, it outperforms both SP-alg and CSLP-alg.

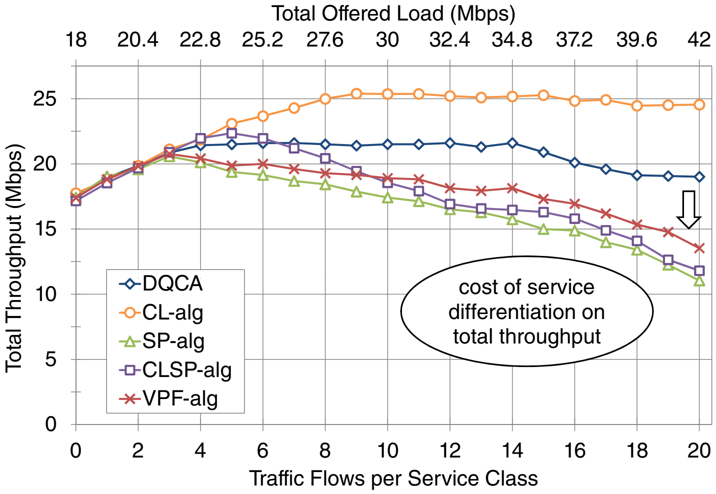
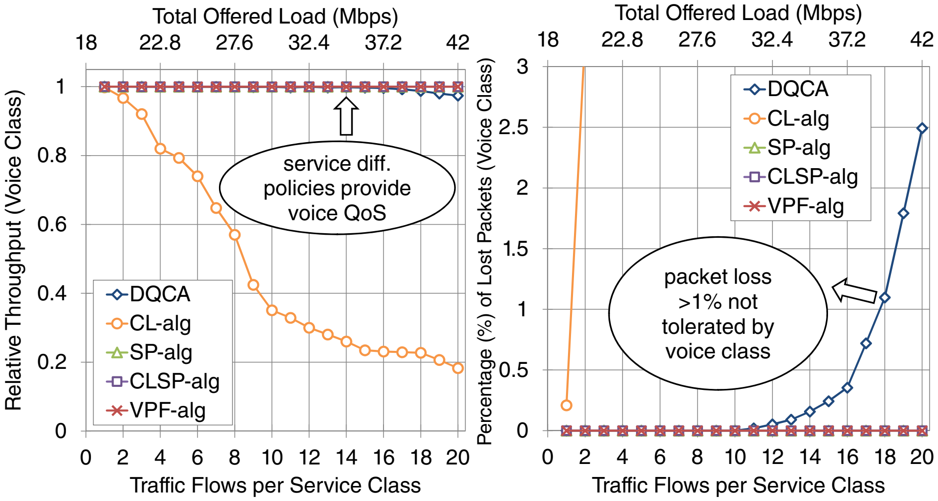


Figure 5.31: Total system throughput

Even though a direct comparison with the corresponding system throughput results of study case 2 (Figure 5.25) is not possible, given that simulation conditions are not the same in terms of traffic load and number of users, it is interesting to note that the overall performance of the four CL algorithms is not the same. Before, when only voice and data services were present, the strict opportunistic CL-alg and CLSP-alg showed a clear difference in throughput performance compared to the other schemes. In this case, however, the service differentiation policy adopted in different degrees by SP-alg, CLSP-alg and VPF-alg seems to take its toll on the system throughput. The difference is that now four service classes are defined and the ones that are assigned higher priorities (e.g., the voice and video services) also happen to generate a lower amount of traffic load with respect to the data services (best-effort and background). Therefore, prioritizing QoS-demanding applications in this scenario has an impact on the maximum achieved system throughput.

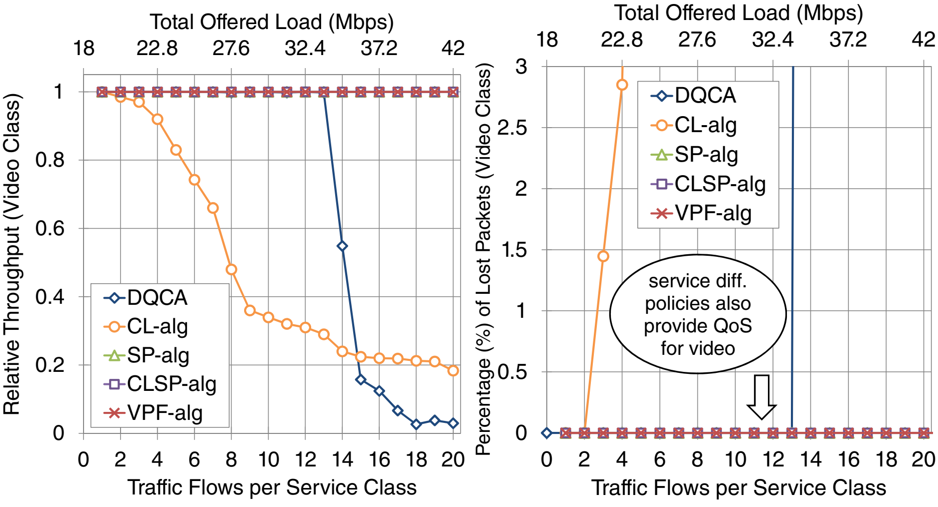


(a) Relative throughput **(b)** Percentage of lost packets

Figure 5.32: Throughput performance for the voice class

The above remark is made clearer by observing the relative throughput per service class. Starting with the most delay-sensitive voice service, Figure 5.32 (a) illustrates the relative throughput experienced by the voice users. The three algorithms that employ service differentiation, SP-alg, CLSP-alg and VPF-alg, achieve a relative throughput of 1 for voice. DQCA also has a close to one throughput that slowly deteriorates when more than 16 traffic flows per service class are present. In order to determine whether the slight drop in DQCA throughput is within the voice QoS requirements, the percentage of lost voice packets has been plotted in Figure 5.32 (b). It can be discerned that voice packets begin to get discarded due to delivery with delays exceeding 150 ms from $N_{tf} > 10$ traffic flows and for $N_{tf} > 18$ the QoS requirement of less than 1% packet loss is not met. Finally, the CL-alg is apparently not suitable for the voice service in this scenario, given the rapid decline observed in the achieved relative voice throughput. The packet loss percentage is very high, exceeding 3% for only as few as $N_{tf} = 2$ traffic flows.

The relative throughput and the percentage of lost packets for the second in priority video service class have been plotted in Figures 5.33 (a) and (b), respectively. Again, service differentiation proves critical to the satisfaction of QoS constraints, as shown by the performance of SP-alg, CLSP-alg and VPF-alg that attain relative throughput of 1 with all packets delivered in time. DQCA fully supports up to 13 video traffic flows, but after that point the video performance deteriorates dramatically. This sudden drop is also reflected on the total throughput (Figure 5.31, for $N_{tf} = 13$ to 15). The CL-alg video performance is slightly better that the voice performance, with the algorithm being able to support up to 3 video traffic flows.



(a) Relative throughput **(b)** Percentage of lost packets

Figure 5.33: Throughput performance for the video class

By closely contrasting the two subplots of Figure 5.33 it can be observed that the throughput degradation is to a high degree but not entirely caused by the discarded packets. Take for example the performance of DQCA for $N_{tf} = 20$. The relative throughput is approximately 0.03 whereas the percentage of lost packets is a little over 86%. This means that the 3% of the generated video traffic is successfully transmitted, the 86% is transmitted out of time and consequently discarded, and the remaining 11% of the traffic is being accumulated at the users' buffers due to the heavy congestion of the system.

The best-effort traffic has very relaxed QoS constraints and is the third service class in order of priority. The relative throughput performance has been plotted in Figure 5.34. The strict service differentiation schemes, SP-alg and CLSP-alg, yield a relative throughput of 1 for up to $N_{tf} = 13$ traffic flows and then deteriorate rapidly, with CLSP-alg having a slightly better performance. DQCA and the mild-priority VPF-alg can fully support 9 best-effort flows. From that point, VPF-alg deteriorates progressively and has the worse overall performance for this service class whereas DQCA throughput decreases more mildly and for $N_{tf} \geq 17$ it becomes the second most efficient scheme. Finally, the CL-alg maintains a throughput of 1 for up to $N_{tf} = 7$ flows only, but then performance deteriorates at a slow rate and eventually surpasses the other schemes.

Finally, the relative throughput for the background traffic that has no delay constraints is depicted in Figure 5.35. Bear in mind that a constant background

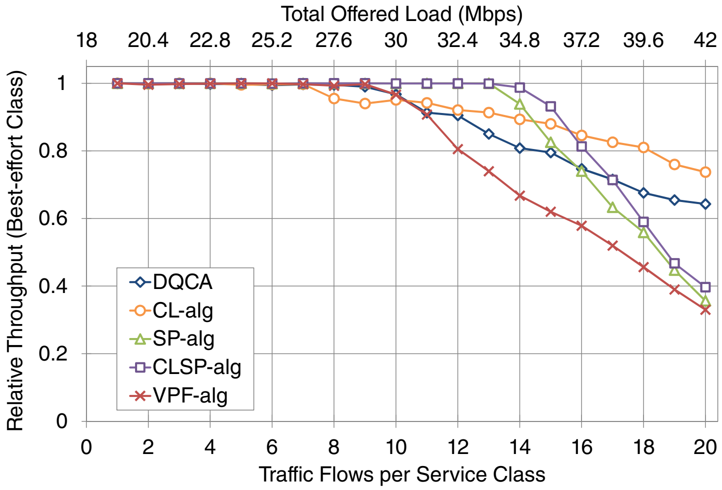


Figure 5.34: Relative throughput for the best-effort class

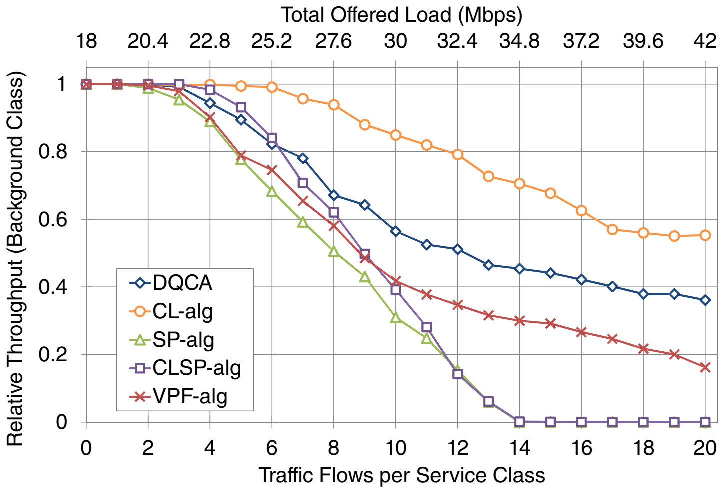


Figure 5.35: Relative throughput for the background class

traffic load of 18 Mbps is considered, generated by 10 users. Therefore, this plot actually shows how the background service performance is affected by the different scheduling algorithms, as more traffic flows (or users) belonging to the other three service classes are added to the system. Without doubt, the best performance is achieved by the CL-alg with a throughput a close to 1 for up to $N_{tf} = 6$ traffic flows that, then, drops gradually down to 0.55 for $N_{tf} = 20$. DQCA comes second with a considerable difference (e.g., around 35% lower than CL-alg for $N_{tf} = 20$), followd by VFP-alg (around 70.7% lower than CL-alg for $N_{tf} = 20$). The strict priority SP-alg and CLSP-alg have the worst performance and fail to deliver any background traffic beyond $N_{tf} = 13$.

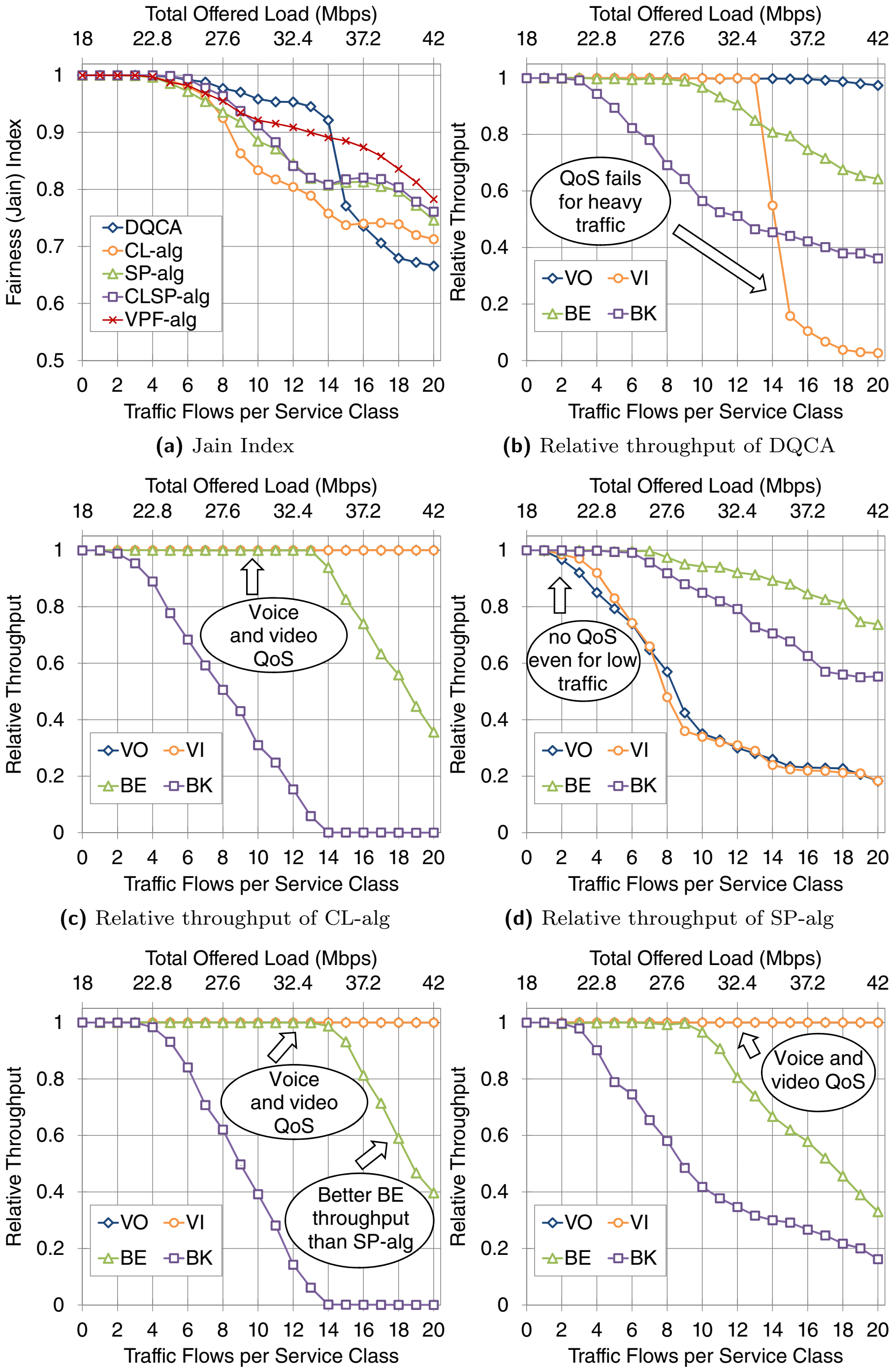
The relative throughput results for each service class emphasize the inevitable trade-off between QoS provisioning and equal resource allocation among different flows of traffic. The overall system fairness can be measured by calculating the Jain factor which is depicted in Figure 5.36 (a). Figure 5.36 also plots the relative throughput achieved by DQCA and the four CL schemes, CL-alg, SP-alg, CLSP-alg and VPF-alg (plots (b) to (f) respectively). These are the same throughput results presented before, but organized in a different way to demonstrate how each CL algorithm allocates resources among the four service classes.

DQCA is the fairest scheme up until it reaches saturation, for more than $N_{tf} = 14$ traffic flows per service class. After that point, a steep drop of the Jain index is observed, that can be explained by the sudden decrease in the relative throughput of the video class. The lowest fairness is observed for the CL-alg, caused by the lack of QoS provisioning for the delay-sensitive services. SP-alg and CLSP-alg have a close fairness performance, with the latter scheme being slightly better due to the higher relative throughput achieved for the best-effort and background services. Finally, the VPF-alg, in this scenario, achieves a very balanced performance that combines QoS for the voice and video classes without completely compromising the data services. In terms of fairness, it comes second after DQCA for low and medium traffic loads and becomes the most fair scheme as the traffic grows.

The mean packet delay per service class has been considered next, starting with the voice class performance depicted in Figure 5.37. Two different calculations of the mean delay are offered. In plot (a), the represented delay metric refers only to the packets received within the QoS imposed delay limitation of 150 ms, with any packets received outside this delay constraint considered lost. This is the usual delay representation adopted for all delay statistics presented in this chapter and should be interpreted along with the percentage of lost packets for the voice class, shown in Figure 5.32 (b). Alternatively, all received packets, including the discarded ones, have been included in plot (b).

The algorithms that prioritize voice, SP-alg, CLSP-alg and VPF-alg, achieve by far the best performance with the mean delay not exceeding 5 ms. Since no packets are lost when these algorithms are employed, there is no difference between the two delay plots, (a) and (b). For DQCA, the mean voice delay gradually rises as the traffic grows but, overall, it remains at a relatively low level below 75 ms. A slight difference between the two plots is observed for $N_{tf} > 18$ traffic flows, reflecting the progressive increase in the percentage of lost packets. Finally, a mean delay of approximately 35 ms is measured for the CL-alg in plot (a). At first, CL-alg seems to perform better than DQCA, thanks to the fact that transmissions are more likely to take place at higher rates. Nevertheless, this is not a realistic conclusion, given the very high percentage of lost voice packets that exceed by far the 1% constraint (as seen in Figure 5.32). Hence, the mean delay for CL-alg is more accurately represented in plot (b), where a steep increase in delay that exceeds 100 ms for $N_{tf>5}$ can be clearly observed.

Similar conclusions can be derived for the mean delay of the video class, depicted in Figure 5.38, with only non-discarded packets taken into consideration this time.



(a) Jain Index

(b) Relative throughput of DQCA

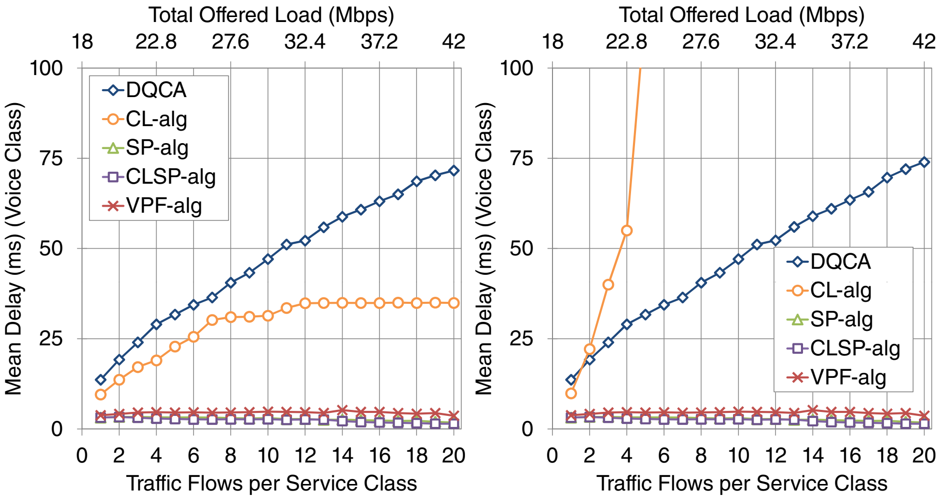
(c) Relative throughput of CL-alg

(d) Relative throughput of SP-alg

(e) Relative throughput of CLSP-alg

(f) Relative throughput of VPF-alg

Figure 5.36: Fairness and relative throughput for DQCA and the CL algorithms



(a) Only non-discarded packets considered (QoS constraint of 150 ms maximum delay)

(b) All received packets considered

Figure 5.37: Mean Delay for the voice class

SP-alg and CLSP-alg continue to guarantee the delivery of video packets with a minimal delay, considerably below the tolerated maximum of 300 ms. VPF-alg yields higher delay for the video service compared to the voice service, but still does not exceed 25 ms. Delay performance is worse for DQCA, with a steep increase experienced for more than 13 traffic flows. As before, the delay of CL-alg shows a relatively mild increase without exceeding 75 ms, however it should not be forgotten that this metric only corresponds to the non-discarded video packets which, for high traffic, constitute approximately 30% of the total received video traffic.

The mean delay curves for the best-effort and the background classes are shown in Figure 5.39 (a) and (b), respectively. It is interesting to notice how the same delay level can be achieved by each algorithm for different amounts of traffic load. For example, for a mean delay of approximately 500 ms, the CL-alg supports $N_{tf} = 6$ traffic flows per service class. DQCA and VPF-alg support three additional traffic flows whereas SP-alg and CLSP-alg yield a 500 ms delay for 8 additional flows with respect to CL-alg.

Finally, as anticipated, the delays for the background traffic are much higher and for that reason the logarithmic scale has been adopted in Figure 5.39 (b). In general, the delay increases significantly as the traffic load grows. The CL-alg performs better with a maximum mean delay of 11 s whereas the SP-alg and CLSP-alg have the poorest performance with maximum mean delays of approximately 25 s. Note, however, that for fewer traffic flows, the delay for the CLSP-alg is much lower compared to the other schemes with the exception of the CL-alg (e.g., for $N_{tf} = 4$ CLSP-alg gives a delay of around 500 ms whereas the delay for the SP-alg and the VPF-alg is approximately 2.7 s).

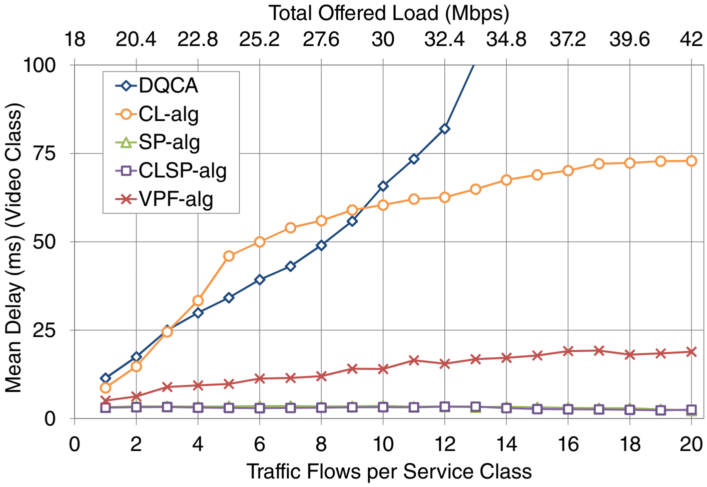
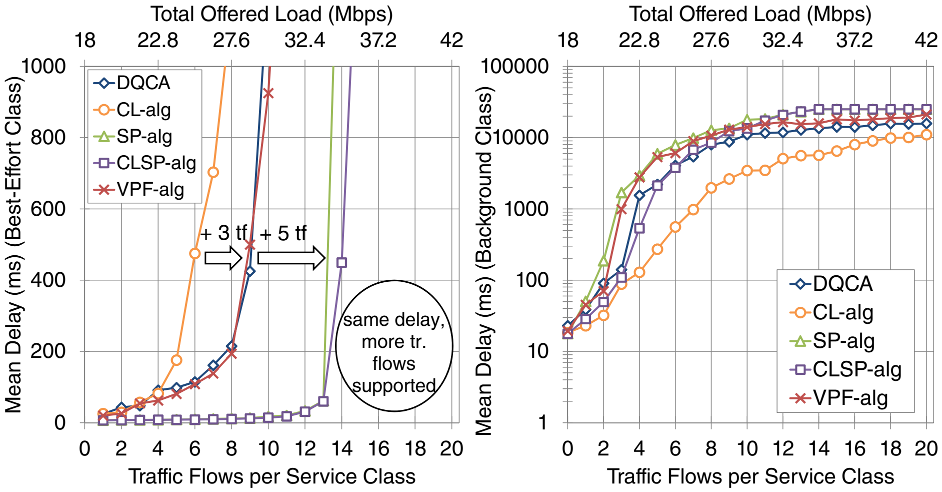


Figure 5.38: Mean Delay for the video class

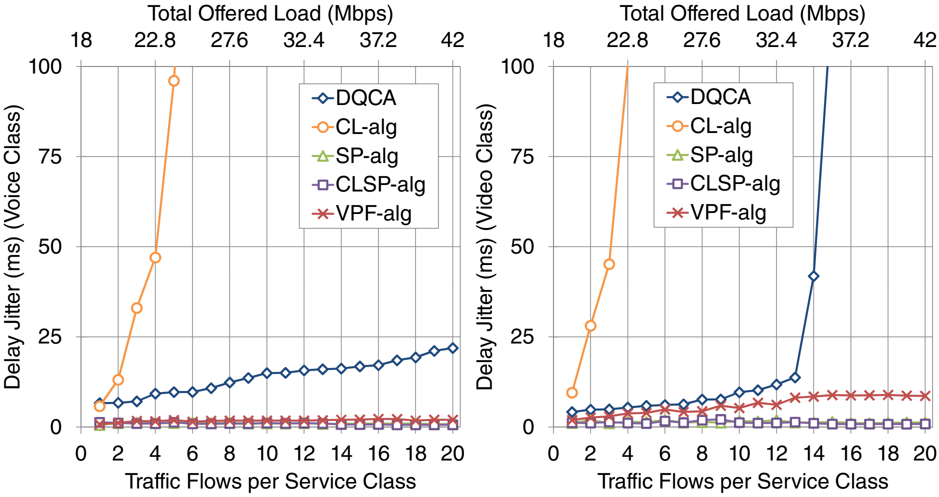


(a) Best-effort class

(b) Background class

Figure 5.39: Mean Delay for the data classes

To complete the delay analysis, the delay jitter for the voice and the video service classes has been plotted in Figure 5.40 (a) and (b), respectively. The jitter for the SP-alg and the CLSP-alg is almost zero, meaning that all the received voice and video packets are delivered with the same amount of delay. The jitter for the VPF-alg is also very small, and does not exceed 3 ms for voice and 10 ms for video traffic. For DQCA the voice jitter increases with a slight slope and is kept below 25 ms whereas the video jitter rapidly rises after $N_{tf} = 13$, which is exactly when its relative throughput performance drops dramatically. Clearly, the jitter is considerably larger for the CL-alg.



(a) Voice class **(b)** Video class

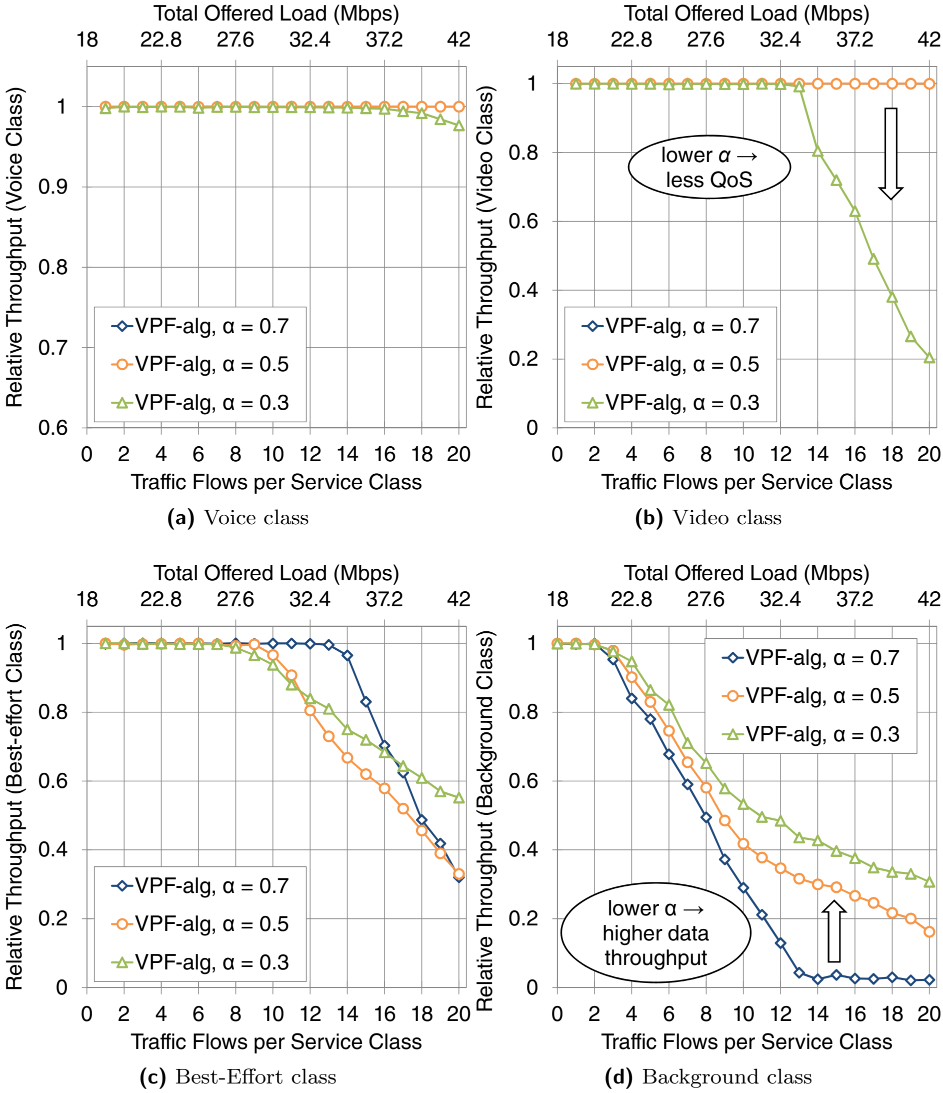
Figure 5.40: Delay jitter for the multimedia classes

The presented results in this section lead to some interesting conclusions on the behavior of DQCA and the four CL scheduling algorithms. DQCA is a fair, balanced protocol that does not take into consideration any QoS requirements. As a result, it performs well when the traffic load is relatively low but fails to provide QoS guarantees when traffic grows. CL-alg is focused on maximizing the system throughput but also ignores QoS demands. Therefore, it enhances significantly the performance for delay tolerant classes (such as the data best-effort and background traffic) but is not suitable for networks with heterogeneous multimedia traffic.

On the other hand, SP-alg gives priority to delay-sensitive service classes and therefore has a near-optimum throughput performance with low delays for voice and video applications. This, however, has an impact on the best-effort and background classes, whose performance drops significantly and eventually reaches zero. The CLSP-alg employs opportunistic scheduling while maintaining the service priority scheme, yielding higher throughput and lower delays with respect to the SP-alg for all classes. However, when the traffic load is very high, the background traffic class is also led to starvation (zero throughput).

Finally, the VPF-alg combines service differentiation with a mild opportunistic scheme that takes into account the age of the packets in the transmission queue. Its performance depends on the selection of the tunable parameter α. For $\alpha = 1$ the VPF-alg performance becomes identical to the SP-alg and for smaller values of α it approximates (but cannot reach) the performance of the CL-alg. For $\alpha = 0.5$, which is the case evaluated in this section, the VPF-alg ensures QoS performance for the voice and the video traffic and in addition overcomes the starvation problem. The improvement of the background traffic performance is achieved at the cost of best-effort throughput degradation (with respect to the other algorithms). Nevertheless, the system fairness is increased since all classes get access to the channel (non-zero throughput).

To better illustrate the latter point, the VPF-alg throughput has been evaluated for three values of the parameter α, in particular for $\alpha = 0.3$, 0.5 and 0.7. The relative throughput for the voice, video, best-effort and background service classes has been represented in Figure 5.41, in plots (a) to (d), respectively. Smaller values of α correspond to a more opportunistic scheduling policy that yields higher data throughput but cannot always guarantee QoS. On the other hand, larger values of α mean that the service differentiation plays a more important role.



(a) Voice class

(b) Video class

(c) Best-Effort class

(d) Background class

Figure 5.41: Relative throughput for different values of the VPF-alg parameter α

5.4 An Update Channel State Information (CSI) Mechanism

5.4.1 The Problem of Inaccurate CSI

In the channel-aware schemes proposed and evaluated in the previous sections, the AP performs an estimation of the available transmission rate of the users by measuring the SNR each time an ARS is correctly received. The CSI is stored at the AP and is fed back through the FBP, to be employed by the respective user for the data transmission and to be included in the scheduling decisions of the opportunistic scheduling CL-based algorithms.

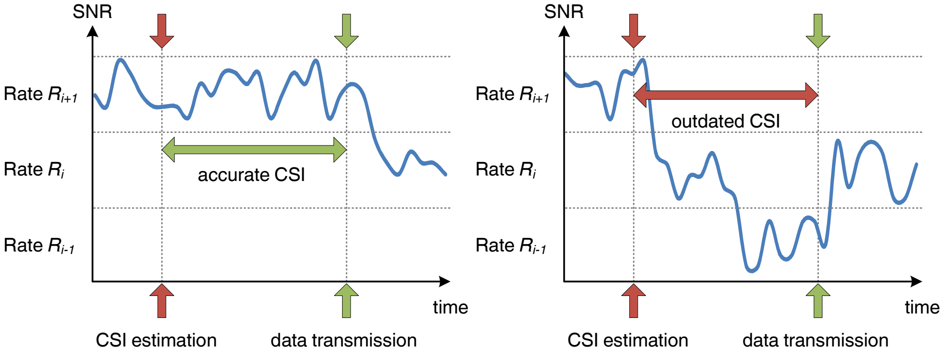
A possible weakness of this mechanism is that the estimated rate is not employed immediately by the user and, as a result, there is a chance that the estimated value may not reflect the channel condition at the time of data transmission. In principle, once a user successfully transmits an ARS and receives the estimated rate from the FBP, it enters the DTQ and waits until is is scheduled for transmission. This waiting time depends on several factors, but mainly on the serving time required by the preceding users and the scheduling policy (FIFO or CL-based) that determines the transmission order.

Assuming that the wireless channel varies slowly with time, the estimated rate can be considered accurate, despite the elapsed time from the CSI acquisition to the data transmission. In general, however, this is not the case and the channel condition of the users is likely to change while they are waiting in the DTQ. An example is given in Figure 5.42. In plot (a) the channel condition is maintained relatively stable during the waiting time of the user in the DTQ and as a result, the data transmission takes place with an accurately estimated rate.[3] On the contrary, in plot (b), the channel varies faster with time (or alternatively, the user waiting time in the DTQ is longer) and by the time the data transmission occurs, the obtained CSI does not reflect the actual channel condition.

The consequences of possessing inaccurate CSI are twofold:

- First, the users attempt transmission with a rate that does not reflect the actual channel condition. In the best case scenario, the actual channel condition will be better that the previously estimated one, meaning that the employed transmission rate will be below the channel capacity. In this case, the data packets will be correctly transmitted but at a rate that is lower than the optimum. On the other hand, in the worst case scenario, a degradation of the channel conditions will take place and the employed transmission rate will not be supported by the channel. As a result, there will be an increased probability of transmission errors, causing packet retransmissions and a significant drop in the system performance.

[3] Even though channel variations are present, on average they remain within an SNR range that corresponds to a particular transmission rate for a desired error performance.



(a) Accurate CSI **(b)** Outdated CSI

Figure 5.42: Example of outdated CSI

- Second, if channel-aware CL-based policies are employed, the scheduling decisions will be suboptimal if based on inaccurate CSI. The performance enhancement gained by opportunistic scheduling is lost if, for example, the high-rate user scheduled for transmission turns out to possess a worse channel condition than expected.

Summarizing, due to the time-varying nature of the wireless channel, the CSI acquired through the link adaptation mechanism can become outdated. This can affect the performance of DQCA and has an even stronger impact on the channel-aware CL algorithms proposed in the previous section. In continuation, a simple mechanism will be proposed to alleviate this problem through periodic CSI updates.

5.4.2 The CSI Update Mechanism

The aim of the proposed update scheme is to provide a low-complexity mechanism for the acquisition of valid CSI with minor modifications to the DQCA protocol. To this end, a special DQCA frame called Update frame is defined during which the AP performs link estimation for all the users that are waiting in the DTQ and recalculates their available bit rates.

As mentioned repeatedly throughout the thesis, the standard DQCA frame consists of the CW the data slot and the FBP. The Update frame, on the other hand, has only two parts, depicted in Figure 5.43.

The first part can be thought of as an extended CW where the users waiting in the DTQ retransmit an ARS in order to enable the reevaluation of their link condition by the AP. However, unlike the probabilistic access employed in the CW of the standard DQCA frame, the ARS transmissions within the Update frame follow a deterministic order dictated by the position of each user in the DTQ. In other words, the DTQ users (whose total number is expressed by the counter TQ)

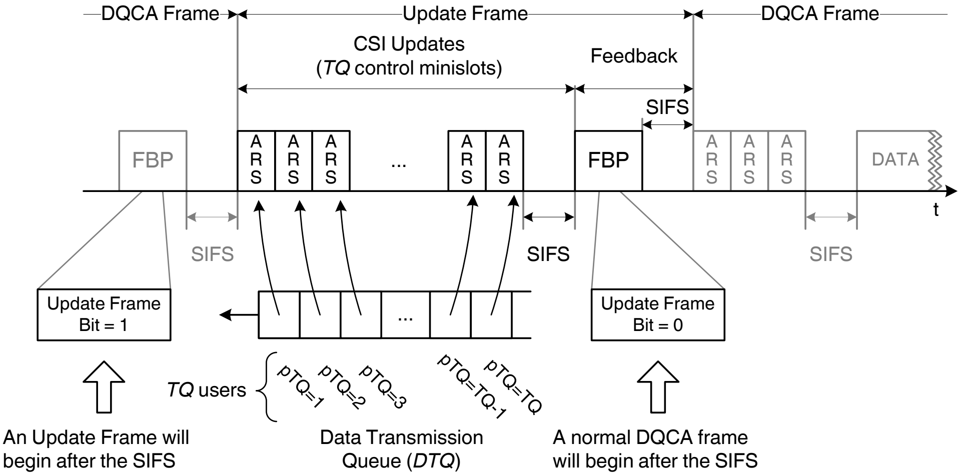


Figure 5.43: Structure of the Update Frame

transmit sequentially an ARS in order of their pTQ counter. As a result, the first part of the Update frame consists of TQ control minislots in which collision-free ARS are sent. In this occasion, the role of the ARS is not to ask for channel access but to allow the AP to perform link estimation and update the available rates of all users.

In the second part of the Update frame a FBP is broadcasted by the AP, meaning that no data transmission takes place. This FBP is a modified version of the FBP employed in the normal DQCA frames and its structure can be seen in Figure 5.44. The main difference is that the modified FBP (instead of containing information on the outcome of the control minislots and the data slot) includes a vector with the recalculated transmission rates of the users in the DTQ, obtained through the SNR measurements on the received ARS. As in the case of the channel-aware CL algorithms of Section 5.2, the vector elements are sorted by the position of the users in the DTQ. Upon the reception of the FBP, each user extracts the corresponding rate from the vector and replaces its previous, outdated rate value.

For the implementation of the update mechanism, a one-bit update flag is also added in the FBP to indicate the type of the following frame. Setting this flag to the bit '1' indicates that after a SIFS from the transmission of the FBP, an Update frame will be initiated and the DTQ users are expected to transmit an ARS in the predefined order (as shown in the example of Figure 5.43). The modified FBP transmitted at the end of the Update frame sets the update flat to '0', so that the normal DQCA operation is resumed without requiring further notification.

The Update frames are inserted periodically between standard DQCA frames. The frequency of their occurrence is expressed by the parameter Data-to-Update Ratio (DU ratio), defined as the number of standard DQCA frames transmitted between two consecutive Update frames. For instance, a DU ratio of 2 means that an Update frame is sent every 2 standard DQCA frames, as shown in the example

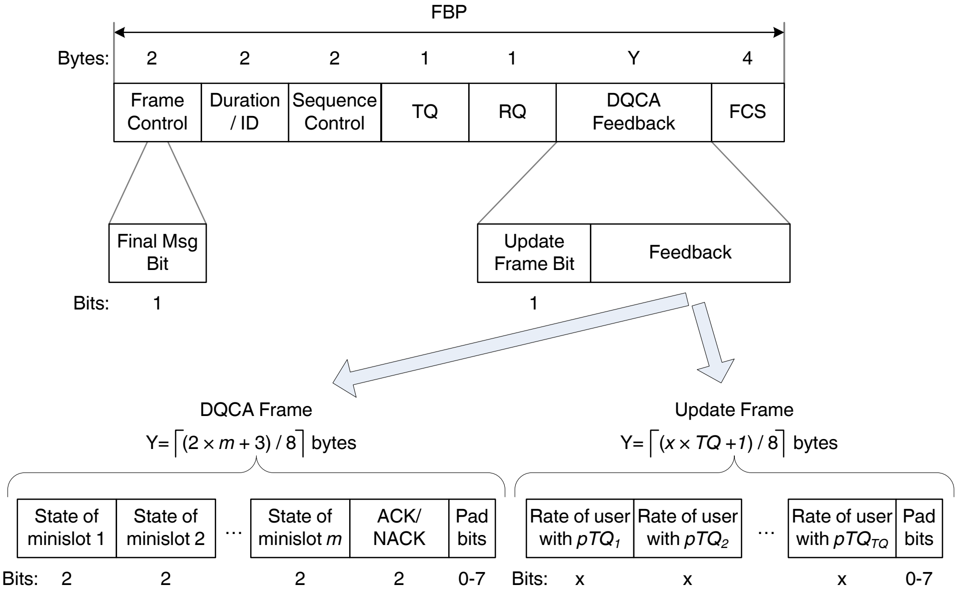


Figure 5.44: The modified structure of the FBP

of Figure 5.45.

The DU ratio is a configurable system parameter that should be selected depending on the channel circumstances. Smaller DU ratio values indicate a high CSI update frequency, required to reflect the state of fast changing channels, whereas larger DU ratio values correspond to relatively sparse updates suitable for more static channels. There is a clear trade-off in the selection of the DU ratio, since frequent updates reduce the likelihood of outdated CSI but introduce a significant control overhead. On the other hand, too few updates may not guarantee valid CSI, thus failing to overcome the problematic situations explained in Section 5.4.1. The DU ratio may have a fixed value of may be dynamically selected by the AP on runtime, by appropriately setting the update flag bit in the FBP.

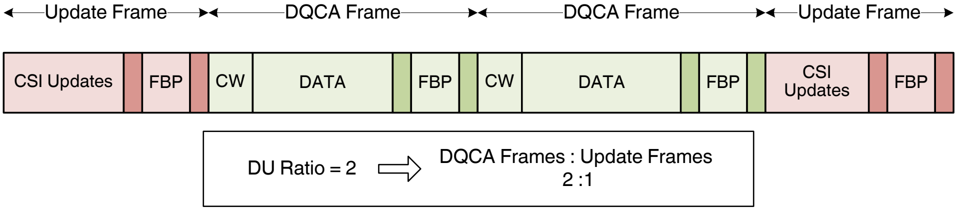


Figure 5.45: Example of Data-to-Update Ratio (DU Ratio=2)

Further discussion on the efficiency of this update mechanism and the selection of the DU ratio value will take place in the following section, where performance results will be presented in detail.

5.4.3 Performance Evaluation of the CSI Update Mechanism

This section examines the impact of outdated CSI on the system performance and evaluates the proposed update mechanism. The update mechanism has been evaluated on the DQCA-based CL-alg that gives priority to users with higher available transmission rates (described in detail in Section 5.2.1). A network of $N = 20$ users has been assumed, that generate Poisson-generated data traffic of various packet sizes L. Without loss of generality, the IEEE 802.11b channel model has been employed, with four available rates of 1, 2, 5.5 and 11 Mbps. The most significant parameters employed in this study case have been included in Table 5.8.

Table 5.8: Summary of simulation parameters for the performance evaluation of the CSI update mechanism

Parameter	*Value*
Number of users	$N = 20$
Number of service classes	$P = 1$ (Best-effort)
Traffic generation	Poisson msg arrivals, average $\kappa = 10$ packets/msg
Packet size	varying, $L_{BE} = [512, 1000, 2312]$ bytes
Evaluated Schemes	CL-alg enhanced by the CSI update mechanism
ARS	$10\mu s$
CL Overhead (per frame) [a]	$\mathrm{CL}_{o,CL-alg} = 2 \times TQ$ bits, for the IEEE 802.11b channel
Update Frame duration	$ARS \times TQ = 10 \times TQ\mu s, TQ = [0, N]$

There are two sources of additional overhead with respect to DQCA. First, for the implementation of the CL-alg the FBP, sent at the minimum rate of 1 Mbps in order to ensure reliable transmission, must contain a vector with the rate R of each user waiting in the DTQ. Since there are four possible rate values, 2 bits are sufficient for the representation of the rate set, resulting to an overhead of $2 \times TQ$ bits[4], rounded up to an integer number of bytes. The second source of overhead is the update frames whose frequency depends on the DU ratio. The update frame consists of TQ control slots during which all users in the DTQ transmit an ARS, followed by a FBP with the new rate estimations.

The first plot (Figure 5.46) shows how the system throughput is affected by the use of outdated CSI, considering a coherence time of $\tau_c = 100$ ms and different packet sizes. The solid lines correspond to the ideal scenario in which the available CSI always reflects the true channel condition at the moment of data transmissions.

[4] The integer TQ represents the number of users waiting in the DTQ and lies within $[1, 20]$ in this scenario.

The dashed lines represent the case in which the available CSI is inaccurate. By the term inaccurate, it is not meant that the channel state detection was erroneous but that the acquired CSI has become outdated due to the elapse of the coherence time while the users were waiting for transmission in the DTQ. As a result, the transmission rate employed by the users may not correspond to the optimal rate supported by the current link state. In this case, the following outcomes are considered:

- If the employed rate is lower that the optimal rate, the packet is successfully received but no rate update is performed until the completion of the message. In other words, all packet transmissions maintain the same rate even though a higher rate could be supported by the channel.
- If the employed rate is higher that the optimal rate, the packet is considered lost (i.e., received with errors). In the next DQCA frame the user selects the immediately lower rate from a predefined rate set and retransmits the packet. This procedure is repeated until a valid rate is found and the packet is received correctly.

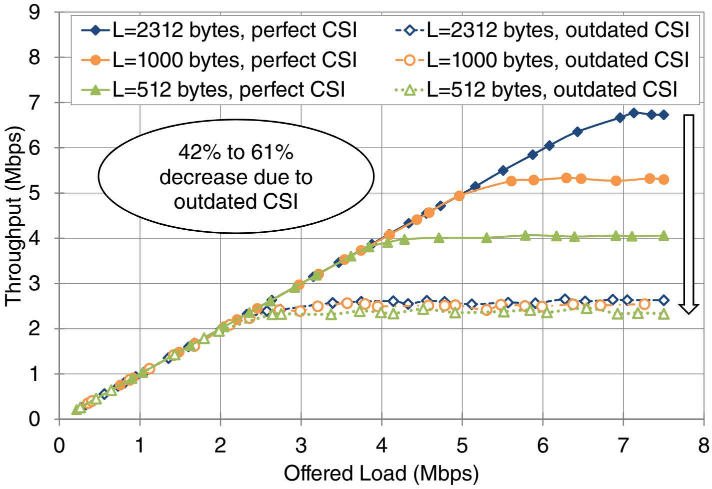


Figure 5.46: Impact of the outdated CSI on the system throughput ($\tau_c = 100$ ms).

As shown in Figure 5.46, the impact of inaccurate CSI on the system performance is considerable. The maximum throughput achieved as the offered load grows suffers a significant degradation when scheduling is based on outdated CSI. In particular, a 42% decrease is observed for small packets of 512 bytes, with the figure further dropping to 61% for packets of 2312 bytes. The problem is more aggravated when larger packets are used, since the associated transmission times are longer.

To alleviate this problem, the proposed update mechanism has been applied. The maximum achieved throughput for the three packet sizes has been plotted in Figure 5.47. The results have been obtained for the fast changing channel considered in Figure 5.46 (with $\tau_c = 100$ ms), and for a slower channel ($\tau_c = 150$ ms). The achieved throughput varies as a function of the frequency of the update frames, expressed by the DU ratio. Small values of the DU ratio correspond to very frequent updates that burden the system with excessive overhead. On the other hand, a high DU ratio results to sparse updates that may not adequately match the channel variability. An optimum region for the selection of the DU ratio lies between the two extremes. For longer packets, e.g., for $L_d = 2312$ bytes, the best results are obtained when the DU ratio is within 3 to 7, meaning that one update frame is sent every 3 to 7 standard DQCA frames. Outside this narrow region, throughput decreases rather steeply, as the updates become less often. For smaller packets the optimum region is wider. For example, for $L_d = 512$ bytes, the maximum throughput is reached for a DU ratio of 10 and is maintained up to a DU ratio of 34.

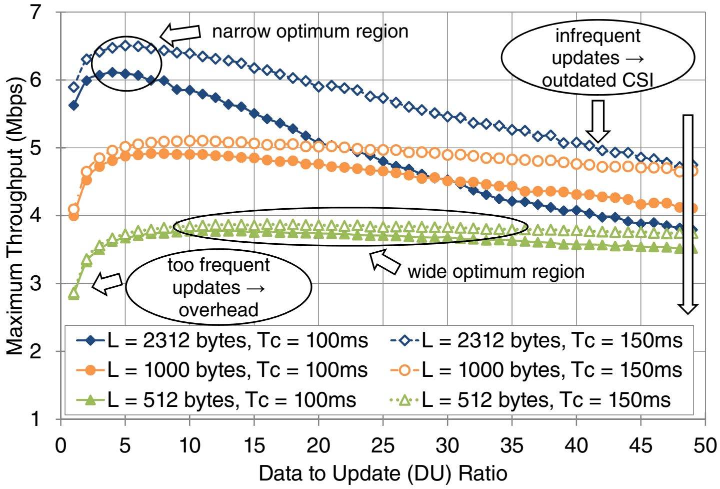


Figure 5.47: Maximum achieved throughput as a function of the update frame frequency (DU ratio)

As far as the coherence time is concerned, a better throughput performance and a wider range of optimal DU ratios are obtained for slower channels. The difference is slight for smaller packets but becomes more pronounced for large packets, in which case a suitable selection for the DU ratio plays a more important role. Nevertheless, an interesting lesson learned from the results of Figure 5.47 is that there is a clear gain to be earned with the use of the proposed update mechanism, even when the DU ratio is not optimally chosen. A comparison with Figure 5.46 shows that for $\tau_c = 100$ ms and without update frames the maximum throughput ranges from 2.4 to 2.6 Mbps approximately, for the different packet sizes in ascending order. When update frames are employed with an optimal frequency, throughput raises to 3.7 Mbps for packets of 512 bytes up to 6.1 Mbps for packets of 2312 bytes. What is more important, throughput does not drop below 3.5 Mbps for any value of DU ratio within the considered range (i.e., for a DU ratio within [1,50]).

As mentioned before, even though the maximum achieved throughput is higher for bigger packet sizes, performance seems to decline at a higher rate with the DU ratio, as opposed to the milder decline observed for smaller packets. This can be explained by considering that the DU ratio expresses the frequency of frame updates with respect to standard DQCA frames but does not reveal the average time between consecutive updates. To clarify this point, the average interval between consecutive update frames has been plotted in Figure 5.48. This time is proportional to the duration of the standard DQCA frames that are transmitted between consecutive updates, which, in turn, depends on the packet size L_d and the transmission rate R. For example, for the case of $\tau_c = 100$ ms, a DU ratio of 30 corresponds to an update time interval of approximately 35 ms for $L_d = 512$ bytes and 117 ms for $L_d = 2312$ bytes. In the first case, there are roughly three frame updates within a single coherence time interval of $\tau_c = 100$ ms, whereas in the latter case the coherence time may elapse without a single CSI update. In other words, while a DU ratio of 30 may be convenient for packets of 512 bytes (and in fact lies within the optimal frequency region), the same DU ratio is not adequate for longer packets.

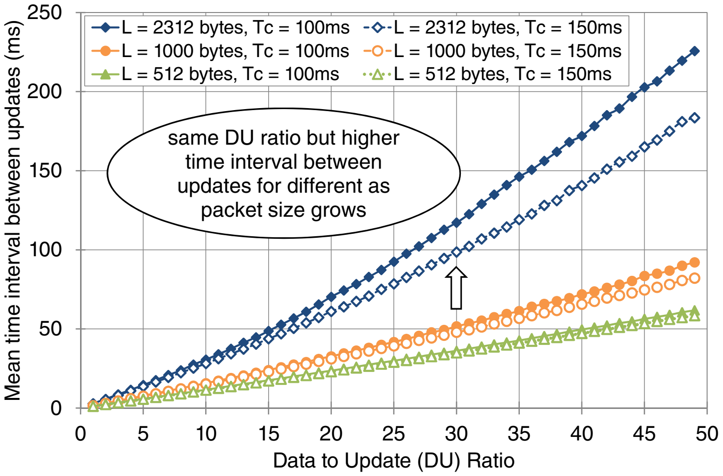


Figure 5.48: Mean time interval between consecutive update frames

Finally, Figures 5.49 and 5.50 display the performance enhancement in terms of throughput and mean delay, respectively, that can be gained with the use of the update mechanism, for the case of $L_d = 2312$ bytes and $\tau_c = 100$ ms. As a reference, the ideal case where perfect CSI is always known without any additional overhead has been plotted as an upper performance bound and the worst case scenario where no updates are available has been given as a lower bound. Three values have been considered for the update frequency:

- The optimal frequency value, which in this case corresponds to a DU ratio equal to 5.
- A case of extremely frequent updates with a DU ratio of 1, where one update frame is sent for every standard DQCA frame.
- A case of infrequent updates with DU ratio of 30.

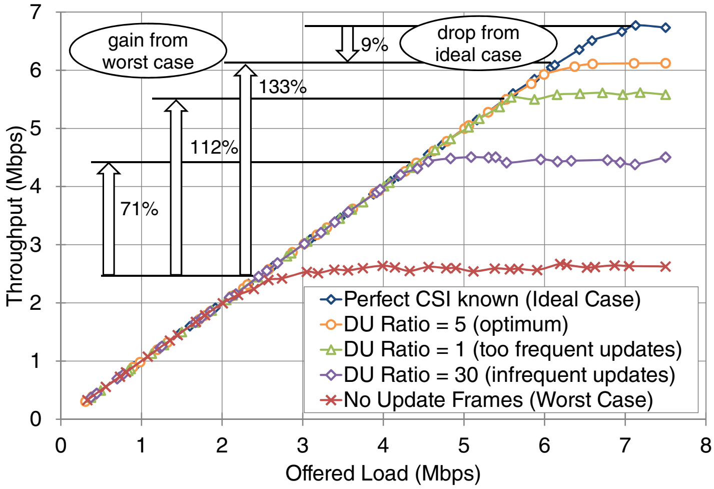


Figure 5.49: Throughput enhancement with the use of the update mechanism

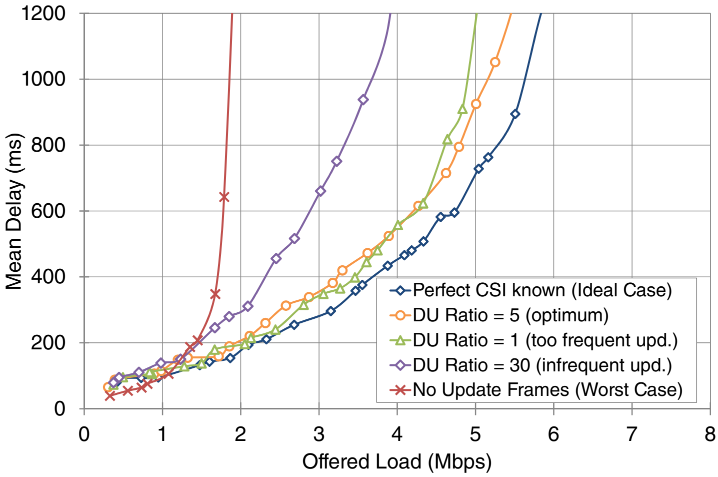


Figure 5.50: Delay enhancement with the use of the update mechanism

Again, the gain from the use of the proposed scheme is remarkable. The smallest improvement is obtained in the case of infrequent updates with DU ratio 30 (every 100 ms, as shown in Figure 5.48). Despite that, the achieved performance has a gain of approximately 71% with respect to the lower bound. With an optimal selection of DU ratio, the performance gain can be increased up to 133%, with throughput being only 9% below the ideal upper bound. Hence, even though the ideal maximum cannot be reached, due to the overhead added by the transmission of the update frames, the price to pay is reasonable. Moreover, it can be observed that between more frequent and less often updates, the first option is preferable. Similar improvements are observed in the delay performance, with the mean delay for the different DU ratios of the update mechanism being much lower compared to the worst case scenario and tending towards the ideal performance, when the DU ratio is optimally selected. Concluding, it can be said that in order to reap the profits of opportunistic channel-aware scheduling, accurate CSI must be available and the proposed mechanism constitutes an easy and viable solution to the problem of outdated channel reports.

5.5 Conclusions

The CL design paradigm breaks the traditional layering principle of network architecture in an effort to optimize some functions of the system and thus improve the overall performance. This chapter has discussed the integration of CL enhancements in the DQCA MAC protocol to offer additional functionalities.

Four CL-based scheduling algorithms have been proposed:

1. the strict opportunistic CL-alg that always schedules the user with the highest available transmission rate.
2. the strict service-aware SP-alg that assigns transmission priorities depending on the service class of each traffic flow, thus always favoring the most delay-sensitive applications.
3. the CLSP-alg that combines the two previous policies to achieve service-aware opportunistic scheduling.
4. the most balanced VPF-alg that determines the scheduling order based on a priority function that can be flexibly selected.

These four algorithms have been evaluated with the help of extensive simulation results under three different scenarios, leading to the following conclusions:

- Homogeneous scenario with data-only traffic. Service differentiation does not apply in this case and, therefore, only the opportunistic schemes CL-alg and VPF-alg have been evaluated, with the following results:
    1. The CL-alg always maximizes the (total) system throughput since the sole scheduling criterion is the available transmission rate of each user. This is the best option under a scenario where all users share similar average channel statistics. In this channel scenario, the performance of the CL-alg is fair, given that all users have the same chance to transmit on average.
    2. On the other hand, in a scenario where some users have worse average channel conditions than others, fairness becomes an issue for CL-alg. The low rate users will only be allowed to transmit if by chance their channel condition improves or if the system traffic load is sufficiently low

(hence, in the absence of other users, low rate users will be scheduled for transmission). In any case, there is a considerable risk of completely depriving low rate users from transmission.

3. On the contrary, the VPF-alg overcomes the problem of starvation by employing a more flexible priority function definition that does not only depend on the available transmission rate but on other factors such as the relative position of the users in the DTQ. Hence, users with low rates will get a chance to transmit after some waiting time. The cost for this fairer treatment is an inferior performance with respect to CL-alg, even though performance improvement with respect to basic DQCA is achieved.

- Heterogeneous scenario with voice and data traffic. In this case, traffic that belongs to two service classes, voice and data, is considered.

  1. In this scenario, service-aware policies make a difference in QoS provisioning. SP-alg and CLSP-alg meet the stringent QoS constraints of the voice class whereas the other schemes fail to maintain the packet loss rate within the QoS-defined limits.
  2. The performance of the VPF-alg can be tuned by appropriately setting the adjustable parameter α.

- Heterogeneous scenario with four traffic classes, real-time voice and a video applications and two more delay tolerant data classes.

  1. The presented results clearly show the performance trade-offs with respect to each service class. The finite system resources must be distributed among the four service classes. Unavoidably, schemes that prioritize multimedia applications (voice and video) and satisfy their QoS requirements provide fewer resources to data applications.

Finally, this chapter has proposed an update mechanism to provide up-to-date CSI, based on the introduction of special Update frames exclusively dedicated to channel state measurement. There is a trade-off between the system performance and the number of Update frames since more frequent updates provide more accurate CSI but require the exchange of additional control information. The impact of the frequency of these special frames on the performance has been studied in order to help in the selection of the appropriate value, depending on the channel variation characteristics. Through simulations it has been shown that especially for long data packet sizes, there is a need for more frequent updates in order to avoid retransmissions due to channel errors. In addition, updates are also required for smaller coherence times that correspond to rapid changing channels. The update mechanism has been evaluated in combination with the Cl-alg and has provided significant performance enhancement with respect to the case where no channel updates are available. In any case, the proposed mechanism can also be combined with other channel-aware schemes such as the CLSP-alg and the VPF-alg.

Chapter 6

Multiuser MAC Schemes for IEEE 802.11n Wireless Networks

6.1 Introduction

Given the widespread deployment of WLANs in the recent years and the increasing requirements of multimedia applications, the need for high capacity and enhanced reliability has become imperative. MIMO technology and its single receiving antenna version, MISO, promise a significant performance boost and have been incorporated in the emerging IEEE 802.11n standard [2].

Multiple antenna transmission techniques such as spatial multiplexing and transmit beamforming are used to provide rapid and robust point-to-point wireless connectivity. On the other hand, due to the inherent diversity of the MIMO channel, it is possible to achieve simultaneous point-to-multipoint transmissions and serve multiple users at the same time, through the same frequency. The MIMO multiuser transmission concept where data streams are assigned to different users can increase the overall system capacity when compared to single-user MIMO transmission where all streams are dedicated to just one user [77]. The two different setups are illustrated in Figure 6.1.

Even though IEEE 802.11n has been designed with MIMO technology in mind, its main focus is on maximizing throughput in point-to-point transmissions, through spatial multiplexing and mechanisms such as frame aggregation. Neither the standard nor the majority of related work consider any MAC mechanisms for multiuser scheduling, thus leaving a significant MIMO capability unexploited. As accurately pointed out in [78], there is a need for low-complexity multiuser transmission schemes, especially for downlink communications.

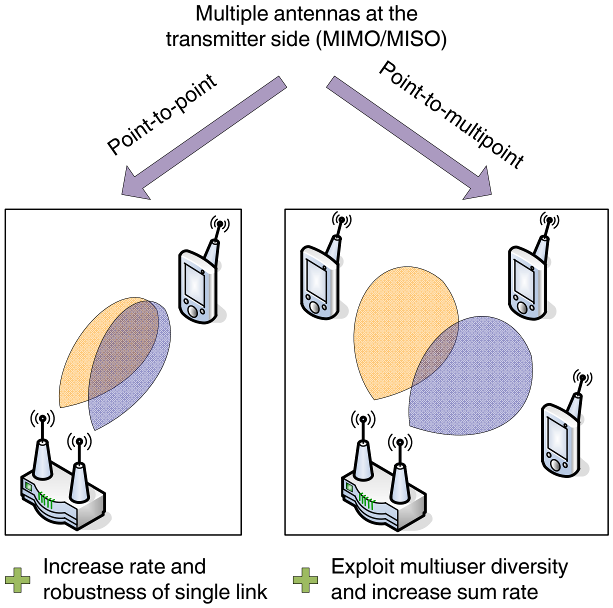

Multiple antennas at the transmitter side (MIMO/MISO)
Point-to-point
Point-to-multipoint
Increase rate and robustness of single link
Exploit multiuser diversity and increase sum rate

Figure 6.1: Point-to-point versus point-to-multipoint links

This chapter is dedicated to the investigation of solutions for the incorporation of multiuser capabilities in IEEE 802.11n-based WLAN systems by using CL information, while maintaining backward compatibility with the standard. The main contribution is the design of a number of opportunistic channel-aware multiple antenna MAC schemes that handle multiuser downlink transmissions and explore the advantages that can be gained by exploiting multiuser diversity.

The remaining part of this chapter is divided into eight sections. Section 6.2 discusses the problem statement and presents the considered setup. Section 6.3 highlights some PHY-related issues and presents the underlying beamforming transmission technique. The description of the proposed multiuser MAC schemes given in Section 6.4, followed by an analytical model for the theoretical calculation of their throughput performance in Section 6.5. Section 6.6 provides the performance evaluation of the proposed schemes and discusses the obtained trade-offs. Finally, Sections 6.7 and 6.8 are dedicated to the presentation of future investigation lines and some general conclusions.

6.2 Problem Statement and System Setup

As indicated in the literature review, presented in Section 2.5.2, it can generally be said that most contributions on multiuser transmission schemes focus on particular aspects of the problem and simplify the rest. Usually, when the focus is laid on the PHY layer transmission techniques, practical mechanisms for the channel access and the feedback acquisition are not considered, whereas multiuser MAC schemes often fail to consider PHY layer implementation issues. For example, some schemes optimize resource allocation but ignore feedback mechanisms and others minimize the required feedback but assume a dedicated control channel and a less sophisticated scheduling policy.

The aim of the work presented in this chapter is to introduce a multiuser MAC mechanism that handles in a joint manner the processes of channel access, scheduling, channel estimation and feedback acquisition, in conjunction with a low-complexity beamforming technique at the PHY layer. The proposed schemes have been designed in the context of a downlink communication channel in an infrastructure WLAN in which multiple antennas are available at the transmitter side. Without loss of generality, a MISO scenario with single-antenna users has been considered, even though the presented analysis can be also applied to MIMO systems with multiple-antenna users.

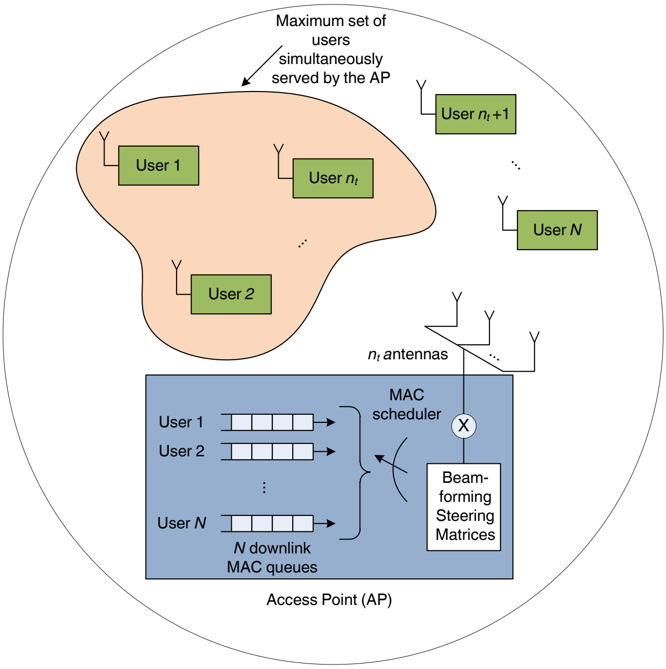


Figure 6.2: Scenario setup

The considered setup is illustrated in Figure 6.2. The proposed schemes can be considered as a downlink transmission phase, initiated by an AP equipped with n_t antennas ($n_t \geq 2$) in a system with N single-antenna users. By exploiting the MIMO/MISO spatial signal processing capabilities and employing an appropriate transmission technique, the AP can serve up to n_t users at the same frequency and time. Nevertheless, in order to extract multiuser diversity gain, the pool of served users should exceed the number of transmitting antennas (i.e., $N > n_t$).

Transmitting multiple downlink packets simultaneously, however, is feasible only when there is no interference among the selected users, or in a more realistic case, when the interference is relatively low. Hence, the AP must have some knowledge of the channel to select the most appropriate set of users for each transmission. These issues must be handled by the MAC layer in a practical way, as it will be described in detail in the following sections.

6.3 MIMO/MISO Multiuser Physical Layer

This section will provide a brief description of the channel model and the multiuser transmission technique used at the PHY layer. This theoretical background is necessary for the proper understanding of proposed MAC schemes that are the main contribution of this chapter. In general, the IEEE 802.11n MIMO specification with OFDM has been considered as the base for the PHY layer, with some modifications that will be explained in this section.

6.3.1 MIMO/MISO Channel

With the use of OFDM, the frequency selective MIMO/MISO channel is transformed into a number of frequency flat channels. In particular, a block-fading model is considered for the channel which remains constant during the coherence time and changes between consecutive time intervals with independent and identically distributed complex Gaussian entries $\sim \mathcal{CN}(0,1)$. This model represents the IEEE 802.11n channel model B in NLOS conditions [79], assuming that there are no time correlations among the different blocks and that the channel impulse response changes at a much slower rate than the transmitted baseband signal.

In the considered MISO downlink scenario, the channel between the AP that is equipped with n_t antennas and the ith single-antenna user (out of N total users with $N > n_t$) is described by a $1 \times n_t$ complex channel matrix $\mathbf{h}_i(t)$. Let $\mathbf{x}(t)$ be the $n_t \times 1$ vector with the transmitted signal to all the selected users in a particular transmission sequence and $y_i(t)$. Then, the received signal for the i^{th} user can be expressed as

$$y_i(t) = \mathbf{h}_i(t)\mathbf{x}(t) + z_i(t) \tag{6.1}$$

where $z_i(t)$ is an additive Gaussian complex noise component with zero mean and $E\{|z_i|^2\} = \sigma^2$ is the noise variance. The transmitted signal $\mathbf{x}(t)$ encloses the independent data symbols $s_i(t)$ to all the selected users with $E\{|s_i|^2\} = 1$. A total transmitted power constraint $P_t = 1$ is considered and for ease of notation, time index is dropped whenever possible.

6.3.2 Multibeam Opportunistic Beamforming (MOB)

Multibeam Opportunistic Beamforming (MOB) is a low-complexity transmission technique for multiple-antenna broadcast channels [80]. MOB requires the presence of multiple antennas at the transmitter side and one or more antennas at each receiving user, meaning that it can be applied to MISO or MIMO scenarios. Its goal is to exploit multiuser diversity by finding a set of orthogonal users that can be simultaneously served on orthogonal beams, while maintaining the interference low. The key advantage of this transmission scheme is that it only requires partial CSI at the transmitter side in terms of the user received SNIR, making it very suitable for multiuser downlink communications.

The main steps of MOB are illustrated in Figure 6.3. It should be mentioned that these steps describe the main concept behind the MOB scheme without entering into implementation details. These will be more thoroughly addressed in Section 6.4 where the description of the proposed multiuser MAC schemes will take place. At the beginning of each transmission sequence, the AP forms n_t random orthogonal beams, equal to the number of its transmitting antennas (plot (a)). The users measure the SNIR related to each beam, select the highest measured SNIR value to the AP (plot (b)). In turn, the AP selects the best user for each beam and initiates the downlink data transmission (plot (c)). The scheme presented in [80] involves the opportunistic transmission by the users with the highest instantaneous SNIR for each beam, although MOB can also be combined with different scheduling policies.

Through this low-complexity processing based on the instantaneous SNIR values, the MOB scheme achieves a high system sum rate by spatially multiplexing several users at the same time. In the best case where n_t users are selected for downlink transmission, the transmitted signal $\mathbf{x}$ can be expressed as

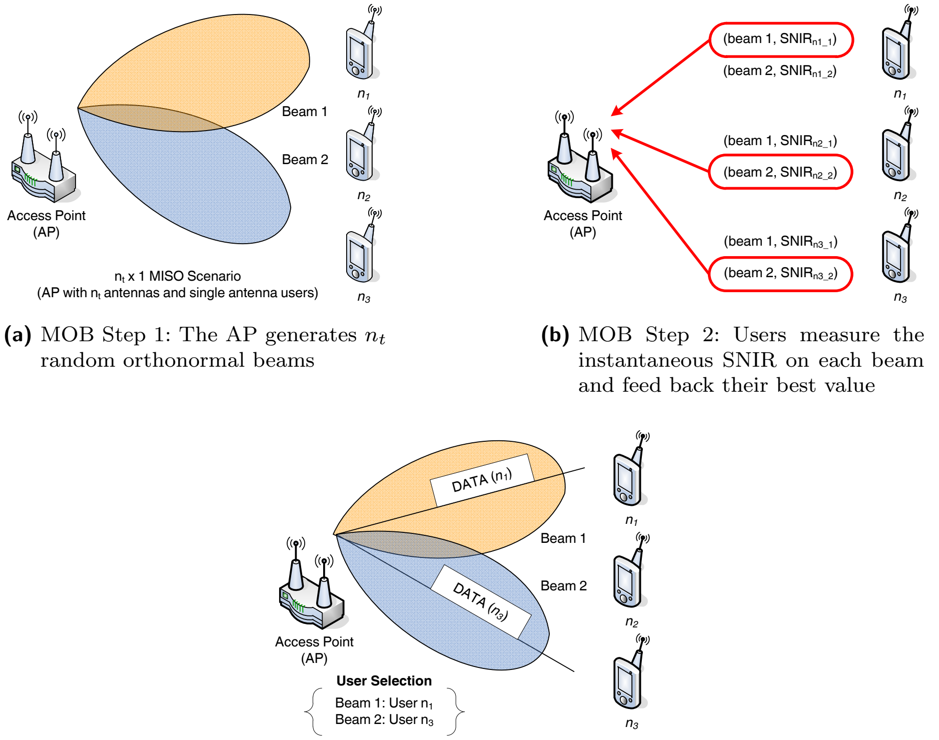
$$\mathbf{x} = \sqrt{\frac{1}{n_t}} \sum_{k=1}^{n_t} \mathbf{b}_k \, s_k \tag{6.2}$$

where s_k are the data symbols that correspond to the kth selected user, $\mathbf{b}_k$ is the assigned unit-power beam and the square root term is employed for total power constraint.

Although the beams are orthogonally generated, some of this orthogonality is lost in the propagation channel [80]. Consequently, some interference is generated by each beam on non-intended users. The SNIR formulation for the kth user that is served by the vth beam is

$$SNIR_{k,v} = \frac{\frac{1}{n_t} \left| \mathbf{h}_k \mathbf{b}_k \right|^2}{\sigma^2 + \sum\limits_{u \neq v}^{n_t} \frac{1}{n_t} \left| \mathbf{h}_k \mathbf{b}_u \right|^2} \tag{6.3}$$

where a uniform power allocation is considered. The numerator is the received power from the desired beam, while the denominator represents the noise plus the interference power from the other beams.



(a) MOB Step 1: The AP generates n_t random orthonormal beams

(b) MOB Step 2: Users measure the instantaneous SNIR on each beam and feed back their best value

(c) MOB Step 3: The AP maps best users on beams and begins downlink transmission

Figure 6.3: Basic steps of MOB transmission technique

As the number of users N grows, the AP can search for users in a larger pool, thus increasing the probability of finding a set of n_t users that do not interfere a lot among themselves [80]. Obviously, having $N \approx n_t$ results in an interference limited system, but for more practical values, such as $n_t = 2$ transmit antennas and $N \geq 10$ users, this scheme is efficient and has been shown to obtain higher performance with respect to single user opportunistic beamforming [81], [82].

The IEEE 802.11n PHY layer specification does not contemplate multiuser transmissions, even though it supports beamforming as a means to achieve higher data rates in point-to-point communications. Since the MOB scheme is practically a random beamforming transmission technique, it can be easily implemented within the standard without any further requirements in terms of hardware. The only necessary modification is to set accordingly the values of the beamforming steering matrices defined in the standard in order to form the random orthonormal beams.

6.4 Multiuser MAC Schemes

The MOB technique is a low-complexity transmission scheme that can be easily implemented at the PHY layer to provide multiuser downlink communications. In a practical system, however, the beamforming scheme must be accompanied by a set of MAC layer functions to collect the necessary feedback information and handle the additional challenges that stem from simultaneous multiuser transmissions. This section will present three MAC layer schemes that modify the IEEE 802.11n MAC protocol to account for the demands and restrictions of the MOB technique. The required modifications are easy to implement within the IEEE 802.11n standard and are backward compatible with the legacy single user transmission, in the sense that MOB and legacy users can coexist in the system.

Since the proposed MAC schemes aim to support the MOB transmission technique, they provide a common set of functions, graphically shown in Figure 6.4. These functions provide a practical MAC layer implementation to complement the three steps of the MOB scheme, namely the generation of the orthonormal beams, the acquisition of CSI feedback and the multiuser downlink transmission. In continuation, it is convenient to first present the common framework that applies to the three proposed schemes before proceeding with their detailed description that will focus on their differences in terms of complexity and efficiency.

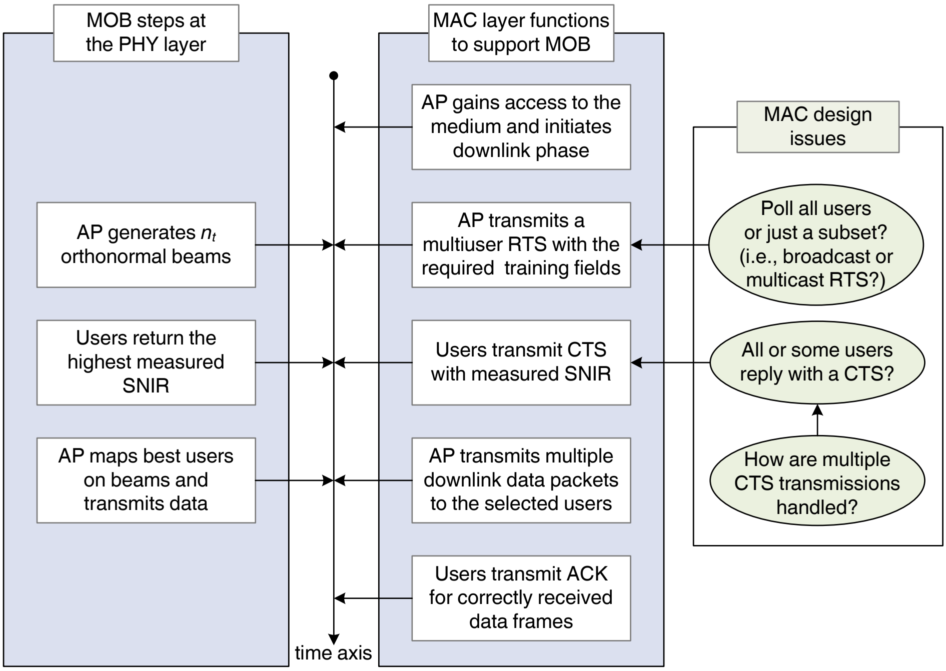

MOB steps at the PHY layer
AP generates n_t orthonormal beams
Users return the highest measured SNIR
AP maps best users on beams and transmits data
time axis
MAC layer functions to support MOB
AP gains access to the medium and initiates downlink phase
AP transmits a multiuser RTS with the required training fields
Users transmit CTS with measured SNIR
AP transmits multiple downlink data packets to the selected users
Users transmit ACK for correctly received data frames
MAC design issues
Poll all users or just a subset? (i.e., broadcast or multicast RTS?)
All or some users reply with a CTS?
How are multiple CTS transmissions handled?

Figure 6.4: MAC layer functions to support the MOB transmission technique

As illustrated in Figure 6.4, the common functions provided by the MAC layer schemes are:

- *The initiation of the downlink phase.* The proposed multiuser schemes constitute a downlink phase that is always initiated by the AP, so for the sake of simplicity the backoff mechanism defined in the IEEE 802.11 specification is not employed in this study. Generally, in a scenario with both uplink and downlink transmissions, the AP would have to follow the backoff rules to gain access to the medium before initiating the downlink phase.

- *The generation of a multiuser RTS frame.* The initiation of the downlink phase is marked by the transmission of a modified RTS frame that basically serves two purposes:

    1. it is a call for participation in the downlink phase that may be addressed to a subset or to all the associated users (i.e., multicast or broadcast). The employed receiver address included in the RTS is a point of differentiation between the proposed schemes and will discussed later in this section.
    2. it acts as a sounding frame that will enable the receiving users to measure the SNIR on each of the n_t generated beams. For this reason, the PHY layer preamble of the RTS contains a number of HT-LTFs (High-Throughput Long Training Fields), as defined in IEEE 802.11n standard. Apart from the training fields, the main body of the RTS frame is transmitted conventionally (i.e., on a single beam).

    The structure of the modified RTS frame is shown in Figure 6.5. The length of the PHY layer preamble of the RTS frame is determined by the number n_t of spatial streams (i.e., orthonormal beams and subsequently antennas). For a single-antenna transmission, a PHY layer header of 28 μs is introduced, whereas for every additional spatial stream an extra HT-LTF of 4 μs required. The description of the PHY header fields is given in Table 6.1 and more details can be found in the IEEE 802.11n specification [2].[1] The length of the MAC header mainly depends on the receiver address field. When a single receiver address is employed, the MAC header has a length of 20 bytes. Nevertheless, some of the proposed MAC schemes include multiple destinations in this address field, as it will be further clarified later.

- *The transmission of a CTS frame by the downlink users.* Once the users receive the RTS frame and estimate their channel quality, they reply with a CTS frame that (unless otherwise stated) contains the best measured SNIR value and an integer identifier that corresponds to the respective beam. The structure of the modified CTS frame is shown in Figure 6.6. Assuming single-antenna users, a 28 μs PHY layer preamble is required, whereas the MAC header complies with the IEEE 802.11n specification, with the addition of an

[1] The PHY layer header structure presented in this section has been based on the IEEE 802.11n greenfield operation mode meant for IEEE 802.11n-only compatible stations. If compatibility with legacy devices is desired, the PHY layer headers should be modified accordingly, as indicated in Clause 20 of the IEEE 802.11n specification [2].

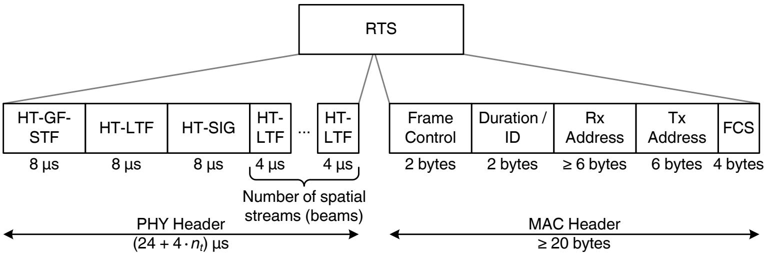


Figure 6.5: Structure of the modified RTS frame

Table 6.1: Elements of the PHY layer header for the Multiuser MAC schemes

Element	*Description*
HT-GF-STF	High-Throughput (HT) Greenfield Short Training Field
HT-LTF1	First HT Long Training Field
HT-SIG	HT SIGNAL Field
HT-LTF	HT Long Training Field

extra 1-byte field that contains the CSI information (i.e., the SNIR and the beam identifier).[2] Two design issues arise at this point. The first is whether a CTS should be transmitted by every polled user or a limit should be posed to the number of CTS replies, for example by filtering out users with very bad channel conditions. The second issue concerns the transmission order of the CTS frames by the multiple users which can be either deterministic, thus collision-free, or random (probabilistic) that will likely result to collisions among simultaneously transmitted CTS. These two issues will be handled in different ways by the proposed MAC schemes, as it will be discussed later.

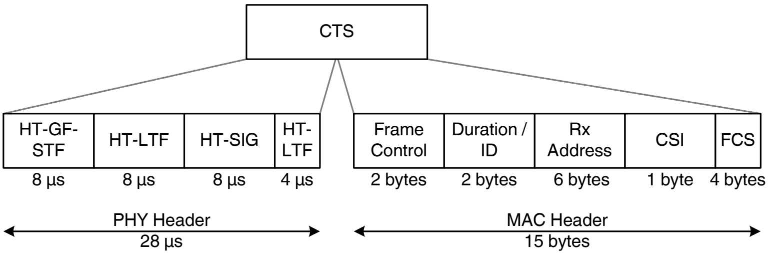


Figure 6.6: Structure of the modified CTS frame, including CSI feedback

- *The transmission of multiuser data frames by the AP.* Once the AP collects the feedback information included in the CTS frames it assigns the user with

[2] In this work, it has been assumed that a SNIR quantization scheme has been employed so that the CSI field can be sufficiently represented by 1 byte.

the highest measured SNIR on each beam (at most one user per beam) and transmits a maximum of n_t data packets simultaneously. The data packets employ the channel over the same time, frequency and code but are transmitted over different beams. This can be supported by the IEEE 802.11n standard, by exploiting the multiplexing capabilities of multi-antenna systems. This is actually an important shift from current systems where the simultaneous transmission of multiple packets in the same medium leads to collision and packet loss. Link adaptation is also employed and the transmission rate on each beam is determined by the measured SNIR.

- *The transmission of ACK frames.* The users signal the correct reception of a data frame by transmitting an ACK. In the proposed schemes, the multiple (up to n_t) ACK frames are transmitted sequentially, following the mapping of the users onto the beams.

In the remaining part of this section, the three proposed MAC layer schemes will be described in detail.

6.4.1 Mu-Basic Scheme

The first and simplest scheme is called *Mu-Basic* and is a straightforward adaptation of the IEEE 802.11 mechanism to support downlink multiuser transmission. This scheme is based on the principle that at most n_t users can be served simultaneously by an AP equipped with n_t transmitting antennas that generate an equal number of orthogonal beams. Hence, in the beginning of the transmission sequence, the AP randomly selects n_t users from the downlink message buffer and transmits a multidestination RTS frame that includes the respective n_t receiver addresses, as illustrated in Figure 6.7.

The figure focuses on the receiver address field, since the remaining part of the RTS frame follows the structure indicated previously (Figure 6.5). The order in which the addresses are listed serves two purposes. First, it indicates the order in which CTS frames are to be sent in order to avoid collisions. Second, the address list is used to implicitly map the polled users to the beams. The users that receive the RTS frame check whether their address is in the list and wait for a predefined time before sending a CTS, which includes the SNIR measurement that corresponds to the assigned beam.[3] Note that in this case, the users do not reply with the best SNIR value since the beam assignment is predefined by the AP.

The AP proceeds to the simultaneous transmission of the n_t data packets after selecting the transmission rate for each beam, according to the corresponding SNIR measurement that indicates the link quality. The users acknowledge the data reception by sequentially sending an ACK frame. An example of the transmission

[3] Since each CTS slot is of a fixed duration (i.e., a SIFS time and the time required for the transmission of the 15 byte CTS with the minimum available transmission rate) and assuming negligible propagation delays, each user can determine when to initiate the CTS transmission.

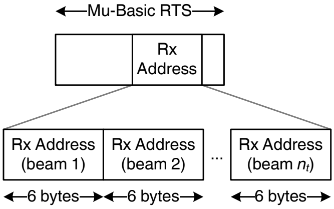


Figure 6.7: The modified RTS frame for the Mu-Basic scheme

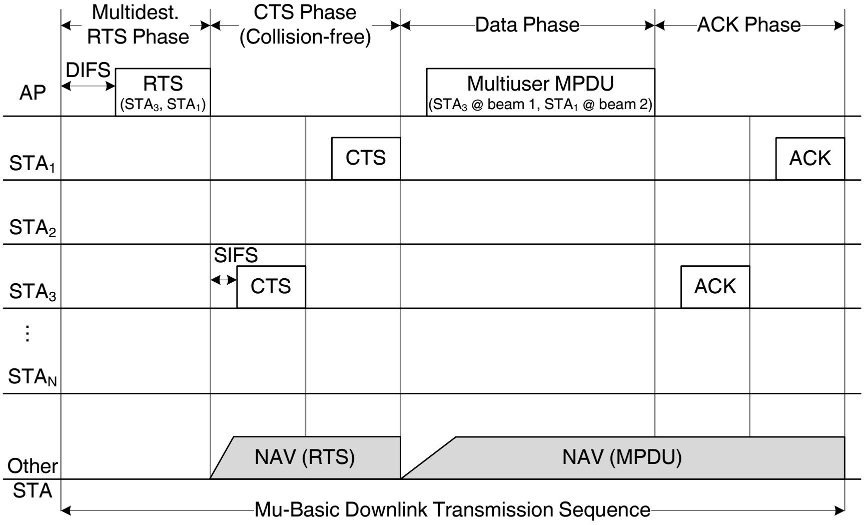


Figure 6.8: Transmission sequence example for the Mu-Basic scheme

sequence according to the Mu-Basic scheme is given in Figure 6.8. In this example, there are $n_t = 2$ antennas at the AP, so two users are randomly selected for transmission (STA$_3$ and STA$_1$).

To avoid collisions by users that do not participate in the process, the NAV mechanism can be employed (described in Section 2.3.2). For this reason, the time from the transmission of the RTS until the end of the CTS phase is marked in the duration field of the RTS frame (Figure 6.5). The remaining time of the frame sequence, from the end of the CTS phase until the transmission of the last ACK, is indicated in the respective duration field of the data packet MAC header. Hence, non-participating users can set their NAV timer upon the RTS reception and can later update it when the header of a data packet is decoded.

Mu-Basic is easy to implement since it is a simple polling scheme initiated by the AP. Its performance will serve as a benchmark for the evaluation of the two more advanced multiuser schemes that will be presented next. In the considered case the destination users are randomly selected, however different criteria could also be applied to prioritize users with specific demands (e.g., with delay sensitive applications). Mu-Basic requires some additional overhead in the RTS frame as multiple receiver addresses must be included, but has the shortest possible CTS phase, since the number of received CTS frames is equal to the n_t served users (it would not make sense to receive feedback from less than n_t users if all the parallel streams were to be employed). On the other hand, multiuser diversity is not exploited since the users are scheduled without any consideration of their channel quality. Thus, the user selection and the beam assignment processes are not optimally done. As a result, the interference among the scheduled set of users may be high, leading to transmissions at low data rates (i.e. interference controlled system).

6.4.2 Mu-Opportunistic Scheme

In an effort to exploit multiuser diversity and transmit opportunistically to the best set of users, according to the principles of the MOB transmission technique, the *Mu-opportunistic* scheme has been proposed. This scheme provides a mechanism for the AP to acquire the CSI of all users before reaching a scheduling decision, in order to optimize user selection and beam allocation. To this end, in the beginning of the transmission sequence, the AP polls all users with available data for downlink transmission. For the sake of simplicity, it will be assumed that the system is under saturation and there is always downlink traffic for each of the N system users.[4] Hence, the AP transmits a multidestination RTS frame that includes the N receiver addresses of all the network, as illustrated in Figure 6.9.

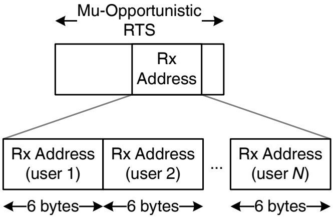


Figure 6.9: The modified RTS frame for the Mu-Opportunistic scheme

The users measure the SNIR on all the beams and include the maximum SNIR value in the CTS, along with an integer identifier of the beam that yielded that value. As before, CTS packets are transmitted in a collision-free manner, following the order of the address list in the RTS. After receiving all the feedback, the AP assigns each beam to the user with the highest SNIR and proceeds to the downlink data transmission. If a beam is not selected by any user then it is not used for transmission, even though this is not likely to happen very often for a large number of active users and a time-varying channel. Correct data reception is marked by the transmission of ACK frames that are sent sequentially, according to the beam allocation order (the user served on the first beam replies first, and so on). An example of the transmission sequence according to the Mu-Opportunistic scheme is

[4] The non saturation case will be examined later in this chapter.

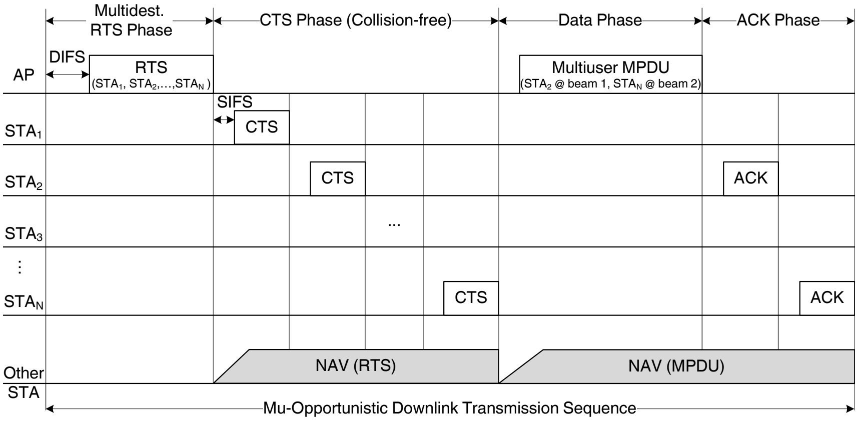


Figure 6.10: Transmission sequence example for the Mu-Opportunistic scheme

given in Figure 6.10. In this example, there are $n_t = 2$ antennas at the AP and N users with available data. The AP receives N CTS frames and then selects the best set of users (STA_2 and STA_N, in the example) for the downlink data transmission.

The Mu-Opportunistic fully exploits multiuser diversity since it opportunistically schedules users with good channel conditions and with low mutual interference (i.e., users with high SNIR values measured on different beams). The weakness of this scheme is that it introduces significant overhead, mainly due to the long CTS phase, and the trade-off between overhead and efficiency becomes critical, especially as the number of users N grows.

6.4.3 Mu-Threshold Scheme

The Multiuser Threshold-Selective algorithm (*Mu-Threshold)* is the third proposed multiuser MAC layer scheme. It maintains the opportunistic scheduling policy of selecting a set of users with high rates and low mutual interference but also aims to limit the additional control overhead. In order to achieve these objectives, it introduces two major changes with respect to the Mu-Opportunistic scheme:

- Instead of the deterministic, collision-free CTS transmissions, Mu-Threshold introduces a CTS contention phase during which users compete with each other within a predefined number of slots. Generally, even though collisions among CTS frames are likely to occur, the number of slots is smaller than the total number of users, thus reducing the length of the CTS phase.
- In order to reduce the CTS collision probability, the algorithm imposes a SNIR threshold so that only users with a relatively good channel are allowed to participate in the feedback process. Even though the idea of threshold application is not new, the novelty lies in the inclusion of this concept on a feasible MAC scheme for a multiuser MIMO scenario.

The frame exchange sequence of the Mu-Threshold scheme is initiated with the broadcast transmission of an RTS by the AP. The advantage of this configuration is that it calls all the users to participate in the CTS contention phase by employing a single 6-byte destination address instead of a long address list, as shown in Figure 6.11. Without doubt, this setup is meaningful under a saturation scenario in which the AP has always packets to transmit to all the associated users. This consideration is made to facilitate the evaluation of the full potential of the Mu-Threshold scheme, given that opportunistic downlink schemes are mostly needed under high-traffic conditions. In non-saturation conditions, the Mu-Threshold scheme could be applied with a minor modification. In this case, the AP would have to periodically set up multicast groups with the subset of active users (i.e., those who are waiting to receive downlink data) and use a multicast instead of a broadcast address. This point will be further discussed later in this chapter, in Section 6.6.4.

After the RTS transmission, a CTS contention phase of m slots is initiated, with m being a system parameter subject to optimization. The slots have a predefined length, equal to a SIFS duration plus the time required for the transmission of the 15 byte CTS with the minimum available transmission rate. Depending on whether the maximum SNIR measured by a user is above or below the threshold, the user is either allowed to participate in this phase, or forced to remain silent until the beginning of a new frame sequence. Those allowed to participate select randomly a slot with equal probability and transmit a CTS containing the maximum measured SNIR and the corresponding beam identifier. Whenever multiple users select the same slot a collision occurs and the involved CTS frames are considered lost (the capture effect is not considered, even though it could increase the effectiveness of the proposed scheme). A slot can also remain empty if no user selects it for transmission.

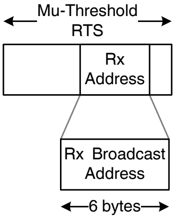


Figure 6.11: The modified RTS frame for the Mu-Threshold scheme

The next stage of the algorithm depends on the outcome of the contention phase. If no CTS has been correctly received (due to either collisions or lack of user participation because of the SNIR threshold value) no data is transmitted and a new contention phase is initiated.[5] User synchronization has been assumed, so that a collision in the mth slot only affects the involved CTS packets and does not have any effect on transmissions in the remaining slots of the contention phase. Thus, if at least one CTS is received, transmission of downlink data packets can take place. As before, the AP assigns the best user on each beam, based on the feedback

[5] Different policies could be implemented to avoid the presence of empty frames (e.g., transmission to a randomly selected user or to a user with a long waiting time using a basic rate) but will not be considered in this work.

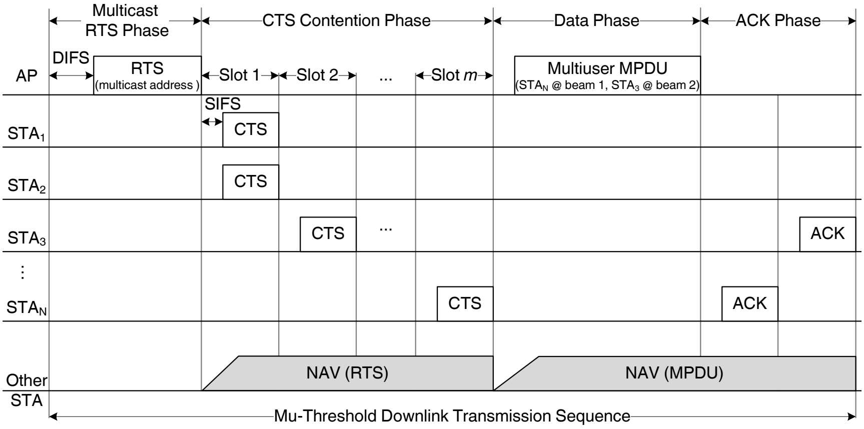


Figure 6.12: Transmission sequence example for the Mu-Threshold scheme

information collected by the received CTS frames and transmits a maximum of n_t data packets simultaneously. Note that, unlike the contention phase where collisions among CTS frames can occur, the transmission of data is collision-free. Finally, the users acknowledge the data reception by sequentially sending an ACK frame, following the order of the user the mapping onto the beams.

An example of the transmission sequence according to the Mu-Threshold scheme is given in Figure 6.12. In this example, there are $n_t = 2$ antennas at the AP and N users with available data that compete in m CTS slots (with $m < N$ in general). Some users may select the same slot and collide (e.g., STA$_1$ and STA$_2$), others may transmit a CTS successfully (e.g. STA$_3$ and STA$_N$) and finally a number of users will refrain from this phase due to their unfavorable channel conditions.

An important decision is the selection of the SNIR threshold that serves two purposes: it reduces the number of contending users, thus decreasing the probability of CTS collisions, and it filters out those users with harsh channel condition, resulting to transmissions with higher data rates. Nevertheless, selecting a high threshold could cause adverse effects such as starvation if the majority of users experience low link quality. The threshold is determined by the AP and it is made known to the users during an initial association phase (alternatively, it could be included in the RTS packet, thus increasing its size by a few overhead bits). It is also possible to design a dynamic scheme that will adapt the threshold value at runtime depending on measured channel statistics.

The number of the CTS contention slots m is another important parameter that depends on the number of participating users which, in turn, is determined by the total number of users N, their channel condition and the selected threshold. An interesting observation is that, since the duration of each CTS slot is fixed, the duration field of the RTS packet (that indicates the length of the CTS phase) implicitly reveals the number of contention slots m. Therefore, the AP can let the users know the value of m without requiring an additional control field.

6.5 Analytical Model for the Throughput Evaluation of the Mu-Threshold Scheme

A mathematical model has been developed to calculate the throughput performance of the Mu-Threshold algorithm as a function of the number of users N, the number of contention slots m and the rate threshold r_γ. It has been assumed that the system is saturated and there are always downlink packets for all associated users. The number of transmitter antennas at the AP has been set to $n_t = 2$, for two reasons. First, it seems to be the most practical setup up to date in existing WLAN systems (i.e., in the majority of IEEE 802.11n and pre-n commercial products) and second, it permits a more intuitive interpretation of the analytical model. The analysis could be extended to a larger number of antennas but this would significantly increase the computational complexity of the presented results.

The proposed model can be used to determine the system parameters that maximize performance. For example, assuming that the channel is known, the best combination for the threshold and the number of slots can be calculated for a given number of users. Another possible application could be to determine the optimum number of users and implement a traffic control policy by adjusting the users that participate in the downlink process. Table 6.2 summarizes the main variables employed in the model description.

Table 6.2: Main parameters for the Mu-Threshold throughput analysis

Symbol	*Description*
n_t	Number of antennas (and antenna beams) at the AP ($n_t = 2$)
m	Number of contention CTS slots
r_w	Data transmission rate (bps), $w \in [1, R]$
r_γ	Rate threshold, $\gamma \in [1, R]$
N	Number of users
n	Users participating in a given contention phase ($n \leq N$)
s	Users surviving a given contention phase ($s \leq n$)
b	Users assigned on the first beam ($b \leq s$)
i	type of frame (i.e. outcome of the data phase) (i=0,1,2 for empty frame, single and double user transmission)

To enhance readability, the analysis is divided into three parts. The first part discusses the channel statistics, the second presents the general formulation of the throughput expression and the third part discusses in detail the calculation of the probabilities of having different outcomes in each frame. As it will be explained in more detail next, depending on the outcome of the contention phase, there are three possible frame types: empty frame, when no data is transmitted; single frame, when a single user is selected by the AP; and double frame, when two users are scheduled on the two available beams. Finally, with some minor modifications, the model can also be employed for the calculation of the average throughput of the Mu-Opportunistic scheme. This issue will be discussed in the fourth and last part of this section.

6.5.1 The Channel Distribution

Consider an AMC scheme that offers R available rates $\{r_1, \ldots, r_R\}$, in ascending order. Each rate can be used for transmission when the measured SNIR of the particular link lies within a predefined SNIR range. Obviously, the SNIR of a link is time-varying and depends on the instantaneous channel conditions, but the probability of a user SNIR being in a given range can be statistically estimated, as the channel distribution is known. Following the calculations in [81] for the MOB system, the approximate Cumulative Distribution Function (CDF) of the SNIR value y that is measured at the received is

$$F(y) = \left[1 - \frac{e^{-(y n_t \sigma^2)}}{(1+y)^{n_t - 1}}\right]^{n_t} = \left[1 - \frac{e^{-(2y\sigma^2)}}{(1+y)}\right]^2 \tag{6.4}$$

as each user feeds back the maximum SNIR value with respect to the $n_t = 2$ beams and σ^2 is the noise variance.

The probability $P_r(r_w)$ of a user having available a particular rate r_w, with $w \in [1, R]$, is equal to the probability of having a SNIR within a specific predefined range, below y_{w+1} and above y_w and can be calculated with the user of the CDF as

$$P_r(r_w) = F(y_{w+1}) - F(y_w). \tag{6.5}$$

A SNIR threshold y_γ is defined so that only users with a higher SNIR value can participate in the contention phase. Equivalently, it can be said that a corresponding rate threshold r_γ is imposed and users with $r_w \geq r_\gamma$ (with $w \geq \gamma$) can contend for access.

6.5.2 Average Throughput Calculation

The average throughput $S(m, N, r_\gamma)$ for m slots, N users and a threshold of r_γ is defined as the average number of transmitted bits per frame $\bar{x}$ divided by the average frame duration $\bar{t}$

$$S(m, N, r_\gamma) = \frac{\bar{x}(m, N, r_\gamma)}{\bar{t}(m, N, r_\gamma)} \tag{6.6}$$

where

$$\bar{x}(m, N, r_\gamma) = \sum_{i=1}^{2} i \cdot l \cdot \left(\sum_{w=\gamma}^{R} P_f(i, m, N, r_\gamma, r_w) \right) \tag{6.7}$$

and

$$\begin{aligned}\bar{t}(m,N,r_\gamma) = {} & t(0,m)\cdot P_f(0,m,N,r_\gamma)\\ & + \sum_{i=0}^{2}\sum_{w=\gamma}^{R} t(i,m,r_w)\cdot P_f(i,m,N,r_\gamma,r_w).\end{aligned} \tag{6.8}$$

The terms included in the above equations will be explained next. The index i expresses the three possible frame types: $i = 0$ indicates an empty frame in which no data transmission has taken place; $i = 1$ corresponds to single transmission of a data packet of length l bits; finally $i = 2$ indicates a double transmission frame where two users have simultaneously transmitted data packets on the two available beams, corresponding to $2 \cdot l$ transmitted bits.

The average transmitted bits $\bar{x}$ can be calculated by multiplying the transmitted bits per frame type by the probability that the particular frame type will occur, for all data rates that are above or equal to the threshold. This probability of having a frame of type i transmitted with a rate of r_w, for a given number of slots m, users N and threshold r_γ is denoted by $P_f(i,m,N,r_\gamma,r_w)$ and its calculation is based on the following factors:

- the SNIR distribution of the users and the probability of them being above the threshold.
- the outcome of the contention phase of m slots and the number of users that survive (by successfully transmitting a CTS).
- the opportunistic selection of the best user for each beam from the subset of surviving users.

The expression for the $P_f(i,m,N,r_\gamma,r_w)$ and its derivation are given in the next section.

The term $t(i,m,r_w)$ in (6.8) expresses the total transmission time of a frame sequence of type i when rate r_w is used and is calculated as

$$t(i,m,r_w) = t_{data}(i,r_w) + t_{ovh}(i,m) \tag{6.9}$$

where $t_{data}(i,r_w)$ is the transmission time of the data packet for a frame of type i and $t_{ovh}(i,m)$ is the control overhead.

The data transmission time is given by the following expression:

$$t_{data}(i,r_w) = \begin{cases} 0, & \text{if } i = 0 \text{ (empty frame)}\\ t_{data}(r_w), & \text{if } i > 0 \text{ (single/double frame)}\end{cases} \tag{6.10}$$

For a non-empty frame, the t_{data} can be easily calculated for a known packet size l and a given transmission rate r_w. For an empty frame ($i = 0$), the t_{data} will be equal to zero and the total frame duration will only consist of the overhead.

The overhead time t_{ovh} is also a known quantity for each frame type and by considering the frame sequence depicted in Figure 6.12, can be calculated as follows:

$$t_{ovh}(i,m) = \begin{cases} DIFS + RTS + m \cdot (CTS + SIFS), & \text{if } i = 0 \text{ (empty frame)} \\ t_{ovh}(0,m) + 2 \cdot SIFS + t_{o,data} + ACK, & \text{if } i = 1 \text{ (single frame)} \\ t_{ovh}(0,m) + 3 \cdot SIFS + t_{o,data} + 2 \cdot ACK, & \text{if } i = 2 \text{ (double frame)} \end{cases} \tag{6.11}$$

where t_o data contains the PHY and MAC headers introduced in the data frame.[6] In the case of an empty frame, the overhead time $t_{ovh}(0,m)$ consists of the transmission time of an RTS frame and the duration of the contention window of m slots, with each slot consisting of a CTS frame and a SIFS. The same overhead is introduced in the case of non-empty frames and, in addition the data overhead t_o, the transmission time of the ACK frames (one or two, depending on whether a single or double transmission has taken place, respectively) and any additional SIFS must be included.

Note that in the case of an empty frame ($i = 0$), the frame duration and the probability P_f are independent of the transmission rate and the index r_w in (6.8) is dropped for convenience.

6.5.3 Frame Type Probabilities

To calculate the probability $P_f(i, m, N, r_\gamma, r_w)$ of having a frame of type i one must consider the implementation steps of the Mu-Threshold algorithm. First, only a fraction n of the total N users, those with an available rate of $r_w \geq r_\gamma$ (with $w \geq \gamma$), are allowed to participate in the contention phase. As the channel statistics are known, the probability that exactly n out of N users have a rate above the threshold r_γ, can be calculated with the use of the SNIR CDF in (6.4), as

$$P_{select}(n, r_\gamma) = \binom{N}{n} \big(1 - F(y_\gamma)\big)^n \big(F(y_\gamma)\big)^{N-n}. \tag{6.12}$$

with y_γ the SNIR value that corresponds to the rate threshold r_γ.

Those n users that pass the threshold selection phase will contend for channel access by transmitting a CTS in one of the m system slots. If a slot is selected by exactly one user, then the contained CTS is successfully received and the respective user is said to have survived the contention phase. We define the probability $P_{survive}(s, m, n)$ of having exactly s users surviving the contention phase of m slots, when there are n participating users (i.e. users that will transmit a CTS in the current CW with probability 1). This is a combinatorial problem known as the "assignment of n packets in m cells" . It considers all the possible assignments

[6] The exact values for the time calculations can be determined by consulting the IEEE 802.11n specification [2]. The main parameters considered in this thesis will be further give in Section 6.6.1 where the system setup will be discussed.

of the n users in the m slots (including the cases where multiple users select the same slot) and then calculates the probability of having exactly s slots with success (i.e., selected by a single user) whereas the remaining $m-s$ slots are empty or have suffered a collision. This problem has been analyzed in [83] and the final expression for $P_{survive}(s,m,n)$[7] is

$$P_{survive}(s,m,n) = \\ = \begin{cases} \frac{(-1)^s m! n!}{m^n s!} \sum\limits_{j=s}^{min(m,n)} \frac{(-1)^j (m-j)^{n-j}}{(j-s)!(m-j)!(n-j)!} & , s \in [0, min(m,n)] \\ 0 & , \text{otherwise.} \end{cases} \tag{6.13}$$

Note that the above formula is defined for values of $s \in [0, min(m,n)]$, since the number of successful slots s in a frame cannot exceed the number of participating users n, or the total number of slots m. In addition, $P_{survive}$ returns a zero value for several combinations of s,m and n within the defined range, thus indicating impossible outcomes. For instance, consider the case of having $n = 2$ participating users competing in $m = 3$ slots. Given that both users will attempt to transmit a CTS, they will either select the same slot and collide, leading to zero successful slots ($s = 0$), or they will select different slots leading to two successful outcomes ($s = 2$). Under these circumstances, the event of having exactly one successful slot ($s = 1$) is not possible and hence $P_{survive}(s = 1, m = 3, n = 2) = 0$.

The probability $P_f(i, m, N, r_\gamma, r_w)$ is calculated for the three different frame types. Figure 6.13 illustrates these three possible outcomes and summarizes the conditions that lead to each case. These conditions will be the base for the calculation of $P_f(i, m, N, r_\gamma, r_w)$ that is given next.

If no user has a rate above the threshold ($n = 0$), or no user survives the contention phase ($n > 0$ but $s = 0$), an empty frame will follow. Thus, the probability of having an empty frame is

$$\begin{aligned} P_f(i = 0, m, N, r_\gamma) = \\ &= P_{select}(0, r_\gamma) \\ &+ \sum_{n=1}^{N} P_{select}(n, r_\gamma) \cdot P_{survive}(0, m, n). \end{aligned} \tag{6.14}$$

A single transmission frame occurs when there is at least one surviving user ($s \geq 1$) and all the surviving users select the same beam. Hence, the probability of

[7] The result of this expression, as well as the time duration for each frame type, could be precalculated and included in a lookup table in the case of runtime application of the theoretical model.

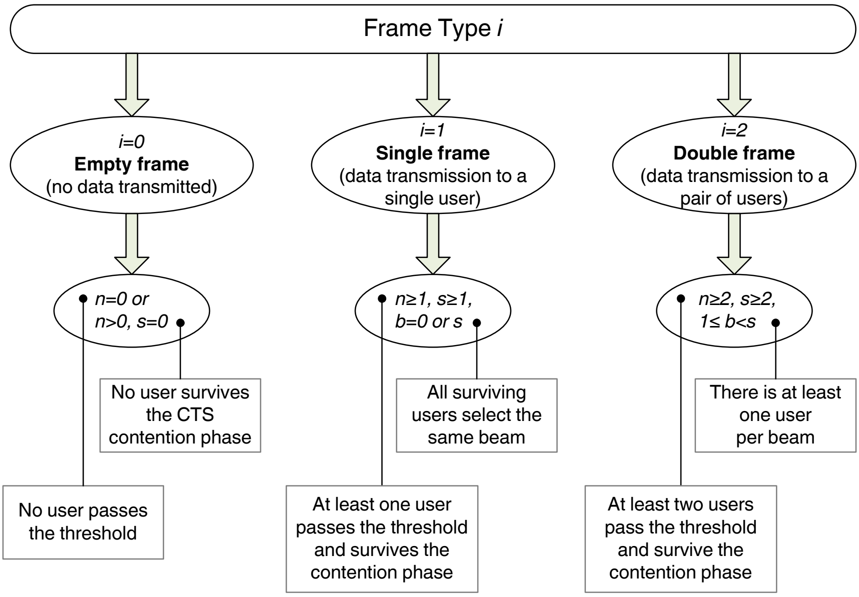

Frame Type *i*
i=0
Empty frame
(no data transmitted)
i=1
Single frame
(data transmission to a single user)
i=2
Double frame
(data transmission to a pair of users)
n=0 or n>0, s=0
n≥1, s≥1, b=0 or s
n≥2, s≥2, 1≤ b<s
No user survives the CTS contention phase
All surviving users select the same beam
There is at least one user per beam
No user passes the threshold
At least one user passes the threshold and survives the contention phase
At least two users pass the threshold and survive the contention phase

Figure 6.13: Possible frame type outcomes in Mu-Threshold scheme

having a single transmission frame with rate r_w is

$$\begin{aligned} &P_f(i=1,m,N,r_\gamma,r_w) = \\ &= \sum_{n=1}^{N}\left(P_{select}(n,r_\gamma)\cdot\sum_{s=1}^{min(m,n)}\left[(P_{survive}(s,m,n)\right.\right. \\ &\left.\left.\cdot\ Pr\Big\{s \text{ users on same beam}\Big\}\cdot P_{r1}(r_w,s)\right]\right) \end{aligned} \tag{6.15}$$

where $P_{r1}(r_w,s)$ is the probability that rate r_w is used for transmission.

The system considers two available beams ($n_t=2$) and each user may be assigned to a beam with an equal probability of 0.5. The probability $P_b(b,s)$ of having b out of s users assigned on the first beam (and hence $s-b$ users on the second) can be expressed as

$$P_b(b,s) = \binom{s}{b}\cdot\left(\frac{1}{2}\right)^b\cdot\left(\frac{1}{2}\right)^{s-b} = \binom{s}{b}\cdot 2^{-s}. \tag{6.16}$$

It can be easily derived that the probability of having all users selecting the same beam (either the first or the second) is equal to $P_b(s,s)+P_b(0,s)=2^{1-s}$.

Since the scheme is opportunistic, the surviving user with the highest rate will be selected for transmission. In other words, the transmission rate will be r_w if there is at least one surviving user with this rate while there is no user with a rate above

r_w. Hence, the probability $P_{r1}(r_w, s)$ that r_w is the maximum available rate among s surviving users and can be calculated as

$$\begin{aligned}
&P_{r1}(r_w, s) = \\
&\quad = Pr\Big\{(\text{at least 1 user with } r_w) \cap (\text{no user with } r > r_w)\Big\} \\
&\quad = Pr\Big\{(\text{at least 1 user with } r_w) | (\text{no user with } r > r_w)\Big\} \\
&\qquad \cdot Pr\Big\{\text{no user with } r > r_w\Big\} \\
&\quad = \Big(1 - Pr\Big\{\text{all users with } r < r_w\Big\}\Big) \\
&\qquad \cdot Pr\Big\{\text{all users with } r \leq r_w\Big\} \Rightarrow
\end{aligned}$$

$$P_{r1}(r_w, s) = \left(1 - \left[\frac{\sum\limits_{v=\gamma}^{w-1} P_r(r_v)}{\sum\limits_{v=\gamma}^{w} P_r(r_v)}\right]^s\right) \cdot \left[\frac{\sum\limits_{v=\gamma}^{w} P_r(r_v)}{\sum\limits_{v=\gamma}^{R} P_r(r_v)}\right]^s . \tag{6.17}$$

Note that the rates of all surviving users are greater or equal the rate threshold r_γ.

So far, the calculation of the probability P_f for the cases of $i = 0$ and $i = 1$ has been presented. We will now proceed to the third case of having a double transmission frame ($i = 2$) that occurs when there are at least two surviving user ($s \geq 2$) and at least one user is assigned per beam (i.e not all users on the same beam). The transmission rate on each beam will be equal to the highest rate available among the users assigned on that beam. Although different rates may be used on each of the two beams, the total frame sequence duration is determined by the lower rate (i.e. the longest transmission of the two). We define $P_{r2}(r_w, s)$ as the probability that the frame duration is determined by rate r_w, given that both beams are used for transmission. Then

$$\begin{aligned}
&P_f(i = 2, m, N, r_\gamma, r_w) = \\
&\quad = \sum_{n=2}^{N} P_{select}(n, r_\gamma) \cdot \sum_{s=2}^{min(m,n)} P_{survive}(s, m, n) \cdot P_{r2}(r_w, s)
\end{aligned} \tag{6.18}$$

with the probability $P_{r2}(r_w, s)$ given by

$$P_{r2}(r_w, s) = \sum_{b=1}^{s-1} P_{r2\_cond}(r_w, b, s) \cdot P_b(b, s). \tag{6.19}$$

In this equation $P_b(b, s)$ is the probability of having b users on the first beam, calculated by (6.16). Then, $P_{r2\_cond}(r_w, b, s)$ is the conditional probability that the frame duration is determined by rate r_w, when b out of s users are assigned to the first of the two beams (and $s - b$ to the second). This probability can be calculated with the help of (6.17) as

$$
\begin{aligned}
&P_{r2\_cond}(r_w, b, s) = \\
&\quad = Pr\{\text{beam 1: rate } r_w \text{ used}|b \text{ users}\} \\
&\qquad \cdot Pr\{\text{beam 2: } r > r_w|s-b \text{ users}\} \\
&\quad + Pr\{\text{beam 2: rate } r_w \text{ used}|s-b \text{ users}\} \\
&\qquad \cdot Pr\{\text{beam 1: } r > r_w|b \text{ users}\} \\
&\quad + Pr\{\text{beam 1: rate } r_w \text{ used}|b \text{ users}\} \\
&\qquad \cdot Pr\{\text{beam 2: rate } r_w \text{ used}|s-b \text{ users}\} \Rightarrow
\end{aligned}
$$

$$
\begin{aligned}
&P_{r2\_cond}(r_w, b, s) = \\
&\quad = P_{r1}(r_w, b) \cdot \sum_{v=w+1}^{R} P_{r1}(r_w, s-b) \\
&\quad + P_{r1}(r_w, s-b) \cdot \sum_{v=w+1}^{R} P_{r1}(r_w, s) \\
&\quad + P_{r1}(r_w, b) \cdot P_{r1}(r_w, s-b).
\end{aligned}
\tag{6.20}
$$

Equations (6.14), (6.15) and (6.18) can be used in (6.7) and (6.8) to calculate the system throughput.

6.5.4 Analytical Model for the Throughput Evaluation of the Mu-Opportunistic Scheme

So far, an analytical framework for the throughput calculation of the Mu-Threshold scheme has been provided. This section will discuss a number of modifications that can be made to the model, in order to extract the throughput calculation of the Mu-Opportunistic scheme. In fact, the Mu-Opportunistic scheme is simpler to model, since it does not employ a SNIR threshold and ensures a collision-free CTS phase.

The channel statistics given in Sections 6.5.1 are also valid in the case of the Mu-Opportunistic scheme, whereas the throughput formulation in Section 6.5.2 can be employed by setting $m = N$ (i.e., a fixed number N CTS slots) and dropping the threshold. Hence, equation 6.6 for the average throughput can be rewritten as:

$$S(N) = \frac{\bar{x}(N)}{\bar{t}(N)} \tag{6.21}$$

with the average number of transmitted bits per frame:

$$\bar{x}(N) = \sum_{i=1}^{2} i \cdot l \cdot \left(\sum_{w=1}^{R} P_f(i, N, r_w) \right) \tag{6.22}$$

and the average frame duration:

$$\bar{t}(N) = t(0) \cdot P_f(0, N) + \sum_{i=0}^{2} \sum_{w=1}^{R} t(i, r_w) \cdot P_f(i, N, r_w). \tag{6.23}$$

Clearly, the frame duration times should be calculated appropriately, using the frame sequence structure shown in Figure 6.10 as a reference.

The main modifications of the analytical model concern the calculation of the frame type probability $P_f(i, N, r_w)$ due to the following reasons:

- The Mu-Opportunistic scheme does not consider a threshold and all users are called to participate in the CTS phase. Nevertheless, even for transmission with the minimum available rate r_1, a minimum SNIR threshold condition must be satisfied. Hence, the probability P_{select} of having n out of N users participating in the CTS phase (eq. 6.12) can be rewritten as:

  $$P_{select}(n) = \binom{N}{n} \big(1 - F(y_1)\big)^n \big(F(y_1)\big)^{N-n}. \tag{6.24}$$

  where instead of the threshold, the lowest available transmission rate r_1 is employed and $F(y_1)$ the CDF of the minimum SNIR value y_1 calculated by equation (6.4). In other words, all users that have a rate $r_w \geq r_1$ will participate in the contention phase. Practically, unless a particularly harsh channel is considered, all N users are likely to satisfy the minimum threshold condition for the majority of the time.

- There are no collisions during the CTS phase, therefore all participating users (n out of N) will survive (i.e., their feedback will reach the AP). As a result, the probability $P_{survive}$ is greatly simplified from eq 6.13

  $$P_{survive}(s, m = N, n) = \begin{cases} 1, s = n \\ 0, \text{ otherwise.} \end{cases} \tag{6.25}$$

With the above considerations in mind, new expressions for the probability $P_f(i, N, r_w)$ are obtained, based on the scenarios illustrated in Figure 6.14.

An empty frame ($i = 0$) can only occur if all N users fail to satisfy the minimum SNIR condition to employ the lowest rate r_1. Hence:

$$P_f(i = 0, N) = P_{select}(0) = F(y_1)^N \tag{6.26}$$

A single transmission frame occurs when there is at least one participating user ($n \geq 1$) and all the participating users select the same beam. Hence, the probability

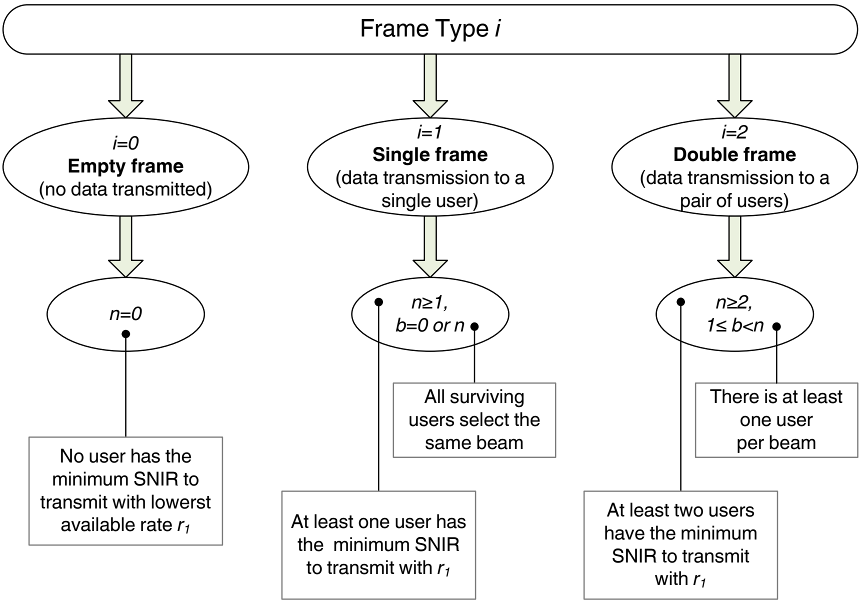

Frame Type *i*
i=0
Empty frame
(no data transmitted)
i=1
Single frame
(data transmission to a single user)
i=2
Double frame
(data transmission to a pair of users)
n=0
n≥1, b=0 or n
n≥2, 1≤ b<n
No user has the minimum SNIR to transmit with lowerst available rate *r₁*
All surviving users select the same beam
At least one user has the minimum SNIR to transmit with *r₁*
There is at least one user per beam
At least two users have the minimum SNIR to transmit with *r₁*

Figure 6.14: Possible frame type outcomes in Mu-Opportunistic scheme

of having a single transmission frame with rate r_w is

$$\begin{aligned} P_f(i = 1, N, r_w) &= \\ &= \sum_{n=1}^{N} P_{select}(n) \cdot Pr\Big\{n \text{ users on same beam}\Big\} \cdot P_{r1}(r_w, n) \\ &= \sum_{n=1}^{N} P_{select}(n) \cdot 2^{(1-n)} \cdot P_{r1}(r_w, n) \end{aligned} \tag{6.27}$$

where $P_{r1}(r_w, n)$ is the probability that rate r_w is used for transmission and is given by equation (6.17) (by setting the threshold index $\gamma = 1$).

Finally, the probability of having a double transmission frame is given by:

$$\begin{aligned} P_f(i = 2, N, r_w) &= \\ &= \sum_{n=2}^{N} P_{select}(n) \cdot P_{r2}(r_w, n) \end{aligned} \tag{6.28}$$

with $P_{r2}(r_w, n)$ as the probability that the frame duration is determined by rate r_w, calculated from the equation (6.19).

6.6 Performance Evaluation

6.6.1 Simulation Setup

This section will focus on the performance evaluation of the proposed multiuser schemes and will also demonstrate the validity of the analytical models for the throughput calculation presented in Section 6.5. Apart from the theoretical values, simulation results have been obtained with the help of a custom-made link layer simulation tool implemented in C++. The motivation behind the development of the simulator has been the flexibility in the MAC design and the possibility of incorporating a detailed MISO channel model, which could not be easily included in existing network simulators such as ns-2. The developed software tool employed the channel and traffic models that will be described in this section to provide a simulation of all the steps of the multiuser transmission sequence.

The simulation setup considers an infrastructure downlink network that consists of an AP with $n_t = 2$ transmitting antennas and $N = 10$ single-antenna users (MISO scenario). An ideal AMC that ensures error-free data transmission has has been assumed at the PHY layer, given that the rate for each transmission is selected according to the link quality, as expressed by the SNIR.

Channel Model

A channel model that represents the IEEE 802.11n channel model B in Non-line-of-sight (NLOS) conditions has been considered [79]. As mentioned in Section 6.3, a block-fading model with independent and identically distributed complex Gaussian entries $\sim \mathcal{CN}(0, 1)$ has been considered, with a noise variance of 0.1.[8] Each block corresponds to the duration of a frame sequence and no correlations have been assumed among the different blocks. This model has been employed to generate a SNIR matrix whose form is depicted in Figure 6.15.

Each line of the matrix contains the SNIR values of the N users on the $b = 2$ beams during a given frame sequence t_j. These entries have been fed to the C++ simulator to determine the channel condition of each user on a frame-by-frame basis. The SNIR limits employed to determine the available transmission rate of each user are given in Table 6.3 [84].

Four different scenarios have been considered, characterized by four channel implementations (i.e., different SNIR matrices) denoted by Ch_A, Ch_B,Ch_C and Ch_D. The average link quality varies for each channel, with Ch_A corresponding to the most unfavorable conditions and Ch_D representing a channel with high quality links. For reference, the average user SNIR for channels Ch_A to Ch_D is 15dB, 17dB, 20dB and 25dB, respectively.[9] According to Table 6.3, the average user rate for each

[8] Without loss of generality, a relatively low noise variance has been used. Higher values would lead to different numerical results but without affecting the behavior of the evaluated MAC schemes.

[9] These calculations consider the maximum SNIR value of each user on the two available beams,

$$\mathrm{SNIR}_{\mathrm{matrix}} = \begin{array}{c} \\ \left[\begin{array}{ccccccc} & & \multicolumn{2}{c}{\overbrace{\hspace{6em}}^{\text{User } N_2}} & & & \\ y_{N1,b1}(t_1) & y_{N1,b2}(t_1) & y_{N2,b1}(t_1) & y_{N2,b2}(t_1) & \cdots & y_{NN,b1}(t_1) & y_{NN,b2}(t_1) \\ y_{N1,b1}(t_2) & y_{N1,b2}(t_2) & y_{N2,b1}(t_2) & y_{N2,b2}(t_2) & \cdots & y_{NN,b1}(t_2) & y_{NN,b2}(t_2) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ y_{N1,b1}(t_T) & y_{N1,b2}(t_T) & y_{N2,b1}(t_T) & y_{N2,b2}(t_T) & \cdots & y_{NN,b1}(t_T) & y_{NN,b2}(t_T) \\ & & \underbrace{\hspace{3em}}_{\text{beam 1}} & \underbrace{\hspace{3em}}_{\text{beam 2}} & & & \end{array}\right] \end{array}$$

SNIR entries for every frame sequence

Figure 6.15: Representation of the matrix with the user SNIR values

Table 6.3: SNIR thresholds

Rate (Mbps)	*SNIR (dB)*
0 (no transmission)	$\leq$-8
6	-8 to 12.5
9	12.5 to 14
12	14 to 16.5
18	16.5 to 19
24	19 to 22.5
36	22.5 to 26
48	26 to 28
54	>28

scenario will be 12, 18, 24, and 36 Mbps, respectively. Since the channel realizations are random, the available rate for each user at every time instance will oscillate around the mean value (with the same variance for all users), through the block fading channel defined in Section 6.3.

Traffic Model

The results presented in this chapter have considered saturated traffic conditions, with a constant flow of downlink traffic for all users always available at the buffers of the AP. The rationale behind this assumption has been to evaluate the maximum gain that can be extracted from downlink transmissions, which requires the system to operate under heavy traffic load. Unless otherwise stated, the size of the data

which are the values employed in Mu-Opportunistic and Mu-Threshold schemes. In the case of Mu-Basic, where a user is not always scheduled on its preferred beam, the average SNIR and rate for each channel model would be lower.

Table 6.4: Simulation parameters

Parameter	*Value*
Number of antennas (AP)	$n_t = 2$
Number of antennas (Users)	$n_r = 1$
Downlink Users	$N = 10$
SIFS	16 μs
aSlotTime	9 μs
PHY Header (AP)	28 μs
PHY Header (Users)	32 μs
MAC Header	40 bytes
RTS (Mu-Basic)	$14 + 6 \cdot n_t$ bytes
RTS (Mu-Opportunistic)	$14 + 6 \cdot N$ bytes
RTS (Mu-Threshold)	20 bytes
CTS	15 bytes
DATA	2312 bytes
ACK	14 bytes
Bandwidth	20 MHz

packets has been fixed to 2312 bytes. All control frames are transmitted at the lowest rate (i.e., at 6 Mbps) to ensure correct reception. The IEEE 802.11n frame format has been adopted at the MAC layer, with the modifications proposed in Section 6.4 for each multiuser scheme. A summary of the simulation parameters is given in Table 6.4.

Definition of performance metrics

The performance metrics employed in this section are briefly described next:

- *Throughput* is defined as the rate of transmitted bits per second and is calculated as the ratio of the total number of successfully transmitted useful data bits to the duration of the simulation experiment (average throughput). Unless otherwise stated, the throughput metrics presented in this section will refer to the average system throughput, which is the aggregate downlink throughput corresponding to all N users.

- *End-to-End Delay* refers to the total delay related to the transmission of each packet, calculated as the sum of the queuing, the access and the transmission delay, defined as follows:

  – The *Queuing Delay* is the waiting time of a packet in the system, from its generation until it reaches the head of the corresponding MAC layer buffer. In the case of saturated traffic, where an infinite amount of packets is available, this term is not evaluated and the End-to-End delay consists only on the two other metrics, the access and the transmission delay.

- The *Access Delay*, measured from the moment a packet arrives at the head of the MAC queue of a particular user until the initiation of the corresponding MAC layer transmission sequence (i.e, the sequence that will result in the transmission of the data packet, typically beginning with an RTS).
- The *Transmission Delay*, defined as the elapsed time from the beginning of the transmission sequence until the completion of the data packet transmission, including the time required for the transmission of ACK packets.

Reference Schemes

Finally, in the performance evaluation of the proposed multiuser schemes, some MAC layer algorithms are employed as a reference. These reference schemes will be briefly explained next:

- **Su-Basic** is a downlink version of the legacy, single-user IEEE 802.11 DCF MAC. The only difference from the standard is that the backoff mechanism is not employed in order to provide a fair comparison with the other downlink schemes (as explained in Section 6.4, the backoff mechanism is not employed in the downlink phase since all frame sequences are initiated by the AP). The frame sequence of the Su-Basic scheme is formed by the typical RTS/CTS/DATA/ACK frame exchange. The downlink user is randomly selected by the AP and its address is included in the RTS frame. Summarizing, the Su-Basic can be considered as the single-user version of the proposed Mu-Basic scheme (Section 6.4.1) where two users are randomly selected for downlink transmission by the AP.

- **Mu-Ideal** is an ideal opportunistic multiuser scheme in which the users with the highest SNIR values are scheduled on each beam. In other words, the same scheduling objective as in the Mu-Opportunistic scheme (Section 6.4.2) is targeted. The difference is that, in the Mu-Ideal scheme, it has been assumed that the AP has a perfect knowledge of the channel condition and can select the best set of users without any additional overhead. Hence, the frame sequence is similar to the one depicted in Figure 6.8 for the Mu-Basic scheme, but with the best set of users polled by the RTS. Clearly, this scheme is not practical, since some mechanism for the CSI acquisition must be available at the AP. The Mu-Ideal serves as an upper bound for the performance proposed multiuser schemes.

- **Su-Ideal** can be considered as a single-user version of the Mu-Ideal algorithm. In this case, only a single downlink transmission takes place (through the usual

RTS/CTS/DATA/ACK frame exchange) but, in, addition, the user with the best channel conditions is selected by the AP in each frame. Again, perfect CSI is assumed at the AP with no additional overhead cost. The Su-Ideal serves as an upper bound for the performance that can be achieved by single user transmissions in the considered system setup.

- **Su-Opportunistic** is the last reference scheme that is a single-user equivalent of the Mu-Opportunistic scheme. All the downlink users are polled by a multidestination RTS that has the format shown in Figure 6.9. Then, the users reply with N sequentially transmitted CTS, in a predefined order determined by the order of the address list in the RTS. Once the CTS frames are received, the AP selects opportunistically the best (and single) user for transmission.

6.6.2 Potential Gains from Multiuser Transmissions: Evaluation of the Mu-Basic Scheme

The first set of presented results focuses on the performance of the Mu-Basic scheme, the simplest multiuser algorithm where users are randomly selected by the AP. Throughput and mean end-to-end delay plots are presented in Figures 6.16 and 6.17, for the four channel models, Ch_A to Ch_D and a packet size of $L = 2312$ bytes. The presented results have been obtained through simulations. Apart from Mu-Basic, the performance of three reference schemes is also plotted, in order to examine different performance trade-offs and gains. In particular, the following comparisons are made:

- Mu-Basic versus Su-Basic, in order to determine the gain from multiuser transmissions when random scheduling is employed.
- Mu-Basic versus Mu-Ideal, in order to determine the potential multiuser diversity gain that can be obtained through ideal opportunistic scheduling.
- Mu-Ideal versus Su-Ideal, in order to determine the gain from multiuser transmissions when ideal opportunistic scheduling is employed.

The remaining of this section will address there performance trade-offs in detail.

First of all, Mu-Basic is compared to the legacy, single-user IEEE 802.11g MAC, which is denoted by the name Su-Basic. In both schemes, the AP selects randomly the downlink users that are to be served in each transmission sequence. The difference lies in the fact that in Su-Basic a single user is served in the typical RTS/CTS/DATA/ACK frame exchange whereas in Mu-Basic two users are simultaneously served in the modified transmission sequence illustrated in Figure 6.8.

The presented results show that multiuser transmissions do not guarantee a performance improvement with respect to single-user transmissions. On the contrary, under relatively harsh channels, such as Ch_A and Ch_B, where the average supported

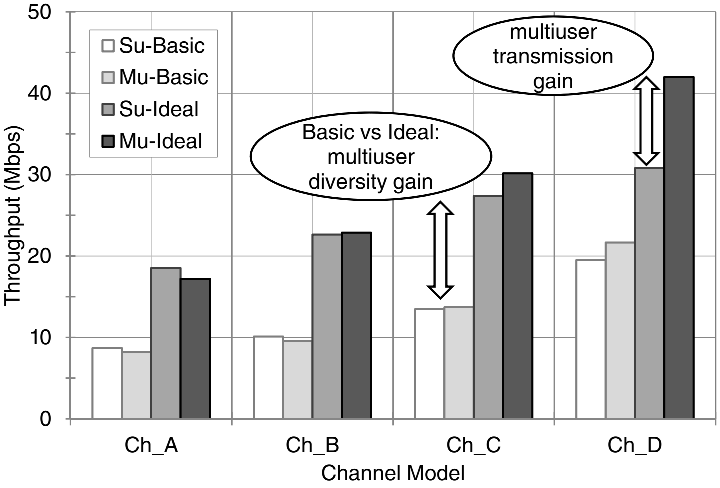


Figure 6.16: Throughput performance bounds ($L = 2312$ bytes)

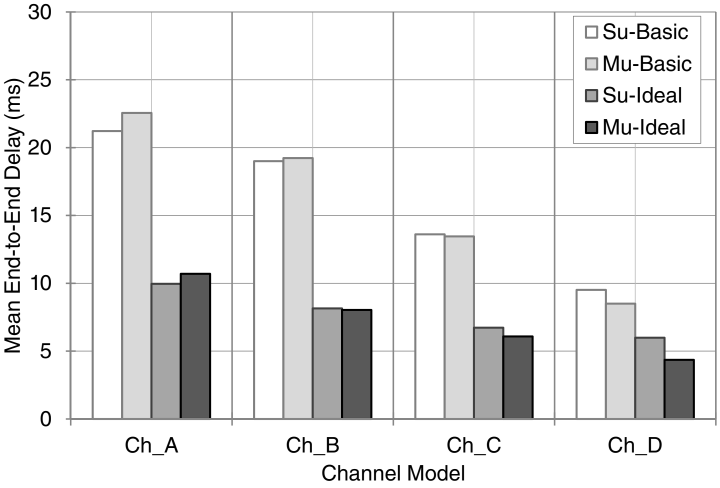
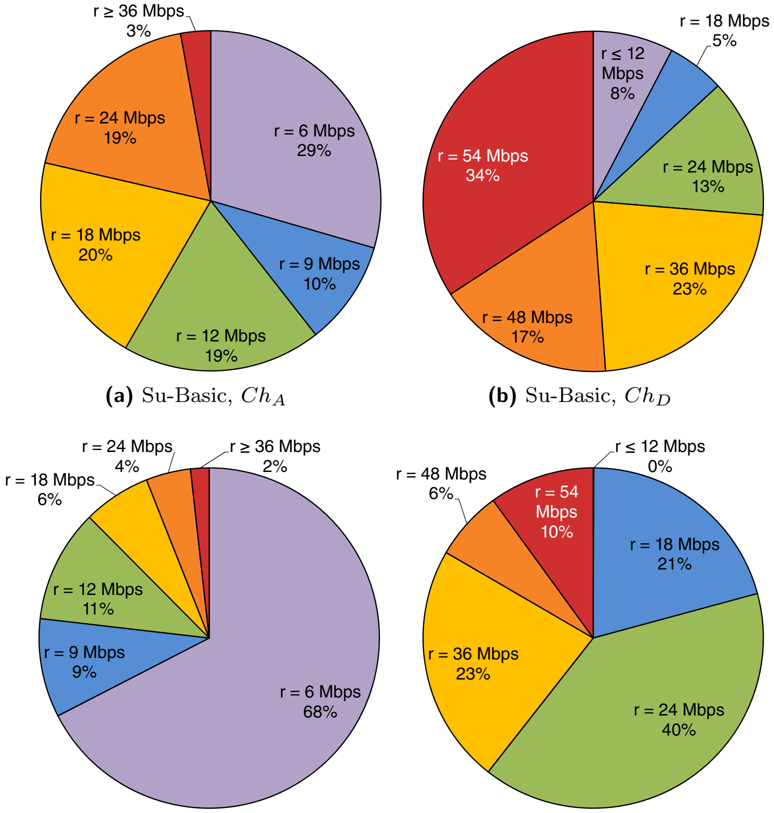


Figure 6.17: End-to-End delay performance bounds ($L = 2312$ bytes)

transmission rates are low, the Su-Basic scheme performs slightly better than the Mu-Basic scheme. The performance of both schemes is similar for the case of Ch_C. Finally, only under the good overall link conditions of Ch_D, does the Mu-Basic scheme perform better that the Su-Basic.

The performance of these two schemes can be better understood by examining the distribution of the rates employed for transmission, shown in Figure 6.18. Plots (a) and (b) show the rate percentages of Su-Basic for channels Ch_A and Ch_D, whereas plots (c) and (d) contain the same statistics for Mu-Basic. First of all, it can be clearly seen that higher rates can be supported in the case of Ch_D, resulting to higher throughput and lower delay performances for both schemes.



(a) Su-Basic, Ch_A

(b) Su-Basic, Ch_D

(c) Mu-Basic, Ch_A

(d) Mu-Basic, Ch_D

Figure 6.18: Percentage of frames at the eight available rates ($L = 2312$ bytes)

By comparing the rate distributions of the two schemes for the same channel type, it can be observed that, on average, Su-Basic employs higher rates with respect to Mu-Basic. This can be explained by considering that in Su-Basic there is a single downlink data transmission in each frame sequence. On the other hand, Mu-Basic implements concurrent data transmissions on the two beams which cause some interference to each other, despite being generated orthogonally. As a result, the SNR on single transmission links is generally higher that the SNIR when the two beams are employed.

The advantage of Mu-Basic scheme lies in the simultaneous transmission of two data frames with less control information within the transmission sequence. However, there is a performance trade-off between single transmissions with higher data rates and multiuser transmissions with potentially lower rates due to interference limitations. Figure 6.19 shows an example of how the two schemes can have similar throughput performance, despite the differences in the transmission sequences and the employed rates. In general, multiuser transmissions are more efficient as the channel quality increases, as in the case of Ch_D.

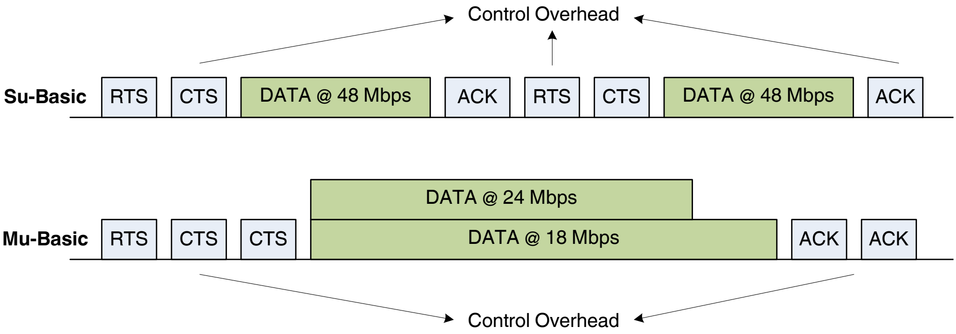


Figure 6.19: Example of the performance trade-off between single and multiuser transmissions

Going back to Figures 6.16 and 6.17, two other schemes have been evaluated, referred to as Su-Ideal and Mu-Ideal. These two schemes are ideal opportunistic implementations of the respective basic schemes: they maintain the same frame sequence structure but assume that the AP selects the user (or set of users) with the highest available transmission rate. These implementations assume perfect knowledge of the channel conditions of all the downlink users by the AP with no feedback cost. The two schemes provide an upper performance bound that cannot be reached by realistic schemes that require control overhead for the CSI acquisition.

The presented results clearly show that both ideal schemes achieve a considerably higher throughput performance with respect to the basic schemes. This is due to the multiuser diversity gain that is exploited by opportunistically selecting the best users for each transmission. Figure 6.20 plots the diversity gain in the throughput performance of Mu-Ideal with respect to Mu-Basic, for three different packet sizes of L =2312, 1500 and 1000 bytes. Bigger packet sizes yield better performance due to the reduced relative overhead information. In addition, even though the gain is different for each channel model, the advantage of opportunistic scheduling is clear, since throughput is practically doubled in most presented cases.

The potential enhancement that can be obtained by the multiple transmissions of packets, instead of the single-user transmission, can be appreciated by comparing the performance of Mu-Ideal and Su-Ideal schemes. This gain has been plotted in Figure 6.21, again for the three different packet sizes. The results show that even though single user transmissions are slightly more efficient for a hostile channel, such as Ch_A, multiuser transmissions yield higher throughput for the other three channel models. This is the same trend that has been observed in the comparison of Mu-Basic and Su-Basic, but in the case of ideal opportunistic scheduling, the obtained improvement is more pronounced with up to 35% gain.

Another observation is that the impact of the packet size on the multiuser transmission gain is not straightforward. In the case of Ch_D, performance is better when bigger packet sizes are employed. However, for the harsher channel models smaller

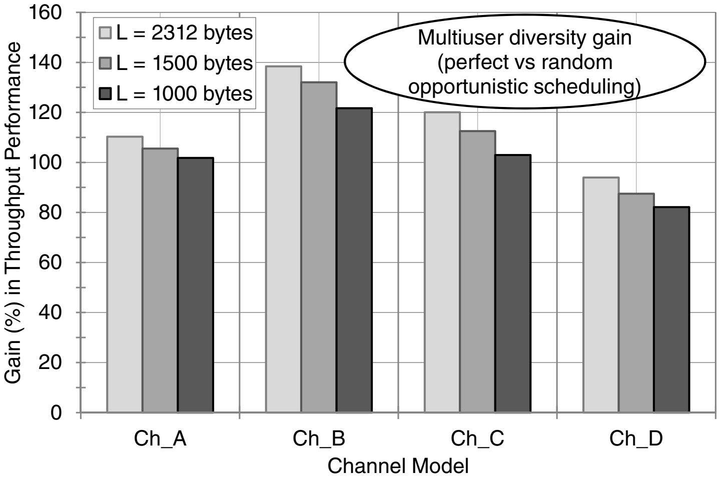


Figure 6.20: Multiuser diversity gain of Mu-Ideal with respect to Mu-Basic

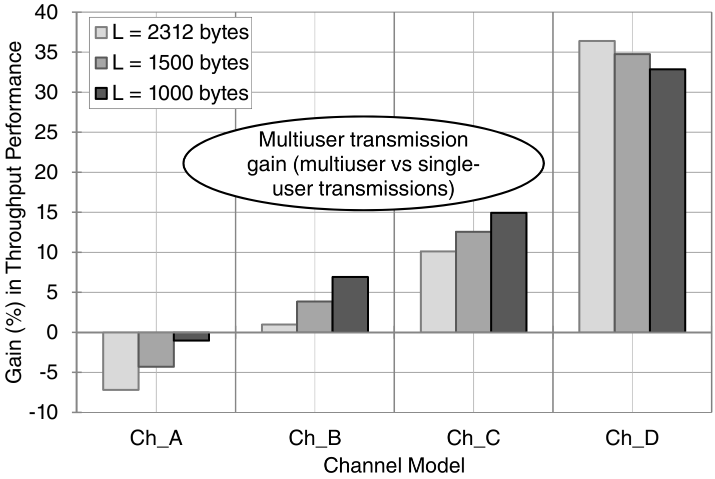


Figure 6.21: Multiuser transmission gain of Mu-Ideal with respect to Su-Ideal

packets seem to be more appropriate. This comes as a consequence of the trade-off explained before, in Figure 6.19, where it was shown that multiuser transmissions at low rates (i.e., with long transmission times) are not very efficient. In this case, the transmission time is not only affected by the rate but also by the packet size. As a result, long packets transmitted at low rates aggravate the performance of the multiuser schemes.

The results presented in this section lead to some interesting conclusions, that can summarized as follows:

- The efficiency of multiuser transmissions depends greatly on the channel conditions of the downlink users. Multiple simultaneous transmissions yield significant gain under good channel links that allow the use of high transmission rates for all participating users. On the other hand, under harsh channels, the gain is reduced and in some cases (especially if opportunistic scheduling is not employed) it could be preferable to dedicate the available resources to single-user transmissions.

- The upper performance bounds provided by the two ideal schemes show that a substantial performance enhancement can be achieved by exploiting the multiuser diversity and transmitting opportunistically to high rate users. Nevertheless, practical implementations of these schemes must be considered, to provide mechanisms for the CSI acquisition required for the opportunistic scheduling. Two possible implementations are the proposed Mu-Opportunistic and Mu-Threshold schemes that will be evaluated in the following section.

6.6.3 Performance Analysis of Realistic Multiuser Schemes: Validation of the Analytical Model

This section will focus on the performance evaluation of the two novel and more advanced multiuser schemes, Mu-Opportunistic and Mu-Threshold and will also demonstrate the validity of the analytical models for their throughput calculations presented in Section 6.5.

Mu-Opportunistic Scheme

Figure 6.22 shows the average system throughput achieved by Mu-Opportunistic scheme for the four channel models, $N = 10$ downlink users and three packet sizes of L =2312, 1500 and 1000 bytes. The theoretical throughput values obtained by the analytical model are represented by lines, whereas the markers correspond to the simulation results. A close match between the theoretical and the simulated values has been obtained.

Employing bigger packet sizes is more efficient as far as throughput is concerned and the best throughput results are achieved for $L = 2312$ bytes. When smaller packet sizes are considered, the cost of the control information has a stronger impact on performance and efficiency is reduced. Performance also depends on the channel model. Better channel conditions, as in the case of Ch_D lead to higher throughput values. The same principle holds for the mean end-to-end delay performance, depicted in Figure 6.23. Delay is higher for the hostile Ch_A but is decreased as the channel quality improves. On the other hand, for the same channel model, delay is lower for smaller packet sizes, since they require less time for their transmission.

Figure 6.24 compares the throughput performance of Mu-Opportunistic scheme with an equivalent single-user scheme called Su-Opportunistic, for $L = 2312$ bytes. In both schemes, the AP acquires the CSI information on all downlink users through N sequentially transmitted CTS. In Su-Opportunistic the AP selects the best user

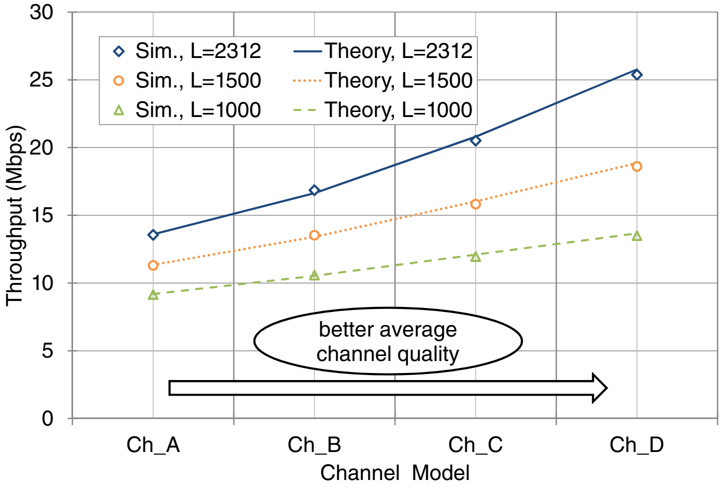


Figure 6.22: Throughput performance for the Mu-Opportunistic scheme, theoretical model versus simulations

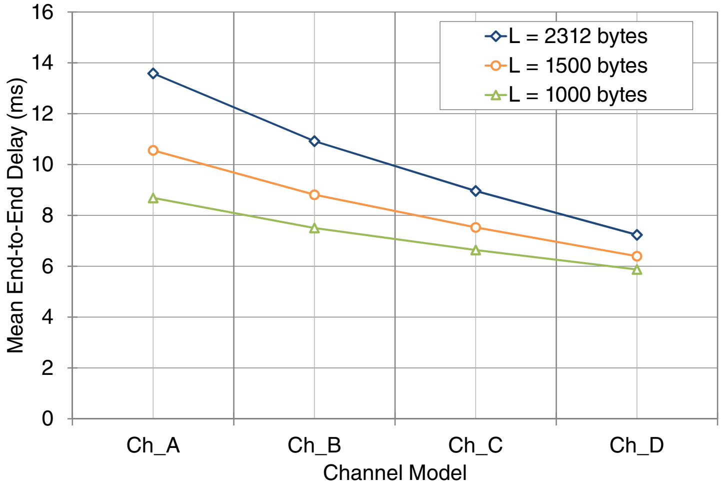


Figure 6.23: End-to-End delay performance for the Mu-Opportunistic scheme

for transmission, whereas in Mu-Opportunistic the best set of users is assigned to the beams. The transmission gain obtained by the multiuser opportunistic scheduling varies with the channel model but is substantial in all cases. A 21.3% improvement is achieved for Ch_A, that reaches up to 72.5% for Ch_D. By contrasting these results with Figure 6.16 where the single and multiuser basic schemes are compared, it can be said that even though the advantage of multiuser versus single-user transmissions is not significant if random scheduling is employed, it becomes evident with opportunistic policies.

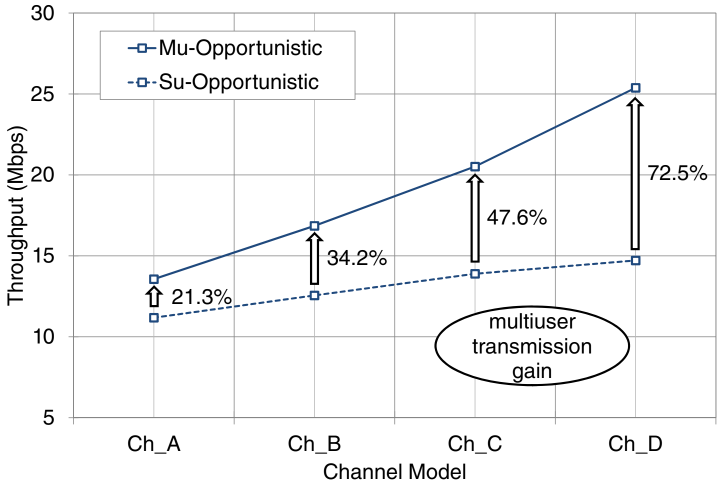


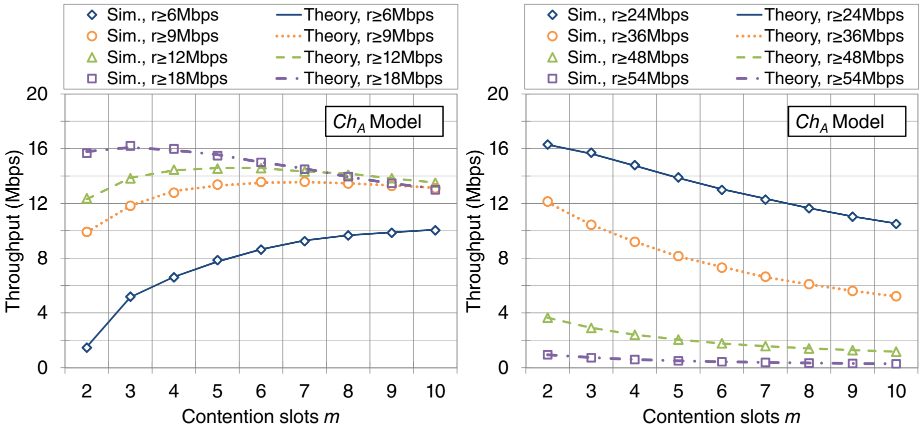
Figure 6.24: Mu-Opportunistic and Su-Opportunistic throughput comparison ($L = 2312$ bytes)

Mu-Threshold Scheme

Continuing with Mu-Threshold, Figures 6.25 to 6.28 show the system throughput for the four channel models Ch_A to Ch_D, respectively. The performance of Mu-Threshold depends on two configurable parameters: the selected rate threshold and the CTS slot number m. Since $N = 10$ downlink users have been considered, the number of slots m has taken values between 2 and 10. For clarity, throughput obtained for each channel model has been divided into two subplots, for lower and higher threshold values (plots (a) and (b), respectively). As before, the theoretical throughput values obtained by the analytical model are represented by lines, whereas the markers correspond to the simulation results. Theory and simulation results match closely in every case.

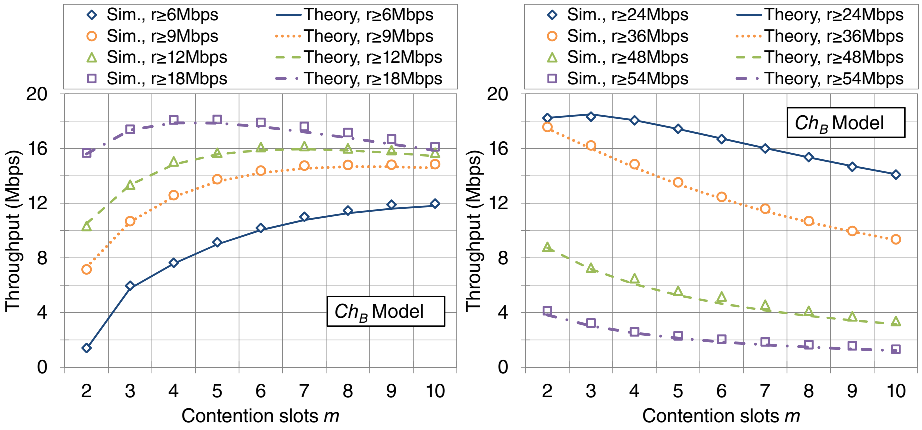
It is interesting to observe the combination of threshold and slot number m that yields the maximum system throughput for each channel model. The two more hostile channels (Ch_A and Ch_B) achieve a maximum throughput for a threshold of 24 Mbps, whereas Ch_C and Ch_D require higher thresholds (36 and 48 Mbps, respectively). This is expected since better channels that support higher average rates must employ higher thresholds to filter out low-rate users from the contention phase. On the other hand, the optimum number of CTS slots is relatively low (2 or 3 slots) for all channels. In other words, few slots are sufficient to handle $N = 10$ downlink users, since the imposed rate thresholds limit the number of users that participate in each transmission sequence.

Generally, the best configuration for the slot number m depends on the selected threshold. Low thresholds lead to a high number of contending users and therefore more slots are needed. On the other hand, for high thresholds, the number of participating users is limited and m can be decreased. Note, also, that performance drops



(a) Low threshold values (6 to 18 Mbps) **(b)** High threshold values (24 to 54 Mbps)

Figure 6.25: Throughput performance for Mu-Threshold, theoretical model versus simulations (Ch_A)

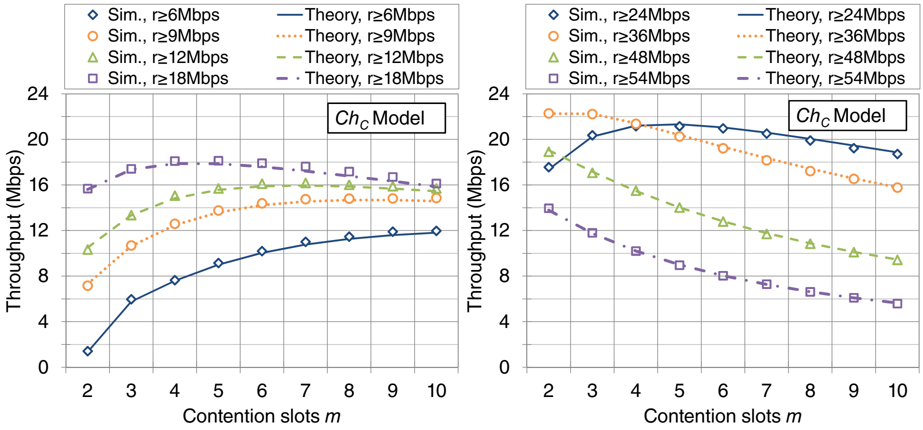


(a) Low threshold values (12 to 24 Mbps) **(b)** High threshold values (36 to 54 Mbps)

Figure 6.26: Throughput performance for Mu-Threshold, theoretical model versus simulations (Ch_B)

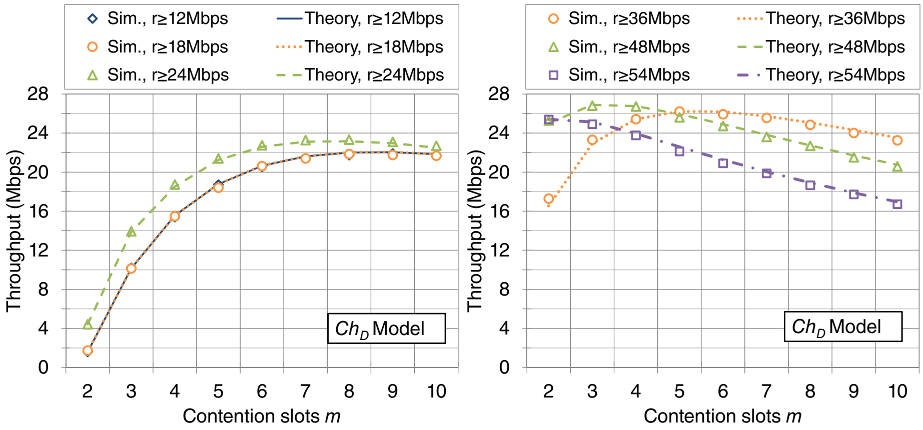
when the threshold is too high, due to the fact that not many users satisfy the SNIR condition, thus resulting to a high occurrence of empty frames (i.e., transmission sequences that consist of the RTS/CTS phase but do not have any DATA and ACK transmission).

The impact of the data packet size on the performance of Mu-Threshold has also been evaluated. Figure 6.29 presents the obtained throughput and delay performance for three different packet sizes, L = 2312, 1500 and 1000 bytes, as a



(a) Low threshold values (6 to 18 Mbps) **(b)** High threshold values (24 to 54 Mbps)

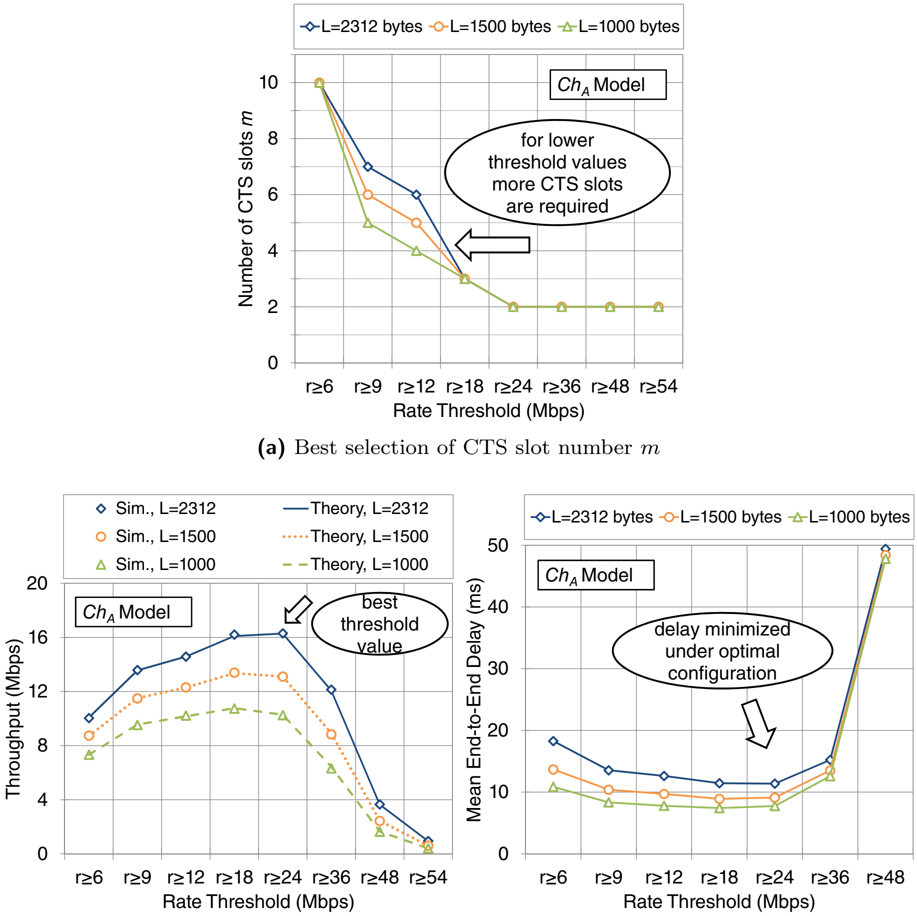
Figure 6.27: Throughput performance for Mu-Threshold, theoretical model versus simulations (Ch_C)



(a) Low threshold values (6 to 18 Mbps) **(b)** High threshold values (24 to 54 Mbps)

Figure 6.28: Throughput performance for Mu-Threshold, theoretical model versus simulations (Ch_D)

function of the employed rate threshold under the channel Ch_A (similar results can be obtained for the other channel models). Plot (a) shows the optimum number of CTS slots m that maximizes performance (i.e., maximizes throughput and minimizes delay) for every threshold value. As before, fewer slots are required as the threshold increases. The obtained throughput and delay performance for these optimal values of m are given in plots (b) and (c), respectively.



(a) Best selection of CTS slot number m

(b) Maximum Throughput

(c) Minimum End-to-End Delay

Figure 6.29: Performance for Mu-Threshold under Ch_A

6.6.4 Performance Comparison of the Multiuser Schemes

This section compares the performance achieved by the proposed multiuser MAC schemes, Mu-Basic, Mu-Opportunistic and Mu-Threshold. The presented results are divided into three parts, first discussing the enhancement obtained through opportunistic scheduling, then exploring the impact of the user number on performance and finally contemplating what happens under a non-saturation regime.

Gains from opportunistic scheduling

Figures 6.30 and 6.31 plot the throughput and mean total delay performance for the four channel models, $N = 10$ users and a packet size of $L = 2312$ bytes. The performance of the non-realistic Mu-Ideal scheme is also depicted, as a reference of the upper bound that corresponds to the considered scenarios.

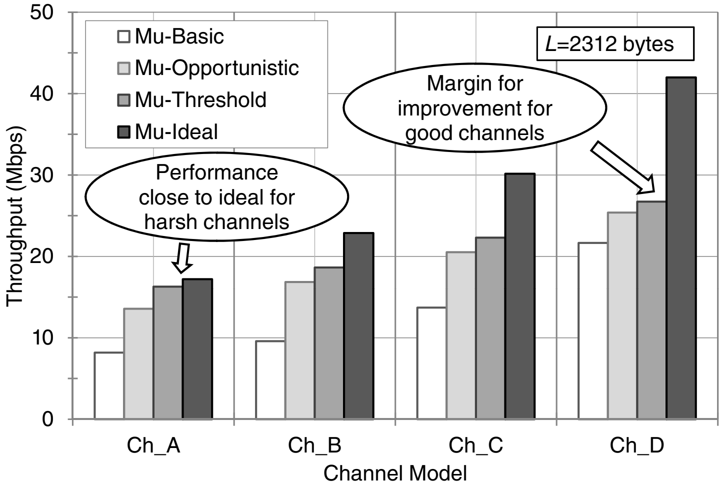


Figure 6.30: Throughput performance comparison

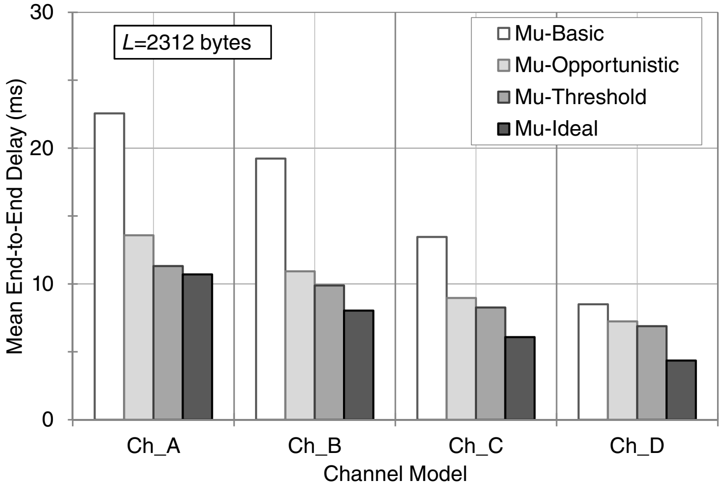


Figure 6.31: End-to-End delay performance comparison

The presented results for the Mu-Threshold have been obtained by considering the best combination of threshold and CTS slot number values. These optimum parameters are summarized in Table 6.5. In general, the channel statistics influence heavily the Mu-Threshold performance and the optimization of the algorithm is not straightforward since different objectives must be met to maximize performance in diverse scenarios. This can be better understood by examining the percentage of empty frames, given in the last column of the table. In the case of Ch_D, this percentage is low, meaning that the majority of frames feature single or double data transmissions. On the other hand, for harsh channels the minislot-threshold combination that maximizes throughput may result to a higher number of empty frames (even up to 50% for Ch_A), thus revealing that it is more efficient, as far as throughput is concerned, to transmit fewer packets but with a higher rate that to transmit in every transmit sequence with lower rates.

Table 6.5: Best configuration for Mu-Threshold scheme

Channel	*Threshold*	*Slots*	*Empty Frames*
	r_γ	m	%
Ch_A	24	2	49.9
Ch_B	24	3	29.2
Ch_C	36	3	30.7
Ch_D	48	3	23.9

The performance of the two opportunistic schemes, Mu-Opportunistic and Mu-Threshold, is bound between Mu-Basic and Mu-Ideal schemes. To illustrate this point, two performance statistics have been calculated in Table 6.6. The first metric reflects the throughput gain of the two schemes with respect to the Mu-Basic algorithm. It can be observed that both schemes improve performance under all the considered channel models by scheduling users with high available transmission rates. However, the exact value of the achieved gain depends on the channel quality. For harsh channels, the improvement is more pronounced. In the case of Ch_A, for instance, a gain of approximately 66 % and 99 % is obtained by Mu-Opportunistic and Mu-Threshold, respectively. On the other hand, when the channel quality is good, as in Ch_D, the need for opportunistic scheduling is less critical. Nevertheless, even in that case, an enhancement of more than 20 % can be achieved.

Table 6.6: Performance statistics for the proposed multiuser schemes

	Throughput gain (%) with respect to Mu-Basic		**Improvement margin (%) with respect to Mu-Ideal**	
Channel	*Mu-Opport.*	*Mu-Thres.*	*Mu-Opport.*	*Mu-Thres.*
Ch_A	65.84	99.21	26.80	5.56
Ch_B	75.76	94.36	35.64	22.66
Ch_C	49.67	62.69	47.03	35.25
Ch_D	17.25	23.47	65.45	57.11

Mu-Opportunistic and Mu-Threshold are two efficient multiuser schemes but there is still a margin for improvement in order to achieve the upper bound set by the Mu-Ideal. The second metric presented in Table 6.6 refers to the the available improvement margin. The three schemes share the principle of opportunistic scheduling, but implement it in different ways. Mu-Ideal assumes perfect CSI knowledge without any additional overhead cost, which is an assumption that does not hold for realistic schemes. Mu-Opportunistic introduces considerable overhead since $N = 10$ CTS packets are sent in each transmission sequence. Finally, Mu-Threshold manages to reduce overhead by employing m control slots, with m usually much smaller than the number of total users N (in the presented example, the best performance throughput has been obtained for no more than $m = 3$ slots). As a result, Mu-Threshold is closer to the Mu-Ideal.

Another interesting observation is that the two practical schemes are closer to the ideal performance under worse channel conditions. In the case of Ch_A, for instance, the improvement margin is 26.8 % for Mu-Opportunistic and only 5.6 % for Mu-Threshold (less that 1 Mbps below the upper throughput bound). The gap between the achieved throughput and the ideal performance opens as the channel conditions improve and in the case of Ch_D both schemes have an improvement margin of more than 50 %. This occurs because the overhead information, consisting of control packets transmitted at the lowest rate, has a greater impact on performance when high data rates are employed.

Table 6.7 gives an estimation of the improvement achieved by exploiting the multiuser diversity. This gain is reflected in the increase of the average data transmission rate compared to the average user rate for each channel model. The average data transmission rate is calculated as the average of the rates employed for the transmission of all data frames. The average user rate is obtained by calculating the average value of the maximum rate at which a user can transmit, if the best beam (i.e., with the higher SNIR) for the particular user is selected. This value depends on the channel model and is indicated in the second column of the table.

Table 6.7: Multiuser diversity gain

Channel	*Avg. User*		*Avg. Tx Rate (Mbps)*		
	Rate (Mbps)	*Mu-Basic*	*Mu-Opport.*	*Mu-Thres.*	*Mu-Ideal*
Ch_A	12	9.73	18.77	27.60	18.77
Ch_B	18	14.37	23.76	30.48	23.76
Ch_C	24	19.01	34.46	44.40	34.46
Ch_D	36	32.41	46.73	51.64	46.73

In the case of Mu-Basic, the average transmission rate is lower than the average user rate. This is a direct consequence of random scheduling and beam allocation: users may be selected for transmission when their channel quality is low, or they may receive increased interference from other simultaneous transmissions due to the suboptimal beam allocation. Mu-Opportunistic, on the other hand, exploits multiuser diversity by assigning the best user on every beam. As a result, most transmissions take place at rates above the average. In the case of Ch_D, for instance, the transmission rate is 46.7 Mbps whereas the average user rate is limited to 36 Mbps. It should be noted that Mu-Opportunistic yields the same average transmission rate as the Mu-Ideal scheme, since both schemes implement the same scheduling policy. Despite providing the same transmission rate, the throughput performance of Mu-Opportunistic is lower than the ideal, due to the additional control overhead required for the CSI acquisition.

Finally, the maximum transmission rate values are achieved by Mu-Threshold. At first glance, is seems puzzling to obtain rates above those of the Mu-Ideal scheme. Nevertheless, this can be explained with the help of the data presented in Table 6.5. By imposing a rate threshold, Mu-Threshold scheme controls the minimum rate that can be employed for transmission. For instance, in the case of Ch_D, the optimum performance is achieved for a threshold of 48 Mbps, meaning that all transmissions have taken place at the rates of 48 and 54 Mbps, thus increasing the average transmission rate. On the other hand, since the average user rate for this channel is 36 Mbps, there is a high possibility that users may not satisfy the threshold condition, resulting to empty frames with no data transmissions. For the best configuration of Mu-Threshold for Ch_D, the percentage of empty frames is approximately 24% of the total frame sequences, as indicated in the last column of Table 6.5.

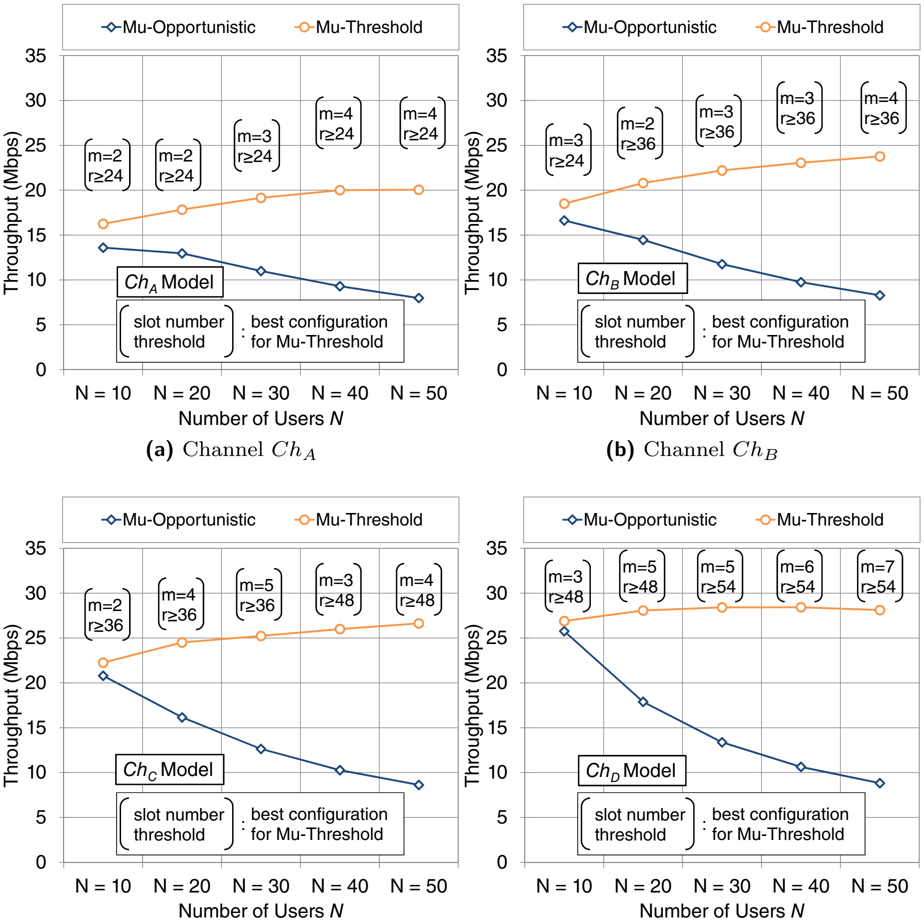
Performance as a function of the number of users

So far, a relatively small number of users, $N = 10$ has been considered. The following set of plots in Figure 6.32 shows the maximum throughput obtained by Mu-Opportunistic and Mu-Threshold as a function of the number of system users N for the four channel models. The best configuration for the Mu-Threshold has been considered and the employed values for the slot number and the rate threshold are also indicated in the figure.

In the case of Mu-Opportunistic, throughput decreases as the number of users grows. This is an unavoidable consequence of the control overhead required for the acquisition of CSI by all the system users. The degradation is more pronounced as the channel improves (e.g., Ch_D) and higher rates are employed for the data transmission (but not the overhead that is sent at the lowest supported rate). The lesson learned from this observation is that when multiple users are present, the Mu-Opportunistic mechanism is not very efficient. As a more viable alternative, the AP could divide the users in smaller groups and poll a user subset in each transmission sequence. This would reduce the multiuser gain but would also limit the introduced overhead.

On the other hand, Mu-Threshold handles multiple users in a more efficient way and throughput is actually improved as the user number increases. Several factors affect this behavior. First, when more users are present, the gain extracted from multiuser diversity also increases, since there is a higher probability of assigning a high-user rate on each beam. Second, unlike the Mu-Opportunistic scheme, the control overhead does not increase linearly with N but depends on the number of CTS slots m.

The selection of the slot number and the rate threshold provides a flexible mech-



(a) Channel Ch_A

(b) Channel Ch_B

(c) Channel Ch_C

(d) Channel Ch_D

Figure 6.32: Throughput performance comparison versus the number of users

anism to control the number of participating users in each transmission sequence. The best configuration depends on the channel distribution but generally the following principles hold:

- More slots are required as the number of user grows, to reduce the collision probability in the contention window. By observing the best configuration for each case, marked in Figure 6.32, it can be said that m generally follows an increasing trend.
- The collision probability can also be reduced by increasing the rate threshold, which results to a smaller number of participating users (but with higher available rates). Again, as the number of users grows, the threshold is progressively raised.

The joint selection of the rate threshold and the slot number is a gradual process. Take for example the case of Ch_D, depicted in Figure 6.32(d). For $N = 10$ users, the best configuration is $m = 2$ and a threshold of 36 Mbps. As more users are added to the system, the number of slots is increased to $m = 5$ for $N = 30$ users, with the threshold remaining the same. Then, for $N = 40$, the threshold is raised to 48 Mbps. Since the higher threshold filters out part of the users, a lower slot number can be afforded ($m = 3$). Finally, to handle 50 users, m again is increased by 1 unit.

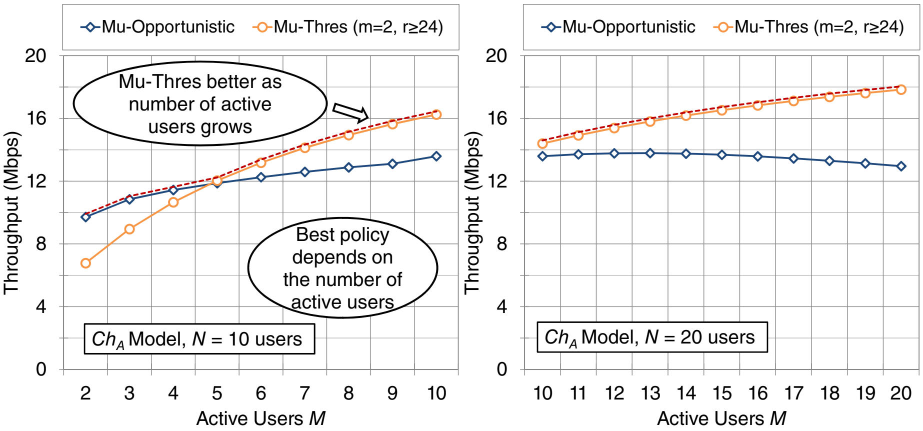
This step by step adaptation of the two configurable parameters allows the system to adjust to more users and attain high throughput values. Nevertheless, the system capacity is limited, so unavoidably performance eventually drops once the maximum throughput is gained. This occurs, for instance, in the case of Ch_D, where the maximum throughput is obtained for $N = 40$ users, for $m = 6$ and a threshold of 54 Mbps.

Non-saturated case

Finally, a non-saturated traffic scenario has been considered. In particular, it has been assumed that the AP is associated with N users but at a given moment it has downlink data packets for only a subset of M active users (with $M \leq N$). According to the Mu-Threshold scheme, the AP initiates the downlink phase by transmitting a broadcast RTS intended for all N associated users. The users then measure the SNIR of the link and reply with a CTS if they meet the threshold condition. However, they have no way of knowing whether they belong to the active set or not; their only criterion for participating in the contention phase is the measured SNIR. Consequently, it is possible that the users that survive the contention phase may not belong to the active set of users, leading to an occurrence of empty frames. This problem is avoided by the Mu-Opportunistic scheme, since a multi-destination RTS frame that includes the addresses of the M active downlink users is employed. Hence, in the case of non-saturated traffic, only the active user set is polled by the AP; the non-active users do not participate in the CTS phase since their address is omitted from the RTS.

Figure 6.33 plots the performance of Mu-Thres and Mu-Opportunistic under non-saturated traffic for the harsh channel Ch_A. In plot (a) the total number of users has been set to $N = 10$ and the number of active users M has been varied from 2 to 10. The obtained results show that when the number of active users is small (for $M \leq 5$), higher throughput is achieved by Mu-Opportunistic. On the other hand, as traffic grows and more active users are present, Mu-Threshold yields the best throughput results. This can be interpreted in the following way. When the AP has downlink data for relatively few users (less than five, in this case) it is more efficient to implement Mu-Opportunistic and poll these users by including their address in the RTS. On the contrary, when more than five out of ten users are active, it is preferable to implement Mu-Threshold and employ a broadcast RTS instead of a long address list, despite the risk of having some empty frames due to the inactive users that may participate in the contention phase.

Figure 6.33 (b) plots the throughput of the two algorithms under the same channel model (Ch_A) for a total number of $N = 20$ users. The number of active users in this case takes values from $M = 10$ to 20 users. It can be observed that, for this configuration, Mu-Threshold is always better than Mu-Opportunistic. Initially, for $M = 10$, the two algorithms have a relatively close performance, but as the active user set grows, the throughput achieved by Mu-Threshold increases steadily and the performance gain is clearly marked. This occurs because the multiuser diversity gain grows as more active users are present and Mu-Threshold is able to extract this gain with relatively low overhead. On the other hand, the length of the address lists and more importantly the duration of the CTS phase of the Mu-Opportunistic becomes prohibitive and has a negative impact on the throughput performance.



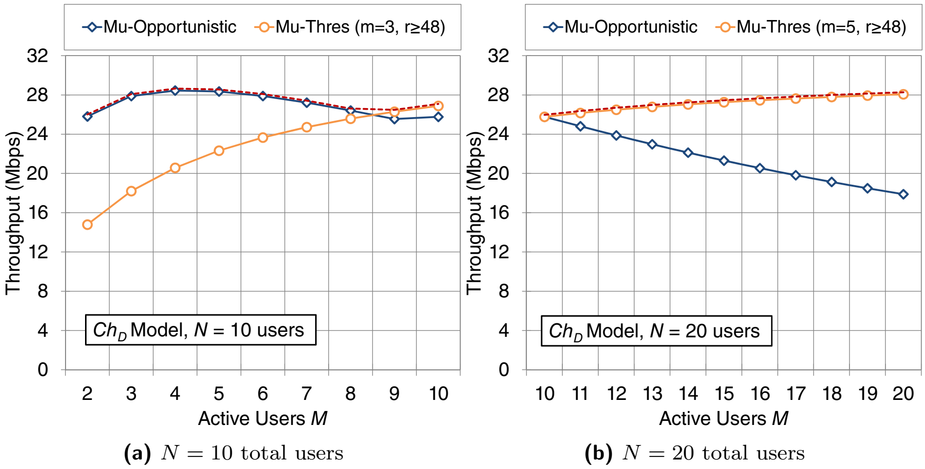
(a) $N = 10$ total users

(b) $N = 20$ total users

Figure 6.33: Throughput for non-saturated downlink traffic under channel Ch_A

Figure 6.34 plots the respective results that correspond to the channel model Ch_D, in which the users have a better average link quality. As a result, higher transmission rates can be supported by the users and the system capacity is generally increased. Nevertheless, the use of high rates for the data transmission is followed by a loss in efficiency due to the transmission of control frames at the lowest available rate (to increase robustness against errors). This is reflected on the performance of the Mu-Opportunistic scheme: maximum throughput is achieved for $M = 4$ users and then performance begins to deteriorate due to the overhead caused by the increasing number of CTS responses by the users.

Conversely, the throughput of Mu-Threshold is very low when few active users are present but keeps increasing as the number of users grows. This is shown in Figure 6.34 (a) where $N = 10$ users are considered. Initially, the throughput achieved by Mu-Threshold is significantly lower with respect to Mu-Opportunistic but shows an increasing slope and eventually Mu-Threshold becomes the most efficient scheme near saturation (for $M \geq 9$). In the case of $N = 20$ users (plot (b)), both schemes obtain similar results when half of the users are active (i.e., $M = 10$) but as the active user set grows Mu-Threshold performance increases while Mu-Opportunistic deteriorates.



(a) $N = 10$ total users

(b) $N = 20$ total users

Figure 6.34: Throughput for non-saturated downlink traffic under channel Ch_D

The main conclusion drawn from the presented results is that when the number of active users in the system is low, Mu-Opportunistic is more efficient since it schedules the best set of users with relatively low control overhead. Under these conditions Mu-Threshold underperforms because the threshold mechanism backfires: the imposed threshold reduces the number of participating users in the contention phase but does not guarantee that the selected users will belong to the active user set. This becomes more pronounced when the channel condition of the users are more favorable (e.g., in the case of Ch_D) since a higher threshold is selected. Nevertheless, when the number of active users exceeds a critical value that mainly depends on the average channel conditions of the users, Mu-Threshold is able to extract the multiuser gain of the channel and its advantage over Mu-Opportunistic becomes clear.

6.7 Further Discussion and Open Issues

The integration of MIMO/MISO technology in WLANs has become a reality with the emerging IEEE 802.11n standard; however there are many open issues regarding MAC protocol design that must be addressed to fully exploit its potential. This chapter has presented a number of multiuser MAC downlink schemes that, combined with a low-complexity transmission technique, can boost the performance of infrastructure IEEE 802.11 based WLANs.

Still, there are many directions for future investigation in in the multiuser MAC layer design. With the proposed opportunistic schemes in mind, some modifications that could lead to further performance enhancements include the following:

- Frame Aggregation mechanisms to reduce the control overhead.
- QoS provisioning that will include delay or fairness metrics in the scheduling decisions.
- Adaptive schemes for the selection of the best multiuser algorithm and for parameter optimization for each scenario.

Some further thoughts on these issues are given in the remaining part of this section.

6.7.1 Frame Aggregation for Multiuser Scenarios

In the proposed multiuser schemes, transmissions to different users can take place at different transmission rates, depending on the link quality of each user with respect to the AP. When two different rates are employed for simultaneous transmissions on different beams and assuming the same packet size, the duration of the transmission sequence is limited by the lower-rate user. An example of this situation is shown in Figure 6.35.

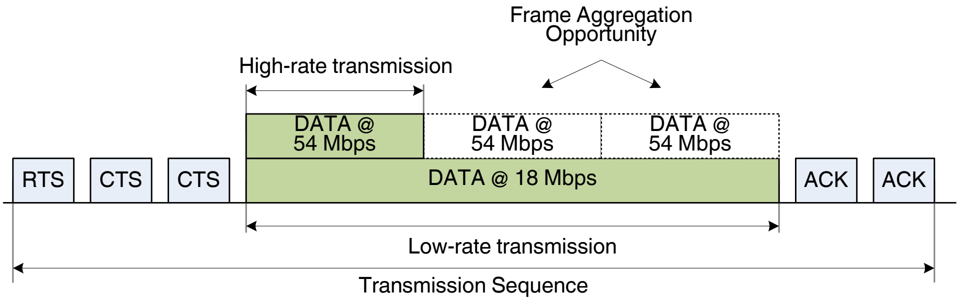


Figure 6.35: Example of frame aggregation

The IEEE 802.11n specification defines two frame aggregation schemes, called A-MPDU and A-MSDU. In the A-MPDU scheme, multiple MAC data frames (MPDUs) are joined in a single PHY layer frame (PSDU). Due to the robust delimiting, errors in one MPDU do not imply the loss of the whole A-MPDU frame. In the A-MSDU, a common PHY and MAC header is used for multiple data units (MSDUs). This scheme is very useful for applications that use many small data frames, such as TCP acknowledgments, but is less robust against errors.

These existing mechanisms could also be applied to the multiuser schemes, along with some smart algorithms for the selection of the most appropriate aggregation size. A possible aggregation objective could be to minimize the gap in the transmission times of data packets at different rates and thus increase the efficiency of the transmission sequence. In the example of Figure 6.35, an aggregated A-MPDU frame of three data packets transmitted at 54 Mbps could be scheduled for the high-rate user, in order to compensate for the longer second user transmission at 18 Mbps.

A more advanced policy could implement multi-destination frames for multiple users scheduled on the same beam. In the MAC schemes presented in this chapter, the number of selected users in each transmission sequence cannot exceed the number of generated beams (or antennas). Nevertheless, the AP collects CSI from multiple users in order to opportunistically select the best set. This information could be further exploited to schedule more than one user on the same beam.

6.7.2 QoS Provisioning

The results presented in this chapter mainly focused on maximizing the system throughput performance, which is the most representative metric under saturation conditions. Nevertheless, especially in non-saturated scenarios, more metrics can be taken into consideration to attain difference performance goals.

The most direct application of QoS provisioning in multiuser downlink transmissions would be to include delay constraints in the scheduling decisions. So far, the sole criterion for user selection has been the link quality of the users in order to perform transmissions at the highest available rates. Nevertheless, the AP could easily modify the scheduling rules to serve users with strict delay deadlines, despite their channel condition. This policy could lead to interesting trade-offs between user satisfaction and global system efficiency.

On a different note, it can be claimed that fairness is always an issue in opportunistic scheduling schemes. In this analysis, a homogeneous network scenario has been considered in which all users have similar channel statistics on average, despite instantaneous fluctuations of their channel conditions. However, users that suffer from channel fading for long time intervals would seldom get the chance to be served, especially under heavy traffic conditions.

The application of QoS and fairness criteria in the case of Mu-Opportunistic scheme is rather straightforward, due to the fact that all system users are sequentially polled by the AP and their CSI is known. As a result, the AP can select the best set of users depending on the performance goal (e.g., the user with delay-sensitive traffic or the one with the smallest share on the allocated bandwidth) and perform the downlink transmission at the maximum available rate.

In the case of Mu-Threshold, more drastic modifications are required since the threshold application limits the participation of low-rate users in the CTS contention phase. If necessary, the AP could still initiate data transmission to these users, however, due to the lack of valid CSI information, the lowest transmission rate should be employed to minimize reception errors. An intermediate solution would be to apply a more relaxed threshold condition along with a p-persistent policy. In this case, low-rate users who would otherwise be excluded from the CTS phase could be allowed to participate with a probability p. This probability could be constant or dynamically selected by the user depending on QoS constraints.

6.7.3 Adaptive Multiuser Scheme

This chapter has presented two opportunistic multiuser schemes, Mu-Opportunistic and Mu-Threshold. The obtained results have shown a general overall improvement of the total system throughput with respect to non-opportunistic transmissions. In most studied scenarios, Mu-Threshold achieved the best performance, when the slot number and threshold parameters were optimally tuned. Nevertheless, the best parameter configuration depended on the several parameters such as the channel distribution or the number of system users.

An interesting line of investigation would be to develop an adaptive algorithm to dynamically adjust the configurable parameters of the multiuser scheme to match the time-varying channel conditions. Different approaches to this issue are available:

- For systems under slow-varying channels, an adaptive scheme could be developed with the help of the analytical models for the throughput calculation, presented in Section 6.5. These models assume some steady state probabilities for the channel distribution, which could be obtained through channel measurements during an initial training phase. Based on these statistics, the optimum configuration could be found through the mathematical model. Clearly, if the mathematical model is to be employed at runtime, some computing power must be available at the AP.

- Another approach is the implementation of a heuristic algorithm to gradually guide the system to the best configuration after a number of steps. Some guidelines on the impact of the slot number and the threshold on the system behavior have been provided in the previous section.

- A combination of the two aforementioned schemes could potentially be the best solution. The analytical model could be employed to obtain some initial estimation of the optimum configuration which would, in turn, be fed as an input to the heuristic algorithm.

6.8 Conclusions

This chapter has presented a novel approach for the integration of multiuser capabilities in IEEE 802.11n based WLANs. On one hand, a low-complexity beamforming technique named MOB has been employed at the PHY layer. The main strength of MOB lies in the fact that it only requires partial CSI information at the transmitter side, in the form of SNIR measurements acquired by the downlink users. Since the IEEE 802.11n specification supports beamforming, MOB can be easily implemented with minor modifications in the beamforming steering matrices.

On the other hand, in order to exploit the potential of the MOB technique in a realistic scenario, it is necessary to design appropriate MAC layer mechanisms to handle multiuser transmissions. In this chapter, three MAC layer schemes have been proposed. The first scheme, Mu-Basic, implemented a simple random scheduling multiuser scheme, meant to serve as a performance reference. Then, two opportunistic schemes have been proposed, Mu-Opportunistic and Mu-Threshold, that enhance performance by extracting the multiuser diversity gain. Table 6.8 provides a comparison of the basic features of these three MAC layer schemes.

Table 6.8: Summary of the proposed multiuser MAC Schemes

	Multiuser MAC Schemes		
Parameter	*Mu-Basic*	*Mu-Opportunistic*	*Mu-Threshold*
RTS Destination	2 recipients	Multicast	Broadcast
CTS collisions	No	No	Yes
CTS Phase Length	2× CTS	N× CTS	m×CTS
Scheduling	Random	Opportunistic	Opportunistic
Mac Overhead	Low	High	Low
Efficiency	Low	Medium	High
Optimization Parameter	-	-	m, threshold

The performance evaluation of the proposed schemes through both mathematical analysis and simulations under four different channel scenarios has led to many interesting observations. The lessons learned can be employed to improve the proposed algorithms but also as more general guidelines in the design of multiuser MAC schemes. The more remarkable conclusions are summarized as follows:

- Employing multiuser transmissions does not always guarantee a performance improvement with respect to single-user transmissions when random scheduling is employed, as shown by the comparison between Mu-Basic and its single user equivalent. A trade-off exists between allocating antenna resources to multiple users, risking increased interference, versus employing all antennas for a more reliable single-user transmission. This trade-off mainly depends on the channel and user distributions of the system.

- Opportunistic scheduling of the best set of users so as to maximize the average transmission rate is an effective way to shift the aforementioned trade-off towards multiuser transmissions. In other words, if users are opportunistically selected depending on their measured channel quality, the gain achieved by multiuser transmissions is significant.

- Multiuser diversity gain increases with the number of system users, since there is a higher probability of finding a high-rate set of users among a larger user pool. On the other hand, more users come with a cost of additional control overhead for the CSI acquisition. Mu-Threshold handles efficiently multiple users by setting appropriately the slot number and the rate threshold parameters. In the case of Mu-Opportunistic scheme, the control overhead increases linearly with the user number and performance eventually drops. On the contrary, the Mu-Threshold scheme requires less control overhead and is shown to be more efficient than a straight-forward approach that explicitly polls all the associated users, under medium and high traffic conditions. Nevertheless, when few users are present or under non-saturated scenarios, the Mu-Opportunistic approach of collision-free sequential CTS transmissions becomes more effective that the probabilistic threshold selection mechanism.

- Under harsh channels, the performance of the proposed multiuser schemes approaches the upper performance bound set by the ideal case of having perfect CSI knowledge with no additional overhead. On the other hand, under more favorable channel conditions, there is still a margin for potential improvement by exploiting multiuser transmissions.

Chapter 7

Conclusions

The design of efficient MAC layer protocols for wireless networks is not an easy task. The continuous growth of the WLAN market and the emergence of new technologies create new application scenarios and service requirements. As a result, it is not always possible to find ubiquitous solutions for performance enhancement, since, an optimum policy under certain network circumstances may be detrimental under a different scenario. Hence, it is often necessary to focus on improving particular aspects of WLANs and identify the performance trade-offs associated with the proposed solutions.

This statement is particularly true in the case of CL design. The layered principle of the OSI protocol stack was meant to optimize each layer protocol operation independently. Breaking this rigid architecture with CL interactions among different layers involves some risks but can also lead to significant performance enhancements. The main motivation of this thesis has been to explore the potential benefits and the possible trade-offs associated with CL-based MAC schemes.

A significant part of this thesis has been focused on the DQCA protocol (Chapter 3), an efficient MAC scheme that avoids data collisions and can achieve near-optimum performance even as the number of users grows. DQCA provides a suitable framework for CL optimization, mainly due to the structure of its frame sequence that is always completed with the broadcast transmission of a feedback packet by the AP. This packet can be employed as the vessel to convey to the users all the necessary feedback information (including CL parameters) collected by the AP.

The feedback mechanism of DQCA has been employed, in the first place, to collect and distribute information on the channel state of each user, acquired by the AP from the user access requests (ARS). This has led to the design of a link adaptation scheme that permits each user to select the transmission rate according to the quality of its wireless link to the AP (Chapter 4). A mathematical framework has been developed for the throughput and mean delay performance evaluation of DQCA with link adaptation and has been validated with the help of computer-based simulations carried out with a custom-made C++ simulation tool.

The study of the DQCA performance under different channel and traffic conditions has revealed several interesting conclusions. First, it has been shown that for a given transmission rate, DQCA obtains a higher channel utilization with respect to the legacy IEEE 802.11 DCF, by operating in a collision-free TDMA-like manner with low control overhead. The gain over the legacy MAC is also achieved when multi-rate channel scenarios are considered and the link adaptation mechanism is applied. DQCA maintains a stable maximum throughput performance regardless of the number of users due to its efficient collision resolution algorithm that can generally operate effectively with as few as $m = 2$ control minislots.

In continuation, four CL scheduling algorithms have been proposed, that alter the FIFO transmission order of DQCA in order to achieve different performance goals (Chapter 5). On the one hand, channel-aware scheduling has been employed to prioritize users with a good channel and deter transmissions by those who suffer from harsh link conditions. Opportunistic policies generally enhance the overall system performance since transmissions take place at higher bit rates. Different levels of priority can be assigned to high-rate users, resulting to different performance trade-offs. For instance, absolute opportunistic schemes, such as CL-alg, maximize throughput can be unfair towards users that experience long fading periods.

On the other hand, service-aware policies have been proposed that take into account QoS requirements imposed by the application layer. QoS provisioning becomes a necessity due to the increasing number of real-time traffic for multimedia applications. SP-alg implements a strict service differentiation scheme that guarantees high performance to the delay-sensitive classes of voice and video traffic, in the expense of the lower priority data services.

The combination of these two policies forms CLSP-alg that provides high performance opportunistic scheduling with QoS provisioning. At last, the VPF-alg offers a more general and flexible framework that can implement different CL scheduling strategies by appropriately selecting the definition of the priority function that determines the transmission order.

The proposed CL schemes have been designed with the aim to enhance throughput performance and provide QoS provisioning. Nevertheless, advanced scheduling comes at a price. The most immediate repercussion is the introduction of additional control overhead for the distribution of CL information among the system users. Another important consequence is that, although transmission priorities may benefit a subset of the system users, they have a negative impact on fairness and may cause significant performance deterioration to the less privileged users.

Hence, the suitability of each algorithm depends on the network setup (channel and traffic statistics of the users) but also on the desired performance goal. If the total system throughput is more important than the individual user satisfaction, then opportunistic schemes can be employed without caution. Under this hypothesis, low rate users would be seen as a bottleneck for the system performance and would be starved from system resources. On the other hand, if fair user treatment is the goal, then the VPF-alg is a more suitable and flexible option.

All the channel-aware CL schemes make scheduling decisions based on information of the condition of the wireless channel of each user. However, the gain achieved through opportunistic scheduling is lost if the channel information does not reflect the actual link condition. Outdated CSI leads to less efficient scheduling decisions and can cause transmission errors and packet retransmissions. The need for accurate CSI has led to the design of an update mechanism for the acquisition of periodic reports on the channel state of the users. This mechanism is particularly important under fast time-varying channels when the cost of the additional control overhead is compensated by the performance improvement of the opportunistic scheduling with valid CSI.

The last major contribution of this thesis is the proposal of solutions for the incorporation of multiuser capabilities in IEEE 802.11n-based WLAN systems, without losing backward compatibility with the standard (Chapter 6). A low-complexity beamforming technique based on the generation of random orthogonal beams has been employed at the PHY layer. The main idea behind the proposed MAC schemes is the same: to find the best set of users that can be simultaneously served by the AP, while maximizing the transmission rate and minimizing the interference. In other words, the CL-based opportunistic scheduling paradigm is again employed, this time in the context of multiuser transmissions. The challenge lies in providing efficient mechanisms for the acquisition of the necessary feedback information, which, in this case consists of the SNIR measurements of the users on the generated beams.

Two novel schemes have been proposed that differ in the implementation of the contention phase during which the downlink users attempt to transmit their channel state to the AP. Mu-Opportunistic is a polling scheme in which the users transmit sequentially a CTS frame with their CSI in a predefined order, thus avoiding collisions but generating a considerable amount of control feedback. On the other hand, in the Mu-Threshold scheme a contention window of reduced size is defined and users attempt CTS transmissions in randomly selected slots. To filter out the user participation and reduce the probability of collisions, a SNIR threshold is imposed and only users with superior link conditions are allowed to participate to the contention phase.

An analytical model for the maximum throughput performance under saturation traffic conditions of both multiuser schemes has been presented and validated through simulations. Comparison with some benchmark reference schemes has shown that considerable throughput enhancement can be obtained. Nevertheless, simultaneous multiuser transmissions do not guarantee a performance improvement. The key is to also exploit multiuser diversity by selecting the best set of users, a goal achieved through the CL-based opportunistic principle. In addition, the advantage yielded by the proposed schemes increases as the channel conditions of the users deteriorate and as the traffic load of the system grows. In other words, the additional complexity of CL scheduling becomes more justified under adverse conditions; on the other hand when the link quality of the users is high and the traffic load is low, simpler scheduling algorithms can be employed.

Overall, this thesis has followed two different roads in the context of MAC layer design. The first has been to improve the basic DQCA mechanism, an efficient but not standard protocol and the second to propose enhancements to the IEEE 802.11n standard for multiuser transmissions in multiple antenna systems. Despite the difference in the two approaches, the main philosophy behind the CL enhancements and the nature of the performance trade-offs share many similarities.

The research contributions presented in this work has opened several new lines for future investigation. Some open issues have already been identified throughout this thesis and guidelines for their approach have been provided.

The main goals for future work with respect to the first part of the thesis on DQCA-based CL design can be summarized as follows:

- So far, the performance of DQCA has been evaluated through an analytical framework and with the help of extensive simulations. An important step forward would be the hardware implementation of the DQCA MAC protocol on a testbed. This would provide further insight on the DQCA operation, it would permit the practical selection and fine tuning of several protocol parameter values and would, without doubt, open the road to many interesting experiments.

- This thesis has described some mechanisms for the recovery of the system in the presence of errors that guarantee the stable protocol operation. The next step would be to proceed to a more complete analysis and evaluation of the DQCA performance under channel errors and interference.

- The study of the coexistence of DQCA with other wireless systems that operate in unlicensed frequency bands, such as IEEE 802.11 and Bluetooth, is an open issue that should be taken into account in future deployments of DQCA-based networks. Some work in the direction of the DQCA/IEEE 802.11 coexistence has already been conducted and discussed in Chapter 3.

- Another interesting line of investigation involves the design of handoff mechanisms for a cellular scenario. This thesis has provided a description of the basic handoff actions that can be implemented in DQCA, including the channel scanning, the discovery and selection of the best AP and the reassociation process. These processes and especially the AP selection can be further enhanced through CL interactions to achieve load balancing and QoS-aware policies. The preliminary results presented in [68] can be used as a starting point for this line of future work.

There are also several open issues regarding the second part of this thesis, focused on multiuser downlink MAC schemes for IEEE 802.11n-based WLAN systems:

- The work on multiuser MAC schemes has been developed in compliance with the IEEE 802.11n specification that supports MIMO but does not consider multiuser transmissions. Recently, the IEEE 802.11ac task group has been working on an amendment aimed to extend the total network throughput beyond the gigabit-per-second barrier by adding, among other things, multiuser capability to the system. The contributions of this thesis can be used as a basis for more innovative solutions that can be aligned with the new draft standard.

- The presented results have shown that the proposed multiuser algorithms can generally improve the total system throughput but there is still a margin for performance improvement. Future work can be focused on the effort to increase efficiency by further reducing the control overhead through mechanisms such as frame aggregation or more intelligent user selection strategies. Different scheduling policies can also be employed for QoS provisioning and fairness among users.

- The analysis provided in this thesis has mostly considered homogeneous channel conditions and saturation traffic. Nevertheless, in a realistic scenario, link quality and downlink traffic may not be the same for all users or they may change over time. In order to fully exploit the potential of the proposed schemes, it would be necessary to design an adaptive algorithm for the selection of the most appropriate multiuser scheme and the optimal configuration of parameters such as the threshold and the CTS slot value in the case of Mu-Threshold.

Concluding, this thesis has advanced the state of the art first by presenting DQCA, a new efficient MAC layer protocol for WLANs, enriched through CL interactions and, second, by introducing multiuser MAC schemes for MIMO systems. The two parts of the thesis have provided valuable lessons on CL design centered at the MAC layer. Even though they have been treated independently throughout this dissertation, it is possible to envision a system where both parts are combined. This joint scenario could consist of a DQCA-based communication in the uplink direction, combined with a multiuser transmission scheme in the downlink. The road ahead lies open for further research following the new lines of investigation that have been identified.

Bibliography

[1] "IEEE Standard for Information technology – Telecommunications and information exchange between systems – Local and metropolitan area networks – Specific requirements – Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications," *IEEE Std 802.11-2007 (Revision of IEEE Std 802.11-1999)*, Dec. 2007.

[2] "IEEE Standard for Information Technology – Telecommunications and information exchange between systems – Local and metropolitan area networks – Specific requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications Amendment 5: Enhancements for Higher Throughput," *IEEE Std 802.11n-2009 (Amendment to IEEE Std 802.11-2007 as amended by IEEE Std 802.11k-2008, IEEE Std 802.11r-2008, IEEE Std 802.11y-2008, and IEEE Std 802.11w-2009)*, pp. c1–502, Oct. 2009.

[3] International Organization for Standardization and International Electrotechnical Commission, "Information Technology – Open Systems Interconnection – Basic Reference Model," *International Standard ISO/IEC 7498* , 1984.

[4] S. Kumar, V. S. Raghavan, and J. Deng, "Medium Access Control Protocols for Ad-Hoc Wireless Networks: A Survey," *Elsevier Ad Hoc Networks Journal*, vol. 4, pp. 326–358, May 2006.

[5] A. C. V. Gummalla and J. O. Limb, "Wireless medium access control protocols," *IEEE Communication Surveys &Tutorials*, vol. 3, pp. 2–15, quarter 2000.

[6] N. Abramson, "THE ALOHA SYSTEM: another alternative for computer communications," in *Proceedings of the Fall Joint Computer Conference (AFIPS '70)*, pp. 281–285, Nov. 1970.

[7] J. Capetanakis, "Tree algorithms for packet broadcast channels," *IEEE Transactions on Information Theory*, vol. 25, pp. 505–515, Sept. 1979.

[8] J. Capetanakis, "Tree algorithms for packet broadcast channels," *IEEE Transactions on Communications*, vol. 27, pp. 1476–1484, Oct. 1979.

[9] B. S. Tsybakov and V. A. Mikhailov, "Free Synchronous Packet Access in a Broadcast Channel with Feedback," *Probl. Information Transmission, I*, vol. 14, pp. 259–280, October - December 1978.

[10] E. Kartsakli, J. Alonso-Zárate, A. Cateura, C. Verikoukis, and L. Alonso, *Contention-Based Collision-Resolution Medium Access Control Algorithms.* Nova Science Publishers Inc., Apr. 2009.

[11] R. Rom and M. Sidi, *Multiple Access Protocols: Performance and Analysis.* Springer Verlag, New York, Apr. 1990.

[12] M. Schwartz and N. Abramson, "The Alohanet - surfing for wireless data [History of Communications]," *IEEE Communications Magazine*, vol. 47, pp. 21–25, Dec. 2009.

[13] F. Aune, "Cross-Layer Design Tutorial." Dept. of Electronics and Telecommunications, Norwegian University of Science and Technology, Trondheim, Norway, Nov. 2004. Published under Creative Commons License.

[14] V. Kawadia and P. R. Kumar, "A Cautionary Perspective on Cross-Layer Design," *IEEE Wireless Communications Magazine*, vol. 12, pp. 2–11, Feb. 2005.

[15] V. Srivastava and M. Motani, "Cross-layer design: a survey and the road ahead," *IEEE Communications Magazine*, vol. 43, pp. 112–119, Dec. 2005.

[16] G. Carneiro, J. Ruela, and M. Ricardo, "Cross-layer design in 4G wireless terminals," *IEEE Wireless Communications Magazine*, vol. 11, pp. 7–13, Apr. 2004.

[17] L. Alonso and R. Agustí, "Optimization of wireless communication systems using cross-layer information," *Signal Processing*, vol. 86, pp. 1755–1772, Dec. 2006.

[18] V. T. Raisinghani and S. Iyer, "Cross-layer design optimizations in wireless protocol stacks," *Computer Communications*, vol. 27, pp. 720–724, May 2004.

[19] M. van der Schaar and S. S. N., "Cross-layer wireless multimedia transmission: challenges, principles, and new paradigms," *IEEE Wireless Communications Magazine*, vol. 12, pp. 50–58, Aug. 2005.

[20] D. Qiao and S. Choi, "Goodput enhancement of IEEE 802.11a wireless LAN via link adaptation," in *Proc. of IEEE International Conference on Communications (ICC 2001)*, pp. 1995–2000, vol.7, June 2001.

[21] Q. Xia, X. Jin, and M. Hamdi, "Cross Layer Design for the IEEE 802.11 WLANs: Joint Rate Control and Packet Scheduling," *IEEE Transactions on Wireless Communications*, vol. 6, pp. 2732–2740, July 2007.

[22] B. Sadeghi, V. Kanodia, A. Sabharwal, and E. Knightly, "Opportunistic media access for multirate ad hoc networks," in *Proc. of ACM 8th Annual International Conference on Mobile Computing and Networking (MobiCom 2002)*, pp. 24–35, ACM, Sept. 2002.

[23] M. Heusse, F. Rousseau, G. Berger-Sabbatel, and A. Duda, "Performance anomaly of 802.11b," in *Proc. of the 22nd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2003)*, pp. 836–843, vol.2, Mar.–3 Apr. 2003.

[24] A. Kamerman and L. Monteban, "WaveLAN II: A High-performance wireless LAN for the unlicensed band," *Bell Labs Technical Journal*, vol. 2, pp. 118–133, Aug. 1997.

[25] G. Holland, N. Vaidya, and P. Bahl, "A rate-adaptive MAC protocol for multi-Hop wireless networks," in *Proc. of ACM 7th Annual International Conference on Mobile Computing and Networking (MobiCom 2001)*, pp. 236–251, ACM, July 2001.

[26] J. Pavon and S. Choi, "Link adaptation strategy for IEEE 802.11 WLAN via received signal strength measurement," in *Proc. of IEEE International Conference on Communications (ICC 2003)*, pp. 1108–1113, vol.2, May 2003.

[27] J. Wang, H. Zhai, and Y. Fang, "Opportunistic packet Scheduling and Media Access control for wireless LANs and multi-hop ad hoc networks," in *Proc. of IEEE Wireless Communications and Networking Conference (WCNC 2004)*, pp. 1234–1239 Vol.2, Mar. 2004.

[28] P. Bhagwat, P. Bhattacharya, A. Krishna, and S. Tripathi, "Enhancing throughput over wireless LANs using channel state dependent packet scheduling," in *Proc. of the 15th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 1996)*, pp. 1133–1140, vol.3, Mar. 1996.

[29] S. Bouam and J. Othman, "A 802.11 multiservices cross-layer approach for QoS management," in *Proc. of IEEE 60th Vehicular Technology Conference (VTC 2004 Fall)*, pp. 2698–2702, vol.4, Sept. 2004.

[30] J. Dunn, M. Neufeld, A. Sheth, D. Grunwald, and J. Bennett, "A practical cross-layer mechanism for fairness in 802.11 networks," *Mobile Networks and Applications*, vol. 11, pp. 37–45, Feb. 2006.

[31] W. Song, M. N. Krishnan, and A. Zakhor, "Adaptive packetization for error-prone transmission over 802.11 WLANs with hidden terminals," in *Proc. of IEEE International Workshop on Multimedia Signal Processing (MMSP 2009)*, pp. –, Oct. 2009.

[32] M. Cao, V. Raghunathan, and P. Kumar, "Cross-Layer Exploitation of MAC Layer Diversity in Wireless Networks," in *Proc. of IEEE 14th International Conference on Network Protocols (ICNP 2006)*, pp. 332–341, Nov. 2006.

[33] P. H. Lehne and M. Pettersen, "An Overview of Smart Antenna Technology for Mobile Communications Systems," *IEEE Communication Surveys &Tutorials*, vol. 2, pp. 2–13, 4th Quarter 1999.

[34] M. Zorzi, J. Zeidler, A. Anderson, B. Rao, J. Proakis, A. Swindlehurst, M. Jensen, and S. Krishnamurthy, "Cross-layer issues in MAC protocol design for MIMO ad hoc networks," *IEEE Wireless Communications Magazine*, vol. 13, pp. 62 – 76, Aug. 2006.

[35] D. Gesbert, M. Shafi, D. shan Shiu, P. Smith, and A. Naguib, "From theory to practice: an overview of MIMO space-time coded wireless systems," *IEEE Journal on Selected Areas in Communications*, vol. 21, pp. 281–302, Apr. 2003.

[36] W. C. Jakes, *Microwave Mobile Communications*. New York, NY: Wiley and Sons, 1974.

[37] N. Seshadri and J. Winters, "Two signaling schemes for improving the error performance of frequency-division-duplex (FDD) transmission systems using transmitter antenna diversity," in *Proc. of IEEE 41th Vehicular Technology Conference (VTC 1993)*, pp. 508 – 511, May 1993.

[38] A. Wittneben, "Basestation modulation diversity for digital simulcast," in *Proc. of IEEE 41th Vehicular Technology Conference (VTC 1991)*, pp. 848–853, May 1991.

[39] A. Wittneben, "A new bandwidth efficient transmit antenna modulation diversity scheme for linear digital modulation," in *Proc. of IEEE International Conference on Communications (ICC 1993)*, pp. 1630 – 1634 vol.3, May 1993.

[40] S. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 16, pp. 1451–1458, Oct. 1998.

[41] V. Tarokh, N. Seshadri, and A. Calderbank, "Space-time codes for high data rate wireless communication: Performance criterion and code construction," *IEEE Transactions on Information Theory*, vol. 44, pp. 744–765, Mar. 1998.

[42] V. Tarokh, H. Jafarkhani, and A. Calderbank, "Space-time block codes from orthogonal designs," *IEEE Transactions on Information Theory*, vol. 45, pp. 1456 – 1467, July 1999.

[43] P. Wolniansky, G. Foschini, G. Golden, and R. Valenzuela, "V-BLAST: an architecture for realizing very high data rates over the rich-scattering wireless channel," in *URSI International Symposium on Signals, Systems, and Electronics (ISSSE 1998)*, pp. 295–300, Sept. 1998.

[44] A. Paulraj and C. Papadias, "Space-time processing for wireless communications," *IEEE Signal Processing Magazine*, vol. 14, pp. 49–83, Nov. 1997.

[45] L. Zheng and D. Tse, "Diversity and multiplexing: a fundamental tradeoff in multiple-antenna channels," *IEEE Transactions on Information Theory*, vol. 49, pp. 1073 – 1096, May 2003.

[46] D. Tse, P. Viswanath, and L. Zheng, "iversity-multiplexing tradeoff in multiple-access channels," *IEEE Transactions on Information Theory*, vol. 50, pp. 1859–1874, Sept. 2004.

[47] C. Anton-Haro, P. Svedman, M. Bengtsson, A. Alexiou, and A. Gameiro, “Cross-layer scheduling for multi-user MIMO systems,” *IEEE Communications Magazine*, vol. 44, pp. 39–45, Sept. 2006.

[48] A. Macedo and E. Sousa, “Antenna-sector time-division multiple access for broadband indoor wireless systems,” *IEEE Journal on Selected Areas in Communications*, vol. 16, pp. 937–952, Aug. 1998.

[49] H. Yin and H. Liu, “Performance of space-division multiple-access (SDMA) with scheduling,” *IEEE Transactions on Wireless Communications*, vol. 1, pp. 611–618, Oct. 2002.

[50] D. Gesbert and M.-S. Alouini, “How much feedback is multi-user diversity really worth?,” in *Proc. of IEEE International Conference on Communications (ICC 2004)*, pp. 234–238, June 2004.

[51] S. Sanayei and A. Nosratinia, “Opportunistic Downlink Transmission With Limited Feedback,” *IEEE Transactions on Information Theory*, vol. 53, pp. 4363–4372, Nov. 2007.

[52] X. Qin and R. Berry, “Opportunistic splitting algorithms for wireless networks,” in *Proc. of the 23rd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2004)*, pp. 1662–1672, vol.3, 2004.

[53] V. Shah, N. Mehta, and R. Yim, “Analysis, Insights and Generalization of a Fast Decentralized Relay Selection Mechanism,” in *Proc. of IEEE International Conference on Communications (ICC 2009)*, pp. 1–6, June 2009.

[54] J. Mirkovic, J. Zhao, and D. Denteneer, “A MAC Protocol with Multi-User MIMO Support for Ad-Hoc WLANs,” in *Proc. of IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2007)*, pp. 1–5, Sept. 2007.

[55] Y.-J. Choi, N.-H. Lee, and S. Bahk, “Exploiting Multiuser MIMO in the IEEE 802.11 Wireless LAN Systems,” *Wireless Personal Communications*, vol. 54, pp. 385–396, Aug. 2008.

[56] W. Xu and G. Campbell, “A near perfect stable random access protocol for a broadcast channel,” in *Proc. of IEEE International Conference on Communications (ICC 1992). Conference record, SUPERCOMM/ICC ’92, Discovering a New World of Communications*, pp. 370–374 vol.1, June 1992.

[57] W. Xu and G. Campbell, “A distributed queueing random access protocol for a broadcast channel,” in *Conference Proceedings on Communications Architectures, Protocols and Applications (ACM SIGCOMM ’93*, pp. 270–278 vol.1, Sept. 1993.

[58] C.-T. Wu and G. Campbell, “Extended DQRAP (XDQRAP): A Cable TV Protocol Functioning as a Distributed Switch,” in *Proc. of the 1st International Workshop on Community Networking Integrated Multimedia Services to the Home, 1994*, pp. 191–198, July 1994.

[59] H.-J. Lin and G. Campbell, "PDQRAP - Prioritized Distributed Queueing Random Access Protocol," in *Proc. of the 19th Conference on Local Computer Networks*, pp. 82–91, Oct. 1994.

[60] C.-T. Wu and G. Campbell, "Interleaved DQRAP with Global TQ," tech. rep., DQRAP Research Group Report 944, Computer Science Department, Illinois Institute of Technology (IIT), Jan. 1995.

[61] L. Alonso, R. Agustí, and O. Sallent, "A near-optimum MAC protocol based on the distributed queueing random access protocol (DQRAP) for a CDMA mobile communication system," *IEEE Journal on Selected Areas in Communications*, vol. 18, pp. 1701–1718, Sept. 2000.

[62] L. Alonso, R. Ferrús, and R. Agustí, "WLAN throughput improvement via distributed queuing MAC," *IEEE Communication Letters*, vol. 9, pp. 310–312, Apr. 2005.

[63] G. Campbell and W. Xu, "Method and apparatus for detecting collisions on and controlling access to a transmission channel," *United States Patent 6292493*, Sept. 2001.

[64] G. Campbell and C.-T. Wu, "Method and apparatus for detecting collisions on and controlling access to a communications channel," *United States Patent 6408009 B1*, June 2002.

[65] J. Alonso-Zárate, E. Kartsakli, C. Skianis, C. Verikoukis, and L. Alonso, "Saturation Throughput Analysis of a Cluster-based Medium Access Control Protocol for Single-hop Ad Hoc Wireless Networks," *SIMULATION Transactions of The Society for Modeling and Simulation International*, vol. 84, pp. 619–633, Dec. 2008.

[66] J. Alonso-Zárate, C. Verikoukis, E. Kartsakli, and L. Alonso, "Coexistence of a Novel Medium Access Control Protocol for Wireless Ad Hoc Networks and the IEEE 802.11," in *Proc. of IEEE International Conference on Communications (ICC 2010)*, pp. 1–5, May 2010.

[67] J. Alonso-Zárate, E. Kartsakli, A. Antonopoulos, L. Alonso, and C. Verikoukis, "Handoff Functions for a Distributed Queuing Collision Avoidance Medium Access Control Protocol for Wireless LANs," in *Proc. of International Conference on Ultra Modern Telecommunications Workshops (ICUMT 2009)*, Oct. 2009.

[68] A. Antonopoulos, J. Alonso-Zárate, E. Kartsakli, L. Alonso, and C. Verikoukis, "Cross Layer Access Point Selection Mechanisms for a Distributed Queuing MAC Protocol," *accepted for publication in the Special Issue on Mobility Management in Future Internet of the Springer Telecommunications Systems Journal*, vol. 84, Feb. 2011.

[69] X. Zhang and G. Campbell, "Performance Analysis of Distributed Queueing Random Access Protocol-DQRAP," tech. rep., DQRAP Research Group Report 93-1, Computer Science Department, Illinois Institute of Technology (IIT), Aug. 1994.

[70] L. Kleinrock, *Queuing systems, Vol. 1.* New York, NY: Wiley Interscience, 1975.

[71] A. Konrad, B. Y. Zhao, A. D. Joseph, and R. Ludwig, "A Markov-based channel model algorithm for wireless networks," *Wireless Networks*, vol. 9, pp. 189–199, May 2003.

[72] P. T. Brady, "A model for generating on-off speech patterns in two-way conversation," *Bell Syst. Tech. J.*, vol. 46, pp. 2445–2472, Sept. 1969.

[73] ITU-T Recommendation G.711, "Pulse code modulation (PCM) of voice frequencies," Nov. 1998.

[74] 3GPP2 TSG-C.R1002-0 v1.0, "cdma2000 Evaluation Methodology Revision 0," Dec. 2004.

[75] Schulzrinne, H. and Casner, S. and Frederick, R. and Jacobson, V., "RTP: A Transport Protocol for Real-Time Applications," *IETF Audio-Video Transport Working Group, RFC 1889*, Jan. 1996.

[76] R. Jain, W. Hawe, and D. Chiu, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," *DEC Research Report TR-301*, Sept. 1984.

[77] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Transactions on Information Theory*, vol. 49, pp. 1691–1706, July 2003.

[78] M. Siam and M. Krunz, "An overview of MIMO-oriented channel access in wireless networks [medium access control protocols for wireless LANs]," *IEEE Wireless Communications Magazine*, vol. 15, pp. 63–69, Feb. 2008.

[79] IEEE 802 11-03/161r2, "TGn Indoor MIMO WLAN Channel Models."

[80] M. Sharif and B. Hassibi, "On the capacity of MIMO broadcast channels with partial side information," *IEEE Transactions on Information Theory*, vol. 51, pp. 506–522, Feb. 2005.

[81] N. Zorba and A. Pérez-Neira, "CAC for Multibeam Opportunistic Schemes in Heterogeneous WiMax Systems Under QoS Constraints," in *Proc. of IEEE Global Telecommunications Conference (GLOBECOM 2007)*, pp. 4296–4300, Nov. 2007.

[82] P. Viswanath, D. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Transactions on Information Theory*, vol. 48, pp. 1277–1294, June 2002.

[83] W. Szpankowski, "Analysis and Stability Considerations in a Reservation Multiaccess System," *IEEE Transactions on Communications*, vol. 31, pp. 684–692, May 1983.

[84] D. Pubill and A. Pérez-Neira, "Handoff Optimization with Fuzzy Logic in 802.11 Networks," in *Proc. of Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2006)*, July 2006.