Flow-level QoS guarantees in IEEE 802.11e-EDCA based WLANs

by

Boris Bellalta

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy (Computer Science and Digital Communication) at the Universitat Pompeu Fabra

Advisor: Miquel Oliver i Riera

December 2006
Abstract

Wireless LANs provide a broadband and mobile access to Internet, allowing multimedia mobile services to become a reality by using an alternative network to actual cellular systems. Simultaneously, the Internet user behavior is evolving to include VoIP and P2P services to the already traditional Web browsing, file transfer or e-mail. However, in this scenario, WLANs show several performance limitations, such as the difficult simultaneous coexistence of TCP and UDP flows, as it was not designed to support these heterogeneous traffic conditions.

This dissertation deals with the study of mechanisms to provide QoS guarantees at flow level in Hotspot WLANs, mitigating and solving the different impairments without losing the basic characteristic of this technology: a distributed random access MAC protocol. To solve these limitations, a novel admission and rate control scheme is proposed, based on the IEEE 802.11e-EDCA standard. This entity, placed at the Access Point takes control on the Hotspot state and manages the distribution of the transmission resources among the active nodes. However, due the non-linear behavior of the EDCA random access MAC protocol, the estimation of the available set of resources (e.g., the available bandwidth) requires complex computations, specially when different non-saturated traffic profiles are considered.

To provide a framework where the MAC behavior can be evaluated, a mathematical model of the DCF/EDCA protocol is developed, being able to capture the non-linear and reciprocal impact between network and traffic flows. Moreover, the model also captures the dynamics of all QoS traffic differentiation parameters, such as: $CW_{\text{min}}, CW_{\text{max}}, AIFS, TXOP$ and different ACK policies. The model is validated and used to understand how a WLAN Hotspot behaves with heterogeneous traffic flows and how EDCA is able to mitigate the performance impairments that appear. The performance of both TCP and UDP traffic flows is analyzed in detail, providing also different models to capture the observed dynamics. These models can be used along with the DCF/EDCA model to evaluate the expected performance of TCP-like or/and UDP-like services, such as Web Browsing or VoIP calls. For the VoIP service, a capacity analysis is provided, focusing on the causes which motivate the lower maximum number of VoIP calls that can be simultaneously active, such as the existent downlink / uplink unfairness and the multi-rate anomaly. Solutions are provided for each case.

Finally, the admission control is presented and several polices are evaluated. Adaptive policies are based on adjusting dynamically the EDCA parameters, for which a novel parameter tuning algorithm is used. The algorithm searches for the optimal EDCA parameter configuration in order to maximize both the number of rigid flows and the best-effort throughput. Results show how the use of the proposed admission control improves significantly the overall Hotspot performance, being a feasible solution for future distributed random access multimedia wireless networks.
Contents

Abstract i
Publications x
Acronym List xii

1 Introduction 1
1.1 Motivation and problem statement . . . . . . . . . . . . . . . . . . . . . . . . . . 2
1.2 Objectives, initial assumptions and approaches considered . . . . . . . . . . . . 3
1.3 Main thesis targets . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4
1.4 Thesis outline . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6

2 Related Work 7
2.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7
2.2 WLAN Hotspot performance . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8
  2.2.1 PHY performance . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8
  2.2.2 DCF (Distributed Coordination Function) . . . . . . . . . . . . . . . . . . . 9
  2.2.3 Flow Level performance in Hot-spot WLANs . . . . . . . . . . . . . . . . . 9
  2.2.4 EDCA for QoS provisioning . . . . . . . . . . . . . . . . . . . . . . . . . . 11
2.3 DCF and EDCA models . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12
  2.3.1 The conceptual model point of view . . . . . . . . . . . . . . . . . . . . . . 13
  2.3.2 Saturated vs un-saturated models . . . . . . . . . . . . . . . . . . . . . . . 13
  2.3.3 EDCA models/enhancements . . . . . . . . . . . . . . . . . . . . . . . . . . 14
2.4 Modeling the WLAN performance of transport protocols . . . . . . . . . . . . . . 15
  2.4.1 Transmission Control Protocol (TCP) . . . . . . . . . . . . . . . . . . . . . 15
  2.4.2 User Datagram Protocol (UDP) . . . . . . . . . . . . . . . . . . . . . . . . 16
  2.4.3 Mixed UDP and TCP flows . . . . . . . . . . . . . . . . . . . . . . . . . . . 16
2.5 Admission and rate control in WLANs . . . . . . . . . . . . . . . . . . . . . . . . 16
  2.5.1 Estimation of a future state . . . . . . . . . . . . . . . . . . . . . . . . . . . 17
  2.5.2 Adjusting the MAC parameters . . . . . . . . . . . . . . . . . . . . . . . . 18
  2.5.3 Admission Control procedures in EDCA . . . . . . . . . . . . . . . . . . . . 20

3 A Hotspot wireless scenario 21
3.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 21
  3.1.1 Channel propagation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 21
3.1.2 User Mobility ....................................................... 22
3.1.3 Medium Access MAC protocol ................................. 22
3.1.4 System parameters ............................................... 23
3.2 Distributed Coordination Function (DCF) ......................... 24
3.3 Enhanced Distributed Channel Access (EDCA) .................. 24

4 Modeling the IEEE 802.11e EDCA .................................. 29
4.1 Introduction ......................................................... 29
4.1.1 Mathematical Models vs Simulation Models .................. 30
4.2 Modeling the DCF for heterogeneous traffic flows ............... 30
4.2.1 A mobile node .................................................... 30
4.2.2 The MAC protocol ............................................... 31
4.2.3 Computation of channel occupation delays .................... 32
4.2.4 Model computation ............................................... 33
4.2.5 Linear approximation to compute the transmission probability . 33
4.3 The complex coexistence of rigid and elastic flows ............... 35
4.3.1 Homogeneous traffic flows ..................................... 36
4.3.2 Heterogeneous traffic flows .................................... 37
4.4 Modeling the EDCA for heterogeneous traffic flows .......... 39
4.4.1 A mobile node with multiple packet transmission - M/G[1,B]/1/K . 39
4.4.2 The M/G[1,B]/K queue .......................................... 39
4.4.3 The MAC protocol ............................................... 41
4.5 Providing traffic differentiation in WLANs ....................... 43
4.6 Searching the optimal EDCA parameters (EDCA*) ............... 45
4.6.1 Optimization Procedure ....................................... 46
4.6.2 Setting different $\rho$ levels ..................................... 47
4.7 A parameter tuning algorithm (EDCA – TA) ..................... 48
4.8 Model applications .................................................. 49
4.9 Conclusions .......................................................... 51

5 TCP and VoIP Performance in Hot-Spot WLANs .................. 53
5.1 Introduction .......................................................... 53
5.2 TCP performance in Hot-Spot WLANs ............................. 53
5.2.1 Downlink TCP flows ............................................. 55
5.2.2 Uplink TCP flows ............................................... 56
5.2.3 Simultaneous Downlink and Uplink TCP flows ............... 57
5.3 Voice over IP ......................................................... 59
5.3.1 Basic performance impairments ................................ 59
5.3.2 VoIP capacity using the Direct Link Protocol (DLP) ........ 62
5.4 Conclusions .......................................................... 65

6 Admission Control in WLANs with heterogeneous traffic ...... 67
6.1 Introduction .......................................................... 67
6.2 The Admission Control algorithm ................................ 68
6.3 Admission Control for VoIP with heterogeneous traffic

6.3.1 Flow level metrics of VoIP calls coexisting with P2P and Web elastic flows

6.3.2 VoIP call admission control in WLANs in presence of elastic traffic

6.4 Admission Control for VoIP in Multi-Rate WLANs

6.4.1 Maximum number of active calls (call capacity)

6.4.2 Admission Control Policies

6.4.3 Considerations for modeling Multi-Rate scenarios

6.4.4 Benefits of adjusting dynamically the VoIP codec rate

6.4.5 Rate-based codec selection

6.5 Conclusions

7 Conclusions

7.1 Next results beyond this thesis

7.2 Future research

7.3 Concluding remarks
List of Figures

1.1 G.279 VoIP downlink throughput in presence of i) downlink TCP flows and ii) uplink TCP flows ................................................................. 3
1.2 Maximum number of VoIP flows in a single-hop ad-hoc network with two rates 4
1.3 Black-box admission control scenario ........................................... 5
3.1 Sketch of the considered scenario .............................................. 23
3.2 The joint effect of \( m \) and \( R \) (a) \( CW_{\text{min}} = 32, m = 5, R = 4 \), (b) \( CW_{\text{min}} = 32, m = 5, R = 7 \), (c) \( CW_{\text{min}} = 32, m = 2, R = 4 \) and (d) \( CW_{\text{min}} = 32, m = 2, R = 7 \) 25
3.3 Time snapshot of the DCF behavior ............................................ 26
3.4 Maximum number of packets transmitted (\( B_{i,j} \)) each TXOP for (a) 2 Mbps, (b) 11 Mbps ................................................................. 26
4.1 Homogeneous case - Aggregate throughput for \( S_1, S_2 \) and \( E_1 \) traffic profiles ................................................................. 36
4.2 Heterogeneous case - (a) Aggregate throughput for \( S_1 \) and \( S_2 \) nodes in scenario 1, (b) Aggregate throughput for \( S_1, S_2 \) and \( E_1 \) nodes in scenario 2 ........................................ 38
4.3 \( M/G^{[1,B]}/1/K \), for values of \( B < i \) ........................................ 41
4.4 \( M/G^{[1,B]}/1/K \), Embedded Markov chain at departure epochs .......... 42
4.5 Aggregate Throughput for \( S_1 \) and \( E_1 \) nodes ............................ 44
4.6 Aggregate Throughput for \( S_1 \) and \( E_1 \) nodes. (a) IEEE 802.11b parameters (b) \( TXOP_{E_1} = 4 \) (c) \( AIFS_{E_1} = 3 \) (d) \( CW_{\text{min},E_1} = 64 \) ........................................ 45
4.7 Throughput (a) 100 Kbps, (b) 150 Kbps, (c) 200 Kbps and Queue Utilization (d) 100 Kbps (e) 150 Kbps (f) 200 Kbps ........................................ 48
4.8 (a) Goodness of fitting (\( \rho_{\text{VO}} \)) (b) Throughput for different \( \rho_{\text{VO}} \) values ................................................................. 49
4.9 Performance comparison between \( EDCA^* \) and \( EDCA^-TA \) .......... 50
5.1 Discrete Markov Chain describing the evolution of the number of backlogged nodes 56
5.2 TCP throughput (Mbps) (a) Downlink (b) Uplink .......................... 57
5.3 Throughput comparison with simultaneous TCP downlink/uplink flows (Mbps) . 58
5.4 VoIP - Throughput (AP and mobile nodes) for two voice codecs (G.711 and G.729) 60
5.5 WLAN scenario with DLP connections ...................................... 63
5.6 Queuing Utilization (a) AP (b) MNs (in both cases, the DLP is used for intra-cell calls) ................................................................. 65
6.1 Call Admission Control in the AP ............................................. 68
6.2 Call Admission Control in the MN ........................................... 69
6.3 Simulation scenarios ............................................. 71
6.4 Scenario 1 (a) VoIP Blocking Probability (b) Elastic Blocking Probability (c) VoIP Throughput (d) Elastic Throughput ............................................. 72
6.5 Scenario 2 (a) VoIP Blocking Probability (b) Elastic Blocking Probability (c) VoIP Throughput (d) Elastic Throughput ............................................. 73
6.6 CTMCs Example. Feasible states and \( CW_{\text{min}} \) values for uplink TCP flows. Dashed lines show the preemption states transitions. ............................................. 76
6.7 Non-preemptive policies (a) \( BP_s \) - Blocking probability VoIP calls, (b) \( BP_{e,u} \) - Blocking probability for elastic uplink flows ............................................. 77
6.8 Non-preemptive policies (a) \( BP_s \) - Blocking probability VoIP calls, (b) \( BP_{e,u} \) - Blocking probability for elastic uplink flows ............................................. 78
6.9 Non-preemptive policies (a) \( BP_s \) - Blocking probability VoIP calls, (b) \( BP_{e,u} \) - Blocking probability for elastic uplink flows ............................................. 78
6.10 Scenario 1 - Preemptive policies (a) \( BP_s \) - Blocking probability VoIP calls (b) \( BP_{e,u} \) - Blocking probability for elastic uplink flows (c) \( BD_{e,u} \) - Dropping probability for elastic uplink flows ............................................. 80
6.11 Scenario 2 - Preemptive policies (a) \( BP_s \) - Blocking probability VoIP calls (b) \( BP_{e,u} \) - Blocking probability for elastic uplink flows (c) \( BD_{e,u} \) - Dropping probability for elastic uplink flows ............................................. 80
6.12 Feasible number of VoIP flows / Impact of rate changes on network performance 82
6.13 Cross-layer information ............................................. 83
6.14 (a) Effect of the data rate/codec change over the system throughput when changing the codec of (a) slow flows (b) fast flows ............................................. 85
6.15 Multi-rate WLAN scenario ............................................. 86
6.16 Example of the considered bi-dimensional MC ............................................. 87
6.17 Impact of \( R_1 \) with 0 changes / second, (a) BP (b) DP (c) \( E[N] \) (d) \( E[MOS] \); with 1/60 changes / second, (e) BP (f) DP (g) \( E[N] \) (h) \( E[MOS] \) ............................................. 88
6.18 Impact of different offered load values on: (a) BP (b) DP (c) \( E[N] \) (d) \( E[MOS] \) ............................................. 89
6.19 Impact of different rate changes values on: (a) BP (b) DP (c) \( E[N] \) (d) \( E[MOS] \) ............................................. 90
## List of Tables

3.1 Sensitivity thresholds for different data rates ........................................... 22  
3.2 System parameters of the IEEE 802.11b specification [1] ........................ 23  
3.3 Default EDCA Parameter Set element parameter values ............................. 23  
3.4 Protocol overheads from upper layers ...................................................... 24  
4.1 Model validation. Traffic profiles ............................................................ 35  
4.2 System parameters of the IEEE 802.11b specification [1] ........................... 35  
4.3 $S_1$ - Homogeneous case. Model validation for several performance parameters . 36  
4.4 $S_2$ - Homogeneous case. Model validation for several performance parameters . 37  
4.5 $S_1$, $S_2$ - Queue utilization, scenario 1 .................................................. 37  
4.6 Conditional collision probability, expected number of slots of each backoff instance and channel probabilities ($S_1$), scenario 2 ........................................ 38  
4.7 $S_1$, $S_2$ - Mutual interactions among flows, scenario 1 ............................. 39  
4.8 System parameters of the IEEE 802.11b/e specification [1], [2] .................... 44  
4.9 Traffic Profiles ......................................................................................... 46  
4.10 Optimization suggested MAC parameters, with $\{CW_{\text{min,i}}, AIFS_N_i, B_i\}$ ...... 47  
4.11 Optimization suggested MAC parameters, with $\{CW_{\text{min,i}}, AIFS_N_i, B_i\}$ ...... 47  
4.12 Parameter selected from algorithm, with $\{CW_{\text{min,i}}, AIFS_N_i, B_i\}$ ........... 50  
5.1 ns2 - Aggregate throughput for persistent TCP flows (Mbps) ....................... 54  
5.2 TestBed - Aggregate throughput for persistent TCP flows (Mbps) ................. 54  
5.3 Typical values of most used codecs in VoIP .............................................. 59  
5.4 VoIP efficiency over WLAN ....................................................................... 60  
5.5 Bandwidth share between uplink and downlink VoIP flows ......................... 61  
5.6 Downlink throughput (at MAC layer) for VoIP calls in presence of TCP flows (Mbps) ................................................................................................. 62  
5.7 Capacity of VoIP calls at 11 Mbps (without using the DLP protocol) ............ 64  
5.8 Number of DLP VoIP calls at different rates .............................................. 64  
6.1 Traffic Characteristics ............................................................................... 71  
6.2 Parameters for CAC evaluation .................................................................. 76  
6.3 Number of calls with a packetization value of of 20 ms .............................. 82
Publications and other contributions

The work presented in this thesis has been partially published in the following documents (sorted by publication date).

- **Journals / Book Chapters**


  4. B. Bellalta, M. Meo, M. Oliver; *Comprehensive analytical models to evaluate the TCP performance in 802.11 WLANs*. Lecture Notes in Computer Science 3970: 37-48, Springer Verlag, 2006.

- **In proceedings of peer-reviewed International Conferences**

  1. B. Bellalta, C. Cano, M. Oliver, M. Meo; *Modeling the IEEE 802.11e EDCA for MAC parameter optimization*. Heterogeneous Networks (HetNets), September 2006, Bradford, UK.


- **Technical Documents**

  1. B. Bellalta, M. Meo, M. Oliver; *Call Admission Control in IEEE 802.11e EDCA-based WLANs. (Initial Steps)*\(^1\). TD(06)12, The 5th COST 290 Management Committee Meeting. TNO Telecom / University of Twente, The Netherlands, February 9-10, 2006.

  2. NEWCOM AP7 partners (Lúcio Ferreira, António Serrador, Boris Bellalta \(^2\), Fehti Filali, Gabriela Galvano, Guenther Liebl, Jordi Perez-Romero, Jose Monserrat, Luís M. Correia, Michela Meo, Péter Fazekas, Tom van Leeuwen, Virginia Corvino.); *Definition of Reference Scenarios for the evaluation of radio resource allocation algorithms*. To be published in January 2007.

\(^1\)It will be published by Springer in the COST action book.

\(^2\)B. Bellalta was the responsible of the Single Cell WLAN scenario, defined between UPF, POLITO and EUROCOM.
3. NEWCOM AP7 Final Deliverable; *Final Report on the activities carried out in Dept.*  

Finally, another group of related publications in national conferences is:

- **Peer-reviewed National (Spain) conferences**
  1. C. Cano, B. Bellalta, M. Oliver; *Control de Admisión y Gestión Adaptativa de los parámetros MAC en redes IEEE 802.11e (WMM)*, XVI Jornadas de I+D en Telecomunicaciones 2006, Madrid (Spain).
  2. B. Bellalta, M. Meo, A. Escudero, M. Oliver; *Modelado de una red IEEE 802.11 con tráfico heterogéneo*. Jornadas de Ingeniería Telemática (JITEL) 05. Vigo (Spain). September 2005.
  3. R. Romance, B. Bellalta, M. Oliver; *Diseño y evaluación de un módulo de control de acceso basado en balanceo de carga para redes WLAN*. XIV Jornadas de I+D en Telecomunicaciones 2004, Madrid (Spain).

These publications are also related with contributions to following research projects:


The pre-prints and other publications / documents not directly related with this thesis can be found in the URL: [http://acimut.upf.es/bbellalta](http://acimut.upf.es/bbellalta)

[^3]: [http://www.i2cat.net/](http://www.i2cat.net/)
**Acronym List**

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC</td>
<td>Access Category</td>
</tr>
<tr>
<td>ACK</td>
<td>Acknowledge</td>
</tr>
<tr>
<td>ADDTS</td>
<td>Add Traffic Stream</td>
</tr>
<tr>
<td>AP</td>
<td>Access Point</td>
</tr>
<tr>
<td>AIFS</td>
<td>Arbitration Interframe Space</td>
</tr>
<tr>
<td>AIFSN</td>
<td>Arbitration Interframe Space Number</td>
</tr>
<tr>
<td>BA</td>
<td>Basic Access</td>
</tr>
<tr>
<td>BER</td>
<td>Bit Error Ratio</td>
</tr>
<tr>
<td>BSS</td>
<td>Basic Service Set</td>
</tr>
<tr>
<td>CAC</td>
<td>Call Admission Control</td>
</tr>
<tr>
<td>CSMA/CA</td>
<td>Call Admission Control</td>
</tr>
<tr>
<td>CTMC</td>
<td>Continuous Time Markov Chain</td>
</tr>
<tr>
<td>DCF</td>
<td>Distributed Coordination Function</td>
</tr>
<tr>
<td>DELTS</td>
<td>Delete Traffic Stream</td>
</tr>
<tr>
<td>DIFS</td>
<td>Distributed Interframe Space</td>
</tr>
<tr>
<td>DLP</td>
<td>Direct Link Protocol</td>
</tr>
<tr>
<td>DSSS</td>
<td>Direct-Sequence Spread-Spectrum</td>
</tr>
<tr>
<td>DTMC</td>
<td>Discrete Time Markov Chain</td>
</tr>
<tr>
<td>EDCA</td>
<td>Enhanced Distributed Channel Access</td>
</tr>
<tr>
<td>EIFS</td>
<td>Extended Interframe Space</td>
</tr>
<tr>
<td>ESS</td>
<td>Extended Service Set</td>
</tr>
<tr>
<td>FER</td>
<td>Frame Error Ratio</td>
</tr>
<tr>
<td>HCCA</td>
<td>HCF Controlled Channel Access</td>
</tr>
<tr>
<td>HC</td>
<td>Hybrid Coordinator</td>
</tr>
<tr>
<td>HCF</td>
<td>Hybrid Coordination Function</td>
</tr>
<tr>
<td>LAN</td>
<td>Local Area Network</td>
</tr>
<tr>
<td>MAC</td>
<td>Medium Access Control</td>
</tr>
<tr>
<td>MPDU</td>
<td>MAC Protocol Data Unit</td>
</tr>
</tbody>
</table>
MN Mobile Node.

QoS Quality of Service.

OFDM Orthogonal Frequency-Division Multiplexing.

P2P Peer-2-Peer.

PHY Physical Layer.

PIFS Point Interframe Space.

PLCP Physical Layer Convergence Protocol.

RTS/CTS Ready To Send / Clear To Receive.

SIFS Short Interframe Space.

SIP Session Initiation Protocol.

SNR Signal-to-noise ratio.

TCP Transmission Control Protocol.

TXOP Transmission Opportunity.

TSPEC Traffic Specification.

UDP User Datagram Protocol.

VoIP Voice over Internet Protocol.

WLAN Wireless Local Area Network

WMM Wireless MultiMedia
Chapter 1

Introduction

The increasing popularity of Wireless Local Area Networks (WLANs) based on the IEEE 802.11 technology [1], due to the ease in installing access facilities and to the affordable price of the equipments, is pushing operators to deploy WiFi WLANs as access networks to offer Internet-based services. Therefore, WLAN cells (or hot-spots) are deployed massively in business, public areas and individual homes, being an alternative wireless access to traditional mobile cellular networks.

Balachandran et al., [3] and Na et al., [4] analyze the traffic composition of several hot-spots (these works date from 2002 and 2004 respectively). The authors report that more than 90% of the total traffic is still TCP-based, mainly due to HTTP transactions; however, it is already remarkable the presence of peer-to-peer (P2P) traffic, which reaches values higher than those obtained by the email or FTP services. Another interesting observation is the asymmetry of the traffic flows, where 85% of the total go from the fixed network to mobile nodes (downlink) and the remaining 15%, which is a significant value, from mobile nodes to the fixed network (uplink).

Nowadays, voice calls over Internet (VoIP) are becoming popular, with service providers growing around the world. Most of these VoIP users already have WLANs as access network to the Internet and therefore, the bidirectional VoIP stream goes through the AP from/to a laptop or a VoIP/WLAN phone. Then, the VoIP flow is competing with typical TCP based flows (like P2P, FTP, e-mail, etc.) to get access to the transmission resources over the wireless channel which causes serious performance problems for the real-time voice data. The same reasoning can be done for other traffic flows with strict delay / bandwidth requirements such as video-services, interactive applications, games, etc.

However, the limitations of the WLAN technology, such as the still limited radio resources, the *fair* (among nodes) random access MAC protocol, the poor channel quality depending on relative position of Mobile Nodes (MN), the anomaly observed when MNs transmit at different speeds, the interference from hidden terminals, etc., make it difficult to cope with the need to provide the required service. Indeed, current WLANs provide only best-effort services without any QoS guarantees, which due to the higher bandwidth provided, multimedia flows such video and voice streaming can perform well under low load conditions but, when traffic intensity increases, the delay and bandwidth of these rigid flows are severely affected, thus degrading the
perceived quality of service.

Solutions are therefore needed for service differentiation at the access and solving the different channel / technology impairments without losing the inherent behavior of the simplicity of the random access protocol, which is one of the basic reasons for the huge WLAN acceptance.

In particular, services can be distinguished in two classes based on the mechanisms employed at the transport layer: i) elastic flows, usually adopted to deliver data services such as file transfer applications, which correspond to traffic carried by TCP and TCP-like protocols, that adapt the traffic generation rate to the network working conditions, attempting in this way to reduce network congestion; ii) rigid flows, adopted by multi-media applications, tend to generate traffic unaware and independently of the network conditions.

Access Points (AP) and wireless cards implementing the EDCA (Enhanced Distributed Channel Access) [2] are already available under the WMM (Wireless MultiMedia) denomination. Based on the differentiation capabilities provided by the EDCA specification, some possible solutions for guaranteeing acceptable QoS levels can be achieved, consisting in optimizing the use of the transmission resources. This process requires an entity, the call admission and rate controller, capable to decide whether new traffic flows can be accepted and what resources they can use in order to guarantee the QoS levels of already active flows.

The flow admission and rate control is an open issue in the IEEE 802.11e standard (only the interfaces are specified) in order for manufacturers to provide their solutions to differentiate among them. However, still there is not an optimal accepted solution.

1.1 Motivation and problem statement

Figure 1.1 shows the VoIP throughput at the AP (downlink) when the WLAN cell is shared by MNs with active VoIP calls and other MNs downloading or uploading data, using the new-RENO TCP protocol. The data rate is equal to 2 Mbps and each VoIP flow use the G.729 codec (8 Kbps from voice codec and 24 Kbps including the RTP, UDP and IP headers). Note the fast degradation of the VoIP throughput with TCP downlink flows and the inoperability of any VoIP call with just a single TCP uplink flow. As a remarkable effect, it is also interesting to observe that, when the AP queue is saturated with VoIP traffic, the interaction with TCP traffic is reduced due to the starvation of TCP flows. Therefore, the presence of TCP traffic in both the downlink (buffer losses) and the uplink (AP starvation) leads to low performance of VoIP calls. These problems have to be solved in order to deploy a successful VoIP service over WLANs.

Another different problem is shown in Figure 1.2 where the distribution of the simultaneous feasible VoIP calls in a Hotspot network is presented (considering the use of the G.711 voice codec) for two types of VoIP calls, the FAST and the SLOW ones, which use a data rate of 11 Mbps and 1 Mbps respectively. As obviously, the total number of active is reduced as FAST calls change to SLOW calls, which means that a single data rate change could provoke that all active calls can not achieve their required bandwidth.

Therefore, several conclusions can be pointed out:

1. Without TCP traffic, the maximum number of calls that an AP can carry is lower than could be expected from the nominal data rate (19 % of efficiency), but this problem in-
increases when TCP-like flows are active simultaneously, even with a single uplink / downlink TCP flow.

2. The link adaptation algorithm (which adapts the data transmission rate to the channel conditions by changing the coding rate and modulation), allows to increase the operational area (coverage) or to minimize channel fluctuations. However, due to the nature of the random MAC protocol, a mobile node which change to a lower rate will reduce its own maximum throughput and also of the other mobile nodes. Therefore, a single data rate change (to a lower value) could provoke an overall network degradation [5].

Thus, a particular motivation of this thesis is to provide an effective solution to these problems, proposing different admission control and MAC parameter tuning schemes. However, beyond these punctual problems, the thesis is motivated by the need to provide a set of rules, algorithms and procedures that will allow future multi-service WLANs to become a reality.

1.2 Objectives, initial assumptions and approaches considered

This thesis has two main objectives

1. evaluate the required EDCA-based mechanisms which improve the WLAN hotspot performance in terms of maximizing a certain cost function. This imply the definition of a scheme / architecture which solves previous stated problems and which could be parameterized to provide different priority and controlled QoS levels.

2. propose an analytical model for EDCA-based WLANs (both ad-hoc and infrastructure)
with heterogeneous traffic flows. The models will be used to evaluate the first objective and to provide a framework for future research in the area.

To achieve the first goal, three different mechanisms are considered: i) admission control, ii) a parameter tuning algorithm and iii) cross-layer schemes. Therefore, the design, analysis and performance evaluation of them are required. A general black-box scheme of these problems and how they are inter-related is shown in Figure 1.3.

Furthermore, major part of current references on admission control for WLANs provide results at packet level [6–9], where the expected gain is shown in terms of a higher throughput / better traffic differentiation capabilities when novel traffic streams (in a deterministic way) access the system. However, there is a lack of works which present flow level metrics such as blocking / dropping probability, average flow response time (duration), etc. of the proposed schemes. Brief results about blocking probability are found in [10–12]. This is important, since as it will be noted in Chapter 6, several unexpected results (which are justified) are found (for example a higher voice blocking probability using an enhanced adaptive scheme, which degrades the rate of TCP flows, than using the more simplest one (and static) scheme).

Therefore, to provide flow level metrics, a framework is built. Due to the practical impossibility to consider a two level (packet and flow scales) simulation as the time scales in which the simulation evolve (from µ-seconds to seconds) will increase enormously the simulation time, a framework based on a mathematical model of the MAC packet level dynamics is used. The model allows us to obtain in a fast way the desired outcomes which feed both a i) flow level simulator or ii) analytical solver based on N-dimensional Markov chains.

### 1.3 Main thesis targets

For a better understanding of all the work developed in this thesis, these are the main contributions:
1. A user-centric analytical model of the IEEE 802.11 DCF / EDCA. Main characteristic of the model is its great flexibility to be parameterized in order to capture different and heterogeneous scenarios. The EDCA model includes the impact of BEB parameters \((CW_{min}, CW_{max})\), \(TXOP\) value and \(AIFS\).

2. Procedure and evaluation of a multi-objective optimization process in order to find the optimal tendencies of the MAC parameters. First results, which assure the QoS of VO flows while maximize the BE throughput, are presented.

3. Flow-level models which are able to capture the dynamics of the rigid and elastic flows in WLANs.

4. A detailed admission control procedure for EDCA-based WLANs and its evaluation in an heterogeneous multimedia scenario. Results show how the proposed solution improves significantly the overall cell performance.

5. Impact of different policies (preemption / no preemption / static / adaptive) in a WLAN cell performance.

6. A cross-layer measurement based scheme for VoIP flows in a Multi-Rate WLAN. The algorithm is included in previous admission control and decides the best MAC parameters / codec combination possible.

7. A general discussion about the reciprocal impact of elastic and rigid flows in WLANs. This is a current hot-topic which has been analyzed deeply through this thesis, providing important conceptual results about how to integrate and manage heterogeneous traffic flows.

Moreover, this work is the basis for further research in the same field, for example, as it will be explained in the Conclusions and Future Work chapter, future lines include:
1. A more detailed EDCA parameter algorithm, including a dynamic fine tuning. The new scheme has also to solve the unfairness among downlink / uplink.

2. A complete AC/MN-based optimization considering a more complex cross-layer paradigm between MAC and other stack layers.

3. Comparison between simple admission control rules and the optimal one in order to develop near-optimal admission control algorithms with low computational efforts.

1.4 Thesis outline

The thesis is structured in seven chapters, including the introduction and conclusions chapter.

The introduction chapter presents several existent performance problems in today’s WLANs, which will be studied in more detail in following chapters. Furthermore, the initial objectives and first hypothesis are outlined, which help to understand the thesis structure. Finally major results and outcomes are remarked.

Second chapter, the related work, provides a thorough overview of the state of the art over the topics covered in this thesis. However, the main results of this thesis are also introduced in order to place them in the overall state of the art.

The Wireless Hotspot scenario chapter provides a concise description of the basic concepts related to the scenario used through all the thesis. This chapter is based on the E-MORANS single-AP scenario [13], which is one of the remarkable outcomes of the thesis.

Modeling the DCF/EDCA protocol chapter presents the analytical model developed for both the DCF and EDCA MAC specifications. Validation results are provided, showing the model accuracy in different traffic scenarios. Moreover, the process to find the optimal parameters is outlined and its results are used to define a MAC parameter tuning algorithm.

Sixth chapter includes a detailed description of the proposed EDCA-based admission control. The admission control is evaluated in an heterogeneous traffic scenario with VoIP and TCP-like traffic (P2P, web) by both analytical and simulation techniques. One of the scenarios considered presents an algorithm (which is added to the admission control scheme) to improve the VoIP QoS performance in multi-rate WLANs.

Finally, last chapter include the conclusions and future research lines.
Chapter 2

Related Work

This chapter is an overview of the state of the art, presenting the most significant works related with this thesis.

2.1 Introduction

The IEEE 802.11 MAC and the PHY layer specifications for the IEEE 802.11b standard in the 2.4 GHz band are described in [1]. For other PHY specifications refer to the amendments [14] (802.11a), [15] (802.11g) or [16] (802.11n)\(^1\). These posterior PHY specifications allow to achieve higher transmission rates by improving the modulation and the channel coding, remaining unchanged the MAC specification. The current version of the MAC protocol does not implement any QoS mechanism, which is solved with the IEEE 802.11e standard [2].

Nowadays, WLANs are already a mature technology with more than 15 years of being a hot topic for research discussions / studies, development and standardization. First real WLANs appear in 1990 with the waveLAN proprietary solution from AT&T [17], which already used a DSSS (Direct Sequence Spread-Spectrum), a data rate of 2 Mbps and a MAC protocol based on the CSMA/CA. The IEEE standard for WLANs appeared in 1997, which concluded an extensive period of research and discussion about the adequate technology to be used at both PHY and MAC levels. Further versions and enhancements consider new PHY specifications in order to increase the data rate, up to 54 Mbps in the IEEE 802.11a (1999) [14] or in the IEEE 802.11g (2003) [15]. Currently, the IEEE 802.11n [16] is becoming a standard and promises data rates greater than 100 Mbps using MIMO (Multiple Input - Multiple Output) PHY techniques. Moreover, in 2005, the IEEE 802.11e [2] standard has been released. It provides a set of QoS traffic capabilities based on a new function called HCF (Hybrid Coordination Function), which provides an enhanced version of both the DCF and PCF, now called EDCA (Enhanced Distributed Channel Access) and HCCA (HCF Controlled Channel Access). This new version include the required functions to differentiate traffic streams and thus, provide different QoS levels. Since 2004/2005, commercial products implementing a limited set of EDCA functions are available, under the name of WMM (Wi-Fi MultiMedia) and certified by the Wi-Fi alliance\(^2\).

\(^1\)Currently, the IEEE 802.11n has not been already standardized.
\(^2\)http://www.wi-fi.org/.
2.2 WLAN Hotspot performance

Currently, WLANs provide a broadband access to Internet (with a bandwidth higher than 1 Mbps), which means a low response delay when using traditional content-based Internet applications such as Web browsing, e-mail, messaging or file transfer (P2P), etc., than other actual wireless technologies such as the 2.5G and 3G. Then, in places where exist a WLAN coverage, this technology can be a competitor against the existent cellular systems.

However, there are several functional (user level) questions about WLANs that have to be understand previously to deploy such service. What is the maximum WLAN capacity (in bps)? Are both transfer directions (downlink and uplink) symmetric? How many nodes can be associated to an AP? With $n$ nodes transmitting simultaneously, how much throughput can be allocated to each one? What is the maximum mobility (node speed) that can be supported by a WLAN? Will each MN receive the same service from the network independently of its position? How many energy consumes a typical WLAN card? It will affect the battery duration of a PDA or notebook? How is possible that a 11 Mbps WLAN only supports very few 64 Kbps VoIP calls? Etc.

To provide an specific answer for each question is out of the scope of this chapter. All of them are related with the random behavior of the DCF MAC protocol, which is part of the WLAN success, but also the cause of its main performance impairments. For example, Choi et al., [18], provide a complete overview of the WLAN performance where major part of these questions are answered. However, to understand them, an initial statement is that the DCF performance is basically related to the: number of nodes associated and traffic profile of each node.

2.2.1 PHY performance

To be able to understand the WLAN performance in terms of throughput or/and delay, first step is to know the raw capacity of the system under study: this is, the data rates provided by the IEEE 802.11 standard.

Through this work the data rate of 2 Mbps has been considered, which is one of the defined in the IEEE 802.11b [1] standard. The selection of this data transmission rate has been motivated for the simulation / numerical limitations more than any other reason as it provides a good tradeoff between simulation duration / required conditions (number of nodes) which have to be considered to evaluate the system.

However, data rates in WLANs goes from 1 Mbps up to values greater than 100 Mbps. Each new standard since 1999 extend the possible data rates, by using the more recent advances on channel coding and modulations. Currently, the IEEE 802.11n [16] is ready to be standardized, achieving data rates of 100 Mbps and further using OFDM (Orthogonal Frequency Division Multiplexing) and MIMO (Multiple Input Multiple Output) diversity techniques. Meanwhile, major actual AP and cards implement the IEEE 802.11b (1-11 Mbps) [1] specification and / or the IEEE 802.11g (1-54 Mbps) [15], both at 2.4 GHz.

Furthermore, each PHY specification considers the use of multiple rates, which are chosen

---

3In fact, these question are answered through the different chapters of the thesis.

4the IEEE 802.11a (1-54 Mbps) [14] at 5 GHz has a very low commercial repercussion.
based on the SNR (Signal-to-noise ratio) values or other metrics like packet losses [19]. The use of different data rates allow that MNs with good channel conditions transmit faster than MNs with bad channel conditions.

2.2.2 DCF (Distributed Coordination Function)

Therefore, since the first operational WLANs appeared, an impressive research effort (which continues nowadays) has been focused in performance studies, both using simulation and analytical models [20–26]. For the case of the DCF, main interesting metrics such as throughput and/or delay are considered. From these works, several conclusions can be summarized:

- The throughput / delay depends on the number of nodes in the network.
- There is a maximum throughput (saturation) that the network can achieve.
- The maximum efficiency of the network (in saturation conditions), is near the 80% [22].

A key factor for the success of the DCF protocol is that it provides a distributed fair access to the channel (the number of opportunities that a MN has to transmit packets to the channel). This means that two MNs carrying two traffic flows of equal bandwidth have the same transmission opportunities and they will observe the same network response. For example, in saturation conditions, all nodes transmit the same number of packets to the channel by time unit $^5$.

Maximum theoretical throughput of DCF is studied by Xiao et al., in [27]. First, considering a single saturated user (the contention delay is zero) and different MAC payload lengths. The results show how the throughput is limited by the long headers and some requirements of the MAC protocol itself, such as the transmission of ACKs for each transmitted packet. Moreover, for control frames such as the ACKs, it is required to be used the basic data rate which is the data rate that can be understood by all MNs in the network, so it is the lower one.

Another interesting characteristic is that with a low number of saturated MNs competing for the channel, the arrival of a new MN will increase the saturation throughput. However, as the number of MNs increases more than a certain threshold, the throughput starts to decrease. The reason is that with a low number of MNs the number of collisions are negligible and a multiplexing gain exists. Conversely, with more MNs the collisions and the subsequent retransmissions become noticeable, limiting the saturation throughput [22, 23].

Initially, the use of the RTS/CTS mechanism was designed to avoid the hidden terminal problem [28, 29]. However it also offers two interesting performance considerations: i) as using the RTS/CTS packets, all collisions involve the same duration, when there are a great number of nodes and the number of collisions are the throughput limitation, the overall time spend in collisions is lower than using the BA, thus, a higher throughput is achieved. Although, ii) for low number of MNs, the extra RTS and CTS packet interchange introduces a higher redundancy than in the BA, which reduces the aggregate throughput [30].

2.2.3 Flow Level performance in Hot-spot WLANs

The number of nodes, the position of each one and the traffic characteristics must be considered in order to determine the WLAN performance in terms of individual / aggregate throughput, $^5$Note that different throughput values can be measured if different packet lengths are used.
average packet transmission delay, packet losses, etc. However, the performance of specific applications (and their associated traffic flows) over WLANs is of interest to answer a basic question: What type of services can be provided? Download / upload files? Video streaming? Interactive games? Internet browsing?

To answer this question, it is necessary to remember that WLAN means Wireless Local Area Network. Therefore, it is expected to be just like a LAN but wireless. Then, it is intended to provide the same type of services than the provided in a traditional ethernet (random access) LAN. However, the first remarkable difference is on the use of the wireless channel which implies a set of new challenges for these new mobile services. A second interesting point is that major part of Internet applications when the WLAN was designed were client / server (e.g. web browsing) where major part of traffic goes from server to clients (downlink). Now, with P2P applications, these traffic profiles are changing as each MN can be a client and a server at the same time [4]. Moreover, VoIP is an emergent reality which has to be also considered. Other type of traffic is the interactive, with games becoming popular. Therefore, the real performance (use perception) is related with the interaction of the traffic profiles, the transport protocols and the WLAN itself: the MAC and PHY protocols.

A very simple classification for the traffic flows can be done by considering the transport protocol used: UDP (rigid) and TCP (elastic) flows. Main difference between them is that for rigid flows the bandwidth / delay requirements remain constant during the time that the flow is alive. Second type, the elastic flows, have as a main characteristic its ability to adapt its rate (bandwidth demand) to the network state without being the transfer delay an stringent requirement [31].

**UDP flows: VoIP**

The performance of UDP flows has been considered to show the behavior of un-saturated WLANs, as it is referred as UDP to all connectionless traffic. However, the introduction of VoIP services is a going to further increase the use of WLANs, being this a hot topic in the research community. In papers by Hole et al., [32], Hedge et al., [33], Clifford et al., [34], Wang et al., [35] or Garg et al., [36], a capacity analysis is presented. Similar conclusions are outlined in all of them, remarking that the access point is the bottleneck of the network. To avoid this problem and increase the capacity, several solutions can be considered from the IEEE 802.11e specifications [2], for example by using the EDCA enhancements [9, 34].

These works study the WLAN performance for VoIP with different voice codecs, different transmission data rates or different access schemes, such as DCF, EDCA and HCF. In spite of these heterogeneity of parameters, a similar conclusion can be extrapolated: a low VoIP capacity due to different impairments, which have to be solved or, at least, improved to achieve an efficient use of the transmission resources. They are:

1. Unfairness between uplink and downlink streams.

2. High protocol overheads.

3. Fast VoIP degradation in presence of TCP flows.

4. Variable capacity due to multi-rate transmissions (link-adaptation).
A detailed explanation of them and the solutions considered are provided in Chapter 5.

**TCP flows**

Internet is still basically used to browse over lots of information, placed in distributed servers through all the world. Most part of this information data is carried by the TCP protocol, from Internet to the end user (i.e., most of the traffic goes in the downlink direction). However, as the P2P paradigm is gaining success among the Internet community to share any type of files, the end user is becoming also a server. Therefore, the information flows from the user to Internet (i.e., uplink direction) are gaining more and more importance, changing the traditional Internet usage [37].

For typical downlink-based TCP traffic in WLANs, an interesting performance result, pointed out by several authors such as Bruno et al., [38] and Miorandi et al., [39], is that the downlink aggregate TCP throughput remains constant independently of the number of MNs. This is motivated by the fact that MNs only content against the AP transmitting ACKs, which only occurs after the reception of a TCP packet. Thus, the number of backlogged nodes with ACKs ready to be transmitted is computed to be about 2 in average. Considering only Uplink TCP flows, several performance results are presented by Pilosof et al., [40] and Leith et al., [41]. Basically, they show how the TCP window size increases up to its maximum value and so, MNs behave like saturated sources. The behavior of simultaneous downlink / uplink flows is analyzed by Bruno et al., [42]. It is shown how the downlink throughput suffers from the contention against the saturated MNs, provoking a very lower downlink TCP throughput and showing the existent unfairness between the downlink and the uplink.

In previous cites, the authors consider long-lived flows. For short-lived flows over WLANs, there are few works. For example, Bottigleliengo et al., [43] considers short-lived TCP flows in a multi-rate scenario. In order to solve the multi-rate anomaly, the authors propose a LLC scheduling which only transmits when the system detects good channel conditions. Miorandi et al., [39] models finite TCP flows but they can not be considered strictly as short-lived.

**Mixed traffic: UDP/VOIP vs TCP**

The main limitation of combining rigid (e.g. VoIP) and elastic flows is caused by the saturated-like behavior of the TCP flows. This means that when elastic flows are active, they try to transmit at maximum rate (always they are trying to transmit a packet) until all the data is completely transmitted. Then, due to the characteristics of the DCF, TCP flows achieve more transmission opportunities than rigid flows, and the performance of the rigid flows becomes unacceptable. This joint behavior has been analyzed by Bruno et al., [44] and Wang et al., [45], but there is still further research required to understand and model the TCP/UDP interactions over WLANs.

**2.2.4 EDCA for QoS provisioning**

An overview about the DCF limitations and how EDCA can solve them is shown by Ni et al., [46], including several results as examples to show the EDCA capabilities (traffic prioritization and differentiation). About the EDCA applications and challenges, an experimental study is
presented by Banchs et al., [47], showing how several EDCA capabilities work in a real test-bed. Moreover, they provide an insight on the future research directions for EDCA.

Several works, such as the presented by Aad et al., [48, 49], Mangold et al., [50], Ksentini et al., [8], Scalia et al., [51], Li et al., [52] and Tinnirello et al., [53] show how EDCA is able to provide different QoS levels by using the defined MAC parameters: BEB, AIFS, TXOP and different ACK policies. A typical scenario is considered in these previous works, where the simultaneous presence of video/voice and data flows allow to compare the performance of the EDCA traffic differentiation capabilities. Moreover, as appointed in [50], a fine-parameter tuning and admission control are required for acceptable QoS provisioning.

The challenge for EDCA is to provide a fair provisioning of access opportunities for each flow based on the traffic requirements and characteristics. For example, a set of particular challenges are:

1. To solve the downlink / uplink unfairness by given more transmission opportunities to the AP.
2. To optimize the TCP/ACK packet transmission in order to maximize the system performance.
3. To minimize the impairments caused by the multi-rate transmission.
4. To facilitate the integration of heterogeneous traffic flows over the same shared channel.

There are other QoS proposals for WLANs which do not consider EDCA, see the work by Lindgren et al., [54] and the overview of QoS schemes presented by Ni et al., [55] and Zhu et al., [56]. Moreover, EDCA can be used also to alleviate the multi-rate anomaly by using different MAC parameter setting based on the cell rate distribution and the traffic load (Kim et al., [57]).

2.3 DCF and EDCA models

Since the first specifications of the IEEE 802.11 [1] appeared, a large amount of research has been done to analytically model the IEEE 802.11 MAC protocol [58, 59], remarking the seminal papers of Bianchi [22], Tay et al. [23] and Cali et al. [60], referred in most of posterior works. All these three works consider saturated traffic conditions, as the authors assume that a MN is continuously transmitting data (always have a packet ready to be transmitted) and focusing in the MAC protocol performance.

In last years, un-saturated or finite load DCF models have been a hot topic in the research community, where each MN is considered to be un-active during a certain percentage of time. Basic un-saturated models are presented by Tickoo et al., [61], Cantieni et al., [62], Malone et al., [25] and Medepalli et al., [26].

With the standardization of the IEEE 802.11e [2], the research efforts have been directed to modeling the EDCA traffic differentiation mechanisms, such different BEB, AIFS and TXOP parameters. For example, see the works by Robinson et al., [63], Clifford et al., [64], Engelstad et al., [65] or Banchs et al. [66, 67]. Moreover, there also works focusing in complex aspects of

\[\text{Also EDCA it is referred as EDCF (Enhanced DCF).}\]
previous models or extending them, such the $TXOP$ dynamics: Lu et al., [68] or Kuppa et al., [69], or the use of different ACK policies, such Tinnirello et al., [53].

The reason of all these research efforts in modeling the DCF/EDCA MAC protocol is related with the fact that there are multiple factors which influence in the WLAN performance and today, still there are open questions to be answered. Moreover, the selection of optimal operating points has been a common problem, being this thesis also a contribution in this area.

Recent papers, from Bianchi and Tinnirello [70, 71], point out that several assumptions have to be revised to model more accurately the DCF and EDCA behavior.

2.3.1 The conceptual model point of view

A first basic classification is the point of view of the model, which can be: i) network centric (for example, see [22, 23, 60] or ii) user centric (for example, see [26, 72]).

Network centric models are based on capture the overall network behavior considering the different events from the network point of view (e.g., a packet transmission, without being relevant which MN has transmitted the packet). First saturated models ([22, 23]) are basically of this group but also can be found un-saturated homogeneous (all MN are equal) such as [65].

About user centric models, there are both saturated [72] or un-saturated (with homogeneous and heterogeneous traffic profiles) [25, 26]. Main difference is that now it is modeled the dynamics of each node, from what the network metrics can be derived (e.g., the network throughput can be obtained by adding the throughput of each node). User centric models allow more simple and logical mathematical expressions.

As expected, both type of models provide exactly the same performance results. However, user-centric based models allow a higher flexibility, including the possibility to consider heterogeneous traffic flows.

The model considered in Chapter 4 is user-centric. Its main characteristics are that, it is able to integrate all the EDCA MAC parameters in a single model and allows to consider the presence of heterogeneous traffic flows.

2.3.2 Saturated vs un-saturated models

In [22, 23, 60], the authors address the performance analysis of the MAC protocol assuming a finite number of saturated sources (i.e., sources always have a packet ready to be transmitted) which compete for the use of the shared channel in a single-hop ad-hoc network. Two basic mathematical assumptions are done: i) the decoupling assumption (i.e., nodes attempt to transmit to the channel independently from each other) is applied to simplify the protocol analysis which can be solved through a fixed point procedure and ii) the attempt rate of a node is a regenerative process and then it can be computed by using the renewal-reward theorem.

The results from these pioneer works can be used to understand how the MAC protocol performs or to obtain measures such as the maximum throughput given a number of nodes in the system. These papers are well complemented by the delay analysis of the MAC protocol presented in [72], where Carvalho et al., obtain the first and second moment of the service time, still in saturation conditions. It is remarkable that the authors of [72], find closed form expressions by linearizing the non-linear equation system between the transmission and collision
probability. In Chapter 4 a similar approach is considered for the un-saturated case.

In recent years, several papers addressed the modeling task of the IEEE 802.11 MAC performance in non-saturation conditions. Two groups of papers can be found, based on a Markovian stochastic analysis, which are extensions of the model presented by Bianchi [22]. For example in [62] and [25], the authors introduce new states in the Markov chain that describe the backoff algorithm, modeling the time in which the mobile nodes are empty (i.e., has no packet ready to be transmitted). Another group of papers is based on the observation that the attempt rate of a node is a regenerative process (which is also valid for the previous set of papers) and it can be computed by using the renewal-reward theorem. This approach is used in [61], [26] and [33].

All of this papers are based on the decoupling assumption which surprisingly works well also in non-saturated conditions and results obtained by both type of models are accurate. One of the benefits of the un-saturated system models is that they allow to analyze the system performance when nodes have different (heterogeneous) traffic profiles. For example in [25] the authors analyze the system under the presence of streaming and elastic flows, respectively carried by UDP and TCP transport layer protocols.

### 2.3.3 EDCA models/enhancements

The EDCA MAC protocol is a simple extension of the DCF to provide traffic differentiation capabilities between flows of different Access Categories. In this context, a similar classification between saturated / un-saturated or network / user-centric models can be done. The traffic differentiation is caused by the use of different MAC parameters for each AC, which gives different channel access priorities. Therefore, EDCA models take as a basis the previous DCF models and extend them with the new features defined in the standard, such as different BEB parameters [9, 63, 73, 74], the multiple packet transmission TXOP [68, 69] different AIFS [63, 64] and different ACK policies [52, 53].

From the expression of the average number of slots picked at each transmission, given by Bianchi [22] or Tay et al., [23], the BEB parameters ($CW_{\min}$ and $CW_{\max}$) are already modifiable in order to allow different combinations. For example, it is already considered in [9]. However, as it has been appointed in [53, 70] the use of lower values of $CW_{\min} / CW_{\max}$ reduces the accuracy of the model as the independence assumption losses validity.

The use of different AIFS provides a stringent traffic differentiation as MNs with long AIFS see how their probability to access the channel is significatively reduced when content for the channel with MNs with lower AIFS values. Moreover, the AIFS-based traffic differentiation is load based, which means that MNs with long AIFS values will see a reduction of transmission opportunities proportional to the load of the contending stations. Three main approaches are considered to model the AIFS-based traffic differentiation: i) extend the Markov chain which models the back-off process to include the extra slots due to AIFS [63], [64] ii) compute the average collision probability (lower for the MNs with lower AIFS) [66] and iii) compute the average transmission probability (lower for the MNs with higher AIFS) [65].

The first approach captures the AIFS dynamic accurately. However, a different Markov chain has to be modeled for each AIFS value, which reduces the model flexibility. For example, in [64] or [63], only two different values of AIFS are considered. Banchs et al., [66] proposes to compute recursively the average collision probability based on the fact that it depends on the
AIFS value of others MNs. The third possibility is introduced by Engelstad et al., [65], who propose to compute the average number of blocked slots that a MN will suffer. From this value, the new transmission probability is computed.

There are very few works that consider the TXOP dynamics in un-saturated models (in the saturated case, always a MN transmit a bulk of packets). For example, works which have proposed the use of bulk service queues are the ones by Lu et al., [68] or by Kuppa et al., [69]. However, it is still required further research work about the impact of the TXOP queue dynamics when it is combined with the other EDCA MAC parameters and with different traffic profiles, being this one of the key considerations on the design of the model presented in Chapter 4.

About the use of different ACK policies, Tinnirello et al. [53] and Li et al., [52] analyze the impact of different packet lengths / burst lengths and compare the performance of the different ACK policies provided in the EDCA standard.

2.4 Modeling the WLAN performance of transport protocols

A fast insight on the performance of TCP and UDP transport protocols in Hotspots WLANs have been pointed out in previous sections (see [44, 75]). However, models which predict their performance are also of interest, as they are useful to plan or to test how different network configurations affect the performance of these transport protocols. Therefore, it is fundamental to capture the interactions between the TCP and MAC protocols.

UDP flows do not react to network changes. Then, traffic models can be simple and only have to describe the packet rate generation (e.g., Continuous Bit Rate (CBR) or Poisson sources). However, due to its complex dynamics, TCP has motivated a lot of research, both at wired and wireless networks.

2.4.1 Transmission Control Protocol (TCP)

One of the first works that analytically addresses the performance of a WLAN with TCP traffic is [38]. The authors consider both the data traffic in the downlink and the feedback traffic in the uplink due to TCP ACKs, modeling both the access point and the mobile nodes as saturated sources. In order to catch the effect that the number of backlogged nodes is not constant and depends on the number of TCP connections, the authors propose the use of a discrete-time Markov chain to obtain the probability that $n$ nodes have an ACK ready to be transmitted, and thus compete with the AP to transmit on the channel. The average system throughput is obtained by means of a time-scale decomposition. A similar approach is used in [39], including the effect of delayed ACK techniques and the presence of finite durations TCP connections. More recently, in [76] this problem has been also considered under heterogeneous radio conditions. Similar assumptions and results are obtained in these three papers.

In the uplink direction, a model is presented by Leith et al. [41] with also similar applicability to the previous. From the same group of authors, other works such as [64] also consider the uplink TCP flow and use a judicious parameter setting to achieve fairness among uplink flows.
For what concerns TCP flows simultaneously active in both directions, Pilosof et al. [40] explain the major observed phenomena but, to the best of our knowledge, one of the first works which treats it analytically is presented by Bruno et al. [77] under the assumption that the TCP advertisement window is equal to one. A similar approach is taken by the same authors in [42] and [44]. The authors show how setting $W = 1$ is possible to provide a fair access among the downlink and uplink.

2.4.2 User Datagram Protocol (UDP)

About modeling UDP flows there are no special works as an UDP flow is simply a un-saturated source without any other required consideration. At flow level, which is important is to take into account that a rigid flows normally have a duration independent of the network state, which correspond, for example, to a call duration.

2.4.3 Mixed UDP and TCP flows

Bruno et al., [42] present a Hotspot model (DCF-based) with both TCP and UDP flows simultaneously. However, this model has a limited applicability in real cases as the authors make several stringent assumptions in order to simplify the problem, such as both TCP and UDP sources are saturated, with packets of equal length and that the TCP window is equal to 1. However, this work is one of the firsts of this quite unexplored area. The difficulties about modeling how TCP and UDP flows sharing a single buffer simultaneously is the reason for this lack of results, which also includes the traditional wired links. For example, see [78], where a single buffer is modeled to evaluate how UDP and TCP interact when different packet dropping policies are considered, such as RED (Random Early Detection) or Tail-drop.

Motivated by the use of different queues in EDCA, other authors have considered the problem of analyzing the performance of TCP and UDP arriving to the same node but using independent buffers, with priority for the UDP queue. However, major part of these works only study the EDCA performance using simulations [79], [80], without providing an analytical model. To the best of our knowledge, a first work on this topic is presented by Harsha et al., in [81].

2.5 Admission and rate control in WLANs

Admission Control has been one of the fundamentals topics in the area of mobile communications. The admission control in cellular networks is mandatory to manage the arrival / departure of voice calls, in order to allocate / deallocate the set of channels (transmission resources) to them. Moreover, the hand-off (or handover) process has to be also considered by the admission control in order to give priority (for e.g., reserving several channels exclusively for hand-off calls), which reduces the dropping probability when a cell has no free resources for the arriving hand-off call. There are an extensive number of references about this topic, see the seminal paper by Hong and Rappaport [82], the survey of Ghaderi et al. [83] and references therein for a general overview of the state of the art. For analytical performance evaluation, refer also to the work of T.S. Randhawa et al. [84] where an overview of techniques to compute the main
performance metrics are explained in detail. Also it is remarkable the work by Ramjee et al. [85], which focus on the optimization of the admission control process.

With the GPRS (General Packet Radio Service), integration of data communications in the second generation mobile networks (i.e., GSM (General Service Mobile)) becomes a reality with a more than acceptable bandwidth for low speed services (such as e-mail, Text-based Web browsing, low-rate monitoring data, etc.). Then, with both voice calls and data flows sharing the same radio resources, novel admission control policies has been developed, see [86]. The irruption of the UMTS (Universal Mobile Telecommunications System), already designed for being a multiservice IP-based solution, has also motivated the adaptation of the previous research to the new system specifications / capabilities, such as the use of CDMA (Code Division Multiple Access) [87].

This movement from voice-only to multi-service cellular networks is motivated by the user expectation of mobile multimedia contents, which goes jointly with the Internet evolution. It is in this context where WLANs appear also as a feasible solution to being a mobile access to Internet. WLANs, in spite of being initially designed for data transfer, are also used for voice and video over IP communications. However, due to the random access MAC protocol of WLANs, in case that the number of nodes and/or the traffic characteristics of the flows were greater than the maximum supported, all active flows will perform incorrectly as they will be saturated. To prevent this situation, an admission control is required to decide if accept / reject new flows and to decide what resources each flow could use (using the different MAC parameter settings). From this point of view, the admission control works in a similar way than in traditional cellular / mobile / broadband networks.

Soon after the apparition of the IEEE 802.11 standard, the IEEE 802.11e task-group was created to work in a QoS extension, being released at the end of 2005. Then, EDCA [2] is the common framework where the WLAN QoS mechanisms can be defined (notice that only the interfaces and the frames to be transferred are defined in the standard). Using it as a reference, multiple QoS proposals for admission control has been presented, for example see [6, 7, 9, 67, 88–90]. In these schemes, admission control is mandatory to prevent that the system always is in a stable state, being responsible to accept / reject new flows or if necessary to drop already active flows.

2.5.1 Estimation of a future state

One of the challenges of the admission control schemes is to predict the future system state using actual system information. This prediction is difficult due the non-linear behavior of the IEEE 802.11 MAC protocol parameters (conditional collision probability, transmission probability, mobile node queue utilization, etc.) with the number of flows and their traffic characteristics.

To make this prediction, there are two basic groups of admission control schemes [10]:

**Model based** (for example, see [6, 9, 89]) which estimate the future system status using mathematical models of the system. Main problem of these schemes is that they require to solve non-linear computational models to be able to predict accurately the future system state, which could be impractical.

**Measurement based** (for example, see [11, 88]) which predict the future system status from
current measures. However, they are reactive as the new flow have to be already active to decide if it can affect negatively the other active flows. A measured based scheme is more suitable for further fine parameter tuning or decision which does not involve to accept / reject new flows.

Anyway, this is a soft classification since the major part of CAC schemes use both models and measured information to achieve their goal. For example Pong et al. [6] uses a combined measurement / model solution. For a more detailed overview of admission control (and QoS) in WLANs, see [55, 56] and references therein.

2.5.2 Adjusting the MAC parameters

A lot of research has been done in adjusting the MAC parameters, focusing in fairness between the downlink / uplink ([91–94], maximize the throughput / minimize the transfer delay ([95–97]) or provide traffic differentiation capabilities ([98, 99]). Other works also focuses in dynamically adjust the MAC parameters to mitigate the multi-rate effect ([62, 100]).

Major part of these proposals (for example: [9, 62, 92, 97–99]) focus in tuning the BEB parameters, specially the \( CW_{\text{min}} \), as they assume that the binary exponentially increase is disabled, or at least it has a minor impact over the overall performances (i.e., \( CW_{\text{max}} = CW_{\text{min}} \)). Medepalli et al., [101] shows how this assumption, improves the short-term fairness and the system stability. Thus, the use of a single back-off stage could be beneficial, specially in unsaturated situations, to provide the desired short-term fairness. About modifying the AIFS, in [91] the authors propose that the AP (DCF-based) will use the PIFS time instead of DIFS. This solution is similar to the suggested in EDCA allowing to the AP to use the defined AIFS values minor 1, so \( AIFS_{AP} = AIFS_{MN} - 1 \), for each AC. Finally, the TXOP parameter is also considered in [11], [8] and [94], where it is adjusted in function of the load of each MN (at the MAC queue). About using a combination of these parameters, an algorithm to adjust them is presented in [93].

Moreover, several authors also link the parameter tuning algorithm with the admission control ([6, 9, 93]), being these solutions a complementary part to improve the system performance.

Then, a first basic classification of these papers is based on when the MAC parameters are updated, there are two possibilities [10]:

1. **Continuous parameter updating (measurement based)** [99]: Major part of previous cited works can be classified here as they estimate several parameters such as the collision rate or the number of contending stations [102] to adjust properly the MAC parameters.

2. **Parameter updating at fixed epochs** [6, 93]: the algorithm is only activated when a flow arrives / departs the system or when one of the active flows observes a change on the channel state.

A second classification is about where the new parameters are computed, two possibilities exist:

\[^{7}\text{This is motivated by the fact that it was considered that only the BEB parameters can be modified in the DCF.}\]
1. **Distributed** ([7, 8]) parameters are adjusted by each MN itself. This group of algorithms are measurement based.

2. **Centralized** ([6, 11]) The new parameters are computed at the AP. Both type, measurement and parameter updating at fixed epochs, are considered.

Due to the orientation of this thesis (it is based on the EDCA standard [2]), only the centralized and parameter updating at fixed points are of interest.

A third classification, already based on EDCA, refers on what parameters are updated each time the process activates. The classification is as follows:

1. **Static Parameters**: This solution is the one considered by actual EDCA implementations as is the most simple: always the default parameters are used. Main drawback of this solution is that it is unable to react to different load or channel situations. However, the standard defined parameters are already a good tradeoff to protect rigid flows while also provides a significantly performance for the elastic ones.

2. **Tuning the AC parameters**: A more efficient way is to adjust the parameters of each AC, so all flows using the same AC view the same changes on the parameters. If the flows which use the same AC have different traffic profiles, the computation of the new parameters has to be done in order to satisfy the highest requirements flow, which might result in a low performance.

3. **Tuning each AC/MN parameters**: The MAC parameters for any AC of any MN are modified in an independent way of others. Then, an optimal parameter tuning for each case can be found. However, this solution is the more flexible and the more complex, as any change on network state affect to all flows, which imply to readjust all parameters another time.

Is of special mention the Banchs et al., [9] work, which is one of the firsts where a joint admission control and a parameter tuning algorithm are presented. The authors propose in an admission control scheme and a parameter tuning algorithm for the $CW_{min}$. The rational behind this assumption is that, if the admission control provides the optimal $CW_{min}$, the fact is that $m > 0$ can tend to suboptimal situations. Then, for each new request, the CAC estimates the collision probability by assuming that all nodes are saturated (thus, the conditional collision probability is the same for all flows) and computes the system achievable throughput. Under the assumption that the transmission probability is proportional to the throughput requested by each flow, the CAC computes the individual achievable throughput. It is remarkable the derivation of optimal $CW_{min}$ values which allows the system to maximize the number of active flows.

However, there are very few works which focus in centralized MAC parameter tuning algorithms. Pong et al. [6] present a MAC parameter tuning algorithm based on reduce the $CW_{min}$ and increase the TXOP (which in this case refers to $MPDU$). However, the authors do not specify clearly how this algorithm tune the parameters (it seems that an iteratively parameter search is proposed). Freitag et al. [93] propose another parameter tuning algorithm which tunes the parameters of each AC in function of the load of each flow and the number of MNs contend- ing for the channel (the similar ideas are used in the works by Ksentini et al., [8] and Ma et al.,
As a positive point, the algorithm is not iterative and a single execution is required each time a change in the number of flows occurs. However, its main drawback is that the TXOP parameter is adjusted by the required load, without considering that, for example, rigid flows require a higher protection than elastic flows, in spite of "requesting" a lower bandwidth.

Previous limitation is solved in the tuning algorithm presented in Chapter 4 where a simple algorithm to tune the EDCA parameters is provided. The algorithm is built to explicitly maximize the best-effort traffic while, at the same time, provides the maximum possible protection for the rigid flows. Another difference with respect to the Freitag’s et al., [93] algorithm is that our proposal is iterative. A similar idea is presented by Jiang et al. [103] but they work differs in that they do not consider implicitly the maximization of the data traffic (as well as they solution guarantees a certain bandwidth to it) and only focus in providing short-term fairness for the data traffic.

2.5.3 Admission Control procedures in EDCA

The IEEE 802.11e standard [2] defines the set of procedures and frame interchange to drive the admission control process among MNs and the AP (where the admission control entity resides). Major part of works such as [6, 8, 10, 11] only focuses in the performance analysis without considering how the admission control can be explicitly implemented. Conversely, this interfaces are used (or it is assumed to be used) by several authors such as [90], where the EDCA admission control is linked with the SIP protocol for VoIP signalling, or [93].

In Chapter 6, a detailed admission control scheme is presented, which includes the different steps to be followed by the proposed algorithm.
Chapter 3

A Hotspot wireless scenario

3.1 Introduction

A single cell AP (Hotspot) scenario is considered in this thesis. In order to fix the main features and parameters considered, a brief description is presented in this chapter. For further information, refer to the different IEEE standards [1, 2].

A wireless cell is the coverage area provided by a single access point (in [1] it is referred as a Basic Service Set, BSS). The coverage area is the geographical area where both the access point (AP) and the mobile stations can communicate using the radio channel with an acceptable minimum quality; this quality can be measured in terms of SNR and other derived metrics such as the Frame Error Ratio (FER) [19]. An Extended Service Set (ESS) contains multiple access points and their coverage areas. All or part of these coverage areas can overlap, so that a mobile station can select the access point to use; these regions are called reassociation or handoff areas.

Typical scenarios with this configuration are found in public areas (like cafeterias, hotels, parks, airports) where users can access the Internet from their notebooks or PDAs; company buildings where workers use WLAN networks to communicate through the email service, message applications or voice over IP; individual users at their homes, etc. In all these scenarios, the WLAN technology provides a certain grade of mobility and a broadband access to Internet at very low cost.

A single AP providing access to a fixed network to n Mobile Nodes (MNs) is considered. The MNs and the AP use the EDCA [2] operation mode and the DSSS PHY specifications in the 2.4 GHz band [1]. Ideal channel conditions are assumed, i.e., no packet is lost due to channel errors or the hidden terminal phenomenon. This scenario is based on the E-MORANS (Extended Heterogeneous Mobile Radio Access Networks Reference Scenarios) WLAN single cell reference scenario [13].

3.1.1 Channel propagation

Different propagation models capture the effects of the 2.4 GHz band in both indoor and outdoor scenarios [104, 105]. These models include the effect of objects like walls, doors, glass, etc. The propagation model considered was found in [105]. It is described by:
Flow level QoS guarantees in IEEE 802.11e-EDCA based WLANs

### Table 3.1: Sensitivity thresholds for different data rates

<table>
<thead>
<tr>
<th>Data rate (Mbps)</th>
<th>Sensitivity (dBm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>-89</td>
</tr>
<tr>
<td>5.5</td>
<td>-91</td>
</tr>
<tr>
<td>2</td>
<td>-92</td>
</tr>
<tr>
<td>1</td>
<td>-94</td>
</tr>
</tbody>
</table>

Table 3.1: Sensitivity thresholds for different data rates

\[
PL(dB) = 40 + 35\log_{10}d + \eta(\sigma) \tag{3.1}
\]

where \(PL\) is the Path Loss, \(d\) is the distance to the AP, \(\eta(\sigma)\) is a normal random variable with mean equal to 0 and deviation \(\sigma\) which models the existing shadowing. The \(\sigma\) value is estimated to be equal to 6 dB [106]. The transmission power is assumed to be constant and equal to 100 mW (20 dBm) and the received power is

\[
P_r(dBm) = P_t(dBm) - PL(dB) \tag{3.2}
\]

The model is used to choose the data rate at which MNs will transmit. The receiver minimum sensitivity is \(-100\) dBm and for each data rate it is shown in Table 3.1 \(^1\). Moreover, a reasonable value for the background noise is \(BN = -91\) dBm [29], which can be used in the following relation:

\[
SNR = \frac{P_r}{BN} \tag{3.3}
\]

#### 3.1.2 User Mobility

The coverage area is modeled by a circle of radius \(R\) meters (limited by the lower sensitivity of the receiver). The \((x, y)\) coordinate system is used where \((x, y) = (0, 0)\) is the center of the circle (where the AP is located).

A random mobility model is considered. A Mobile Node (MN) is characterized by a constant speed \(v\) and a direction. The MN position \(P = (x, y)\) is updated every \(T\) seconds following the corresponding direction, where \(T = 1/v\) is computed from the MN average speed to obtain a 1 meter resolution (this is, \(T\) is the time a MN requires to move 1 meter from this actual position). The MN direction is changed only when the MN reaches the scenario boundary (in this case, the cell border). Then, the MN is reflected and a novel direction is selected randomly. Then, at time \(t\) each MN resides at position \((x(t), y(t))\), with a distance \(d(t)\) from the AP (center of the circle). This distance is computed from: \(d(t) = \sqrt{x^2(t) + y^2(t)}\).

#### 3.1.3 Medium Access MAC protocol

The DCF (Distributed Contention Function) [1] and EDCA (Enhanced Distributed Channel Access) [2] are considered. Both are random access MAC protocols, being EDCA an extended version of the DCF, which is enhanced with traffic differentiation capabilities.

\(^1\)These are the common values observed in different specifications of WLAN cards.
### 3.1.4 System parameters

The system parameters for the DCF are reported in Table 3.2 and the extra parameters for the EDCA in Table 3.3. Moreover, the overhead introduced by upper layers are listed in Table 3.4.

Ideal channel conditions are assumed, i.e., no packet is lost due to channel errors or the hidden terminal phenomenon. In Figure 3.1 a sketch of the considered network is presented. The fixed network is modeled by a simple 100 Mbps full duplex link with a propagation delay of 2 ms in both directions. This link is used to interconnect the fixed nodes (server, end-users) where one end-point of the traffic flows resides. The other end-points are in the mobile nodes, which are linked to the server through the access point.

![Diagram of the considered scenario](image)

**Figure 3.1: Sketch of the considered scenario**

<table>
<thead>
<tr>
<th>AC</th>
<th>AIFSNj</th>
<th>TXOPlimit (ms)</th>
<th>CWmin,j</th>
<th>CWmax,j</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (Background: BK)</td>
<td>7</td>
<td>0</td>
<td>CWmin</td>
<td>CWmax</td>
</tr>
<tr>
<td>1 (Best effort: BE)</td>
<td>3</td>
<td>0</td>
<td>CWmin</td>
<td>CWmax</td>
</tr>
<tr>
<td>2 (Video: VI)</td>
<td>2</td>
<td>6.016</td>
<td>CWmin/2</td>
<td>CWmin</td>
</tr>
<tr>
<td>3 (Voice: VO)</td>
<td>2</td>
<td>3.264</td>
<td>CWmin/4</td>
<td>CWmin/2</td>
</tr>
</tbody>
</table>

Table 3.3: Default EDCA Parameter Set element parameter values
### 3.2 Distributed Coordination Function (DCF)

The IEEE 802.11 medium access control is based on a distributed CSMA/CA protocol [1]. According to the basic access (BA) mechanism, when a node has no packets to transmit and receives a packet from the network layer, the node starts to sense the channel to determine its state, that can be either busy or free. If the channel is detected busy, the node waits until the channel is released. When the channel is detected free for a period of time larger than the DIFS (Distributed Inter-Frame Spacing) duration, a new backoff instance is generated. A backoff instance consists on a counter set to a random value each time it is generated. The random value is picked from a uniform distribution in the range $CW(k) = [0, \min(2^kCW_{\text{min}} - 1, 2^mCW_{\text{min}} - 1)]$, where $k$ is the current attempt to transmit the packet, $CW_{\text{min}}$ is the minimum size of the contention window, and $m$ defines the maximum size of the window. For each packet to be transmitted, $k$ is initially set to 0 and it is increased by one at each failed transmission until a maximum number of retransmissions, called Retry Limit (R), is reached, and the packet is dropped. The counter is decreased by one for each time-slot $\sigma$ in which the channel is sensed free, and, when the countdown reaches zero, the node starts the packet transmission on the channel. If during the backoff countdown the channel is sensed busy, the backoff is suspended until the channel is detected free again. In Figure 3.2 the joint effect of $m$ and $R$ is shown. Notice that the real impact of $CW_{\text{max}}$ is done for larger retry limits.

A collision occurs if two nodes transmit at the same time, i.e., the backoff instances from both nodes reach 0 at the same time. After the data packet is transmitted by the sender, the receiver waits for a SIFS (Short Inter-Frame Spacing) time and sends a MAC layer ACK to acknowledge the correct reception of the data packet. In the case the sender does not receive the ACK frame, it starts the retransmission procedure. After discarding or successfully transmitting a packet, if more packets are ready to be transmitted, the node starts the transmission procedure again. Otherwise, it waits for a new packet from the network layer. In Figure 3.3 an example of the basic access (BA) mechanism with three mobile stations contending to transmit a packet is plotted.

Alternative to the BA mechanism, nodes can employ a RTS/CTS protocol to access the channel, so as to reduce the hidden terminal effect.

### 3.3 Enhanced Distributed Channel Access (EDCA)

The EDCA mode of operation of the IEEE 802.11e is an extension of the DCF with the goal to provide priorities and traffic differentiation in the wireless access. To achieve this traffic differentiation, the medium access control protocol classifies each traffic flow in an Access Category (AC). Four AC are defined, each one associated to one MAC transmission queue. Each

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RTP header</td>
<td>12 Bytes</td>
</tr>
<tr>
<td>TCP header</td>
<td>20 Bytes</td>
</tr>
<tr>
<td>UDP header</td>
<td>8 Bytes</td>
</tr>
<tr>
<td>IP header</td>
<td>20 Bytes</td>
</tr>
</tbody>
</table>

Table 3.4: Protocol overheads from upper layers
AC has its own MAC parameters and behaves independently of others. Letting $AC_{j,i}$ the access category $j$ of the $i$-MN, the basic MAC parameters of each access category are labeled as: Arbitration Interframe Space $AIFS_{i,j}$, Minimum Contention Window $CW_{min,j}$, Maximum Contention Window $CW_{max,i,j}$ and Transmission Opportunity $TXOP_{i,j}$. The default values of each parameter are reported in Table 3.3.

According to the basic access (BA) mechanism, when node $i$ has no packets to transmit and receives a packet from network layer, it sends the packet to the corresponding $AC_{i,j}$ queue. At the same time, the node starts to sense the channel to determine its state, that can be either busy or free. If the channel is detected busy, the node waits until the channel is released. When the channel is detected free for a period of time larger than the $AIFS_{i,j}$ duration, a new backoff instance is generated, which consists on a counter set to a random value. The random value is picked from an uniform distribution in the range $CW_{i,j}(k) = [0, min(2^{k}CW_{min,i,j} - 1, CW_{max,i,j} - 1)]$, where $k$ is the current packet transmission attempt. For each packet to be transmitted, $k$ is initially set to 0 and it is increased by one at each failed transmission until a maximum number of retransmissions, called Retry Limit, is reached, and the packet is dropped.

The backoff counter is decreased by one for each time-slot in which the channel is sensed free,
until the countdown reaches zero, instant in which the node starts the packet transmission on the channel. If, during the backoff countdown, the channel is sensed busy, the backoff is suspended until the channel is detected free again. The $AIFS_{i,j}$ value is computed using a non-negative integer $AIFSN_{i,j}$ specific for each $AC_{i,j}$: $AIFS_{i,j} = SIFS + AIFSN_{i,j} \sigma$ (where $\sigma$ is an empty SLOT duration). Once a node gets the channel, it can transmit up to $B_{i,j}$ MPDU packets ($TXOP_{i,j}$ limit). This limit is expressed in time units (ms) and corresponds to the consecutive time that a node can transmit few (large) or several (small) packets (see Figure 3.4), where $B_{i,j}$ is computed by considering the average duration of the packets transmitted by node $i$.

A channel collision occurs if two nodes transmit at the same time, i.e., a backoff instance from two nodes reach 0 at the same time. After the data packet is transmitted to the channel by the sender, the receiver waits for a SIFS (Short Inter-Frame Spacing) time and sends a MAC layer ACK to acknowledge the correct reception of the data packet. In the case the sender does not receive the ACK frame, it starts the retransmission procedure. After discarding or successfully transmitting a packet, if more packets are ready to be transmitted, the node starts
the transmission procedure again. Otherwise, it waits for a new packet from the network layer. Another EDCA feature is the use of the different ACK policies (no ACK transmission or ACKs aggregation) which can also be used to improve the system performance.

Alternative to the BA mechanism, nodes can employ a RTS/CTS protocol to access the channel, so as to reduce the hidden terminal effect.
Chapter 4

Modeling the IEEE 802.11e EDCA

4.1 Introduction

To model means to build an abstraction of a system in order to get a tool which can be used to test the modeled system in situations that are difficult to be recreated in a real or a simulated world. Therefore, among the desired characteristics of models are: accuracy, complexity/simplicity or to be general/specific, which refer to both the computation requirements and the output itself. Some of these characteristics are complementary but other require a trade-off among them. In this situation, modeling a MAC protocol means to capture the intrinsic dynamics of the mechanisms / rules which characterize it.

Therefore, with respect to the motivations of this thesis, the model presented is intended to:

1. provide a fast evaluation of the DCF/EDCA performance.
2. evaluate the impact of tuning the different DCF/EDCA parameters.
3. study the WLAN flow level performance, which depends of the packet level behavior, as it is also appointed in [76].

This issue has been extensively addressed in the literature, starting with the seminal paper by Bianchi [22]. Bianchi provides a model of the DCF under saturation conditions, i.e., all stations compete always and continuously to transmit a packet to the channel. In this situation, the assumption of independence among MNs and of equal and constant collision probability for all stations, simplifies the mathematical complexity and provides an impressive good accuracy. The main limitations of the Bianchi’s model are the lack of applicability in real scenarios (which normally work under non-saturation conditions), the difficulties to provide user-level metrics and the complexity to parameterize the model when the MNs are not exactly equal (different frame lengths, heterogeneous transmission rates, etc.). To cope with these limitations, a simple user-centric model is proposed where each MN is modeled by a finite length queue with network-dependent service time.
Therefore, in this chapter, two user centric models, one for the DCF and its extension for the EDCA, are presented and validated. The models are built to be simple, but providing a clear description of the inter-relations between the DCF/EDCA MAC parameters, such as $AIFS$, $CW_{\text{min}}$, $CW_{\text{max}}$ and $TXOP$ values, and traffic flow characteristics (packet arrival rate and packet length). As the model is able to capture the reciprocal effects of tuning the MAC parameters, it is used for optimization purposes. Thus, using a multigoal optimization function, the optimal EDCA MAC parameters are evaluated in presence of active elastic ($BE$) and rigid ($VO$) flows. The desired goals are set to maximize the elastic throughput while assuring the bandwidth-delay requirements of the rigid flows. The outcomes allow us to define a procedure to tune iteratively the MAC parameters, showing performances close to the optimal ones.

### 4.1.1 Mathematical Models vs Simulation Models

A computer simulation can be more accurate to reproduce the real system behavior than a model (in fact the simulation code is also a model of the system under study). There are no benefits / drawbacks from using a model in front of simulations, as they have to be seen as complementary. Then, solving mathematical models is expected to be computationally more efficient than executing a simulation, but normally with a lower accuracy in the results. For example, in few seconds a computer could solve several mathematical equations which model a system. Conversely, the computer simulation of the same system can require several days in a powerful machine. However, models are build upon simplifications, approximations and/or assumptions about the important characteristics that has to be taken into consideration and how they can be mathematically modeled. So, the model accuracy is limited by the own assumptions done.

Both simulation and models are considered through this thesis, where the simulations are used to validate the mathematical models. It is shown how the models presented are able to predict accurately the behavior of the real system, providing the framework where the WLAN flow level behavior can be studied.

### 4.2 Modeling the DCF for heterogeneous traffic flows

#### 4.2.1 A mobile node

Each mobile node is approximated by a finite length queue with network-dependent service time. Packets with average length $L_i$ arrive to node $i$ with rate $\alpha_i$. Both the time between packet arrivals and the service time are assumed to be exponentially distributed. Therefore, a mobile node (the AP included) is modeled by an $M/M/1/K_i$ queue with $K_i$ as the queue length (including the packet in service) measured in packets. Let $\pi^*$ the equilibrium distribution, which from PASTA is equal to $\pi^0$. Moreover, the departure probability is $\pi^d = \pi^*/(1 - \pi^*_{K_i})$ [107].

The offered traffic load to the MAC layer and the queue utilization for node $i$ are

$$A_i = \alpha_i X_i \quad \rho_i = \alpha_i (1 - P_{b,i}) X_i = 1 - \pi^*_0$$

respectively, where $X_i$ and $P_{b,i}$ are the mean service time and the packet blocking probability. The node throughput is

$$A_i = \alpha_i X_i \quad \rho_i = \alpha_i (1 - P_{b,i}) X_i = 1 - \pi^*_0$$

(4.1)
which depends on the probability to loose a packet that is the probability that the packet is discarded at the queue entrance due to overflow or dropped at the MAC layer because the number of retransmissions have exceeded the retry limit, $R_i$. Then, a packet loss occur with probability

$$P_{L,i} = P_{b,i} + (1 - P_{b,i})P_{d,i}$$

(4.3)

where $P_{d,i}$ is the probability that a packet is dropped at the MAC layer.

By modeling each mobile node using an $M/M/1/K_i$ queue, simple expressions can be obtained to measure the quality of the service observed by a node in terms of blocking probability, average queue length and average transmission delay (including the service time).

$$P_{b,i} = \frac{A_{K_i}}{\sum_{j=0}^{K_i} A_j^i}$$

$$EQ_i = \frac{\sum_{j=0}^{K_i} j A_j^i}{\sum_{j=0}^{K_i} A_j^i}$$

$$ED_i = \frac{EQ_i}{\alpha_i(1 - P_{b,i})}$$

(4.4)

### 4.2.2 The MAC protocol

In previous sections, the considered hypothesis and approximations has been explained in detail. Letting $EB_i$ be the average number of slots selected by node $i$ at each transmission attempt, the steady state probability that the node transmits in a random slot given that a packet is ready in its transmission queue can be computed from

$$\tau_i = \frac{E[Pr(Q_i(t) > 0)]}{EB_i + 1} = \frac{\rho_i}{EB_i + 1}$$

(4.5)

where $EB_i$ is the average number of slots selected by node $i$ at each transmission attempt. For the computation of $EB_i$ has been assumed that at each attempt, the MN sees a constant conditional collision probability $p_i$. Obviously, this is a simple approximation, since the existent correlation among the instantaneous queue occupation of each node, the instantaneous backoff stage and collision probability have not been modeled. In order to address these issues refer to [108], where the number of packets in each mobile node are taken into consideration to derive these metrics.

Node $i$ transmission collides if any other node also transmits in the same slot. Then, the conditional collision probability for node $i$ is

$$p_i = 1 - \prod_{j \neq i} (1 - \tau_j)$$

(4.6)

In order to compute $EB_i$, two different approaches are considered: a stochastic (Markovian) approach [22] and an average analysis [23]. Expressions found in both papers are different but numerically equal. For simplicity, the expression of [23] has been chosen, then $EB_i$ is computed as

$$EB_i = \frac{1 - p_i - p_i(2p_i)^\alpha}{1 - 2p_i} \cdot \frac{CW_{min}}{2} - \frac{1}{2}$$

(4.7)

The effect of the Retry Limit $R_i$ is considered in [109]. However, for operative values of
$p_i < 0.4$, the effect of $R_i$ on the average backoff time at each attempt is almost negligible. Using the conditional collision probability, the dropping probability at the MAC layer is given by the probability that a packet collides $R_i$ times, $P_{d,i} = p_i^{R_i+1}$.

The service time, i.e., the time interval from the instant in which a packet enters in service until it is completely transmitted or discarded, is given by,

$$X_i = (M - 1) \left( EB_i \gamma_i + ET_{ba}^{rts} \right) + EB_i \gamma_i + T_{s,i}$$

(4.8)

where $M$ is the average number of transmissions, $\gamma_i$ is the average slot duration, $ET_{c,i}$ is the average duration of a collision of node $i$ and $T_{s,i}$ is the successful duration of a transmission from node $i$.

The average number of transmissions that a packet undergoes is computed under the decoupling assumption as,

$$M_i = \frac{1 - p_i^{R_i+1}}{1 - p_i}$$

(4.9)

A node freezes its backoff counter every time the channel is sensed busy and releases it after the channel is sensed free for a $DIFS$ period. Therefore, the time between two backoff counter decrements is a random variable which depends on the behavior of the other nodes. By letting $\gamma_i$ be the average time between two backoff counter decrements, or equivalently, the average slot duration, it is obtained

$$\gamma_i = p_{c,i} \sigma + p_{s,i}(ET_{ba}^{rts,*} + \sigma) + p_{c,i}(ET_{ba}^{rts,*} + \sigma)$$

(4.10)

where $ET_{ba}^{rts,*}$ and $ET_{ba}^{rts,*}$ are the average durations of an observed successful transmission or a collision for node $i$ when it is performing a backoff instance.

The probabilities $p_{c,i}$, $p_{s,i}$ and $p_{c,i}$ are related to the channel status in a given slot when a node is in backoff. $p_{c,i}$ is the probability that a slot is observed empty, $p_{s,i}$ the probability that in a slot a successful transmission occurs and $p_{c,i}$ is the probability that a collision occurs. Note that at the end of a successful transmission or a collision, the duration of an empty slot is added, since the backoff counter is only decreased after the channel is sensed empty for the full duration of a slot. These channel probabilities can be computed as

$$p_{c,i} = \prod_{j \neq i} (1 - \tau_j) \quad p_{s,i} = \sum_{z \neq i} \tau_z \prod_{j \neq z \neq i} (1 - \tau_j) \quad p_{c,i} = 1 - p_{c,i} - p_{s,i}$$

(4.11)

### 4.2.3 Computation of channel occupation delays

Once the behavior of the MAC has been modeled, a basic feature is the computation of the duration of the different events. Computationally efficient expressions are proposed. The value of $ET_{c,i}$ is approximated by,

$$\left\{ \begin{array}{l}
ET_{ba}^{c,i} \approx \frac{\sum_{j \neq i} \tau_j \max(T_{s,j},T_{s,i})}{\sum_{j \neq i} \tau_j} \\
ET_{rts}^{c,i} = T_{rts}^c
\end{array} \right.$$  

(4.12)

where the fact that more than two packets collide simultaneously is neglected. Note that if
the RTS/CTS access scheme is used, $ET_{c,i}^{\text{cts}}$ is constant and equal for all nodes.

To compute $ET_{c,i}^{\text{rts}}$, it is considered that the probability that more than two stations collide can be neglected, then

$$
ET_{c,i}^{\text{rts}} \simeq \frac{\sum_{j \neq i} T_{s,j} \left( \tau_j \prod_{u \neq \{i,j\}} (1 - \tau_u) \right)}{\sum_{j \neq i} \left( \tau_j \prod_{u \neq \{i,j\}} (1 - \tau_u) \right)}
$$

(4.13)

and

$$
ET_{s,i}^{\text{rts}} \simeq \frac{\sum_{j \neq i} T_{s,j} \left( \tau_j \prod_{u \neq \{i,j\}} (1 - \tau_u) \right)}{\sum_{j \neq i} \left( \tau_j \prod_{u \neq \{i,j\}} (1 - \tau_u) \right)}
$$

(4.14)

4.2.4 Model computation

Due to the dependence of previous expressions on the queue utilization of each node, $\rho_i$, and the fact that (4.5) and (4.6) form a set of non-linear equations, iterative numerical techniques have been used to solve the model.

4.2.5 Linear approximation to compute the transmission probability

To solve previous non-linear set of equations, a Taylor series approximation is provided. From [22], the transmission probability in a random chosen slot is:

$$
\tau = \frac{2(1 - 2p)}{(1 - 2p)(CW_{\min} + 1) + pCW_{\min}(1 - (2p)^m)}
$$

(4.15)

which is a function of the conditional collision probability $p$, and the BEB (Binary Exponential Back-Off) parameters: $CW_{\min}$ and $m$. The saturated condition implies that each node always has a packet ready to send and each incoming packet is immediately backlogged (the transmission queue is never empty).

In a similar way than in [61], and considering a network with $n$ mobile terminals with a packet arrival rate of $\alpha$ packets per second and a channel service rate $\beta(n)$ packets per second, the queue utilization by a node is $\rho(n)$. Considering a tagged node which transmit in a given slot, a collision occurs if one or more of the remaining $n - 1$ nodes also transmit in this slot. Then, letting $P[NT]$ the probability that a node does not transmit in randomly chosen slot is:

$$
p = 1 - P[NT]^{n-1}
$$

(4.16)

Denoting $P[QE]$ the probability that the queue is empty and $P[QNE]$ the probability the queue is not empty, $P[NT]$ is given by:

$$
P[NT] = P[NT|QE] \cdot P[QE] + P[NT|QNE] \cdot P[QNE] = 1 \cdot (1 - \rho(n)) + \frac{EB - 1}{EB} \cdot \rho(n)
$$

(4.17)

which reduces to:

$$
P[NT] = 1 - \frac{\rho(n)}{EB}
$$

(4.18)
From previous results, combining 4.16 and 4.18, for the non-saturated case, the probability \( \tau(n) \) is given by:

\[
\tau(n) = \rho(n) \frac{2(1 - 2p)}{(1 - 2p)(CW_{\text{min}} + 1) + pCW_{\text{min}}(1 - (2p)^m)}
\] (4.19)

Equations 4.16 and 4.19 form a non-linear system that can be solved using numerical techniques. Bianchi [22] showed that this system has a unique solution. In [72], the authors suggest to find an approximation for the two non-linear system equations by linearizing both equations (they consider the saturated case). The same approximation is applied to the non-saturated case. Let \( \gamma(n) = 1 - \tau(n) \), the probability that a node does not transmit in a randomly chosen slot time, i.e.,

\[
\gamma(n) = \frac{(1 - 2p)(CW_{\text{min}} + 1) - 2\rho(n) + pCW_{\text{min}}(1 - (2p)^m)}{(1 - 2p)(CW_{\text{min}} + 1) + pCW_{\text{min}}(1 - (2p)^m)}
\] (4.20)

The First order Taylor series expansion of \( \gamma(p) \) at \( p = 0 \) is given by:

\[
\gamma(n, p) = \frac{(CW_{\text{min}} + 1 - 2\rho(n)) + pCW_{\text{min}}(1 - (2p)^m)}{(CW_{\text{min}} + 1)^2} + \frac{2p(n)CW_{\text{min}}}{(CW_{\text{min}} + 1)^2} p
\] (4.21)

Considering the saturation case, thus \( \rho(n) = 1 \), the same value that the presented in [72] is obtained. If \( \rho(n) = 0 \), then \( \gamma(n, p) = 1 \) and \( \tau(n) = 0 \) (as there are no input traffic the probability that a node transmit in a slot is zero).

\[
\gamma(n, p) = 1 - \frac{2\rho(n)(CW_{\text{min}} + 1)}{(CW_{\text{min}} + 1)^2} + \frac{2p(n)CW_{\text{min}}}{(CW_{\text{min}} + 1)^2} p
\] (4.22)

which, in terms of \( q = 1 - p \) becomes:

\[
\gamma(n, q) = 1 - \frac{2\rho(n)CW_{\text{min}}}{(CW_{\text{min}} + 1)^2} q - \frac{2\rho(n)}{(CW_{\text{min}} + 1)^2}
\] (4.23)

Considering \( \tau(n, q) = 1 - \gamma(n, q) \)

\[
\tau(n, q) = \frac{2p(n)CW_{\text{min}}}{(CW_{\text{min}} + 1)^2} q + \frac{2\rho(n)}{(CW_{\text{min}} + 1)^2}
\] (4.24)

For simplification, with \( CW_{\text{min}} = 31 \) and \( \rho(n) \), second term of Equation 4.24 is lower than \( 1 \cdot 10^{-3} \) and can be approximated by 0. The probability that no node is transmitting in a randomly chosen slot is \( q \):

\[
q = (1 - \tau(n, q))^{n-1} = \left(1 - \frac{2\rho(n)CW_{\text{min}}}{(CW_{\text{min}} + 1)^2} q\right)^{n-1}
\] (4.25)

Because \( 2 \cdot CW_{\text{min}} \cdot \rho(n) / (CW_{\text{min}} + 1)^2 << 1 \) and \( 0 < q < 1 \), equation 4.25 is approximated by:

\[
q \approx \left(1 - (n - 1) \frac{2\rho(n)CW_{\text{min}}}{(CW_{\text{min}} + 1)^2} q\right)
\] (4.26)

Arranging terms and isolating \( q \):
Flow level QoS guarantees in IEEE 802.11e-EDCA based WLANs

<table>
<thead>
<tr>
<th>Traffic Flow</th>
<th>Bandwidth</th>
<th>Frame Length</th>
<th>Retry Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>elastic (E1)</td>
<td>max.available</td>
<td>1500 Bytes</td>
<td>7</td>
</tr>
<tr>
<td>streaming type 1 (S1)</td>
<td>100 Kbps</td>
<td>400 Bytes</td>
<td>7</td>
</tr>
<tr>
<td>streaming type 2 (S2)</td>
<td>250 Kbps</td>
<td>700 Bytes</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 4.1: Model validation. Traffic profiles.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_{\text{data}}$</td>
<td>2 Mbps</td>
<td>$R_{\text{basic}}$</td>
<td>1 Mbps</td>
</tr>
<tr>
<td>DIFS</td>
<td>50 µs</td>
<td>$CW_{\text{min}}$</td>
<td>32</td>
</tr>
<tr>
<td>SIFS</td>
<td>10 µs</td>
<td>$CW_{\text{max}}$</td>
<td>1024</td>
</tr>
<tr>
<td>SLOT ($\sigma$)</td>
<td>20 µs</td>
<td>$m$</td>
<td>5</td>
</tr>
<tr>
<td>EIFS</td>
<td>364 µs</td>
<td>ACK</td>
<td>112 bits @ $R_{\text{basic}}$</td>
</tr>
<tr>
<td>RTS</td>
<td>160 bits @ $R_{\text{basic}}$</td>
<td>CTS</td>
<td>112 bits @ $R_{\text{basic}}$</td>
</tr>
<tr>
<td>MAC header</td>
<td>240 bits @ $R_{\text{data}}$</td>
<td>MAC FCS</td>
<td>32 bits @ $R_{\text{data}}$</td>
</tr>
<tr>
<td>PLCP preamble</td>
<td>144 bits @ $R_{\text{basic}}$</td>
<td>PLCP header</td>
<td>48 bits @ $R_{\text{basic}}$</td>
</tr>
<tr>
<td>Retry Limit (R)</td>
<td>7</td>
<td>$Q$ (Queue length)</td>
<td>50 packets</td>
</tr>
</tbody>
</table>

Table 4.2: System parameters of the IEEE 802.11b specification [1]

\[ q \approx \frac{(CW_{\text{min}} + 1)^2}{(CW_{\text{min}} + 1)^2 + 2(n-1)\rho(n)CW_{\text{min}}} \] (4.27)

Previous approximation leads to the next expression for $p = 1 - q$ in a finite load scenario:

\[ p \approx \frac{2(n-1)\rho(n)CW_{\text{min}}}{(CW_{\text{min}} + 1)^2 + 2(n-1)\rho(n)CW_{\text{min}}} \] (4.28)

The values of $p$ (collision probability) and $q$ (probability that $n-1$ nodes do not transmit in a randomly chosen slot) are coherent with the definition of $\rho(n)$. For values of $\rho(n) = 0$, the transmission queue of all the mobile nodes is empty and the probability that no node transmit in a given random slot is $q = 1$. Thus, the collision probability is $p = 0$. In the other hand, for $\rho(n) = 1$, both $q$ and $p$ depend on the value of $CW_{\text{min}}$. This approximation for $p$ is consistent because as $n$ tends to infinity $p$ has a maximum value of 1.

### 4.3 The complex coexistence of rigid and elastic flows

In order to validate the model and analyze the performance of the IEEE 802.11 MAC protocol, a single-hop scenario with three different types of flows, whose characteristics are shown in Table 4.1. The network comprises $n+1$ nodes including the access point, each node uses the BA access scheme and carries a single traffic flow. Rigid/Streaming type 1 (Rigid/Streaming type 2) flows are referred with $S1$ ($S2$) and it is used $E1$ to refer to elastic flows. The systems parameters considered are shown in Table 4.2.

Analytical results are compared against simulations performed using the ns2 package [110]. However, a detailed simulator of the IEEE 802.11 MAC protocol using the COST (Component Oriented Simulation Toolkit) simulation package [111] has been also built and verified that it provides equivalent results with respect to ns2 but allowing a higher flexibility to monitor the dynamics of the MAC parameters.
4.3.1 Homogeneous traffic flows

A first validation is done considering that all nodes in the network have the same traffic profile (S1, S2 or E1). Figure 4.1 shows the predicted and simulated aggregate throughput against the number of flows for the three traffic classes specified in Table 4.1. The analytical and simulation models bring to very close results, showing the accuracy of the model.

As the elastic results are obtained from saturated nodes, equivalent results are obtained in [22, 23]. For the unsaturated traffic flows, in Tables 4.3 and 4.4 the values of other parameters are reported, such as the conditional collision probability $p_i$, the queue utilization $\rho_i$, the average queueing delay $ED_i$, and packet losses $P_{L,i}$. The model captures the non-linear dynamics of these parameters, specially the complex transition from the unsaturated to the saturated conditions. Note that under saturation conditions, parameters like the conditional collision probability are equal and independent of the traffic load. Differences between the model and the simulations (the model is pessimistic) are mainly motivated by the assumption of an exponential packet length distribution in the model that is constant in simulation.
4.3.2 Heterogeneous traffic flows

To define an heterogeneous scenario, a configuration where two types of rigid/streaming flows, S1 and S2, compete for the channel in presence of elastic flows, E1, was chosen. Two basic sub-scenarios are considered:

1. A variable number of S1 flows \( n_{s,1} \) and a fixed number of nodes with S2 flows \( n_{s,2} = 2 \) (scenario 1).

2. A variable number of S1 flows \( n_{s,1} \), a fixed number of nodes with S2 flows \( n_{s,2} = 2 \) and a fixed number of nodes with E1 flows \( n_{e,1} = 2 \) (scenario 2).

Figure 4.2 reports the throughput for the three types of traffic flows in both scenarios. It can be underlined that the model can capture the point where both S1 and S2 flows fail to achieve their bandwidth requirements. In Table 4.5 the queue utilization of a node is compared with simulation results (scenario 1). Note that the model provides pessimistic values but matches the dynamics of the queue utilization.

In Table 4.6 are included: the conditional collision probability, the expected number of slots of the backoff instance before a transmission attempt, and the channel probabilities observed by a S1 flow.

In the first column of Table 4.7 it is included the queue occupation for S1 nodes in the homogeneous scenario. Note how the introduction of \( n_{s,2} = 2 \) S2 flows causes an increment of the queue utilization for S1 flows. Therefore, a clear interaction from S2 flows exists and it is added to the own interaction between S1 flows. The total interaction, which is non-linear with the number of nodes, makes the queue utilization of S1 nodes saturate more rapidly. At the same time, for a fixed number of S2 nodes \( n_{s,2} = 2 \), can be observed how their queue utilization is also correlated with the queue utilization of S1 flows.
Figure 4.2: Heterogeneous case - (a) Aggregate throughput for $S_1$ and $S_2$ nodes in scenario 1, (b) Aggregate throughput for $S_1$, $S_2$ and $E_1$ nodes in scenario 2

<table>
<thead>
<tr>
<th>$n_{s,1}$</th>
<th>$p_i$</th>
<th>$EB_i$</th>
<th>$p_{c,i}$</th>
<th>$p_{e,i}$</th>
<th>$p_i$</th>
<th>$EB_i$</th>
<th>$p_{c,i}$</th>
<th>$p_{e,i}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1785</td>
<td>19.91</td>
<td>0.5214</td>
<td>0.1656</td>
<td>0.1789</td>
<td>19.84</td>
<td>0.8149</td>
<td>0.1643</td>
</tr>
<tr>
<td>2</td>
<td>0.2039</td>
<td>20.95</td>
<td>0.7960</td>
<td>0.1858</td>
<td>0.2037</td>
<td>20.96</td>
<td>0.7892</td>
<td>0.1839</td>
</tr>
<tr>
<td>4</td>
<td>0.2525</td>
<td>23.39</td>
<td>0.7474</td>
<td>0.2221</td>
<td>0.2477</td>
<td>23.91</td>
<td>0.7439</td>
<td>0.2163</td>
</tr>
<tr>
<td>6</td>
<td>0.2890</td>
<td>25.75</td>
<td>0.7109</td>
<td>0.2472</td>
<td>0.2812</td>
<td>26.55</td>
<td>0.7085</td>
<td>0.2409</td>
</tr>
<tr>
<td>8</td>
<td>0.3185</td>
<td>28.07</td>
<td>0.6814</td>
<td>0.2659</td>
<td>0.3088</td>
<td>28.91</td>
<td>0.6833</td>
<td>0.2570</td>
</tr>
<tr>
<td>10</td>
<td>0.3431</td>
<td>30.33</td>
<td>0.6568</td>
<td>0.2805</td>
<td>0.3305</td>
<td>31.37</td>
<td>0.6609</td>
<td>0.2710</td>
</tr>
</tbody>
</table>

Table 4.6: Conditional collision probability, expected number of slots of each backoff instance and channel probabilities ($S_1$), scenario 2

For admission control purposes, it is worth noting that in scenario 1, with $n_{s,1} = 7$ flows, if another $S_1$ flow is accepted, the new accepted flow will perform correctly while the $S_2$ flows will perform poorly. Therefore, if the $S_2$ service degradation is not acceptable, this new $S_1$ should be rejected. Notice also that the two types of flows cannot evaluate independently because the maximum number of flows for each type must be related to the presence of the other type of flows.

Notice that, despite the independence assumption and the use of expectations rather than distributions of the involved variables, the model is remarkably accurate. This means that the impact of these approximations on the steady state behavior of the system is limited. The reason for this limited impact may reside in the joint effect of the randomness introduced in the nodes behavior by the access protocol, the backoff mechanism in particular, and the multiplexing of different sources sharing the same radio resources. While an analytical prove of this limited impact is unfeasible, the comparison of analytical results against results obtained through detailed simulators validates the model.
4.4 Modeling the EDCA for heterogeneous traffic flows

First Access Points (AP) and wireless cards implementing the EDCA (Enhanced Distributed Channel Access) [2] are already commercially available under the WMM (Wireless MultiMedia) denomination. EDCA/WMM provides traffic differentiation by classifying the traffic flows in different Access Categories (ACs) where each AC has its own MAC parameters. These different settings affect the performance of each flow and obviously, the overall WLAN performance.

An extension of the IEEE 802.11 model of previous section is presented. This extension allow us to capture the main EDCA enhancements such as different BEB parameters ($CW_{\min}, CW_{\max}$), TXOP and AIFS. Now, each mobile node is approximated by a finite length queue with bulk and network-dependent service times. Each mobile node is assumed to carry a single traffic flow of category $AC_{i,j}$ (or equivalently, each node $i$ has only one $AC_{i,j}$ active at the same time). For the sake of simplicity, henceforth the subscript $j$ is omitted as a single AC is considered to be active in each MN.

4.4.1 A mobile node with multiple packet transmission - $M/G[1,B]/1/K$

Packets with average length $L_i$ arrive to node $i$ with average rate $\alpha_i$. Both the time between packet arrivals and the service time are assumed to be exponentially distributed. Each node (the AP included) is modeled by an $M/G[1,B]/1/K_i$ queue with bulk services times and queue length $K_i$ (which includes the packet/s in service) measured in packets (Figure 4.3).

4.4.2 The $M/G/1[1,B]/K$ queue

The $M/G/1[1,B]/1/K_i$ queue is used to capture the multiple packet transmission behavior each time the MN$_i$ gets the channel. Several assumptions are done:

1. The number of packets transmitted at bulk $m+1$ is variable and depends on the number of packets remaining at the queue after the $m$ bulk transmission.

2. If after the $m$–th bulk transmission the queue is empty, next transmission always comprises a single packet.

3. The bulk of packets is not dequeued until they are completely transmitted.
Arrival and equilibrium stationary distributions

The steady state probability that \( q \) packets are in the \( MN_i \) queue at any arbitrary instant is denoted by \( \pi_{q,i}^s \). The bulk service time \( X_b^i \), where the super-index \( b \) refers to the length of the burst, is used to compute the departure rates from each state, \( \beta_b^i = 1/X_b^i \). The steady state probabilities are computed by solving the \( \pi_i^s Q_i = 0 \) linear equation system, where \( Q_i \) is the infinitesimal generator matrix for the model of the queue (see Figure 4.3 and equation 4.30 for a particular case with \( B = 3 \)). From the \( PASTA \) property, the arrival distribution equals the stationary distribution, \( \pi_a = \pi_s \). Therefore, from the arrival and equilibrium distribution, several performance metrics are computed

\[
\begin{align*}
\rho_i &= 1 - \pi_{0,i}^s \\
P_{b,i} &= \pi^a_i(K_i) = \pi^s_i(K_i) \\
E[Q_i] &= \sum_{i=0}^{K_i} \pi_i^s \cdot i \\
E[D_i] &= \frac{E[Q_i]}{\alpha_i(1 - P_{b,i})} \\
P_{L,i} &= P_{b,i} + (1 - P_{b,i})P_{d,i}
\end{align*}
\]

(4.29)

where \( \rho_i \) is the queue utilization, \( P_{b,i} \) is the packet blocking probability, \( E[Q_i] \) is the average queue occupation, \( E[D_i] \) is the average queuing delay and \( P_{L,i} \) is the probability to loose a packet, with \( P_{d,i} \) as the probability that a packet is dropped at the MAC layer because the number of retransmissions has exceeded the retry limit.

\[
Q = \begin{pmatrix}
-\alpha & \alpha & 0 & 0 & 0 & \cdots & 0 \\
\beta^1 & -(\alpha + \beta^1) & \alpha & 0 & 0 & \cdots & 0 \\
\beta^2 & 0 & -(\alpha + \beta^2) & \alpha & 0 & \cdots & 0 \\
\beta^3 & 0 & 0 & -(\alpha + \beta^3) & \alpha & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \cdots & \vdots \\
0 & 0 & \beta^3 & 0 & 0 & -(\alpha + \beta^3) & \alpha \\
0 & 0 & 0 & \beta^3 & 0 & 0 & -\beta^3
\end{pmatrix}
\]

(4.30)

Departure stationary distribution

To compute the \( \pi_d^s \) distribution, let \( \xi_m \) be the random variable denoting the number of packets in the queue after the completion of the \( m \)-th bulk transmission. Then, the steady state probability of \( q \) packets after a departure is

\[
\pi_d^q = \lim_{m \to \infty} \text{Prob}\{\xi_m = q\}, \quad \forall q = 0 \ldots K
\]

(4.31)

To find \( \pi_d^q \) the embedded \( \xi_m \) Markov process is considered (Figure 4.4), with space state \( S = \{D_0, D_1, \ldots, D_{K-1}, D_K\} \), where state \( D_q \) means that there are \( q \) packets remaining in the queue. The transition probabilities from state \( D_f \) to state \( D_q \) are

\[
p_{g,f} = \lim_{m \to \infty} \text{Prob}\{\xi_{m+1} = g | \xi_m = f\}, \quad 0 \leq g, f \leq K
\]

(4.32)

The probability of \( h \) packets arrivals during the transmission of a bulk of length \( B \) packets, \( A_{h,B} \), is related to the service time distribution of the burst, \( f_b(B,t) \), which is different for each
$B$ value and is approximatively equal to the sum of $B$ exponential distributed packets, which results in an \textit{Erlang} $-B$ distribution. However, for simplicity, is assumed that all burst are exponentially distributed, so

$$A_{h,B} = \int_{t=0}^{\infty} f_a(j,t)f_b(B,t)dt = \int_{t=0}^{\infty} \frac{(\alpha t)^j}{j!}e^{-\alpha t}\beta^B e^{-\beta^B t}dt$$  \hspace{1cm} (4.33)$$

where $f_a(h,t)$ is the arrival process distribution which is assumed to follow a Poisson process. An example of the transition probability matrix $P$ of the embedded Markov Chain for $B = 2$ and $K = 4$ is shown in Eq. 4.34. Notice that $\pi^d_K = 0$ as the packets are dequeued only after they are transmitted. Thus, the queue is never full at a departure epoch.

$$P = \begin{pmatrix}
A_{0,1} & A_{1,1} & 1 - A_{0,1} - A_{1,1} & 0 \\
A_{0,1} & A_{1,1} & 1 - A_{0,1} - A_{1,1} & 0 \\
A_{0,2} & 1 - A_{0,2} - A_{0,1} & 0 & 0 \\
0 & 1 & 0 & 0
\end{pmatrix}$$  \hspace{1cm} (4.34)$$

Once, the departure distribution is found, by solving the linear system $\pi^dP = \pi^d$, the average number of packets transmitted in a bulk and obviously, the average bulk length in bits or in time units can be computed.

$$E[B_i] = \pi^d_0 + \sum_{k=1}^{K} k \cdot \pi^d_k$$  \hspace{1cm} (4.35)$$

Notice that the derivation of $\pi^d$ is computationally costly. A solution is to approximate $\pi^d$ with $\pi^*$ which leads to worst case results.

![Diagram](image-url)

**Figure 4.3:** $M/G^{[1,B]}/1/K$, for values of $B < i$

### 4.4.3 The MAC protocol

Letting $\zeta_i$ be the average number of backoff slots, including the BEB and AIFS-blocked slots, at each transmission attempt by node $i$, the steady state probability that the node transmits in
a random slot given that a packet is ready in its transmission queue can be approximated by

$$\tau_i = \lim_{t \to \infty} Pr(Q_i(t) > 0) = \frac{1 - \pi_0^i}{\zeta_i + 1} = \frac{\rho_i}{\zeta_i + 1}$$  \hspace{1cm} (4.36)

The $\zeta_i$ parameter includes the blocked slots for each AIFS interval ($AIFS_{ni}$) and the average number of back-off slots selected at each transmission attempt, $EB_i$, which can be computed from the expression presented by Tay et al. [23] as

$$EB_i = \frac{1 - p_i - p_i(2p_i)^{m_i}}{1 - 2p_i} \frac{CW_{\text{min},i}}{2} - \frac{1}{2}$$  \hspace{1cm} (4.37)

or using the expression derived by Bianchi [22]. The expression presented by Wu in [109], that includes the impact of retry limit, can also be used. However, the effects of the retry limit can be neglected in non-saturation conditions.

In order to compute the value of $\zeta_i$ an approach similar to the one in [65] is considered. It is taken into consideration that the number of transmissions observed by a node during its back-off is $p_{tr,i} \cdot EB_i$, which implies an extra number of blocked slots that the node has to wait approximately equal to $p_{tr,i} \cdot EB_i \cdot AIFS_{ni}$, then

$$\begin{cases}
\zeta_i \approx (EB_i + AIFS_{ni}) + p_{tr,i} \cdot EB_i \cdot AIFS_{ni} \\
p_{tr,i} = 1 - \prod_{j \neq i} (1 - \tau_j)
\end{cases}$$  \hspace{1cm} (4.38)

where $p_{tr,i}$ is the probability that at least another node transmits in a given slot.

The service time, i.e., the time interval from the instant in which a packet enters in service until it is completely transmitted or discarded, is given by,

$$X_i^{b_i} = (M_i - 1) \left( \zeta_i \gamma_i + ET_{s_i}^{r|\text{rts}} \right) + \zeta_i \gamma_i + ET_{s_i}^{b_i,ba|\text{rts}}$$  \hspace{1cm} (4.39)

where $M_i$ is the average number of required transmissions, $\gamma_i$ is the average slot duration, $ET_{s_i}^{b_i,ba|\text{rts}}$ is the average duration of a $E[R_i]$ packets burst transmission, which is computed from the $\pi^d$ distribution, using the BA or the RTS/CTS access mechanism, and $ET_{s_i}^{ba|\text{rts}}$ is the average duration for a collision of node $i$. 

---

Figure 4.4: $M/G^{[1,B]}/1/K$ Embedded Markov chain at departure epochs
The value of $ET_{c,i}^{ba||rts}$ is approximated by,

$$\begin{align*}
ET_{c,i}^{ba} &\approx \frac{\sum_{j\neq i} \tau_j \max(T_{s,i}^{ba}, T_{s,j}^{ba})}{\sum_{j} \tau_j} \\
ET_{c,i}^{rts} & = T_{c}^{rts}
\end{align*}$$

(4.40)

where it is neglected the fact that more than two packets collide simultaneously. Note that if the RTS/CTS access scheme is used, $ET_{c,i}^{rts}$ is constant and equal for all nodes.

A node freezes its backoff counter every time the channel is sensed busy and releases it after the channel is sensed free for an $AIFS$ period. Therefore, the time between two backoff counter decrements is a random variable which depends on the behavior of the other nodes. By letting $\gamma_i$ be the average time between two backoff counter decrements, or equivalently, the average slot duration,

$$\gamma_i = p_{e,i} + p_{s,i}(ET_{s,i}^{ba||rts,*} + \sigma) + p_{c,i}(ET_{c,i}^{ba||rts,*} + \sigma)$$

(4.41)

where $ET_{s,i}^{ba||rts,*}$ and $ET_{c,i}^{ba||rts,*}$ are the average durations of an observed successful transmission or a collision for node $i$ when it is performing a backoff instance. By neglecting the probability that more than two stations collide,

$$\begin{align*}
ET_{c,i}^{ba,*} &\approx \frac{\sum_{j\neq i} \sum_{k\neq j} \tau_j \max(T_{s,j}^{ba}, T_{s,k}^{ba})(\tau_j \tau_k \prod_{u\neq \{j,k,i\}} (1-\tau_u))}{\sum_{j} \tau_j} \\
ET_{c,i}^{rts,*} & = T_{c}^{rts}
\end{align*}$$

(4.42)

and

$$ET_{s,i}^{ba||rts,*} \approx \frac{\sum_{j\neq i} ET_{s,j}^{ba||rts} (\tau_j \prod_{u\neq \{i,j\}} (1-\tau_u))}{\sum_{j\neq i} (\tau_j \prod_{u\neq \{i,j\}} (1-\tau_u))}$$

(4.43)

where $ET_{s,j}^{ba||rts}$ is the average duration of a successful transmission from a node $j$. Notice that this average duration is related also to the queue occupation and can be averaged from the $\pi^d$ distribution.

The probabilities $p_{e,i}$, $p_{s,i}$ and $p_{c,i}$ are related to the channel status (empty, successful transmission and collision) in a given slot when a node is performing its backoff.

$$\begin{align*}
p_{e,i} & = \prod_{j\neq i} (1 - \tau_j) \\
p_{s,i} & = \sum_{z\neq i} \tau_z \prod_{j\neq z\neq i} (1 - \tau_j) \\
p_{c,i} & = 1 - p_{e,i} - p_{s,i}
\end{align*}$$

(4.44)

### 4.5 Providing traffic differentiation in WLANs

In order to validate the model and to compare the performance of the $DCF$ and $EDCA$ in a particular scenario, a single-hop scenario with two traffic types has been considered: elastic (maximum achievable bandwidth (saturated node) and frame length equal to 1500 Bytes) and streaming (bandwidth of 100 Kbps and frame length equal to 400 Bytes). The network comprises $n$ nodes, each node uses the BA access scheme and carries a single traffic flow. It is referred with $R$ (or $S1$) to unsaturated (rigid) flows and $E$ is used to refer to saturated (elastic) flows.
Analytical results are compared against simulations performed using a detailed simulator of the EDCA IEEE 802.11e MAC protocol built using the COST (Component Oriented Simulation Toolkit) simulation engine [111]. The considered parameters are shown in Table 4.8.

In Figure 4.5 the aggregate throughput for four $E$ flows when the number of $R$ flows increase is plotted. Each rigid flow uses a TXOP limit of $B_R = 4$ packets, a $CW_{\text{min},R} = 8$ and $A_R = 2$ while the $E$ flows use $B_E = 1$ packets, $CW_{\text{min},E} = 32$ and $A_E = 3$. Notice how the DCF is already saturated for a single $R$ flow while EDCA allow up to 8 rigid flows. Obviously, using EDCA the $E$ throughput is lower.

More results are provided for each one of the different parameters, which are plotted in Figure 4.6, where the aggregate throughput for four $E_1$ flows when the number of $S_1$ (rigid)
flows increase. The goal is to show how there are different possibilities to provide protection to S1 flows with respect to the E1 flows. Note how with a properly parameter tuning the goal to improve the system performance can be achieved. Anyway, these results allow to validate the model as both simulation and analytical outputs match very well. Notice that, in this second group of figures, the $CW_{min}$ is set to 32 and $AIFS = DIFS$ for all traffic classes.

### 4.6 Searching the optimal EDCA parameters $(EDCA^*)$

In this section the optimal performance of the EDCA MAC protocol is analyzed, i.e., the performance achieved using the MAC parameters that maximize a certain cost function. Two types
of traffic are considered: BE (best-effort) and VO (voice), which use the corresponding EDCA access category. The MAC parameters used are from Table 4.8.

### 4.6.1 Optimization Procedure

In order to compute the optimal working parameters, a multi-goal optimization function is used. Four objectives are defined: i) to guarantee the rigid throughput, ii) a queue utilization of rigid flows near $p_{VO} = 0.8$, iii) an average queueing delay (including service time) lower than 150 ms and iv) the maximum allowable elastic throughput. The set of weights are: i) the rigid throughput is considered a hard constraint (1 %), ii) a low level of flexibility for $p_{VO}$ (10 %), iii) a soft requirement for the average queueing delay (10%) \(^1\) and iv) a very soft requirement for the best-effort throughput (75 %)). The optimization problem is solved by using the MATLAB \(^2\) package. The allowed range for MAC parameters is: $CW_{min,BE} = [32, 1024]$, $CW_{min,VO} = [8, 32]$, $AIFS_{BE} = [3, 7]$, $AIFS_{VO} = [2]$, $TXOP_{BE} = [1, 7]$ and $TXOP_{VO} = [1, 7]$.

<table>
<thead>
<tr>
<th>Traffic Flow</th>
<th>Bandwidth</th>
<th>Frame Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>elastic (E)</td>
<td>max.available</td>
<td>1500 Bytes</td>
</tr>
<tr>
<td>rigid (R)</td>
<td>100, 150 and 200 Kbps</td>
<td>200 Bytes</td>
</tr>
</tbody>
</table>

Table 4.9: Traffic Profiles

To obtain the optimal parameters a fixed number of elastic flows equal to $n_e = 3$ and a variable number of rigid flows $n_s$ are considered. Packet lengths and bandwidth values are the shown in Table 4.9. The optimal parameters are shown in Table 4.10. From these results, next tendencies are observed:

1. The $CW_{min,BE}$ and $CW_{min,VO}$ are used only for a "fine-tuning" optimization. These parameters are only significatively modified when all other parameters are at their maximum/minimum values.

2. As the goal is trying to maximize the BE throughput, the $TXOP_{BE}$ starts with higher values. It is decreased as the contention increases ($n_s$ flows increases). However, which is remarkable, prior to decrease this value, the system tries to check whether higher values of $TXOP_{VO}$ are feasible. For example with $B_{VO} = 150$ Kbps and $n_s = 4$, a single increment of $TXOP_{VO}$ allows also to increase $TXOP_{BE}$, regarding the optimal parameters for $n_s = 3$.

3. The $TXOP_{VO}$ starts with lower values and it is increased as contention increases.

4. $AIFS_{BE}$ only is increased when both $TXOP_{BE}$ and $TXOP_{VO}$ are near its minimum/maximum values. It is interesting to see how its influence is related with the traffic load. Thus, increasing the number of VO flows could not require increase the $AIFS_{BE}$ value.

Using the optimal settings, called $EDCA^*$, the system performance is plotted in Fig. 4.7. Notice that using the DCF any rigid flow is able to achieve its throughput requirements as the elastic flows starve the rigid ones. This is solved by using the $EDCA$ (fixed parameters) which

\(^1\) As the queue length of MN is limited to 20 packets, the average delay had a very low influence during the optimization process.

\(^2\) © http://www.mathworks.com
show a clear and strict traffic differentiation between the VO and BE ACs. However, allowing higher $\rho_{VO}$ values than the ones provided by EDCA, the overall performance is improved using the EDCA* parameters. For all three $B_{VO}$ values, using EDCA* there are an increment of one rigid flow over the maximum number of supported rigid flows from EDCA at the cost of reducing the BE throughput only in this point. For any other number of $n_s$ a higher BE throughput is obtained.

### 4.6.2 Setting different $\rho$ levels

The goal over the VO queue utilization provides a threshold between queuing delay and losses for VO flows and the flexibility to allocate extra resources to the BE AC flows. This means that lower objective values for $\rho_{VO}$ will give higher protection to VO flows (less losses and delay) but also a lower throughput for BE flows. Setting a bandwidth of $B_{VO} = 150$ Kbps, with $n_s = 5$ and $n_r = 3$, the parameter evolution for different $\rho_{VO}$ values is described in Table 4.11. Notice that the range of possible $AIFS_{VO}$ values has been increased up to 7 to provide more freedom to the optimization function.

<table>
<thead>
<tr>
<th>Objective $\rho_{VO}$</th>
<th>$B = 150$ Kbps</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$BE$</td>
</tr>
<tr>
<td>0.1</td>
<td>{177, 1, 4}</td>
</tr>
<tr>
<td>0.2</td>
<td>{81, 1, 5}</td>
</tr>
<tr>
<td>0.3</td>
<td>{32, 2, 3}</td>
</tr>
<tr>
<td>0.4</td>
<td>{32, 1, 1}</td>
</tr>
<tr>
<td>0.5</td>
<td>{32, 4, 1}</td>
</tr>
<tr>
<td>0.6</td>
<td>{32, 3, 1}</td>
</tr>
<tr>
<td>0.7</td>
<td>{32, 2, 1}</td>
</tr>
<tr>
<td>0.8</td>
<td>{39, 7, 5}</td>
</tr>
<tr>
<td>0.9</td>
<td>{32, 2, 1}</td>
</tr>
</tbody>
</table>

Table 4.11: Optimization suggested MAC parameters, with \{CW_{min,i}, AIFS_{VO}, B_i\}
4.7 A parameter tuning algorithm (EDCA – TA)

The parameters suggested by EDCA are designed to provide a clear traffic differentiation, which could reduce unnecessarily the performance of BE traffic. To solve this situation, a simple tuning algorithm has been proposed, which, compared with EDCA: i) increases the maximum number of feasible VO flows and ii) improves significantly the throughput for the BE active flows. These gains are expected to be greater as the data rate of the system increases. As the optimization process is computationally costly, a simple algorithm capable of capture the tendencies observed is required.

1. Initialize the parameter vector to the EDCA defined parameters, so $\xi = [BE : (32, 7, 3), VO : (8, 1, 2)]$. Set the goals vector $\mathbf{G} = (S_{VO}, D)$ to the desired values of throughput $S_{VO}$ and delay $D$ (mean and jitter). Go to 2.

2. Using the current parameters $\xi$. Check whether the goals (allowing a 1% of packet losses) are achieved or the maximum number of iterations have been reached. If so, go to Step 4. Else, go to Step 3.
Flow level QoS guarantees in IEEE 802.11e-EDCA based WLANs December 2006

3. Increase/Decrease sequentially (only one change each iteration):

   (a) $TXOP_{VO} \rightarrow TXOP_{VO} + 1$. Update $\xi$. Go to Step 2.
   (b) $AIFS_{BE} \rightarrow AIFS_{BE} + 1$. Update $\xi$. Go to Step 2.
   (c) $TXOP_{BE} \rightarrow TXOP_{BE} - 1$. Update $\xi$. Go to Step 2.

   If the maximum values of all three parameters have been reached, double at each iteration the value of $CW_{min,BE}$ until its maximum value is reached. Update $\xi$. Go to Step 2.

4. Set the working parameters to $\xi$.

   Running the algorithm in the same conditions in what the $EDCA^*$ parameters have been obtained, the MAC parameters resulting are shown in Table 4.12. Notice the similar behavior with the optimal parameters of previous section. Moreover, in Fig 4.9 there are plotted the aggregate throughput and the queue utilization for the $EDCA^*$ and the $EDCA-TA$ parameters. Rigid throughput is perfectly matched by the proposed algorithm as it is the main objective of the $EDCA-TA$ algorithm. However, not significant lower values than the optimal are obtained for the $BE$ throughput. Notice also, the similar evolution of the $\rho_{VO}$ queue utilization.

4.8 Model applications

One of the most interesting features of the DCF/EDCA model is its flexibility to reproduce several situations of interest:

Complex scenarios. As each node can be configured independently, it is easy to model nodes with different functions in the same network (access point, mobile nodes, etc.). This flexibility was difficult to achieve using saturated models due to their reduced parametrization.
Flow level QoS guarantees in IEEE 802.11e-EDCA based WLANs

<table>
<thead>
<tr>
<th>$n_r$</th>
<th>$B_r = 100$ Kbps ${BE, VO}$</th>
<th>$B_r = 150$ Kbps ${BE, VO}$</th>
<th>$B_r = 200$ Kbps ${BE, VO}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>${32, 7, 3}$ ${8, 2, 2}$</td>
<td>${32, 7, 4}$ ${8, 2, 2}$</td>
<td>${32, 7, 4}$ ${8, 2, 2}$</td>
</tr>
<tr>
<td>2</td>
<td>${32, 7, 3}$ ${8, 2, 2}$</td>
<td>${32, 7, 4}$ ${8, 2, 2}$</td>
<td>${32, 6, 4}$ ${8, 2, 2}$</td>
</tr>
<tr>
<td>3</td>
<td>${32, 7, 3}$ ${8, 2, 2}$</td>
<td>${32, 7, 4}$ ${8, 2, 2}$</td>
<td>${32, 6, 4}$ ${8, 3, 2}$</td>
</tr>
<tr>
<td>4</td>
<td>${32, 7, 3}$ ${8, 2, 2}$</td>
<td>${32, 6, 4}$ ${8, 3, 2}$</td>
<td>${32, 5, 5}$ ${8, 4, 2}$</td>
</tr>
<tr>
<td>5</td>
<td>${32, 7, 4}$ ${8, 2, 2}$</td>
<td>${32, 6, 5}$ ${8, 3, 2}$</td>
<td>${32, 2, 7}$ ${8, 6, 2}$</td>
</tr>
<tr>
<td>6</td>
<td>${32, 7, 4}$ ${8, 2, 2}$</td>
<td>${32, 4, 6}$ ${8, 4, 2}$</td>
<td>$\cdots$ $\cdots$</td>
</tr>
<tr>
<td>7</td>
<td>${32, 6, 4}$ ${8, 3, 2}$</td>
<td>${256, 1, 7}$ ${8, 3, 2}$</td>
<td>$\cdots$ $\cdots$</td>
</tr>
<tr>
<td>8</td>
<td>${32, 6, 5}$ ${8, 3, 2}$</td>
<td>$\cdots$ $\cdots$</td>
<td>$\cdots$ $\cdots$</td>
</tr>
<tr>
<td>9</td>
<td>${32, 5, 6}$ ${8, 4, 2}$</td>
<td>$\cdots$ $\cdots$</td>
<td>$\cdots$ $\cdots$</td>
</tr>
<tr>
<td>10</td>
<td>${32, 2, 7}$ ${8, 6, 2}$</td>
<td>$\cdots$ $\cdots$</td>
<td>$\cdots$ $\cdots$</td>
</tr>
<tr>
<td>11</td>
<td>${1024, 1, 7}$ ${8, 8, 2}$</td>
<td>$\cdots$ $\cdots$</td>
<td>$\cdots$ $\cdots$</td>
</tr>
</tbody>
</table>

Table 4.12: Parameter selected from algorithm, with $\{CW_{min, i}, AIFS_i, B_i\}$

![Figure 4.9: Performance comparison between EDCA* and EDCA – TA](image)

**Multi-rate capabilities.** The extension of the model to multi-rate networks is rather straightforward, as can be assigned to each flow/node a different value of $R_{data}$.

**Admission control and resource scheduling.** As the model catches the relationships be-
between flows, it can be used to evaluate different resource strategies, such as those based on setting the $CW$ value, and test the overall effect of this setting on the network performance.

4.9 Conclusions

In this chapter some issues related to the provision of integrated services through WiFi access networks have been discussed. The popularity of WiFi access and the common use of a variety of different services is making this topic the more and more crucial for operators in the field.

In particular, the performance of the IEEE 802.11 MAC protocol (DCF/EDCA) was investigated by means of a novel user-centric model that was validated against simulation results. The model was used to analyze the case of heterogeneous traffic scenarios, in which elastic and streaming traffic, share the common radio resources. The interaction between the two classes of traffic was highlighted.

The parameters suggested by EDCA are designed to provide a strict traffic differentiation, which could reduce unnecessarily the performance of $BE$ traffic. To solve this situation, a simple tuning algorithm has been presented which, compared with EDCA: $i$) increases the maximum number of feasible VO flows and $ii$) improves significantly the throughput for the $BE$ active flows.
Chapter 5

TCP and VoIP Performance in Hot-Spot WLANs

5.1 Introduction

There are two different families of transport protocols, the TCP and the UDP, used for the major part of Internet applications. These protocols are independent of the physical or link layer technology, in order to be compatible between any two nodes connected to Internet, allowing them to communicate. However, how they interact with the MAC/PHY layer is a key question as the user performance is highly affected by their joint dynamics, specially for the case of the TCP protocol. Moreover, this is specially critical in the wireless channel because the transport protocols have not been designed for a channel with such error rates. Then, in order to evaluate the performance that a user will obtain from a WLAN, to be able to model this interaction between the transport layer and the random MAC protocol is of crucial importance.

5.2 TCP performance in Hot-Spot WLANs

The interaction between TCP flows and the IEEE 802.11 MAC protocol is analyzed in this section. The goal is to provide comprehensive models capable of predicting the TCP performance (throughput) in WLAN Hot-Spot networks with persistent elastic flows. Using the DCF/EDCA MAC model as a basis, several modeling strategies are analyzed to catch the main interactions between TCP and MAC protocols. Results obtained from these models are compared among themselves and against simulations, showing their accuracy and the superior simplicity with respect to previously published work, such as [38] or [41].

The models have been developed to provide a fast estimation of the TCP throughput when solving resource scheduling problems in WLANs. However, the model can be also used by hot-spot operators to estimate the TCP achievable bandwidth for different MAC parameters configurations of their Access Points (APs) or Mobile Nodes (MNs).

To evaluate the TCP performance in a Hot-Spot scenario, the ns-2 simulator [110] and a real test-bed based on a Linksys WRT54GS AP with the DD-WRT\(^1\) firmware are used. The ns-2

\(^1\)http://www.dd-wrt.org/
simulator is configured to use persistent new-RENO TCP connections and a TCP MSS of 1460 Bytes. The MAC parameters considered are shown in Table 3.2, with a data rate of 11 Mbps, a basic rate of 2 Mbps and both the AP and the MNs have a queue length of 50 packets. As a test-bed, the iperf\(^2\) tool is used, allowing us to obtain the maximum achievable bandwidth for the TCP flows. These results are compared with those obtained by using the models presented in next section. Both simulation, real and analytical results are derived considering the use of the RTS/CTS handshake. Model and simulation results are evaluated in terms of MAC layer throughput which includes the upper layers overhead, however, test-bed results provide directly the TCP throughput. To clarify the notation, all parameters related with the AP are referred with the subscript \(d\) (downlink) and with the subscript \(u\) (uplink) to the parameters of the MNs. The number of elastic flows in the downlink (uplink) will be denoted by \(n_{e,d}\) \((n_{e,u})\).

First are considered only downlink or uplink TCP flows. The throughput achieved, \(S_{tcp}^{e,d}\) and \(S_{tcp}^{e,u}\) respectively, is shown in Tables 5.1 and 5.2. The maximum TCP window size has been fixed to \(W = 1\) and \(W = 42\) (as it is commonly used in the operative TCP versions \[40\]).

Two basic system throughput tendencies can be underlined:

- The increment of the number of downlink TCP flows does not reduce the aggregated throughput since TCP reduces the channel contention \[38\] (the average number of backlogged nodes with feedback traffic is lower than the number of TCP flows).

- As the number of TCP flows in the uplink increase, the aggregate throughput increases since the TCP window of MNs reaches its maximum value (for \(W > 1\)) despite packet losses and the starvation of the downlink ACK flow \[41\].

Some pieces of work previously mentioned show that some unfairness exists among TCP flows in both uplink and downlink directions. Similar results have been observed with significant

\(^2\)http://dast.nlanr.net/Projects/Iperf/
throughput differences among nodes [41]. However, for the sake of simplicity, it is assumed that all flows share fairly the aggregate throughput, then each flow receives $S_{e,z}^{tcp}/n_{e,z}$ bps, with $z = \{d, u\}$.

### 5.2.1 Downlink TCP flows

In the downlink, TCP flows compete, through the AP, with their own feedback traffic sent by the MNs.

A single downlink TCP flow, with maximum window size equal to $W = 1$ has a throughput proportional to $L_{tcp}/RTT$, where $RTT$ is the TCP Round Trip Time. Then, the throughput is computed from

$$S_{e,d}^{tcp}(W = 1) = \frac{L_{tcp}}{1000 \times L_{TCP}} + X_d(L_{tcp}) + X_u(L_{ack}) + 2\delta$$  \hspace{1cm} (5.1)$$

where $X_d(L_{tcp})$ ($X_u(L_{ack})$) is the service time over the WLAN for a TCP (ACK) packet and $\delta$ is the signal propagation delay. Simulation ($S_{e,d}^{tcp}(W = 1) = 0.896$ Mbps) and analytical results ($S_{e,d}^{tcp}(W = 1) = 0.879$ Mbps) show a good match.

With $W = 1$ and a single TCP flow, there is no competition to access the channel between the AP and the MN because the two nodes never simultaneously have a packet ready to be transmitted at the MAC queue. From Table 5.1, as the number of simultaneous TCP flows increases, despite keeping $W = 1$, the AP queue tends to have always a packet ready to transmit, which justifies the assumption to model the access point as a saturated queue [38, 39, 76], which is obviously confirmed for values of $W > 1$. For each received packet, a MN sends the corresponding ACK (the delayed ACKs technique as in [39] has not been considered). Therefore, the number of ACK packets sent by a MN will be $1/n_{e,d}$ per packet emitted by the AP.

To analyze this situation, several approximations (models) are described in next sections.

**All sources are saturated (Model $A_d$)**

In this model, both the AP and the MNs are considered as saturated with data and ACK packets, respectively. This approximation provides pessimistic results because the level of contention suffered by data packets is very high as the AP has to compete with all the MNs.

**A time-scale decomposition (Model $B_d$)**

Introduced by [38] and used also by [39, 76], this model computes the distribution of backlogged nodes (the probability that $n_{e,d}^b$ of the $n_{e,d}$ MNs are backlogged) and the system throughput is computed averaging the throughput obtained with $n_{e,d}^b$ saturated nodes for $n_{e,d}^b = 0 \ldots n_{e,d}$. A novel variant of this model is suggested, where transitions between states are done after any successful transmission in the channel and not only after a successful transmission of the AP as in [38]. In Figure 5.1 the DTMC (Discrete-Time Markov Chain) is shown, which governs the average number of backlogged nodes as a function of the number of TCP flows, $C$. Note that the DTMC changes its state after any successful transmission over the channel, independently

---

3For this single result, a $R_{data} = 2$ Mbps and a $R_{basic} = 1$ Mbps have been considered. Moreover, the AP is connected to the TCP servers through a 100 Mbps fixed link.
Figure 5.1: Discrete Markov Chain describing the evolution of the number of backlogged nodes

on that it has been done by the AP or a MN. The probability to move from state \( j - 1 \) to state \( j \) depends on the probability that the AP transmits \((1/j)\) and the probability that the packet was sent to a non-backlogged MN \(((C - j + 1)/C)\). The probability to remain in the same state \( j \) is the probability that the AP transmits \((1/(j + 1))\) a packet, which is sent to a backlogged node \((j/C)\). Finally, the probability to move from state \( j \) to state \( j - 1 \) is the probability that a backlogged MN transmits \((j/(j + 1))\). Note that only one ACK is stored in each MN queue.

**Un-correlated ACKs (Model \( C_d \))**

Finally, the ACK arrival process at the MNs is assumed to be Poisson with rate \( \lambda_{\text{ack},u} = \lambda_{e,d}/\gamma n_{e,d} \), where \( \lambda_{e,d} \) is the maximum packet arrival rate at the AP under saturation assuming that the TCP packets are being distributed uniformly among destination nodes. This model does not consider the inherent correlation among the reception of TCP packets and the generation of ACKs; however, it provides very good results. This is motivated by the randomization caused by the backoff algorithm which mitigates this inherent correlation. It is worth noting that the delayed ACK technique can be modeled by simply dividing the value of \( \lambda_{\text{ack},u} \) by the delayed ACK factor \( \gamma \), \( \lambda_{\text{ack},u} = \lambda_{e,d}/(\gamma \cdot n_{e,d}) \).

In Figure 5.2.(a) the downlink throughput obtained by simulation is compared with the outcomes of the three models previously described. Note how model \( A_d \) clearly overestimates the negative effects of the feedback traffic and thus, the throughput obtained is lower than in simulation. Model \( B_d \) provides a very good approximation. Finally, model \( C_d \) also shows very accurate results, despite the assumption of Poisson arrivals for the ACK packets, that corresponds to assuming that there is no correlation with the reception of TCP data packets. These results allow us to validate our models in this new scenario (with TCP traffic).

### 5.2.2 Uplink TCP flows

Several applications need to transmit data from the MN (typically the client) to a node in the fixed network (the server); e.g., files sharing in P2P networks. In this case, TCP flows in the uplink compete among themselves and with the feedback traffic from the AP ACKs.

Leith et al. [41] show the existing unfairness among competing uplink TCP flows and, in order to evaluate their performance, propose an analytical model for the uplink TCP throughput.
They address an ACK prioritization scheme at the AP using the EDCA to reduce the inherent asymmetry of the WLAN, which is given by the different access opportunities to transmit TCP data packets. Their model assumes that the MNs are saturated and compute the transmission probability of the AP as the probability that the AP observes a successful transmission in the channel, i.e., $\tau_d = p_{s,d}$, with $p_{s,d}$ computed as in Chapter 4. This first model is called Model $A_u$.

In Figure 5.2.(b) the results obtained by simulation for $W = 1$ and $W = 42$ are compared with the results obtained by previous Model $A_u$ and an additional model which simply assumes that the AP is also saturated (Model $B_u$). Both models provide good accuracy.

Figure 5.2: TCP throughput (Mbps) (a) Downlink (b) Uplink

### 5.2.3 Simultaneous Downlink and Uplink TCP flows

When the AP carries simultaneous downlink / uplink flows, the performance of downlink flows is severely affected, showing values of downlink throughput near 0 when the TCP congestion window is larger than $W = 1$ (Figure 5.3.(a)). This is caused by the starvation of TCP downlink packets in the AP queue. The AP queue is shared by both ACKs and data packets while the MNs only send either TCP data packets or ACKs (in the system there are $n_{e,u}$ MNs sending TCP data packets and $n_{e,d}$ sending ACKs). The results confirm those in [40] about the different behavior of the TCP window for the uplink and downlink flows. Pilosof et al. argue that the TCP window for uplink senders reaches the maximum value, even with high ACK losses at the AP buffer, while downlink flows struggle with low window values (0 − 2 packets) caused by frequent timeouts due to data packet drops. Moreover, for $W = 1$, uplink and downlink throughput is the same. This result was also appointed in [77], where it is also shown that by setting $W = 1$ fair access to the channel can be provided to the TCP flows.

Modeling the presence of simultaneous bidirectional flows is very complex due to the interaction in the AP queue of the two types of packets: ACK and TCP downlink data packets. A simple approximation (Model $A_b$) which captures the main tendencies observed in the simula-
tions is suggested:

- The downlink queue is always saturated. The average packet length transmitted by the AP is computed from $EL_d = \phi_d L_{TCP} + \phi_u L_{ACK}$ where $\phi_d$ and $\phi_u$ are the probability that a packet sent by the AP is a data or an ACK packet.

- These probabilities are computed from: $\phi_d = 1 - \phi_u$ and $\phi_u = E\tau_{e,u} n_{e,u} / (E\tau_{e,u} n_{e,u} + \tau_{e,d})$. The $E\tau_{e,u}$ parameter is the MN expected uplink slot transmission probability (considering only the nodes with uplink TCP data packets) and $\tau_{e,d}$ is the AP slot transmission probability.

- MNs with uplink TCP data packets are always saturated.

- Nodes with uplink ACKs have $\lambda_{ack,u} = \lambda_{e,d} / n_{e,d}$, where $\lambda_{e,d} = \phi_d / X_d (EL_d)$.

At the AP queue ACKs controlled by the transmission opportunities of MNs compete with data packets of downlink flows. As the number of uplink flows increases, since MNs have more transmission opportunities than the AP, the number of ACKs in the AP increases. Thus, the downlink flows suffer for both contending for buffer space in the AP, and for contending on the access to the channel, which causes that downlink TCP flows tend to starve. To solve this limitation, a properly MAC parameter tuning can be done, for example increasing the $CW_{min}$ parameter of uplink TCP flows, which reduces their slot transmission probability $\tau_{e,u}$.

Results for different values of $CW_{min}$ are also plotted in Figure 5.3. These results suggest that a cross-layer solution between transport and MAC layer could improve the overall performance.

![Figure 5.3: Throughput comparison with simultaneous TCP downlink/uplink flows (Mbps)](image-url)
Voice over IP (VoIP) has raised an interest in recent years due to its usefulness as an easy and cheap alternative to traditional telephony in many environments. In spite of this and all the research on this area, there are still unsolved issues concerning the quality of VoIP calls, especially over wireless networks, where the specific network characteristics cause more jitter, delay and packet losses than would be acceptable for widespread voice transmission.

Voice communication using WLAN technology as access network could be a promising alternative to traditional cellular networks (2G, 3G). Currently, roaming problems between WLAN coverage areas have to be solved to provide a continuous service to the user. However, novel proposals to interconnect and manage WLAN cells using common fixed infrastructure operators are yet a promising reality [112]. Moreover, three major technological issues of the IEEE 802.11 MAC protocol itself have to be solved or improved to achieve an efficient use of the transmission resources.

1. High protocol overheads.
2. Unfairness between uplink and downlink streams.
3. Fast VoIP degradation in presence of TCP flows.

A criterium to determine the maximum number of VoIP calls that can be transported by a network (also called VoIP capacity) given the desired voice quality in terms of bandwidth, delay, losses, can be found in [79]. For a good quality, the average delay have to be less than 150 ms with losses less than 3%. A medium quality is achieved with delays between 150 and 400 ms and packet losses less than 7%. Finally, a poor voice quality corresponds to delays higher than 400 ms and losses higher than 7%. Considering that the WLAN is only one hop of the whole path between the two end points, for a conservative good design, the quality target should be set to at least one-third of the maximum recommended values.

In Table 5.3 the basic characteristics of the most frequently used voice codecs for VoIP are summarized. The average throughput is plotted in Figure 5.4 for the G.711 and G.729 voice codecs. The AP is the bottleneck of the system, which limits the VoIP capacity.

<table>
<thead>
<tr>
<th>Codec</th>
<th>G.711</th>
<th>G.723.1</th>
<th>G.726-32</th>
<th>G.729</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bit Rate (Kbps)</td>
<td>64</td>
<td>5.3/6.3</td>
<td>32</td>
<td>8</td>
</tr>
<tr>
<td>Framing Interval (ms)</td>
<td>20</td>
<td>30</td>
<td>20</td>
<td>2x10</td>
</tr>
<tr>
<td>Payload (Bytes)</td>
<td>160</td>
<td>20/24</td>
<td>80</td>
<td>10</td>
</tr>
<tr>
<td>Packets/sec</td>
<td>50</td>
<td>33</td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>

Table 5.3: Typical values of most used codecs in VoIP

5.3.1 Basic performance impairments

Protocol Overheads

Taking into account the parameters defined by the IEEE 802.11b standard [1], summarized in Table 3.2, the maximum number of voice calls without contention can be computed, i.e., the
Figure 5.4: VoIP - Throughput (AP and mobile nodes) for two voice codecs (G.711 and G.729)

<table>
<thead>
<tr>
<th>Codec</th>
<th>G.711</th>
<th>G.723.1</th>
<th>G.726-32</th>
<th>G.729</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max number of calls: $C_{voip}$ (no contention)</td>
<td>4</td>
<td>9/9</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Efficiency (no contention: $\eta = C_{voip} \cdot B_{voice}/R_{DATA}$)</td>
<td>12.8%</td>
<td>2.37%/2.85%</td>
<td>8%</td>
<td>2.4%</td>
</tr>
<tr>
<td>Max number of calls (contention)</td>
<td>4</td>
<td>7/7</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 5.4: VoIP efficiency over WLAN

channel is ideally shared among voice calls, and with contention, see results in Table 5.4. From the results, can be conclude that the contention to access the channel reduces the maximum number of calls but it is not the main limiting factor. It is clear that the main problem is the large overhead introduced by the higher layer protocols. A technique to solve this situation is header compression such as ROHC (RObust Header Compression, RFC 3243).

Unfairness

As the access point carries the same data as the whole set of the mobile nodes, it has to attempt to transmit $n$ times more than each mobile node. Therefore, it is desirable that the transmission attempts of the AP are $n$ times greater than the transmission attempts of a single mobile node, or equivalently $\tau_d = n\tau_u$, to achieve a fair access to the channel (each node access to the channel proportionally to the traffic volume it has to send).

As a measure of the system fairness, can be compared:

$$w_d = \frac{\tau_d}{n_u\tau_u + \tau_d} \quad w_u = \frac{n_u\tau_u}{n_u\tau_u + \tau_d}$$

(5.2)

Considering the G.729 voice codec, the first two columns of Table 5.5 show $w_d$ and $w_u$ versus the number of mobile nodes. As the number of voice calls (or mobile nodes) increases, the AP
Table 5.5: Bandwidth share between uplink and downlink VoIP flows

<table>
<thead>
<tr>
<th>Voice calls (G.729)</th>
<th>AP ($w_d$)</th>
<th>Mobile Nodes ($w_u$)</th>
<th>AP ($w_d$)</th>
<th>Mobile Nodes ($w_u$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
</tr>
<tr>
<td>2</td>
<td>0.482</td>
<td>0.517</td>
<td>0.501</td>
<td>0.498</td>
</tr>
<tr>
<td>3</td>
<td>0.466</td>
<td>0.533</td>
<td>0.504</td>
<td>0.495</td>
</tr>
<tr>
<td>4</td>
<td>0.450</td>
<td>0.549</td>
<td>0.506</td>
<td>0.493</td>
</tr>
<tr>
<td>5</td>
<td>0.428</td>
<td>0.571</td>
<td>0.510</td>
<td>0.489</td>
</tr>
<tr>
<td>6</td>
<td>0.352</td>
<td>0.647</td>
<td>0.514</td>
<td>0.485</td>
</tr>
<tr>
<td>7</td>
<td>0.217</td>
<td>0.782</td>
<td>0.515</td>
<td>0.484</td>
</tr>
<tr>
<td>8</td>
<td>0.210</td>
<td>0.789</td>
<td>0.516</td>
<td>0.483</td>
</tr>
</tbody>
</table>

unfairness grows, resulting in a fast saturation of the AP queue that limits the system capacity. To solve this problem, a simple solution consists in updating the $CW_{\min}$ value of mobile nodes each time a new call arrives at the system. By computing the value of $CW_{\min}^*$ for each mobile node, assuming $m = 0$, it is obtained

$$CW_{\min,u}^* = \frac{n_s \rho_u (CW_{\min,d} + 1)}{\rho_d} - 1$$  \hspace{1cm} (5.3)

Notice, how the relation between both $CW_{\min}$ is given by the relation between the queue utilization of MNs and APs,

$$\frac{CW_{\min,u}^* + 1}{CW_{\min,d} + 1} = n_s \frac{\rho_u}{\rho_d}$$  \hspace{1cm} (5.4)

The fairness obtained by applying this solution is showed in the two last columns of table 5.5. Note that the AP gets equal or more transmission opportunities than the uplink mobile nodes.

This solution can also be combined with the one presented in [34], where the TXOP mechanism of EDCF [2] is used to provide fairness. However, in both cases a limited gain is obtained in terms of capacity increment. Finally, another solution, similar to the use of TXOP, is presented in [35]. In this case, several voice packets are encapsulated in only one multicast packet which is sent to all mobile nodes, where each one gets its own data.

**Interaction with TCP flows**

As the AP queue is shared by all downlink streams, the VoIP packets have to compete for the buffer space with all the other flows, that can be streaming (UDP) or elastic (TCP) flows (both data and ACKs destined to a mobile node). TCP downlink traffic tends to saturate the MAC queue [38] which cause high losses to the VoIP flows. Moreover, any uplink flows reduce, proportionally to its bandwidth, the transmission opportunities gained by the AP. The negative effect over the downlink traffic is increased if these uplink flows are TCP-based which also tend to saturate the mobile nodes queue [40].

The negative influence of the TCP traffic is clearly shown in Table 5.6. Note the fast degradation of the VoIP throughput with TCP downlink flows and the inoperability of any VoIP call with just a single TCP uplink flow. It is also interesting to observe that, when the AP queue is saturated with VoIP traffic, the interaction with TCP traffic is reduced due to the
starvation of TCP flows. Therefore, the presence of TCP traffic in both the downlink (buffer losses) and the uplink (AP starvation) leads to low performance of VoIP calls. These problems have to be solved in order to deploy a successful VoIP service over WLAN.

In the downlink, a simple classification/prioritization scheme can be used (using the different Access Categories (AC) defined in the EDCA [2] or, for example using the dual queue proposed in [80]) where the TCP and UDP packets occupy separated buffers with priority for the VoIP packets. However, the main problem is with the TCP uplink flows because each node acts independently from the others. The only possible solution is setting different MAC parameters (such as \( AIFS [113], \text{CW}_{\min} [9], \text{TXOP} [8] \)) to each mobile node in order to reduce the interaction of these TCP flows on the VoIP calls. Another possibility is to apply some ACK policies such as No ACK or Block ACK [53].

**Multi-rate anomaly**

Multi-rate, due to link adaptation, IEEE 802.11 environments are quite problematic for VoIP traffic [5], with the rate changes of some of the flows affecting the transmission of all others, which causes an unacceptable voice quality degradation. In Chapter 6 this problem is analyzed more deeply, providing a feasible solution.

### 5.3.2 VoIP capacity using the Direct Link Protocol (DLP)

One of the novel features of the IEEE 802.11e is the Direct Link Protocol (DLP), which allows direct STA-to-STA transfer without relying on the Access Point (AP). Main objective is to alleviate the traffic load carried by the AP, that is the main limitation on the overall WLAN performance and the most critical to deploy VoIP services (see Figure 5.4).

DLP (Direct Link Protocol) allows Mobile Node (MN) to Mobile Node direct communication among stations with direct visibility and placed in the same Access Point (AP) coverage area. Using the DLP, the traffic generated between two end points does not require to use the AP to rely, resulting in a better use of the transmission resources. Therefore, DLP is an interesting solution for VoIP calls with origin and destination in the same coverage area. The capacity of a WLAN hotspot offering VoIP services for both intra-cell and out-cell calls is evaluated, with the goal to compare the capacity gains obtained by using the DLP protocol for intra-cell calls. Results show the significant improvements that can be achieved in spite of lower data rates used for direct link connections.

With a single Access Point (AP) a coverage area up to few hundred meters could be achieved.
Several scenarios, such as factory departments, require frequent phone calls among people in the same floor but in different offices. An already existent solution, is to use wireless VoIP phones which share the same infrastructure than the traditional wireless access to Internet for web-browsing, e-mail or file transfer. Therefore, it is a simple and cheap solution. However, the maximum number of calls that an AP can carry is lower than could be expected from the nominal data rate, specially in the case that the two call end-points are in the same coverage area. In this situation, a possible solution to improve the WLAN capacity, is to allow a direct link between the two call end-points, if direct communication is possible.

To enable the direct link between MN, a setup protocol is followed. First, one of the MNs sends a DLP request to the AP, which relays it to the other MN. If this MN accepts the direct link, it responds with a successful DLP response, which is relayed to the first MN. Once the direct link is initiated, both MN exchange information such as the rate set to be used. The other EDCA features, as TXOP limit, AIFS or different BEB parameters remain as defined for each Access Category (AC). In Fig. 5.6 the considered scenario is depicted. The voice codec used is the G.711, which sends packets of length 200 Bytes (including the RTP/UDP and IP headers) each 20 ms.

In infrastructure WLANs the AP has to carry all data from Internet directed to MNs. Thus, for higher traffic load, the AP will not be able to transmit all the required data, becoming the network bottleneck. For TCP-based traffic flows, the congestion is mitigated by the TCP flow control mechanisms, but for UDP flows, such as VoIP traffic, congestion (saturation) at the AP will cause losses and/or an increment of the transfer delay, resulting in an unacceptable performance for voice.

For VoIP calls where one end-point is placed outside the cell, the AP has to carry n VoIP
downlink flows, where \( n \) is the number of calls and of MNs with an active call and which the AP has to fight against to be able to transmit a packet to the channel. A more critical situation is found when the two call end-points are located in the same cell. In this case, the AP carries \( 2 \cdot n \) VoIP downlink flows and has to compete to access the channel with also \( 2 \cdot n \) MNs. Therefore, for intra-cell calls, the AP carries more traffic and has to compete with more MNs. A clear example is shown in Table 5.7 where the capacity, in terms of calls, is plotted for out-cell calls and intra-cell calls. Notice how the maximum number of active calls is reduced to half.

To alleviate this situation, the utilization of the DLP is considered for intra-cell calls. In Table 5.8, the maximum number of VoIP calls using the DLP is shown for different data rates (assuming that all DLP MN use the same data rate). Notice that, in spite of possible rate degradations due to the distance among MN using a direct Link, which imply a lower data rate, the capacity is already higher unless the low rate of 1 Mbps is used. From these results, it seems necessary to introduce a decision entity which uses the cell state information to allow or deny the usage of direct communication for a given VoIP call.

\[
\begin{array}{c|c|c}
\text{DATA RATE} & \text{Number of in-cell VoIP calls} \\
11 & 9 \\
5.5 & 8 \\
2 & 6 \\
1 & 3 \\
\end{array}
\]

Table 5.8: Number of DLP VoIP calls at different rates

In Figure 5.6 the queue utilization of the AP (Sub-Figure 5.6.(a)) and of the MNs (Sub-Figure 5.6.(b)) using DLP are respectively shown. Notice how the AP saturates rapidly when the number of VoIP calls increases, especially with the out-cell calls, as they increase the load that the AP has to manage. Nevertheless, the system is also vulnerable (albeit to a lesser degree) to the increment of intra-cell calls, which increase the channel contention. Notice that MNs with intra-cell (DLP) calls, in spite of the AP being already saturated, are able to continue the call correctly.

In spite of the advantages of direct intra-cell communications at the MAC layer, an intrinsic problem exists with SIP-based calls in 802.11e scenarios. Nowadays, most VoIP communications take place through commercial VoIP operators, which tend to route all signaling messages through a number of centralized servers for authentication and accounting purposes. Hence, even if a call is, from the user perspective, intra-cell, it frequently has to be routed through the Internet, rendering MAC layer intra-cell direct communication impossible. This problem may even arise for the bulk of data packets, e.g. if NAT-traversal mechanisms are also present. In order to take full advantage of the 802.11e possibilities, the AP should implement a cross-layer proxy role. Whenever the MAC setup protocol is started, the AP should identify the service being requested (a VoIP call in this case) and decline participating in the signaling or data path. This is a complex mechanism, which might only be adequate in serverless architectures, like e.g.
the emerging P2P versions of SIP.

5.4 Conclusions

A study of the TCP protocol dynamics in WLANs has been presented. Using the provided knowledge, different modeling techniques has been compared, resulting in a set of different TCPoWLAN models, which can be instrumental to the performance evaluation of admission control schemes, and the radio resource dimensioning and management strategies for WLAN hot-spots.

The VoIP capacity in WLANs using the both infrastructure and ad-hoc (Direct Link Protocol (DLP)) configurations is evaluated. The VoIP performance problems have been pointed out, explaining the possible solutions or at least how to mitigate them. Moreover, results show the convenience of using the DLP for in-cell calls which could imply a significant gain in terms of number of VoIP calls, in spite of lower DATA rates for STA-to-STA direct data transfer. However, the use of DLP has several signalling drawbacks, which can be taken in consideration when this protocol option is enabled.
Chapter 6

Admission Control in WLANs with heterogeneous traffic

6.1 Introduction

In Chapter 1 two situations were described: \( i \) heterogeneous traffic flows and \( ii \) the multi-rate anomaly due the link-adaptation process. Here, it is shown how the proposed admission control entity provides a feasible solution for them. However, these two situations are only simple examples and the admission control presented can be considered of general application.

The Admission Control (CAC) in WLANs is an entity which uses a set of rules (policy) to manage the assignation of transmission resources to both AP and MNs. The CAC process is activated at any time instants where the system state changes, at arrival / departure time of flows, when the channel conditions or/and any of the flow characteristics change. For example, in multi-rate wireless networks, when a flow changes its data rate due to the use of a link adaptation algorithm, the overall distribution of network resources changes (the network capacity) and the admission control function must be reinvoked in order to manage the new situation.

In this chapter, a detailed Admission Control EDCA-based mechanism is presented, which includes the MAC parameter tuning algorithm described in previous Chapter 4. The considered admission control is model-based as it uses the EDCA model previously presented to compute all system performance metrics. Notice that the model includes the MAC QoS enhancements such as the transmission opportunity (TXOP), the use of different inter-frame spaces (AIFS) and different values of the back-off instance (\( CW_{\text{min}}, CW_{\text{max}} \)), allowing the admission control entity to decide if the new flow can be accepted and to select the suitable MAC parameters. Therefore, the Admission Control entity uses an adaptive MAC parameter tuning algorithm which decides the best instantaneous parameter configuration in order to: \( i \) guarantee the QoS requirements for rigid flows (such as VoIP) and \( ii \) maximize the elastic flow throughput (web browsing and P2P transfers). Results, compared with the standard EDCA and DCF, show clearly how our proposal overdue them, providing a better service performance to rigid traffic flows with a controlled balance for the elastic ones.

The chapter is structured in two main topics. In the first one, the coexistence of VoIP calls with TCP flows is evaluated by means of analytical and simulation techniques. Although
simulation results are more detailed than the analytical ones (due to the assumptions considered for the analytical models), both are coherent and show how the proposed mechanism overcomes the current version of the EDCA, being a flexible and generalist scheme that can achieve near optimal results. Second part focuses in the possible problems derived from the use of a link adaptation scheme with multiple rates for VoIP services. The possible capacity reduction is discussed and several adaptive schemes, distributed and/or centralized, are proposed to mitigate this problem. Clearly, the adaptive schemes improve the system capacity (in terms of calls) and reduces the blocking / dropping probability at the cost of using VoIP codecs with a lower rate and therefore, a possible lower MOS. However, a properly selection of codec by analyzing delay and loss metrics allow to maximize both the average cell MOS and the cell capacity.

6.2 The Admission Control algorithm

In this section, a detailed EDCA model-based admission control is presented. It is based on the specifications from the IEEE 802.11e standard [2] and includes the MAC parameter tuning algorithm presented in Chapter 4. The provided scheme defines the action to be taken by the admission controller and the MNs when a new flow starts / departs or the WLAN state changes. The Call Admission Control entity is assumed to be located at the AP.

In Figure 6.1 the algorithm followed by the AP is depicted. The algorithm operates as follows:
when a new ADDTS request arrives to the admission control, the system searches for a successful combination of MAC parameters, which satisfy the QoS requirements of both the new arriving flow and all the already active. Then, if a MAC parameter combination exists, an ADDTS response with status SUCCESS is sent to the requesting MN and the new MAC parameters are announced into the next beacon frame. Otherwise, the CAC sends an ADDTS response with status SUGGEST and the suitable TSPEC (if exists) or refuses the request sending a RETRY response, with the field TSDelayElement equal to 0, suggesting to the MN to compute the delay before to retransmit another ADDTS request. Notice that the CAC consider the case where up to $D$ BE/BK flows can be dropped by sending to the corresponding MN an ADDTS DROP packet.

If a DELTS request arrives, the CAC deletes the corresponding TSPEC and executes the parameter tuning algorithm in order to properly select the new MAC parameters (notice that probably, this MAC combination is already stored and so, no extra computation is required).

When a new flow starts in a MN, a new TSPEC is built and it is sent to the AP into an ADDTS request. Then, the MN waits for the ADDTS response, which can be SUCCESS, SUGGEST or RETRY. If a SUCCESS response is received, the MN starts the data transmission until the flow finishes or it is dropped by the AP. If a SUGGEST response is received, the MN evaluates the suggested TSPEC and decides whether to accept it or not. In case of accepting the suggested TSPEC, a new ADDTS request containing the suggested TSPEC is sent to the AP. Otherwise, the MN schedules the retransmission of the same ADDTS after a random delay. The same procedure is followed if a RETRY response is received. The details are shown in Figure 6.2.

![Flow level QoS guarantees in IEEE 802.11e-EDCA based WLANs December 2006](image)

Figure 6.2: Call Admission Control in the MN
6.3 Admission Control for VoIP with heterogeneous traffic

The simultaneous coexistence of VoIP calls with TCP flows is complex. Basically, the presence of TCP traffic in both the downlink (buffer losses) and the uplink (AP starvation) leads to low performance for VoIP calls.

These problems have to be solved in order to deploy a successful VoIP service over WLANs. In the downlink, a simple classification/prioritization scheme can be used (four Access Categories (AC) are defined in the EDCA [2]) where the TCP and VoIP packets can occupy separated buffers with priority to the VoIP packets. However, main problems raise in presence of uplink TCP flows because nodes act independently from each others. The only possible solution is setting different MAC parameters (such as $AIFS$ [113], $CW_{min}$ [9] or/and $TXOP$ [8, 33]) at each mobile node in order to reduce the interaction of these TCP flows with the VoIP calls.

It has been pointed out that the use of EDCA is more suitable for WLAN networks where rigid (e.g., VoIP) and elastic (BE) flows coexist. However, in order to avoid the system saturates, thus causing an overall WLAN degradation, admission control is needed to block new flows, or drop already active ones, when the system cannot allocate enough resources for the new flows.

One basic desirable feature of an admission control mechanism is the ability to adapt the rate of elastic flows in order to release resources which can be allocated to a new arriving rigid flow (and the opposite procedure when a rigid flow departs). In EDCA-based WLANs this rate control can be done by setting different MAC parameters for each AC or AC/MN. Currently, the adequate parameter selection still requires further research as there are no general tuning criteria. For example, a simple parameter tuning algorithm can be found in [11].

First, detailed simulation results are presented. In this case, the admission control uses the $EDCA - TA$ MAC parameter tuning algorithm presented in Chapter 4. Later, the scenario is simplified in order to model it analytically. In that case, a more simple algorithm based only on tuning the $CW_{min}$ parameter for TCP flows is considered. However, as expected, similar conclusions are obtained in both cases. Moreover, in the second case, a preemptive policy is also considered.

6.3.1 Flow level metrics of VoIP calls coexisting with P2P and Web elastic flows

A simulator of the scenario presented in Chapter 3 has been developed using the COST simulation toolkit [111]. The simulator is based on two modules: i) an admission control module that receives and processes the admission requests and ii) the MN module, which implement the previous described admission control algorithm.

The goal is to analyze the performance of the proposed Admission Control algorithm (EDCA-TA) and compare its performance with the obtained using the DCF and EDCA. For the sake of simplicity, only the BE (P2P or Web) and VO (voice) AC are considered to be active and each MN has only a single flow active. Moreover, the AP suggesting TSPEC capabilities are disabled and so, always that a new flow has no free resources to satisfy its requirements, it will be blocked.
Traffic profiles and parameters

Three types of traffic flows are considered: HTTP, P2P and VoIP. The traffic characteristics of each type are shown in Table 6.1. Furthermore, to test the Admission Control algorithm two scenarios are defined (see Figure 6.3):

1. The number of nodes with VoIP (bi-directional rigid flows) and downlink P2P and Web flows are fixed to 4 and 2 respectively, while the number of nodes with P2P uplink flows is increased from 0 to 15. As has been shown, the uplink elastic flows cause the worst performance degradation to rigid flows (specially to the downlink ones). Therefore, it is showed how the EDCA and the EDCA-TA are able to solve the rapid performance degradation of the rigid flows.

2. In the second scenario the number of VoIP flows are increased from 0 to 25. The number of downlink/uplink P2P and downlink Web flows is maintained constant and equal to 5 and 2 respectively. Here, the goal is to observe if the system is able to allocate enough resources for the new rigid flows and how to manage the bandwidth reduction for the elastic ones.

In both scenarios, all the flow arrival rates follow a Poisson process. The VoIP flow length is exponentially distributed. Both P2P and Web flow lengths depend on the bandwidth assigned, which changes dynamically with the time (i.e., the elastic bandwidth is function of the network state). The considered MAC parameters are shown in Tables 3.2 and 3.3. However, a single
Figure 6.4: Scenario 1 (a) VoIP Blocking Probability (b) Elastic Blocking Probability (c) VoIP Throughput (d) Elastic Throughput

Performance results

In Figure 6.4 the results obtained in Scenario 1 are shown. Notice that the blocking probability of VoIP flows using the EDCA-TA is lower than the one obtained using the EDCA. Obviously, in both cases, a clear gain is obtained compared with the DCF. To assess this improvement EDCA-TA reduces the instantaneous elastic throughput by adjusting the affordable bandwidth of BE flows by means of tuning the AC_BE and AC_VO MAC parameters. However, the average BE throughput is also higher than the obtained by using both EDCA and DCF in the uplink. Moreover, it is also not significantly reduced in the downlink, in spite of the inherent unfairness between uplink / downlink traffic streams (the uplink flows have more transmission opportunities than the downlink ones as all of them are carried by the AP). Moreover, the prioritization of VoIP flows using EDCA and EDCA-TA is compensated with a higher blocking probability for BE flows, higher than the obtained by using the DCF.

Figure 6.5 shows the results obtained in Scenario 2. Notice how using the DCF the system
Figure 6.5: Scenario 2 (a) VoIP Blocking Probability (b) Elastic Blocking Probability (c) VoIP Throughput (d) Elastic Throughput

is unable to accept any VoIP call as it is already saturated by BE flows. However, using both EDCA and EDCA-TA, a clear VoIP blocking probability gain is obtained, which is translated into a higher throughput for VoIP calls. As expected, using EDCA-TA the VoIP performance is clearly improved due to the higher flexibility to adapt the BE instantaneous rate, as was mentioned before.

With regards to the elastic flows, notice that the average throughput reduction due to the use of EDCA and EDCA-TA (for values of $n_{voip} > 7$) with respect to DCF. Before that point, EDCA-TA improves the average BE throughput with respect to the other access schemes. Thus, there is an important gain in terms of blocking probability and throughput for both VoIP and BE flows. For values of $n_{voip} > 20$, EDCA-TA shows lower BE throughput as this is the cost to achieve a lower VoIP blocking probability.

It is also important to note the higher blocking probability for BE flows when the EDCA parameters are used, higher than the ones obtained by using DCF and the EDCA-TA.

Results clearly show that the overall cell performance is increased and a better use of the transmission resources is obtained as: i) the requirements of sensitive traffic are significantly improved and ii) the average best-effort throughput is maximized in a wide range of possible situations.
6.3.2 VoIP call admission control in WLANs in presence of elastic traffic

An analytical study, similar to the previous one, has been carried out. Now, a single class of elastic flows is considered but the other assumptions are similar. However, the MAC parameter tuning algorithm is simpler than the previous one, as only changes the $CW_{min}$ parameter of MNs with an elastic flow active, increasing or decreasing their probability to transmit. Both studies are complementary but it is remarkable that similar results are obtained.

To differentiate TCP and VoIP downstream flows, the AP uses two different AC queues: voice queue (AC_VO) for VoIP packets and best-effort queue (AC_BE) for TCP packets. The service prioritization for the voice queue is given only by a short $AIFS$ as has been considered that the $CW_{min}$ and the $TXOP$ burst parameters are equal for both queues with values 32 and 1 respectively. It is referred with $\rho_{s,d}$ to the AC_VO queue utilization and with $\rho_{e,d}$ to the AC_BE queue utilization. To simplify the analysis, it is assumed that TCP packets are served only when the AC_VO queue is empty, then the probability to transmit downstream TCP packets is $1 - \rho_{s,d}$. Note that, since the upstream feedback traffic is proportional to the downstream TCP traffic, a minimal impact of uplink TCP ACKs over the VoIP packets is assumed.

The VoIP and TCP upstream flows use respectively the AC_VO and AC_BE queues of mobile nodes. They use bursts of size $B = 1$, different and fixed values of $AIFS$ and different values of the $CW_{min}$ parameter (variable for TCP uplink flows and determined dynamically by the admission control entity). Let $\Phi_{CW}$ be the set of all possible $CW_{min}$ values that can be used by uplink BE flows, with $\Phi_{CW} = \{32, 64, 128, 256, 1024\}$. When the CAC receives a new request of a VoIP call or a TCP uplink flow, it computes the suitable $CW_{min}$ value for the new and remaining active uplink BE flows and broadcasts the new $CW_{min}$ value in next beacon frame. If there are no VoIP flows in the system, it is assumed that all nodes and the AP use the default value of $CW_{min} = 32$. The considered MAC parameters are shown in Tables 3.2 and 3.3. However, a single data rate of 2 Mbps and a single basic rate of 1 Mbps are considered and both the AP and the MNs have a queue length of 20 packets.

Policies

Four policies are considered:

1. AC-S-NP (static, non-preemptive), according to which all flows always use a fixed value of $CW_{min}$ equal to 32.
2. AC-S-P (static, preemptive), according to which all flows always use a fixed value of $CW_{min}$ equal to 32. However, elastic flows can be dropped in order to allocate their resources to the new arriving VoIP call.
3. AC-A-NP (adaptive, non-preemptive), according to which all flows always use the previous described $CW_{min}$-based algorithm.
4. AC-A-P (adaptive, preemptive), according to which all flows always use the previous described $CW_{min}$-based algorithm. However, elastic flows can be dropped in order to allocate their resources to the new arriving VoIP call.
Flow level QoS guarantees in IEEE 802.11e-EDCA based WLANs

December 2006

It is expected that the AC-A provides better performance than the AC-S as the number of coexisting VoIP calls and TCP flows grows.

Modeling a cell

Under the assumption of exponential distributions of flow arrivals and departures, the system can be described by a three-dimensional Continuous Time Markov Chain (CTMC). In order to solve the model, the CTMC can be broken into two bi-dimensional CTMCs. The first CTMC (CTMC_A) comprises the situation where the VoIP calls compete with uplink TCP flows and second CTMC (CTMC_B) the situation where downlink TCP flows compete with uplink TCP flows.

The partial results of both CTMC are averaged using the approximation that with probability \( \rho_{s,d} \) the system works in the situation described by CTMC_A and with probability \( 1 - \rho_{s,d} \) the system behavior can be modeled by CTMC_B. Notice that \( \rho_{s,d} \) is the AP AC_VO queue utilization.

While the number of BE flows can grow to infinity, the maximum number of streaming flows is limited by the bandwidth requirements of the voice calls to \( N_{th}^{voip} \). In order to solve these infinite bi-dimensional CTMCs, the state space must be truncated. Without loss of generality, a realistic minimum bandwidth \( B_{min}^{e} \) required for a BE flow is introduced, which gives a maximum number of \( N_{th}^{e,u} (N_{th}^{e,d}) \) uplink (downlink) BE flows. The CTMCs state space is described by

\[
S_A = \{ (n_{e,u}, n_s) | S_{e,u}^{tcp}(n_{e,u}, n_s)/n_{e,u} \geq B_{min}^{e}, S_{e,d}^{tcp}(n_{e,u}, n_s) > 0.97 \cdot n_s B_e \} \\
S_B = \{ (n_{e,u}, n_{e,d}) | S_{e,u}^{tcp}(n_{e,u}, n_{e,d})/n_{e,u} \geq B_{min}^{e}, S_{e,d}^{tcp}(n_{e,u}, n_{e,d})/n_{e,d} \geq B_{min}^{e} \}
\]

(6.1)

Notice that if the preemptive policy (P) is considered transitions between state \( S_A(n_{e,u}, n_s) \) and \( S_A(n_{e,u} - d_{e,u}, n_s + 1) \) exist, where \( d_{e,u} \) is the number of dropped uplink TCP flows. For the detailed procedures to compute blocking and dropping probabilities in multi-service wireless networks, refer to [84].

For both voice and best-effort flows the user population is considered to be infinite with steady state arrival rates \( \lambda_{e,u} \) and \( \lambda_{e,d} \) for BE flows and \( \lambda_s \) for VoIP calls. The BE flow duration is a function of the bandwidth observed by the BE flows and the flow length (amount of data to be transmitted) \( FL_e \), with departure rate equal to \( \mu_{e,x} = S_{e,x}^{tcp}(\cdot)/(n_{e,x} FL_e) \), while VoIP calls have a fixed average duration equal to \( 1/\mu_s \). An example of both the CTMCs is depicted in Figure 6.6. It can be observed how by properly tuning the \( CW_{min,BE} \) parameter the number of feasible states grows. The traffic flow models of Chapter 5 had been used for the VoIP sources and TCP flows.

In all cases, the flow arrival and departure rates follow a Poisson process, the flow length has an exponential distribution and, for the sake of simplicity, \( B_{min}^{e} \) is computed to allow a maximum number \( N_{th}^{e,u} = N_{th}^{e,d} = 10 \) of active elastic flows in the system in all situations.

Non-preemptive policies

Results are shown as the comparison of the performance achieved by the WLAN using a non-adaptive CAC scheme called simple CAC, according to which all flows always use a fixed value of \( CW_{min} \) equal to 32, and using the proposed CAC scheme, called adaptive CAC. Table 6.2
Table 6.2: Parameters for CAC evaluation

<table>
<thead>
<tr>
<th>Scenario 1</th>
<th>Scenario 2</th>
<th>Scenario 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_s$ (flows/second)</td>
<td>variable</td>
<td>0.0083</td>
</tr>
<tr>
<td>$1/\mu_s$ (seconds/flow)</td>
<td>240</td>
<td>240</td>
</tr>
<tr>
<td>$\lambda_{e,d}$ (flows/second)</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>$\lambda_{e,u}$ (flows/second)</td>
<td>0.5</td>
<td>variable</td>
</tr>
<tr>
<td>$FL_e$ (Mbits)</td>
<td>1, 2</td>
<td>1, 2</td>
</tr>
</tbody>
</table>

The results obtained using the adaptive CAC can be better understood if the following aspects are considered: i) the number of feasible states of the CTMC grows (i.e., the number of coexisting uplink TCP flows and VoIP calls grows), ii) the adaptive CAC reduces the instantaneous throughput of TCP uplink flows, increasing their latency, that is the time they are active in the system, and iii) the adaptive CAC is not preemptive and does not block already existing flows, it only reduces the rate of uplink TCP flows until the $CW_{min} = 1024$ value is reached.

In Figures 6.7 the effect of increasing the VoIP call arrival rate is investigated, in terms of blocking probability for voice calls, uplink and downlink TCP flows. Using the adaptive CAC scheme the blocking probability for VoIP calls is significantly lower for a reasonable range values of $\lambda_s$. The reason for the higher blocking probability of the adaptive CAC for greater values of $\lambda_s$ is motivated by the no-preemption characteristic as the system tends to be always occupied with uplink TCP flows when a new call request arrives. For the uplink TCP flows, the proposed scheme always show better results. The gain in both VoIP calls and uplink TCP flows is compensated by a higher blocking probability of TCP downlink flows.

In Figure 6.8 the arrival rate of VoIP calls is kept constant to test the sensibility of the novel scheme to variations of the traffic intensity of the TCP uplink flows. For both CAC schemes the blocking probability increases as the traffic intensity of TCP uplink flows increases. However,
under the *adaptive CAC* the blocking probability exhibits substantially lower values, which tend to a constant value equal to 0.1, which shows the desirable insensibility property. For the uplink TCP flows, when $\lambda_{e,u} > 0.75 \text{ flows/s}$, the blocking probability for uplink TCP flows is larger than for the *simple CAC*.

Finally, in Figure 6.9, the impact of the elastic flow length $FL_e$ is evaluated. As expected, as the elastic flow length increases, the blocking probability of VoIP calls also increases. The blocking probability for uplink TCP flows is more sensitive if the *adaptive CAC* is used, showing higher values than the *simple CAC* for large $FL_e$. This result is also motivated by the rate reduction (reduction of the departure rates of the $CTMC_A$ for uplink TCP flows) which increases the normalized traffic load of TCP flows.

In all situations, the higher number of accepted VoIP calls using the *adaptive CAC* results in an increment of the VoIP throughput and then, a higher utilization ($\rho_{s,d}$) of the downlink $AC\_VO$ queue, decreasing the transmission opportunities of the TCP downlink traffic.

**Preemptive policies**

Results are shown as the comparison of the performance achieved by the WLAN using a non adaptive CAC scheme called *AC-S* (static), according to which all flows always use a fixed value of $CW_{\min}$ equal to 32, and using the proposed CAC scheme, called *AC-A* (adaptive). Moreover, two preemption policies are considered for both schemes: Non-Preemption (*NP*) and Preemption (*P*), where uplink TCP flows can be dropped to allocate their bandwidth resources to new arriving VoIP flows. Two scenarios are considered:

- **Scenario A:** *Arrival rate increment of VoIP calls.* VoIP calls arrive to the system with rate $\lambda_s$ and have an average duration equal to $1/\mu_s = 240 \text{ seconds}$. Both downlink and uplink TCP flows have flow length equal to $FL_e = 1$ or $FL_e = 2 \text{ Mbits}$, and arrive to the
Figure 6.8: Non-preemptive policies (a) \( BP_s \) - Blocking probability VoIP calls, (b) \( BP_{e,u} \) - Blocking probability for elastic uplink flows

Figure 6.9: Non-preemptive policies (a) \( BP_s \) - Blocking probability VoIP calls, (b) \( BP_{e,u} \) - Blocking probability for elastic uplink flows
system with $\lambda_{e,d} = 0.5$ and $\lambda_{e,u} = 0.5$ flows / second, respectively.

- Scenario B: Arrival rate increment of uplink TCP flows. VoIP calls arrive to the system with rate $\lambda_s = 54$ calls / hour and have duration equal to $1/\mu_s = 240$ seconds. Both downlink and uplink TCP flows have flow length equal to $FL_e = 1$ or $FL_e = 2$ Mbits, and arrive to the system with $\lambda_{e,d} = 0.5$ and variable $\lambda_{e,u}$ flows / second, respectively.

Results obtained using the $AC-A$ can be better understood if the two following aspects are taken into consideration: i) the number of feasible states of the CTMC grows (i.e., the number of coexisting uplink TCP flows and VoIP calls grows) and ii) the $AC-A$ reduces the instantaneous throughput of TCP uplink flows, increasing their latency, that is the time they are active in the system.

First, the effect of increasing the VoIP call arrival rate is investigated in terms of blocking ($BP_s$) probability for VoIP calls and blocking ($BP_{e,u}$) and dropping ($DP_{e,u}$) probability for uplink TCP flows. Results are shown in Figure 6.10

By using preemption, the blocking probability for VoIP calls is equal for both the $CAC$ schemes and it only depends on the VoIP traffic load. This is motivated by the dropping of $d_{e,u}$ uplink TCP flows each time a new VoIP call arrives to the system and there are available resources for the new VoIP call but they are used by the $d_{e,u}$ uplink TCP flows. However, the gain achieved in terms of $BP_s$ is compensated with the dropping probability of uplink TCP flows. It can be seen that $AC-S$ shows a clear tendency to decrease the dropping probability as $\lambda_s$ increases, showing better $DP_{e,u}$ than the $AC-A$. Moreover, the use of preemption also increases the blocking probability of uplink TCP flows for all $\lambda_s$ values in both, $AC-S$ and $AC-A$ schemes. Notice, also, how the BE flow length has a lower impact when the $AC-A$ is used with respect to the case of $AC-S$.

If preemption is not used, $AC-A$ performs notably better for lower values of $\lambda_s$ in terms of $BP_s$. However, for higher $\lambda_s$ values, the $BP_s$ of $AC-A$ tends to be higher than using $AC-S$. This is motivated by the system tendency to be always occupied with uplink TCP flows, as their rate has been decreased, and there are no the possibility to drop them when a new VoIP call request arrives. However, for the uplink TCP flows, the proposed adaptive scheme always show better results.

In order to test the sensitivity of the novel scheme to variations of the traffic intensity of the uplink TCP flows, the arrival rate of VoIP calls is kept constant and the TCP uplink flow offered load is changed. The results are plotted in Figure 6.11

Using preemption, the VoIP blocking probability is, as expected, insensitive to variations of uplink BE load. The cost, similar to the previous case, is a higher probability of dropping BE TCP flows. In terms of $BP_{e,u}$, both schemes show a higher blocking probability than without preemption. In that case, the $AC-A$ exhibits substantially lower values for $BP_s$ than the $AC-S$ and slightly higher than the one obtained using preemption. The impact of the BE flow length is clearly noticeable in $BP_s$, especially using the $AC-S$ scheme.

In all situations, the higher number of accepted VoIP calls using the $AC-A$ results in an increment of the VoIP throughput and then, a higher utilization ($\rho_{s,d}$) of the downlink $AC_VO$ queue, decreasing the transmission opportunities of the TCP downlink traffic, which increases also the TCP downlink blocking probability.
The benefits of preempt uplink TCP flows are shown clearly in terms of VoIP calls blocking probability. However, the cost of dropping BE TCP flows has to be quantified in order to decide if it is acceptable.
6.4 Admission Control for VoIP in Multi-Rate WLANs

Link adaptation (multi-rate transmission) is one of the IEEE 802.11 PHY/MAC key features [1]. Each mobile node is able to select its own PHY parameters (modulation and channel codec) in order to optimize the bit transmission over the noise/fading prone channel. However, when a mobile node reduces its data rate, the saturation bandwidth also reduces, affecting all active flows independently of whether they observe good/bad channel conditions [5].

Therefore, sporadically rate changes occur for one or more mobile nodes, produced by such effects as increase in distance between the two wireless single-hop end-points, presence of walls when entering a room, atmospheric meteorites (rain, etc.). A rate change on one transmitter affects not only its corresponding data flow, but it also impacts dramatically all the other active calls in terms of packet loss and jitter and produces a general degradation on the performance of the system, with active flows being dropped and reducing the free space for new ones, which result in higher blocking probabilities, provoking an overall suboptimal allocation of the network resources.

This can be observed in Figure 6.12, where the distribution of simultaneous feasible rigid flows is plotted (considering that each flow has a 80 Kbps bandwidth) for two different data rates, the FAST and the SLOW ones, which use a data rate of 5 Mbps and 1 Mbps respectively, each MN carries a single flow and the set of MNs form a single-hop network. Notice how the total number of active flows reduces as FAST flows change to SLOW ones. For example, with 12 flows active, if 4 flows change to the SLOW rate, the new state becomes unfeasible, provoking that the active flows can not achieve their required bandwidth.

To fix this, a solution to mitigate the multi-rate problem with VoIP calls is proposed and evaluated: a combined admission control and rate-based codec selection, which shows a good trade-off between blocking and dropping probabilities and speech quality (MOS).

6.4.1 Maximum number of active calls (call capacity)

In Table 6.3, the VoIP capacity (in number of calls) is shown, for both 11 and 1 Mbps rates\(^1\). Notice that at 1 Mbps rate, the G.711 codec achieves exactly one half of the capacity of the G.729. However, at 11 Mbps, the difference is only of two calls for the G.711. This simple result shows an interesting behavior of the random access WLANs,

1. At low channel data rates, the difference on bandwidth requirements from both voice codecs is the key parameter which defines the system capacity.

2. At high rates, the capacity bound is given by the number of mobile nodes accessing the channel (and so, reducing the transmission opportunities from the AP).

6.4.2 Admission Control Policies

To mitigate the multi-rate problem several policies can be implemented to provide an effective solution. To choose among all policies, the trade-off between complexity and performance gain has to be studied. Examples of policies are:

\(^1\)This values are obtained using the $CW_{\min,u}^*$-based adaptation proposed in Chapter 5 and the EDCA parameters shown in Table 3.3, with $TXOP = 4$ (measured in packets)
• **Lower rate threshold**: To block all calls which start at a rate lower than a given threshold rate. The main drawback of this policy is that it reduces the AP coverage area for VoIP calls, requiring more APs to cover the same region with the problems associated. Moreover, hand-offs due to mobility will be more frequent.

• **Using the lowest possible rate**: Another possibility to avoid the multi-rate problem and extend the coverage area to its maximum is to set the data rate of all VoIP calls to the lowest in the cell, independently of the channel conditions. From this point of view, the multi-rate problem is avoided but the capacity is also significantly reduced, achieving a default low performance.

• **Rate-based codec selection**: In order to maintain a fair relationship between the bandwidth required by the codec and the data rate used, \( \frac{\text{bandwidth}}{\text{rate}} \), this policy aims to use a lower bandwidth codec when low rates are used and, conversely, higher bandwidth codecs with fast data rates. This policy is the one considered in this chapter.

• **Dynamic measurement-based codec selection**: This policy has been proposed by Sfairopoulou
et al. [114] and consists on changing the VoIP codecs only when several QoS metrics (i.e., information carried by the RTCP protocol) indicate that the QoS has been degraded. It is expected that it provides a better performance than other policies but with a higher complexity cost.

For all policies, it is common that for any change from a SLOW rate to a FAST rate, the resulting state is always feasible. Therefore, dropping calls only occurs when a MN using a FAST rate changes to a SLOW rate. Then, if the novel state is not feasible, two possibilities exist:

1. Drop the FAST call which has changed to SLOW rate as it does not have space in the system.
2. Drop up to $D$ FAST calls.

Although the first choice seems the more logical, results are presented for the two cases.

In Figure 6.13, an example of the cross-layer information used by the different policies are shown. For example, joining the RTCP (Real Time Control Protocol) information at application layer and the transmission rate from MAC layer.

6.4.3 Considerations for modeling Multi-Rate scenarios

Being the packet service time the effective delay since a packet (bulk of packets) is (are) ready to be transmitted (head of the queue) until it is (they are) released from the queue, as it has been correctly delivered or discarded due to a maximum retry limit $R_i$. Considering ideal channel conditions in terms of BER (Bit Error Ratio), a packet is not correctly received only if it suffers a collision. Letting $M_i$ be the average number of transmissions required by a packet from $MN_i$, 

\[ M_i = R_i \]
the expected number of blocked slots due to the back-off and AIFS, $\gamma_i$ the average slot countdown delay, $T_{s,i}^B$ the duration of a $B_i$ packets burst transmission and $ET_{c,i}$ the average duration of a collision involving node $i$, the service time is (see Chapter 4):

$$X_i^{B_i} = (M_i - 1) (\zeta_i \gamma_i + ET_{c,i}) + \zeta_i \gamma_i + T_{s,i}^B \tag{6.2}$$

Then, modeling the multi-rate situation in both ad-hoc or infrastructure networks can be done by simply parameterizing the EDCA model, setting different data rates for each MN or the AP. Then, the service time in a multi-rate WLAN will be affected by:

1. The duration of the transmission of the other nodes (they will be longer if they use lower rates).
2. The duration of the own transmission which also depends on the transmission rate.

Moreover, a simple assumption is done to compute the data rate used by the AP: the radio link in each direction behaves equally. So, it means that the same data rate is used for transmission from the AP or from the MN. Then, a load-weighted average has been considered to compute the rate used by the AP. It is the average of the rates (proportional to the offered load directed to each node) of the MNs.

6.4.4 Benefits of adjusting dynamically the VoIP codec rate

The $ns-2$ [110] simulator with the enhancement of a SIP patch to include the standard SIP agents (Proxy, UserAgent, and DNS) has been used to perform the basic SIP operations. The patch was obtained from National Institute of Standards and Technology (NIST) [115] and has been adapted for the $ns-2$ WLAN scenario with the goal to control the codec of each call while the call is in progress. The basic $ns-2$ 802.11 MAC module has been used and the experiments were performed for different channel rates, from 1 to 11 Mbps. Both ad-hoc and infrastructure scenarios were tested.

Figure 6.14 is obtained from [114]. In it the throughput degradation (packet losses) of active rigid flows can be observed when several of the active flows change their data rate to a lower one (point (1)). In this situation all flows are affected and none performs correctly. In both sub-figures, a simple rate change of three flows has dramatic consequences in all active flows. A codec change for some flows that have changed their transmission rate (“slow” flows, Sub-figure a)) or some of the flows remaining at the higher data rate (“fast” flows Sub-figure b)) (point (2)) returns the system throughput to a stable situation. Note that in the second case, the system throughput (point (3)) has to decrease more to achieve the attainable system bandwidth. This means that more codec changes are required for fast than for slow calls.

6.4.5 Rate-based codec selection

Let $C$ be the set of VoIP codecs and $R$ the set of WLAN supported channel rates, with $\nu^* = \max(C)/\max(R)$ as the optimal codec / rate ratio\(^2\). The rate-based codec selection is defined

\(^2\)It is theoretically possible to find a codec with higher bandwidth requirements and lower achieved MOS than a contender. Indeed, such codecs have been proposed in practice. However, once a better performing
as the rule to choose the codec \( c \in C \) given a rate \( r \in R \) which \( \frac{c}{r} \geq \nu^* \) which allows the maximum number of active calls. Notice that by choosing \( \frac{c}{r} < \nu^* \) the system harms the call more than required.

In order to be able to evaluate the system using analytical techniques, a particular case of the previous problem is analyzed. Only two codecs and two transmission rates are considered, choosing the codec by the simple proportional criterion: the higher bandwidth codec is used when the MN transmits at FAST rate and the lower bandwidth codec in the other case. In spite of the simplicity of the situation, it is enough to understand how an adaptive scheme could improve the overall cell performance and justify the use of adaptive codec policies which allow both to maximize the voice quality and the system capacity.

To decide which is the best policy (action) that the admission control has to implement, a certain cost function has to be maximized, which in our case takes into account these parameters:

1. \( E[MOS] \) (Mean Opinion Score).
2. \( E[N] \) Average number of active calls.
3. Blocking (BP) and dropping (DP) probabilities.

Notice that these metrics are complementary and show the different trade-offs available in choosing the adequate constellation in terms of codec usage, depending on the importance given to each metric. These trade-offs can be summarized as follows: it is expected that the higher bandwidth codec (G.711) has a better MOS than the G.729. However, as the maximum number of active calls is greater using the G.729 than the G.711, in the same conditions, the blocking probability is expected to be lower using the G.729. Hence, the results provided here justify in option is found, the older codec becomes useless. Hence, in our argumentation, an ordered list of codecs along a bandwidth-MOS axis it is assumed.
which conditions the use of more bandwidth-efficient codecs, allowing a higher number of active calls, is better than to block or drop calls at the price of providing worse voice quality.

**Cell model**

In Figure 6.15 the considered scenario is shown. The AP coverage area $A$ and radius $R$ is composed of two regions: FAST and SLOW areas. The FAST area, $A_F$ is defined by a circle of radius $R_1$ and the rest is defined as the SLOW area, $A_S$ ($A = A_F + A_S$). Mobile Nodes are uniformly distributed through the coverage area, so with probability $P_F = A_F/A$ the MN is in the FAST region and with $P_S = 1 - P_F$ in the SLOW region. The call arrival rate $\lambda$ is defined in call / $m^2$ (calls per square meter). Therefore, the total arrival rate of FAST calls is $\lambda_F = A_F \cdot \lambda$ and $\lambda_S = A_S \cdot \lambda$. All calls have the same average duration $1/\mu$, which is assumed to be exponentially distributed.

A MN changes from FAST to SLOW rate if it is in the FAST region ($P_F$) and moves to the SLOW region with rate $\gamma_{F\rightarrow S} = P_F \cdot \gamma$ changes / second. Conversely, a MN moves from SLOW to FAST region with rate $\gamma_{S\rightarrow F} = P_S \cdot \gamma$. Notice that if $R_1 = 0$ or $R_1 = R$, there are no rate transitions. Here, the value of $\gamma$ is considered as an abstract parameter but it can be easily linked with node mobility.

A MN which changes its transmission data rate must send a renegotiation request to the AP with the new data rate. Then, the admission control evaluates if the new state is feasible and if so, there are no action and a positive response is sent to the MN. If the state is not feasible, the admission control can chose among several actions,

1. Drop this or another set of calls.
2. Suggest a new codec for this or another set of calls.

This scenario is modeled using a two-dimensional Markov chain where the $i$ dimension represents the FAST rate calls and the $j$ dimension the SLOW rate calls. Notice the transitions between rates ($\gamma$). In case of transitions which incur in dropping one or more calls, the $\theta$ parameter is a variable policy which takes values between $[0, 1]$. Values of $\theta \rightarrow 1$ imply that a call...
which changes its rate from FAST to SLOW is dropped and values of $\theta \rightarrow 0$ imply that up to $D$ other FAST calls are dropped. In Figure 6.16 a particular example of the two-dimensional Markov chain is depicted. The considered EDCA parameters are shown in Tables 3.2 and 3.3 with a $TXOP_{limit}$ for $AC_{VO}$ fixed to 4 packets, and only two data rates: 11 and 1 Mbps.

Three codec combinations are evaluated:

1. FAST: G.711 and SLOW: G.711. Always the G.711 codec is used.
2. FAST: G.711 and SLOW: G.729. The codec change is done each time a MN changes its rate.
3. FAST: G.729 and SLOW: G.729. Always the G.729 codec is used.

![Figure 6.16: Example of the considered bi-dimensional MC](image)

Performance metrics such blocking probability ($BP$), dropping probability ($DP$), average cell MOS ($E[MOS]$) and average number of active calls $E[N]$ are computed and used to understand the system behavior. The blocking and dropping probabilities are computed as presented in [84].

**Numerical results**

Solving numerically the two-dimensional Markov chain, the presented performance metrics can be computed. Considering a cell of $R = 100$ meters, the analysis focuses in three basic points:

1. Effect of $R_1$.
2. Effect of the load for $R_1 = 75$ m.
3. Effect of the mobility for $R_1 = 75$ m.

In the first scenario the effect of $R_1$ is analyzed. Considering a call arrival rate equal to one call each 30 seconds (resulting in $\lambda = 1.061 \mu$ calls / $m^2$) and two different $\gamma$ values, 0 and 1/60 changes / second. Basic result of this first experiment is that, irrespective of rate changes and the load, the cell performance is related mainly to the area of the FAST and SLOW regions.
Figure 6.17: Impact of $R_1$ with 0 changes / second, (a) BP (b) DP (c) $E[N]$ (d) $E[MOS]$; with $1/60$ changes / second, (e) BP (f) DP (g) $E[N]$ (h) $E[MOS]$

(second and third experiment will extend this result). Notice how both blocking probabilities (Figure 6.17) are similar, being slightly lower the blocking probability when the rate changes are considered. This is justified by the higher dropping probability in that case (Figure 6.17(f)).
Then, as there are a lower number of active calls (as the dropping probability has the effect of increasing the system departure rate) and the offered load is lower (ratio between $\lambda/\mu$), the blocking probability is lower.

What is remarkable is that the dropping probability only takes considerably higher values in the range $R_1$ comprised between 60 and 90 meters, since in this area the major part of new calls are FAST but the probability of changing to SLOW rate is still significant. Figure 6.17 also shows the average number of active calls, which presents lower values in the presence of rate changes, and the average cell MOS. Note how as the FAST area increases, the MOS also grows. This result is very intuitive, since there is more bandwidth available for each call.

Hence and summarizing, with increasing $R_1$, the average number of active calls increases, the number of dropped calls increases and the MOS also increases.

Now, the effect of increasing the load of the system is evaluated for the case of 1/60 changes / second and $R_1 = 75$ meters. The load goes from 0.1061 to 10.61 $\mu$ calls / $m^2$. Results are plotted in Figure 6.18. As expected, all parameters increase proportionally to the load. However, notice the different effect of the $\theta$ parameter: With values of $\theta = 1$ the system shows higher dropping probabilities, but all the other metrics are better. Moreover, the MOS gain using $\theta = 1$ is also remarkable. Consequently, in principle it is preferable to drop the call that suffered the rate change than any other, except if the dropping probability is a critical factor in our decision.

Finally, the impact of mobility is also considered. A traffic load of 1.061 $\mu$ calls / second and rate changes from 0 to 1 changes / second in a WLAN cell with $R_1 = 75$ meters are considered. In Figure 6.19 the performance results are shown. Notice the interesting result which tells us that increasing the frequency of rate changes, the average number of active calls decreases,
but at the same time the blocking and dropping probability also decrease. This seems pretty counterintuitive, but the explanation runs as follows: Since the time that a call remains in the system before being dropped is shorter (increasing the system departure rate), the offered load is lower, consequently showing lower blocking and dropping probabilities.

### 6.5 Conclusions

Different solutions to provide QoS to VoIP calls in WLANs have been presented, focusing in two hot-topics: i) the coexistence of both TCP and VoIP traffic in the same coverage area and ii) the implications of the multi-rate anomaly in VoIP calls due to the use of link adaptation schemes.

There are other aspects to be considered, being these results a solid starting point for future research in this area. However, the main conclusion of this chapter is that the associated problems to the VoIP over WLANs must be solved in order to achieve a successful future deployment of this class of services, as the impairments affect severely the VoIP performance.

The suggested solution is to use a cross-layer admission control which uses information of the PHY/MAC and application layers in order to optimize the overall network behavior. Then, the admission control is able to suggest the different layer parameters to achieve the performance optimization goal. For example, a properly set of MAC parameters allow to maximize the system capacity.

Moreover, a proper selection of the VoIP codecs allows the system to react to the random variations in capacity due to transmission rate changes, which are motivated by the wireless
channel and/or the user mobility. A simple decision criterion is presented, based on the ratio between the VoIP codec bandwidth requirement and the channel rate. This procedure, called rate-based codec selection, is based on a fairness principle: \textit{the consumed network resources by a MN has to be proportional to its transmission rate}. It is an interesting solution to improve the user Quality of Experience (QoE) [116] with VoIP services over WLANs.
Chapter 7

Conclusions

The aim of this thesis was to study how heterogeneous traffic flows (both, TCP-like and UDP-like) coexist in WLANs and what mechanisms have to be implemented to provide a properly use of the transmission resources. To achieve this goal, several steps had been followed: i) modeling the DCF/EDCA MAC protocol, ii) understanding and modeling the flow level behavior of TCP and UDP-like protocols over the IEEE 802.11 MAC and iii) proposing feasible solutions to the different observed impairments, based on the admission and rate control capabilities of the IEEE 802.11e EDCA.

A detailed study of past and actual works has been carried out, which has been the basic knowledge over which this thesis has been built. These background knowledge has considered i) theoretical, descriptive aspects or already presented results of the WLAN technology, ii) numerical techniques to solve both the non-linear equation systems (such as the fixed point approach), the multi dimensional Markov chains (the Successive Overrelaxation Gauss-Seidel method) [117] and the constrained multi-objective optimization and iii) several basics of stochastic processes (and its modeling repercussions) [118] such as the regenerative / renewal theory, the implication of using memoryless distributions and other different advanced aspects in modeling queuing systems (for example, the bulk-service times queue).

One of the main contributions of the thesis is the DCF/EDCA MAC model presented. To the best of our knowledge, this is the first EDCA model which captures the dynamics of all traffic differentiation parameters, noting the use of an $M/G[1.B]/1/K$ queue to model the TXOP behavior. Moreover, the model has a great flexibility as any combination of the $CW_{min}$, $CW_{max}$, $AIFS$ and $TXOP_{limit}$ values can be tested, which can not be always applied to the other models found in the literature. Obviously, this is achieved by relaxing the accuracy of several approximations, such as using the independence assumption or the considered AIFS blocked slots model. However, in spite of these approximations, very close outcomes compared with simulations are still obtained, which results in an excellent trade-off among flexibility and accuracy.

The DCF/EDCA MAC model is used to evaluate how elastic and rigid flows coexist over current IEEE 802.11 based networks. Results show clearly how with the presence of elastic flows, the WLAN cell tends rapidly to saturation, limiting the maximum number of rigid flows. This difficult coexistence can be mitigated by using the EDCA traffic differentiation capabilities.
Moreover, the flexibility of the EDCA model is used to try different parameter settings in order to find the optimum MAC parameters which, at the same time, maximize both the number of rigid flows and the best-effort (elastic) throughput. This optimization study results in the definition of a simple but effective algorithm to tune the EDCA MAC parameters.

The behavior of the TCP and UDP transport protocols over the IEEE 802.11 MAC protocol has been also evaluated. For UDP flows, the VoIP service has been considered, since nowadays it is a hot topic in research, focusing the study in capacity aspects (maximum number of simultaneously active voice calls) in different typical situations such as with coexisting TCP flows or in multi-rate scenarios. An unexpected low capacity is noticed (compared with the channel transmission rate and the bandwidth required by a VoIP call). Solutions for the different impairments which cause this low capacity are provided. As for the TCP protocol, a detailed study of how it behave over a random access MAC protocol is presented, including the different possible modeling assumptions that are considered. These models are developed with the objective to be used in analytical studies of admission control and radio resource management in WLANs.

Last contribution is the definition and specification of an EDCA-based admission control scheme. The proposed admission control incorporates the adaptive use of MAC parameters in a per node basis as an implicit part of the WLAN resource management process, computing the most suitable set of MAC parameters each time that the traffic load or the channel state changes. The improvements of using the proposed admission control are demonstrated for the case of heterogeneous traffic flows (the coexistence of TCP flows and VoIP calls in a single cell scenario) and for the case of a multi-rate WLAN hotspot scenario, where the use of an adaptive codec selection improves the overall performance. For both cases, different policies are considered, static and adaptive, including both preemptive and non-preemptive options. The adaptive ones show impressive better results that the ones achieved by using the current EDCA (static) with only minor changes in the open- firmware of the standard. Therefore, it is expected that this type of solutions can be incorporated into APs in the future.

Finally, it is noticeable that there are very few works providing a flow level analysis to evaluate the admission control in WLANs. This is because it is difficult to simulate / analyze the flow level behavior, due to the non-linear and reciprocally variable capacity of the IEEE 802.11 random access MAC protocol, which in this thesis has been solved efficiently by developing the MAC model. However, the flow level analysis becomes fundamental as there has been found unexpected performance results that remain invisible from a simple packet analysis performance.

7.1 Next results beyond this thesis

Following the contributions of this thesis, there are several points which can be further extended from the work presented,

Update the EDCA model to include different Access Categories at each MN ([119])

In order to allow that all MNs can have active more than a single AC at the same time. This will also improve the approximation considered in Chapter 6 to manage both VoIP and TCP traffic at the AP.

\footnote{A similar study could be done for any other type of rigid traffic.}
Different ACK policies, ARQ protocols and TXOP transmission over prone-channels ([120–122]) The joint optimization of all these parameters is expected to improve notably the overall WLAN PHY/MAC performance.

Different and multiple types of rigid / streaming traffic classes ([123,124]) An extended scenario based on different and multiple rigid traffic classes, such as video-conferencing, VoIP, TV broadcast, etc., including also different types of elastic flows.

Infrastructure-based MAC parameter tuning algorithm ([93,94]) The algorithm presented in this thesis does not consider implicitly the uplink / downlink unfairness. Therefore, a more detailed algorithm is expected to be able to improve further the overall WLAN performance.

A model for simultaneous TCP / UDP flows ([44,77]) The distribution of backlogged nodes when the AP transmit randomly TCP and UDP packets can be also modeled by an embedded Markov chain similar to the used when only downlink TCP packets are sent.

Multidimensional model for VoIP over Multi-rate scenarios ([114]) A new dimension of the Markov chain for each codec/channel-rate combination will provide a general scenario where the evaluation of the different alternatives presented would be possible, especially the measurement-based codec selection, which has not been already evaluated at flow level.

7.2 Future research

This thesis has explored a broad list of topics rigourously. Starting from the contributions presented, several continuation lines are open and are of interest for future research, which can be based on the results provided here. In this section, a list of the main research lines is presented.

This thesis has an important implicit part of mathematical background that is itself interesting. Then, the redefinition of several modeling assumptions or re-build them completely starting from another point of view [71] are of interest in order to increase their accuracy or to find closed-form [125] expressions which simplify the computational costs.

However, from an engineering point of view, several research points are remarkable:

Performance of simple admission control algorithms ([83]) It has been shown that there are several conditions which do not allow the use of simple CAC algorithms (it is called a simple CAC algorithm the set of rules which do not consider the non-linear relationships between network and traffic characteristics). In that case, an estimation error is assumed and the quantification of this error by testing different simple algorithms in heterogeneous load and channel conditions is of interest to properly choose one of them. The benefits of using these simple algorithms in real implementations can be measured in terms of minor computational costs.

Adjusting MAC parameters based on active measurements This is not a novel topic as there are several works that consider this (for example [88,99,126]). These mechanisms are
based on simple measured information (such as collision probability) which is translated to a properly value of the MAC parameters. By including this adaptive option in the MAC parameter tuning process, progressive finest performances can be achieved.

**User Mobility** To test how different user mobility patterns affect the overall cell performance. The user mobility could be combined with the use of multiple rates and heterogeneous traffic flows [28, 114].

**Extended coverage using multiple APs** ([127, 128]) To study the impact of an extended coverage area using multiple APs. In that case, the hand-off or reassociation process must be considered. Therefore, different policies to accept / drop hand-offs (for example the use of channel guard policies) can be of interest. Another interesting topic is load balancing to distribute the traffic load among different APs when possible.

**Heterogeneous Wireless Networks** ([129, 130]) Call admission control and load balancing in heterogeneous wireless networks (e.g., WLAN / UMTS). For example, instead of a VoIP call codec change, it could be possible to hand-off from the WLAN to any other mobile network such as UMTS.

**Multi-hop / Mesh-Networks** ([128, 131]) Multi-hop networks allow to increase the coverage areas by using other nodes as a relays. Conversely, the estimation complexity of the transmission resources and their distribution among nodes increase significatively.

**Cross-layer optimization** ([121, 122, 132]) Several IEEE 802.11e protocols can be improved by a cross-layer optimization (for example between application, transport, MAC and PHY layers).

### 7.3 Concluding remarks

Last years WLANs have become one of the first actors in wireless technologies, coping a lot of attention from different points of view, which comprise users, business people, researchers, etc., who see in that technology a new opportunity to offer / receive mobile services without the participation of the established bigger mobile operators. Moreover, this broadband mobile access combined with the change on the Internet traffic profiles, such as the VoIP, P2P and other multimedia services, will be one of the fundamentals axes of an advanced Internet.

Despite that the research on resource (e.g. bandwidth) optimization has motivated a lot of efforts since the first communication systems appeared, the continuous evolution of the technology makes this topic always actual. Thus, a properly use of the transmission resources means that a solution to the problem of maximizing both the overall resource set utilization (or the revenue associated) and the individual user / service requirements (user satisfaction) has to be found. Notice that the possible better solution, always imply a tradeoff among the different parts, in our case the user / service requirements and the system capacity. Moreover, a simple solution would be desired, with fair and clear rules, which can be applied to the real world.
Bibliography


97


[121] A. Graell, B. Bellalta, and M. Oliver. A novel ARQ Protocol for IEEE802.11a Based on Rate-Compatible Codes. IEEE VTC’04 Fall. Los Angeles (USA), September 2004.


