Service differentiation in Wireless Local Area Networks

by

Suong H. Nguyen

Submitted in total fulfilment of
the requirements for the degree of

Doctor of Philosophy

Centre for Advanced Internet Architectures
Swinburne University of Technology
Australia

2012

Swinburne University of Technology

Australia

Abstract

Service differentiation in Wireless Local Area Networks

by Suong H. Nguyen

Wireless Local Area Networks (WLANs) have significantly developed as a means to provide Internet access at many places. With the rapid development of new applications, traffic over WLANs becomes more and more diverse. In particular, there are different types of traffic with different Quality of Service (QoS) requirements. This raises the need for the provision of service differentiation in WLANs.

To support QoS in WLANs at the Medium Access Control sublayer, most of previous proposals as well as the default parameter setting of the Enhanced Distributed Channel Access (EDCA) mechanism defined the IEEE 802.11e standard are based on prioritization, which defines several access classes (ACs) where a higher priority class receives better service in all aspects than a lower priority class. These prioritization-based QoS proposals are known to create an incentive for selfish users to choose the class with the highest priority to gain a higher share of bandwidth, which may lead to the degradation of the whole network. The existing solutions to eliminate this incentive are either complicated or impractical to implement.

In contrast, I seek to provide service differentiation without prioritizing one class over another, that is, there is no ordering of the classes such that one gets better performance in all respects than the later ones. I do this by choosing ACs such that some parameters are less aggressive whenever others are more aggressive.

The proposed scheme to provide QoS in this thesis has many advantages over prior proposals in that it is simple to implement, compatible with the 802.11e standard and robust against selfish users. The properties of the proposed scheme with

selfish users are investigated, using a game framework that requires a model of IEEE 802.11e EDCA.

For that purpose, I also propose a novel model of 802.11e EDCA WLANs with heterogeneous traffic. The proposed model is more tractable and more accurate than previous models of the same scope, by capturing several aspects ignored in the previous models. The accuracy of the proposed scheme is confirmed by comparing with ns-2 simulations for a wide range of parameter settings. Based on the proposed model, the asymptotic analysis of delay distribution is provided.

Acknowledgments

This thesis was completed with the support of many people, to whom I would like to express my sincere thanks.

First and foremost, I would like to give special thanks to my supervisors, Prof. Hai Le Vu and A/Prof. Lachlan Andrew, for their great supervision during the past three and a half years. All the work described in this thesis was finished under their guidance. They supported me with valuable research ideas, constructive comments, efficient research methodology as well as constant encouragement and endless patience. I have gained much knowledge in doing research from their instruction, which will be very useful for my future career.

I greatly appreciate the generous support of Prof. Grenville Amitage for all students in the Centre of Advanced Internet Architecture (CAIA). Also thanks to CAIA, where I spent all my three and a half years of Ph.D candidature and made friends with many friendly, helpful and knowledgable researchers and students, with whom discussion has been very interesting and useful for my own knowledge.

I am grateful to all my friends, specially, Dr. Thuy Nguyen, Dr. Hai Do, Tuyet Tran, Imrul Hassan, Radika Veera Valli, Anjali Munjal, Tuan Nguyen and his wife Anh Phan, for all the emotional support, entertainment, and caring they provided. Thanks for treating me like a family member.

Last but not least, this thesis would not have been possible without the strong support, encouragement and love of my parents and my brothers. They have always been there for me to reply on throughout my time of PhD candidature. It is to them that I dedicate this work.

Declarations

This is to certify that

(i) the thesis comprises only my original work,

(ii) due acknowledgement has been made in the text to all other material used,

(iii) the thesis is less than 100,000 words in length, exclusive of table, maps, bibliographies, appendices and footnotes.

Signature_____

Date_____

Contents

List of Figures

List of Tables

List of Abbreviations

AC	Access Category
ACK	Acknowledgement
AIFS	Arbitration Inter-Frame Space
AIFSN	Arbitration Inter-Frame Space Number
AP	Access Point
ATM	Asynchronous Transfer Mode
BEB	Binary Exponential Backoff
CBR	Constant Bit Rate
ccdf	complementary cumulative distribution function
CDF	Cumulative Distribution Function
CFP	Contention-free Period
CSMA	Carrier Sense Multiple Access
CSMA/CA	Carrier Sense Multiple Access with Collision Avoidance
CTS	Clear-to-send
CW	Contention Window
DCF	Distributed Coordination Function
DIFS	Distributed Inter-Frame Space
DRR	Deficit Round Robin
DS	Differentiated Services
EDCA	Enhanced Distributed Channel Access
EIFS	Extended Inter-Frame Space
HCCA	HCF Controlled Channel Access
HCF	Hybrid Coordination Function
IEEE	Institute of Electrical and Electronic Engineers

IETF	Internet Engineering Task Force
IntServ	Integrated Services
IP	Internet Protocol
MAC	Medium Access Control
MPLS	Multiprotocol Label Switching
NAV	Network Allocation Vector
PCF	Point Coordination Function
PHB	Per-hop Behavior
PHY	Physical
PIA	Proportional Incentive Adjusted
PIFS	PCF Inter-Frame Space
QoS	Quality of Service
r.v.	random variable
RSVP	Resource Reservation Protocol
RTS	Request-to-send
SIFS	Short Inter-Frame Space
SNR	Signal-to-Noise Ratio
STA	Station
TCP	Transmission Control Protocol
TID	Traffic Priority Identifier
TXOP	Transmission Opportunity
UDP	User Datagram Protocol
UP	User Priority
VCG	Vickrey-Clarke-Groves
VoIP	Voice over Internet Protocol
WFQ	Weighted Fair Queueing
WLAN	Wireless Local Area Network
WRR	Weighted Round Robin

List of Notations

SS	the set of saturated sources.
$\mathbb{U}$	the set of unsaturated sources.
$\mathcal{P}$	the set of players in a game.
n	the number of access classes in my proposed scheme of QoS provision.
B_k	an access class with the index k ($k = 1, \cdots, n$).
s	a subscript denoting a particular saturated source ($s \in SS$).
s_k	a subscript denoting a saturated source using an access class B_k.
u	a subscript denoting a particular unsaturated source ($u \in \mathbb{U}$).
x	a subscript denoting an arbitrary source ($x \in SS \cup \mathbb{U}$).
N_j	the number of sources of type j ($j \in \{s, u\}$).
τ_j	the attempt probability of a source of type j ($j \in \{s, u, x, s_k\}$), or of a particular player $j \in \mathcal{P}$.
$\tau_j(a)$	the attempt probability of a particular player $j \in \mathcal{P}$ under an action profile a.
p_j	the collision probability of a source of type j ($j \in \{s, u, x, s_k\}$), or of a particular player $j \in \mathcal{P}$.
$p_j(a)$	the collision probability of a particular player $j \in \mathcal{P}$ under an action profile a.
$\mathbb{E}[.]$	the average value of a random variable.
K	retransmission limit.
m	the maximum number of times a source doubles its contention window due to collision.
Y	the duration of a virtual slot.

Y_u	the duration of a slot observed by an unsaturated source u.
σ	the duration of an idle slot.
λ_u	the arrival rate of an unsaturated source u.
η_j	a r.v. denoting the number of packets sent per burst of a source of type j ($j \in \{x, s, u\}$), which is used Chapter 3.
η_k	a real number denoting the ratio of TXOP limit of class B_k to that of class B_1, which is used in Chapter 4.
r_j	the maximum number of packets can be fit in the TXOP limit of a source of type j ($j \in \{s, u\}$).
W_j	the minimum contention window of a source of type j ($j \in \{s, u\}$), or of a particular player $j \in \mathcal{P}$.
W_{B_k}	the minimum contention window of class B_k.
T_{px}	the transmission time of a packet from the source x.
T_{ACK}	the transmission time of an ACK packet.
T_x^s	the (random) time that a burst sent by a source x occupies the channel if it is successfully transmitted.
T_x	the deterministic value of T_x^s conditioned on the burst containing only one packet.
D_u	the (random) access delay of an unsaturated source u.
b_u	the probability a packet of source u arriving at an empty source finds the channel idle.
$T_{res,u}$	the residual time of the busy period during which a burst from an unsaturated source u arrives.
Q_u	a random variable representing the queue size of an unsaturated source u.
S_s	the throughput in packets/s of a saturated source s.
S_{s_k}	the dimensionless throughput of a saturated source using the class B_k.

$S_j(a)$	the dimensionless throughput of a particular player $j \in \mathcal{P}$ under an action profile a.
C_{s_k}	the throughput in seconds/slot of a saturated source using the class B_k.
$C_j(a)$	the throughput in seconds/slot of a particular player $j \in \mathcal{P}$ under an action profile a.
T_u^c	the (random) duration of the longest packet involved in a collision involving source u.
G	the probability that no sources transmit in a given slot.
L_u	the probability the first packet in a burst from an unsaturated source u is discarded due to exceeding retransmission limit.
U_{xj}	a uniformly distributed random variable representing the number of backoff slots a source x has to wait in the j-th backoff stage.
ϵ_k^0	the difference in CW_{min} of class B_k between the proportional scheme and the PIA.
ρ_u	the queue utilization of a source u.
μ_u	the service rate of a source u.
l_x	the payload of a packet from a source x.
$\mathcal{T}$	the TXOP limit of class B_1.
A_i	the action space of a player i in a game.
A	a general action space of any player.
$A0$	the action space of a player under the proportional scheme.
$A1$	the action space of a player under the PIA scheme.
a_i	the action chosen by a player i.
$a = \{a_i\}_{i \in \mathcal{P}}$	a vector containing the action of every player in the game.
$a_{(X;\cdot)}$	an action profile $\in \{a \in A^{N_s} : a_1 = X\}, \quad \forall X \in A.$
$a_{(X;\cdot;Z;\cdot)}$	an action profile $\in \{a \in A^{N_s} : a_1 = X \text{ and } a_j = Z\}, \quad \forall X, Z \in A.$

Chapter 1

Introduction

1.1 Motivation

The concept of service differentiation comes from the fact that different individuals have different demands/requirements to be satisfied. For example, for some people, the top criterion for a product is good quality while for some others, good looking is their top criterion. Specifically, in communication networks, service differentiation has been an important issue to study, especially when there have been more and more applications developed with different requirements of Quality of Service (QoS). For example, there is a diversity of applications over the Internet from interactive ones (i.e. Voice over Internet Protocol (VoIP)) which require low delay to large transfers (i.e. file/movie download) which requires high throughput [111]. It is important to provide service differentiation because it affects the level of user satisfaction.

A traditional way to provide service differentiation is by prioritization. This means that there are multiple service classes defined with an increasing order of QoS among them, which reflects that some classes will be treated better in all aspects than other classes. Those with higher requirement can choose the service class of the higher priority. An example of this is the design of Internet plans of an Internet Service Provider. In particular, these plans have different capacity and speed, which can be ranked in an increasing order of both capacity and speed. In communication networks, prioritization-based QoS provision means that among all service classes, the highest priority class will receive the highest throughput and lowest delay while the lowest priority class will get the lowest throughput and highest delay.

However, prioritization is not always a good solution for service differentiation.

Consider the example of Internet plans. It is possible that some people may prefer an Internet connection with higher capacity and lower speed (i.e. those with the habit of movies/files downloading) while some may want to have a connection with lower capacity but higher speed (i.e. those with the habit of web browsing/chatting/facebooking). Similarly, regarding QoS provision in communication networks, delay-sensitive traffic (e.g. VoIP) requires their packets to be transmitted as soon as possible but usually has low rate. In contrast, throughput-intensive traffic may be willing to wait a little longer as long as its throughput improves as a result. Those imply that providing service classes in which one class is better in all aspects than another may not be the best.

Instead, providing service differentiation should be based on the actual requirements of users. In IP networks, this concept has been proposed in [42, 44, 60, 65]. The thesis also adopts this to provide service differentiation but in another context, Wireless Local Area Networks (WLANs).

WLANs have become very popular as a means of Internet access at home or in public areas such as hotels, offices, shops and airports. According to the WiFi Alliance [116], shipments of Wi-Fi devices reached nearly 1.1 billion in 2011 and are expected to double by 2015. While the traditional mobile device category of Wi-Fi such as smartphones and laptops continues to shine, it is predicted that the growth of Wi-Fi enabled devices in other categories also take off, as shown in Table 1.1 [116].

Table 1.1: Predicted growth of Wi-Fi devices between 2011-2016 [116].

Categories	**Examples**	**Percent**
Automotive applications	Infotainment systems, navigation, traffic monitoring	109%
Health, fitness and medical applications		39%
Smart meters and automation products		25%

A WLAN is a communication network concentrated in a geographical area that interconnects a variety of devices and enables communications among them using

radio waves as the transmission medium. WLANs can provide almost all the functionality offered by wired Local Area Networks, but without the physical constraints of the wire itself [49]. Since their appearance, WLANs have gained popularity at an unprecedented rate due to their simple, flexible and cost-effective technology [99, 158].

The leading standard for WLANs is the Institute of Electrical and Electronics Engineers (IEEE) 802.11, which ensures a high level of interoperability of products from different equipment suppliers. It adopts the standard 802 Logical Link Control protocol but provides physical (PHY) layer and medium access control (MAC) sublayer which are optimized for wireless communications [158].

WLANs were originally designed with a mandatory multiple access mechanism to support best effort traffic and an optional feature to support real-time traffic using a centralized polling mechanism in the first IEEE 802.11 standard [3]. However, this optional feature is not part of Wi-Fi Alliance's interoperability standard; hence, it is rarely implemented in hardware devices. This means that all traffic types are treated in the same way. Nowadays, with the diversity of applications over Internet, traffic over WLANs is also diverse, with different QoS requirements for different traffic types as shown in Table 1.2. With the lack of QoS support, IEEE 802.11 experienced serious challenges in meeting the demands of multimedia services and applications. Then, to improve user satisfaction, it is important to properly consider the issue of QoS provision in WLANs.

Table 1.2: QoS requirements of different types of application [111].

Application Type	**Examples**	**Requirements**
Interactive	VoIP, Video Conference	Low latency, small jitter, and small throughput variations
Short Web Transfers (<100KB)	Web search, Social networking	Low latency
Medium Sized Transfers (100KB-5MB)	Music/Photo transfer	Low latency
Large Transfers (>5MB)	Movie downloads, Software Updates	High throughput

Then, an amendment to enhance QoS at the MAC sublayer for WLANs, IEEE 802.11e, was ratified in 2005 [7], which defines a distributed medium access scheme called Enhanced Distributed Channel Access (EDCA). Service differentiation in EDCA is provided by defining four access classes (ACs): Background, Best Effort, Video, and Voice. These four ACs have different QoS through being assigned different values for MAC parameters such as CW_{min} and TXOP limit. Note that CW_{min} determines how long a station has to back off before transmitting and TXOP limit specifies how long a station is allowed to transmit without contention per channel access. The default MAC parameter settings for these classes are given in Table 7-37 of the 802.11e amendment [7], which is based on prioritization with the highest priority for voice traffic. Since the release of the 802.11e protocol, there have many proposals to improve its performance in providing service differentiation in various ways such as adapting the MAC parameters of ACs as functions of network load [20, 88, 100, 103, 124] or changing the distribution of backoff values [130]. Note that like the default MAC parameter setting, most of these proposals are based on prioritization, which have been shown to work well when users choose the right class designed for their traffic.

However, these priority-based QoS mechanisms will create an incentive for selfish users (e.g. those try to maximize their performance at the cost of others) to use the access class of the highest priority to gain a higher share of the channel. This can degrade the overall performance of the network [29] and result in no service differentiation [29, 102, 109]. This shows the importance of QoS provision for WLANs with selfish users. The existing solutions to this issue in the literature [29, 45, 102, 109, 110] which deploy additional mechanisms such as policing or pricing are either complicated or impractical to implement.

The above analysis raises an open question: "Is there a scheme that will provide service differentiation, which has the following features:

- robust against selfishness,

- easy to implement, and
- compatible with the standard?"

This thesis will focus on answering this question.

1.2 Thesis statement

The thesis's view is that a simple solution to service differentiation provision which is robust against selfish users is to provide "different but fair" services for different types of traffic by scaling two MAC parameters defined in the IEEE 802.11e standard: CW_{min} and TXOP limit appropriately. Game theory is a useful tool to study the incentives of selfish users and network modeling is required to investigate network performance. This statement poses the following challenges.

- How can the effect of CW_{min} and TXOP limit on providing service differentiation be quantified?
- What should be the right ratio of CW_{min} and TXOP limit for different access classes to guarantee the property of "different but fair" services?
- Which approach should be used to prove that the proposed scheme is actually robust against selfish users?

The thesis will present how to address the above questions to obtain a simple scheme which provides service differentiation and is robust against selfish users.

1.3 Contributions and novelty

The thesis makes the following three main contributions.

First, to quantify the effect of two MAC parameters, CW_{min} and TXOP limit, on service differentiation, the thesis contributes a novel model of IEEE 802.11e EDCA WLANs with services differentiated by these two parameters in a network of

heterogeneous traffic (e.g. multiple sources of different rates and packet sizes with different QoS requirements). Compared with the existing models covering these above features, the proposed model in Chapter 3 is novel in the following aspects.

- The model is more tractable than existing models of the same scope. In particular, the model is based on the renewal reward theorem proposed in [77] which is more tractable and requires less computation than most of the existing models which use the Markov chain-based approach originally proposed in [21].

- To the best of my knowledge, the proposed model is the first to provide a closed form distribution of the number of packets an unsaturated source (e.g. source with queue utilization less than 1) transmits per channel access. (Note that a source can send multiple packets per TXOP limit (hereafter called a "burst") if it has enough packets and the total duration of these packets is at most TXOP limit.) I also show that the bursts can be approximated by a geometric random variable clipped to TXOP limit. This information cannot be drawn from existing models due to their complexity.

- The thesis shows that the residual time of an ongoing transmission from other stations seen by a burst of an unsaturated source arriving during that transmission has big effect on the accuracy of the delay model under some considered scenarios (e.g. large TXOP limit). The proposed delay model captures this feature. In contrast, it has not been taken into account in the existing models.

- The delay model also captures the probability that a packet arrives at an empty buffer. I find that under specific scenarios, it can improve the accuracy of the model up to 25%. This is currently ignored in the existing delay models.

Based on the proposed model, the following work has also been achieved.

- A simple method is proposed to approximate the distribution of the access delay, in contrast with the complex methods used in the existing models such as the numerical inversion of z-transform.

- The thesis provides the derivation of a lower bound on the number of saturated users in the network so that the queueing delay of unsaturated sources will be infinite regardless of the traffic load from unsaturated sources, given that all stations use the same minimum contention window, CW_{min}. Surprisingly, I find that this bound only depends on CW_{min}.

Second, instead of providing priority-based service differentiation which requires complex mechanisms to correct the incentive of selfish users in choosing an access class, I seek to provide service differentiation without prioritizing one class over another, that is, there is no ordering of the classes such that one gets better performance in all respects than the later ones. In Chapter 4, I do this by choosing ACs such that some parameters are less aggressive whenever others are more aggressive, which is motivated by the observations obtained using the proposed model of 802.11e EDCA. The properties of my proposed scheme with and without selfish users are studied analytically by using a game theoretic framework. My proposed scheme has many advantages over prior proposals as follows.

- It improves service for both delay-sensitive and throughput-sensitive traffic and provides the correct incentives for selfish users (e.g. application writers who optimize their code based on measured performance using all the available services).

- It allows easy implementation: a single set of 802.11e MAC parameters provides tradeoff between throughput and delay over the range of load studied. This implies that my scheme can be implemented in infrastructure or adhoc modes.

Third, I find by simulations that under specific scenarios when there is big variability of packet sizes in the network, the collision probability of small packets from unsaturated sources is no longer the same on its first and subsequent attempts. This violates the mean field approximation of constant collision probability used in

most previous models of 802.11 WLANs and the proposed model above, which may affect the accuracy of those models. Therefore, in Chapter 5, I have investigated this phenomenon to find its cause and conditions when it is observed. Besides, to quantify how much the collision probabilities on the first and subsequent attempts are different, and how much improvement in model accuracy can be obtained by modeling this correctly, I also extend the proposed model above to capture that effect. As a result, I find that correctly modeling this can improve the accuracy of collision probability of unsaturated sources up to 30%, which implies its importance for work requiring the accurate capture of collision probability. As an example of such work, the extended model is used to optimize energy consumption of a station by minimizing its collision probability.

1.4 List of publications

All the work in this thesis has been published in [95, 96, 97, 98] which are also listed below.

- S. H. Nguyen, H. L. Vu and L. L. H. Andrew, "Packet size variability affects collisions and energy efficiency in WLANs, " in *Proc. IEEE Wireless Communications and Networking Conference*, pp. 1-6, 2010.
- S. H. Nguyen, L. L. H. Andrew and H. L. Vu, "Service differentiation without prioritization in IEEE 802.11 WLANs," in *Proc. IEEE Local Computer Networks*, pp. 109-116, 2011.
- S. H. Nguyen, H. L. Vu and L. L. H. Andrew, "Performance Analysis of IEEE 802.11 WLANs With Saturated and Unsaturated Sources, " *IEEE Transactions on Vehicular Technology*, vol. 61, no. 1, pp. 333–345, 2012.
- S. H. Nguyen, H. L. Vu and L. L. H. Andrew, "Service differentiation without prioritization in IEEE 802.11 WLANs," accepted to *IEEE Transactions on Mobile Computing*, 2012.

Chapter 2

Background and Literature Review

This chapter will provide a summary of the related literature to my work, including the current models of IEEE 802.11e EDCA WLANs, the existing approaches to provide service differentiation in WLANs and how they cope with the issue of selfish users in choosing an access class. To help the readers understand the related work, I will first provide a summary of the main features of the IEEE 802.11 standards for WLANs.

2.1 WLANs and IEEE 802.11 standards

The dominant standard for WLANs in the market is IEEE 802.11; therefore, my work focuses on this standard, the main features of which will be presented in this section.

2.1.1 IEEE 802.11 architecture

In the IEEE 802.11 architecture, two main components are the stations (STAs) and the access point (AP). Stations are devices with wireless network interface cards such as laptops, personal computers or smartphones. The AP is defined as "any entity that has STA functionality and provides access to the distribution services, via the wireless medium for associated STAs" [8]. The AP allows STAs under its management to connect to wired networks or wireless STAs under other APs' management.

With these two components, there are two main connection modes: ad-hoc mode where all STAs connect to each other directly, and infrastructure mode where STAs

connect to each other through the AP. These modes are illustrated in Figure 2.1. My work focuses on upstream traffic in the infrastructure mode; however, the work can still be applied to ad-hoc mode with insignificant change.

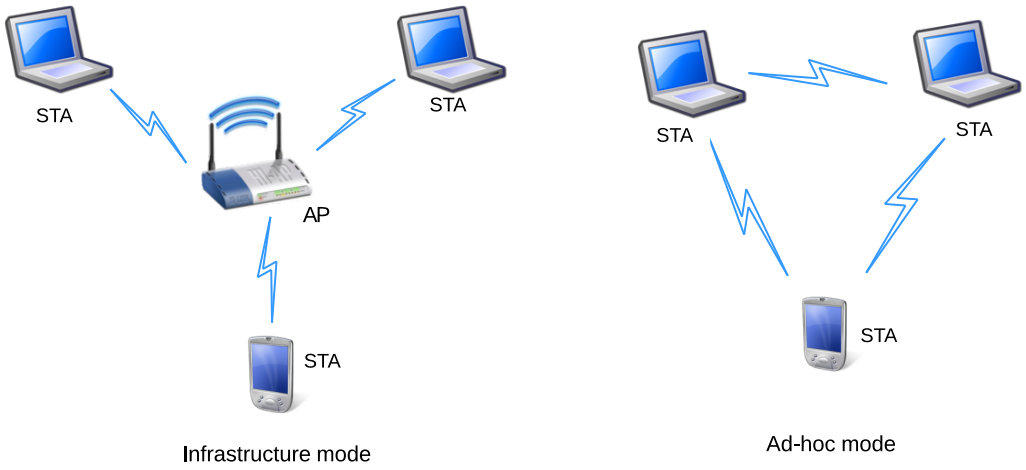


Figure 2.1: The architecture of IEEE 802.11.

2.1.2 IEEE 802.11 standards

The IEEE 802.11 standards define the specifications of the MAC layer and physical layer of WLANs. The first IEEE 802.11 standard was ratified in 1997 and revised in 1999 [3]. Since then, there have been several amendments to this specification which aim to improve the performance of WLANs such as increasing data rates supported at the physical layer and providing QoS at the MAC layer. The latest version of IEEE 802.11 standard, named IEEE 802.11-2012 [10], is published in 2012, which incorporates the previous version and all the amendments. The evolution of the standardization of physical layer and MAC layer are described in the following, with much more emphasis on the MAC sublayer which is the focus of the thesis.

Physical layer

There have been the following four amendments to the IEEE 802.11-1999 standard [3], which improves the technologies of the physical layer to support higher data

rates and hence higher throughput.

802.11b IEEE 802.11b [4] is the amendment to the IEEE 802.11-1999 specification, issued in 1999. It specifies data rates up to 11Mbps in the same frequency band of 2.4 GHz.

802.11a IEEE 802.11a [5] is another amendment to the IEEE 802.11-1999 standard, which is also issued in 1999. It specifies data rates up to 54Mpbs but in another frequency band of 5GHz. Operating in the 5GHz band reduces interference; however, it shortens the transmission range.

802.11g IEEE 802.11g [6] is an amendment to the IEEE 802.11-1999 specification, which was issued in 2003. This amendment allows data rate up to 54Mbps in the same frequency band of 2.4GHz. The 802.11g standard is backward compatible with 802.11b; hence, it may attract more attention from industry than the earlier standardized 802.11a [158]. However, the 2.4GHz has already been used by many home electronic devices such as microwave ovens and cordless phone; hence, the system suffers more interference.

802.11n IEEE 802.11n [9] is an amendment to the IEEE 802.11-2007 standard, which is issued in 2009. The maximum data rate is increased significantly from 54Mbps in 802.11a/g to 600Mbps. It is built on previous standards with additional features to improve network throughput such as the multiple input multiple output technique and the support of 40MHz bandwidth at the physical layer.

Note that the rate of overheads (i.e. PHY overhead and control frames) does not increase by the same factor as that of payload data; hence, the throughput is increasingly dominated by such overheads at high data rates [136]. This means that collision becomes more expensive at higher rate, which increases the improvement observed by the proposed scheme of service differentiation in this thesis.

MAC layer

To improve the performance of WLANs, not only the physical layer but also the MAC sublayer has been improved through a few amendments. The thesis focuses on the performance of MAC sublayer in IEEE 802.11 WLANs; therefore, it is important to understand the operation and features of the MAC sublayer.

IEEE 802.11-1999 The IEEE 802.11 MAC sublayer defines two basic medium access protocols: contention-based distributed coordination function (DCF) and contention-free point coordination function (PCF). DCF mode supports asynchronous transmission and is fully distributed while PCF supports polling-based synchronous transmission and is centralized. An 802.11 WLAN can operate in both DCF and PCF modes; however, DCF mode is mandatory while PCF is optional and rarely implemented. The subsequent amendments for the MAC sublayer are based on DCF and PCF. Therefore, I will provide the detailed description of those as below.

- DCF

  DCF is a distributed medium access protocol based on Carrier Sense Multiple Access (CSMA) with Collision Avoidance (CSMS/CA) and binary exponential backoff (BEB), instead of CSMA with Collision Detection as in the wired network because wireless stations can not listen to channel while transmitting. Carrier sensing in 802.11 DCF is performed at both PHY and MAC layers, which are PHY carrier sensing at air interface and virtual carrier sensing at MAC layer. PHY carrier sensing detects the presence of other STAs by analyzing the received signal strength. At MAC layer, this is done by a station updating the Network Allocation Vector (NAV) in accordance with the Duration field in the MAC header of the received packet. If the station detects a collision, it will set this NAV to be Extended Inter-Frame Space (EIFS). This NAV indicates the amount of time that must collapse until the ongoing transmission from other STAs finishes and the channel is free again to sense

for idle state.

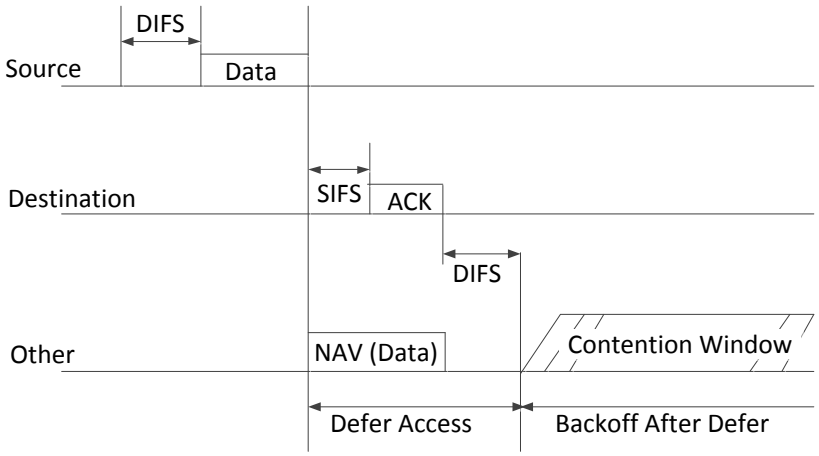


Figure 2.2: The diagram of IEEE 802.11 DCF.

The operation of CSMA/CA in 802.11 DCF is shown in Figure 2.2, the description of which is as follows. When a frame arrives at the MAC sublayer, it will be placed in a transmission queue where the frame at the head of the queue will contend for channel access. Then, the medium access procedure of the DCF can be described as follows. When a frame arrives to an idle source, it senses the channel for a Distributed Inter-Frame Space (DIFS). If it is idle during this whole time, the frame is transmitted immediately (asynchronously). Otherwise, the source waits until the channel is continuously idle for DIFS, and then starts a backoff process. A backoff counter is initialized to a random integer uniformly distributed between 0 and $(CW - 1)$, where CW is the current contention window. For each new frame, CW is initialized to the minimum contention window, CW_{min}, and doubles after each unsuccessful transmission until it reaches the maximum contention window, CW_{max}, after which it remains constant. The backoff counter is decreased by one at every idle slot time, of duration σ, and frozen during periods of channel activity.

Decrementing is resumed after the expiration of a DIFS after a channel activity period ends. When the backoff counter reaches zero, the frame is transmitted. An acknowledgment (ACK) is sent back from the receiver after a Short Inter-Frame Space (SIFS) for every successful frame reception. If an ACK is not received, the source increases CW as described above, and attempts again until the retry limit is reached. After receiving an ACK, the source performs a "post-backoff" process with CW set to CW_{min} before being allowed to restart the above procedure. This prevents back-to-back frame transmission.

To avoid hidden terminal and capture effect problems, the 802.11 standard defines request-to-send (RTS) and clear-to-send (CTS) exchange before transmitting a frame. If RTS is sent but no CTS is received, the station knows that another STA is also transmitting and suspends the frame transmission for some backoff time. The work in this thesis assumes there is no hidden terminal; hence, RTS/CTS is not considered.

Note that in DCF, all stations have the same chance to access to the channel; therefore, no service differentiation is supported.

- PCF

  PCF is designed to support realtime traffic. In particular, when PCF is enabled, a point coordinator (usually the AP) periodically contends for access to the channel. Once it acquires the channel, it polls stations with request for contention-free service (e.g. stations with realtime traffic) and grants them the privilege of transmitting. This implies that service differentiation, which is the focus of the thesis, can be provided with the use of PCF.

  However, PCF has not drawn much attention from either the research or industry community due to the following issues [99, 158].

  – It is difficult for the point coordinator to manage the polling of a large number of active stream without harming the applications using DCF

contention.

- Since the AP needs to contend for the channel using DCF at the beginning of a CFP, the effective period of contention-free polling may vary.
- The hardware implementation of PCF was thought to be too complex.
- PCF experiences substantial delay at low load.
- PCF is centralized and can only be used in infrastructure mode.

Therefore, there still lacks a mechanism to provide QoS in WLANs, which is then addressed in the amendment IEEE 802.11e [7]. However, even if PCF is implemented, the issue of users' incentive to claim to be of realtime traffic to gain a higher share of bandwidth still exists.

IEEE 802.11e This amendment was issued in 2005 to support QoS in WLANs. In particular, it defines several access classes which have different values of several MAC parameters, the detail of which will be provided in Section 2.2.2. The work in this thesis is based on this amendment.

IEEE 802.11n To increase the network throughput, this amendment improves not only physical layer technologies to support high data rates as mentioned above but also MAC layer. In particular, the IEEE 802.11n MAC defines several ways of frame aggregation, which helps to reduce channel waste due to the protocol overhead and hence increases the effective throughput.

2.2 QoS provision

In this section, I will provide a literature review of service differentiation in communication networks, especially at the MAC sublayer of WLANs which is the focus of my work. Service differentiation means providing different QoS for different types of traffic. To see how the techniques to provide QoS in wired networks can be ap-

plicable to wireless networks, it is important to first review the evolution of QoS provision in wired networks.

2.2.1 Evolution of QoS provision

One of the earliest communication networks, the public switched telephone network started building out a worldwide, circuit-switched network a century ago [129]. It was suited to carry realtime traffic such as voice by providing dedicated circuits, which guarantees the QoS measures of voice such as low latency, fixed circuit-based routing, predictable service levels and information-order preservation [129]. A connection is set up before any communication can begin and the admission control is performed to prevent the demand for resources from exceeding the supply [141]. In this kind of networks, when the system capacity is exceeded, any traffic arriving will be dropped or put in a queue.

After that, with the development of computer technologies and the need of interconnecting personal computers with each other and with wide range of resources, data networking began to grow in the 1970s. Then, it took off in the 1980s and exploded in the 1990s [69]. In data networks, packet flows share the same resources and hence contend with each other for these common resources. When there is contention for resources in the network, it is important for resources to be allocated or scheduled fairly. With the increase of new applications, traffic over data networks contains not only best effort but also realtime ones such as voice and video. Then, QoS has become a real issue in data networks. Asynchronous Transfer Mode (ATM) was the first general data-networking technology including a class of service concept at the link layer, which offers different treatment for different traffic types [129]. However, it has been rarely deployed due to its complexity [141]. In the late 1990s, the Internet Protocol (IP) won out as the technology of choice for converged networks which support a combination of voice and data due to its ease of use, ubiquity, and advances in handling realtime traffic [129]. Therefore, it is useful to study how QoS is provided in IP networks, the applicability of which to QoS provision in

wireless networks is then discussed.

QoS models for wireline IP networks

The main models proposed for wireline IP networks are summarized in the following.

IntServ The first QoS model proposed by the Internet Engineering Task Force (IETF) is Integrated Services (IntServ) model. This model proposes three service classes [129, 142], which are

- *Guaranteed service* for applications requiring fixed delay bound;
- *Controlled-load service* for applications requiring reliable and enhanced best-effort service;
- *Best-effort service* for applications requiring no guarantee.

The model uses a flow-based concept coupled with a signalling protocol called Resource Reservation Protocol (RSVP) to set up paths and reserve resources. The signalling protocol guarantees that adequate resources are available (at each hop) for the flow before admitting the flow onto the network. IntServ is implemented by four components: the signalling protocol (e.g. RSVP), the admission control routine, the classifier, and the packet scheduler.

The pros of the IntServ model are [129]

- Conceptual simplicity, facilitating the integration with network policy administration;
- Discrete per-flow QoS, making it architecturally suitable to voice calls;
- Call admission control capabilities, which can indicate to endpoints whether the desired bandwidth is available.

However, the InterServ architecture has the following main cons [129, 142].

- The amount of state information and exchanged signalling messages increases proportionally with the number of flows. This places a huge storage and processing overhead on all network elements. Therefore, this architecture does not scale well in the Internet core. Besides, it might also require a significant bandwidth on large networks.

- The requirement on intermediate nodes are high. They must have RSVP, admission control, multi-field classification, and packet scheduling.

Because of those, IntServ was never deployed in reality [141].

DiffServ Due to the difficulty in implementing and deploying the IntServ model, the Differentiated Services (DiffServ) model was introduced [142]. While IntServ is flow-based mechanism, DiffServ is class-based. DiffServ uses packet markings to classify and treat each packet independently. Different markings correspond to different classes with different services. The model defines packet markings along with specific per-hop behaviors (PHBs) using the Type of Service byte as differentiated service (DS) field in the IPv4 header.

The DiffServ model has the following advantages[129].

- Scalability - this model can scale well due to no state or flow information required to be maintained in nodes.

- Performance - the packet content only needs to be inspected once for classification purpose and hence marking. All subsequent QoS decisions are made based on the value of DS field in the IP header, which helps to reduce processing requirements.

- Interoperability - all vendors are already running IP.

- Flexibility - The DiffServ model does not prescribe any particular feature to be implemented by a network node. Any feature can be used as long as it is consistent with the behavior expectation defined in the PHBs.

The disadvantage of the DiffServ model relative to IntServ is that there is no bandwidth guarantee for packets that belong to a flow and hence no guarantee of services, especially when the network is congested [129, 142]. As the DiffServ architecture creates preferable traffic classes that deliver better service in all relevant parameters than other classes in the system, it requires traffic regulation (i.e. pricing) across service classes to avoid traffic being directed to the high-quality class.

In reality, DiffServ has been deployed by some network service providers [141]. However, it is generally not deployed network-wide and only enabled in a few potential bottlenecks [141]. There have been some efforts to extend DiffServ after it was introduced. The most noticeable ones are Multiprotocol Label Switching (MPLS) support for DiffServ (RFC3270) and Diffserv-aware Traffic Engineering (RFC3564) [141]. The former enables DiffServ to be implemented in a MPLS network, while the latter enables Traffic Engineering to be implemented per DiffServ class.

Note that the IEEE 802.11e standard to support QoS in WLANs can be considered an example of the DiffServ model in that it defines several access classes with different QoS to serve packets of different traffic types.

Hybrid IntservDiffserv This a hybrid model which uses a mix of IntServ and DiffServ as described in RFC2998 [129, 141]. Network operators can use IntServ at the access and edge of the network where bandwidth is limited and scalability is less an issue, and use DiffServ in the core of the network [141]. There is lack of public awareness about this model and there is no deployment [141].

"Non-elevated" IP approach The above QoS models require complex mechanism such as reservation signaling or admission control, policing and pricing, which make them hard to deploy in real networks [65]. Therefore, another approach using *non-elevated services* may be a much better fit for the Internet [65]. These services can be described as "deploy incrementally, with no need for policing, accounting, or significant change to operational practices" [2]. The proposed scheme to provide

QoS in WLANs in the thesis uses this approach and it also has these nice properties.

There have been several QoS proposals using this approach. This approach is different from differentiated or integrated services in that it provides a spectrum of "different but fair services" in which neither service classes can be said to receive better treatment. These proposals are summarized in the following.

- *Best Effort Differentiated Services (BEDS)* is a set of services similar to Best Effort in that QoS provided depends on the network conditions, but differentiated in their tradeoff between packet delay and packet loss probability, which is supported by an architecture and several mechanisms proposed in [42]. In particular, there are two service classes defined: the "loss-conservative" service which is suited for file transfer applications, and the "delay-conservative" service which can benefit VoIP. In their proposal, the "loss-conservative" service has a smaller loss probability but larger delay than the "delay-conservative" service.

- *Equivalent Differentiated Services (EDS)* [44] uses a similar idea of providing different but equivalent services as BEDS by trading off delay versus loss rate. Similar to [42], it is not very clear how to configure service ratios to reflect the absolute service guarantee.

- *Alternative Best Effort (ABE)* is a novel service architecture for IP networks proposed in [60], which relies on the idea of providing low delay at the expense of maybe more loss. With ABE, every packet is marked as either blue or green where the choice of color is made by the application based on nature of its traffic and on global traffic conditions [60]. Green packets are guaranteed a low bounded delay in every router but are more likely to be dropped than blue packets in return. In contrast, the blue traffic is guaranteed to receive at least as much throughput as it would in a flat best effort network. Interactive application with realtime deadlines (e.g. voice) will mark their packets green as long as the network conditions offer large enough throughput while data

applications such as bulk data transfer will seek to minimize overall transfer time and send blue traffic [60].

- *Incentive-Compatible Differentiated Scheduling (ICDS)* is an incentive-compatible framework developed in [65] to provide differentiated services in IP networks, called "Incentive-Compatible Differentiated Scheduling". In particular, it generalizes the idea in ABE [60] to capture more than two classes, in which a service class with lower delay bound will have higher loss probability.

My proposal to provide service differentiation without prioritization in WLANs uses the idea of "different but fair" services in the above works. However, instead of trading delay for loss as in these works, my proposal trades delay for throughput in the absence of loss.

So far I have discussed how end-to-end QoS can be provided in IP networks with the above QoS models. I now discuss how the service differentiation is implemented at each hop, because the thesis considers QoS provision for the uplink in an infrastructure WLAN (e.g. one hop). In data networks, packet flows contend with each other for resources, hence, it is important for resources to be allocated or scheduled fairly according to their QoS weight at each hop. This ensures that there is no flow starving and flows are isolated from bad effects caused by bad sources which inject packets into the network at uncontrolled rate. For this, Weighted Round Robin (WRR) or Weighted Fair Queueing (WFQ) [35], which assign different weights for different traffic types, can be used. Round-robin scheduling is simple, easy to implement, and starvation-free. However, in WRR, to obtain a normalized set of weights, the mean packet size of each flow must be known. Another weakness of WRR is that it cannot guarantee fair link sharing due to allocating the bandwidth on a packet-by-packet basis. The Deficit Round Robin (DRR) scheme [126] can solve the first issue of WRR by taking into account the packet size. However, a drawback of this scheme is that it may not allocate fair bandwidth in short time scales. Better fairness can be achieved with WFQ schemes [46, 105] which approximate the

Generalized Processor Sharing model for packet-based traffic scenario. However, the computational complexity of DRR is proven to be $O(1)$ per packet processing, less than that of fair queueing scheme, $O(log(n))$. The concept of those proposed schemes have been used to provide service differentiation in one-hop WLANs, the detail of which is described in Section 2.2.2.

QoS in wireless networks

So far I have summarized the approaches of QoS provision in wired networks. I will now discuss how QoS provision in wireless networks is different. An important property of wireless networks, which makes them different from wired networks, is their distributed nature. In particular, in wireless networks using random access protocols, stations contend for channel access in a decentralized manner. Hence, the QoS mechanisms proposed for wired networks can not be directly applied without modification. This will be shown in the next section where I discuss in detail about QoS provision in WLANs because this is the focus of the thesis.

2.2.2 Service differentiation in WLANs

Recall that the first version of 802.11 standard do not support service differentiation. Before the official release of 802.11e amendment for QoS enhancement, there have been many proposals to provide service differentiation in WLANs. Those will be presented first in this section, which is followed by the description of 802.11e standard and its existing enhancements.

QoS enhancements for legacy DCF

Approaches to provide service differentiation for the legacy DCF can be classified into priority-based or fair scheduling-based methodology, which are summarized in the following.

Priority-based schemes There are several priority-based proposals to provide service differentiation in WLANs, which will be described below.

- To introduce priorities in 802.11 DCF, [11] proposes three techniques corresponding to three parameters: backoff increase function, DIFS, and maximum frame length.
  - Backoff increase function: Instead of doubling the contention window by two as in DCF, the scheme proposes to change that factor differently for different priority class. The higher priority, the smaller the factor, and hence the higher probability to access the channel. The results show that this scheme works well for User Datagram Protocol (UDP) traffic but does not perform well for Transmission control protocol (TCP) traffic because all TCP ACKs have the same priority.
  - DIFS: Each priority level is assigned a different DIFS. Higher priority is assigned lower DIFS, which causes some slots after a busy period reserved for transmission only from the high priority traffic. This is shown to work for both TCP and UDP.
  - Maximum frame length: Each priority level has a maximum frame length, which is higher for higher priority class. The paper proposes two ways to implement that: (1) drop packets longer than the maximum frame length assigned; (2) fragment packets exceeding the maximum length. This is shown to work for both TCP and UDP.
- *Blackburst* is a scheme proposed in [127] to minimize the delay of realtime traffic. In this scheme, data stations and realtime stations have different access procedures. In particular, data stations still use CSMA/CA and positive acknowledgement as in DCF while realtime stations are scheduled to access channel similar to time division multiple access. When a realtime station first has a frame to send, it waits until the channel is sensed idle for PCF Interframe

Space (PIFS) (PIFS<DIFS) before entering a black burst contention period. In this period, it sends a so-called black burst to jam the channel. The length of this frame is proportional to the time it has to wait until entering black burst period. After that, it observes the channel for a short time to see if there is any other realtime station with longer black burst. If there is any, it will wait for this station to transmit and then start sensing the channel again. If there is not, it will send the frame and then schedule the next frame in t_{sch} seconds in the future. If there are no data stations, realtime stations after transmitting the first packet will not need to contend for the channel again and transmit a frame every t_{sch} seconds. The Blackburst offers very low delay and jitter for realtime traffic. However, it requires modification to the existing 802.11 standard and wastes channel through sending black bursts.

- Deng et al. [36] proposes a scheme which provides four priority levels, based on four combinations of two values of Inter-Frame Space (IFS) and two backoff generation functions. Instead of using DIFS for all stations as in 802.11 DCF, it proposes to use DIFS for the lowest two priority classes and PIFS for the highest two priority classes. Between the lowest two priority classes or the highest two priority classes, the higher priority class generates random backoff in the lower interval than the lower class.

Fair scheduling-based schemes A summary of work based on the fair scheduling-based methodology is presented in the following.

- *Distributed Fair Scheduling (DFS)* is proposed in [137, 138] to improve the fairness issue in the 802.11 DCF standard and provide service differentiation by applying Self-Clocked Fair Queueing [46] in a distributed way. It assigns different weights to different traffic classes where higher priority class has higher weight. Then, the backoff value is determined to be inversely proportional to the weight, which allows traffic of higher priority to access the channel with

higher probability. Besides, to improve fairness, the backoff value is proportional to the packet size. This implies that a station transmits a long packet has to wait longer.

- *Distributed Weighted Fair Queuing (DWFQ)* [19] applies the concept of "weighted fair queueing" into WLANs to distribute the wireless bandwidth among flows proportional to their weight. The higher priority class will have higher weight. To do that, each station will calculate a ratio $L_i = R_i/W_i$ where R_i is its actual throughput and W_i is its weight. Stations will advertise their ratios in the transmitted packet. A station will then compare its ratio with others and adjust the contention window accordingly until it has the same ratio as others.

- *Distributed Weighted Fair Queuing (DDRR)* is proposed in [106] to provide different classes with different throughput requirements, which is based on the DRR mechanism which has the complexity of $O(1)$ compared with $O(log(n))$ of the Self-Clocked Fair Queueing. In this scheme, each station will be allotted a service quantum Q bits every T_i seconds depending on its throughput requirement. Each station will have a Deficit Counter which keeps track of the amount of bandwidth available to that station. The Deficit Counter keeps increasing continuously by Q bits every T_i seconds and decreasing by the size of a frame transmitted by that station. Then, at each time t, the IFS for each traffic class i at a station j is determined by the size of the quantum and the Deficit Counter at time t. The backoff process is removed in this scheme, which explains for its low variation of throughput and delay.

IEEE 802.11e standard

In 2005, IEEE defined a new standard named 802.11e [7] to support QoS at the MAC sublayer. This standard defines a new coordination function called the hybrid coordination function (HCF), which includes two medium access methods: a distributed scheme called enhanced distributed channel access (EDCA) and a centralized one

called HCF controlled channel access (HCCA). These two access schemes are QoS-specific extensions of the original access methods: DCF and PCF.

Similar to PCF, HCCA is more complex and inefficient for normal data transmission [104, 145]. Because of this as well as the simplicity of EDCA to implement [67], EDCA's independence of architecture [13] and with the popularity of the DCF in IEEE 802.11 networks [84], EDCA has received the most attention [39] and is expected to be the dominating access scheme for IEEE 802.11e networks [84]. My work in this thesis is based on EDCA. Hence, the remaining discussion of this section only focuses on EDCA.

EDCA This mechanism provides differentiated and distributed access to the wireless medium for STAs. In particular, there are four AC queues defined in EDCA to support prioritized QoS. When a packet arrives at MAC layer, it is tagged with a traffic priority identifier (TID) based on its QoS requirement. The TID value from 0 to 7 is the user priority (UP), which is identical to the IEEE 802.11D priority tags. The UP value is then mapped to one of four AC queues, as shown in Table 2.1 taken from Table 9-1 of [8].

Table 2.1: Mapping between User Priority (UP) and Access Category (AC) [8].

Priority	UP (same as 802.1D UP)	802.1D designation	AC	Designation (informative)
Lowest	1	BK	AC_BK	Background
	2	–	AC_BK	Background
	0	BE	AC_BE	Best Effort
	3	EE	AC_BE	Best Effort
	4	CL	AC_VI	Video
	5	VI	AC_VI	Video
	6	VO	AC_VO	Voice
Highest	7	NC	AC_VO	Voice

Service differentiation is achieved by varying the following MAC parameters for different ACs [7].

- **The minimum contention window (CW_{min}) and maximum contention**

window ($\mathbf{CW}_{max}$) determine how long a station has to backoff before it can transmit.

- **Arbitration Inter-Frame Space (AIFS)** is used in EDCA instead of DIFS in DCF. AIFS[AC] is given by

$$AIFS[AC] = AIFSN[AC] * \sigma + SIFS. \tag{2.1}$$

where σ is the duration of an idle slot time as defined above, and Arbitration Inter-Frame Space number (AIFSN) is an integer. The difference of AIFSN between two ACs means that higher priority AC is allowed to access channel in some slots where lower priority AC is not allowed. The higher the traffic load, the higher benefit the priority of AIFS gives [22]. In other words, it provides load-dependent prioritization.

- **Transmission Opportunity (TXOP) limit** represents the maximum duration a STA can transmit without contention once it gains channel access, which is used to improve the efficiency of the system. During this time, a STA can transmit multiple frames. These frames can be immediately acknowledged, in which case they are separated by a SIFS, an ACK, and another SIFS. Besides, to improve efficiency further, 802.11e also allows block acknowledgement which acknowledges all frames sent per TXOP using only one frame. Then, frames sent in a TXOP are separated by only a SIFS, which allows more packets to be sent per TXOP than immediate acknowledgement.

Each AC queue works as an independent DCF STAs with its own backoff counter and uses its own backoff parameters. The values of backoff parameters of each AC are advertised by the AP in the Beacon frame at the beginning of each superframe; otherwise, the default EDCA parameters, given in Table 7-37 of the 802.11-2007 standard [8], are used.

In the rest of this thesis, I refer to the service differentiation caused by AIFS

as AIFS differentiation, by CW_{min} and/or CW_{max} as CW differentiation, and by TXOP limit as TXOP differentiation.

Enhancements over 802.11e EDCA

Since the release of 802.11e EDCA, there have been many proposals to improve its performance. As an example, Tadayon et al. [130] proposes to use the gamma distribution instead of uniform distribution when generating backoff value in EDCA. In particular, QoS differentiation is supported by having different backoff distribution shapes of different concentration. Higher priority traffic will have distribution with lower mean and higher concentration. This helps to reduce the collision of higher class. However, this will increase collision probability among stations of the same priority. Moreover, the performance of EDCA has also been improved by considering the impact of network conditions such as network load and channel conditions, as summarized below.

The first factor, network load (i.e. the number of contending nodes in a network) have a clear affect on the priority-based service differentiation. If the MAC parameters of ACs are static, even a node using the highest priority AC may be unable to achieve its minimal required performance. This issue has been addressed in the literature by adapting the values of MAC parameters to the network load. In general, work of this kind can be classified into two types: (1) adapting the MAC parameters based on some implicit measures of the network load such as collision rate and busy rate, which is implemented in a distributed way [88, 100, 103], (2) optimizing the MAC parameters to satisfy a certain goal such as guaranteeing the requirement of realtime traffic, maximizing the admissibility region of real-time traffic or/and maximizing the throughput of data traffic [20, 124].

To address the second issue of varying channel condition in 802.11e, [115] proposes to classify the cause of an unsuccessful transmission, either due to collision or bad channel. Then, three parameters, which are fragmentation threshold, persistent factor, and defer countdown can be adapted to improve the service differentiation

of flows. For example, an algorithm called EDCA-LA is proposed in [80], which adaptively adjusts backoff time for each AC by taking the channel conditions into account. The idea behind EDCA-LA is to increase the backoff time when the channel condition is bad and to decrease this time when the channel condition is good [80]. The channel condition is inferred from the current physical transmission rate used by a station. The results show that EDCA-LA outperforms the default EDCA setting at high network load. Although time-varying channel condition is a practical issue, my work in this thesis does not consider that and assumes ideal channel condition (e.g. an unsuccessful transmission is only due to a collision) for tractability, leaving that for future work.

2.2.3 Service differentiation in WLANs with rational users

Note that the default EDCA parameter setting in the 802.11e standard and the existing QoS schemes introduced in the previous section always provide higher priority for realtime traffic. Those will operate as analyzed if users of a certain traffic type choose the right class designed for that traffic. However, when users are "rational" (e.g. users always try to maximize their own performance, also called "selfish" users), the prioritization will create an incentive for users of lower priority traffic to use the class designed for real-time traffic to gain a higher share of resources [29, 102]. This can degrade network performance drastically [29] and QoS differentiation no longer occurs when all data users use the highest priority class [29, 102, 109].

To analyze the effect of rational users, game theory is often used as a useful tool. The analysis using game theory usually gives insight of whether the considered scenario has a stable outcome or not and the property of this outcome if there is any. Then, mechanism design can be used to make users behave in a way which leads to a desired outcome. My work in Chapter 4 applies both game theory and mechanism design in service differentiation provision.

Therefore, in this section, I first provide a short introduction of game theory and mechanism design, together with their applications. Then, I will investigate

how game theory and mechanism design have been applied to analyze and solve the above incentive issue in service differentiation provision in WLANs.

Game theory

Game theory is a collection of analytical tools designed to help us understand the phenomena that we observe when decision-makers interact [89]. More specifically, it provides a mathematical basis for the analysis of interactive decision-making processes. It provides tools for predicting what might (and possibly what should) happen when agents with conflicting interests interact [89]. A summary of the history of game theory can be found in [37].

In general, games can be categorized as non-cooperative and cooperative games [101, 147]. Noncooperative games consider individual players who act selfishly and deviate alone from a proposed solution if it gives them higher benefit, and do not coordinate their moves in groups of players. Cooperative games are concerned with situations when groups of players coordinate their actions. Analysis in cooperative game theory is centered around coalition formation and distribution of wealth gained through cooperation [147]. A detailed description of cooperative games can be found in [48]. The thesis focuses on noncooperative games; hence, I only discuss about the noncooperative game in the remaining of this section.

A noncooperative game can be divided into two categories: a static game or a dynamic game [48]. In a static (or "one-shot") game, the players take their actions only once, independently of each other. Even though, in practice, the players may have made their strategic choices at different points in time, a game would still be considered static if no player has any information on the decisions of others [48]. In a dynamic game, the players have some information about each others' choices and can act more than once, and where time has a central role in the decision-making [48].

According to whether the players have full information of the game's structure or not, a noncooperative game can be classified into two types: complete information

and incomplete information games.

There are two principal ways to represent a game: the Strategic Form and the Extensive Form [37]. The Strategic Form (e.g. game table) provides the representation of the players, their strategies and their payoffs defined below [48]. Meanwhile, the Extensive Form (e.g. game tree) provides the representation of not only the players, their strategies and payoffs but also the order of play in the decision process, the information available to the players at the time of their decision, and the evolution of the game [81]. A static game is usually represented in the strategic form while the most useful representation of a dynamic game is the extensive form.

In this thesis, I am interested in the strategic form game because the strategic form provides an adequate description of the game I will consider. Hence, the rest of this section will present the basic concepts of a strategic form game.

A strategic form game consists of the following three components [37, 48, 89, 101]:

- a finite set of players,
- a set of possible strategies for each player, and
- payoff function for each player.

Players "Players" are the decision makers in the modeled scenario. In the wireless scenario, players are usually the nodes of the network.

Strategies It is important to differentiate between an action and a strategy. An action is the "move" (or decision) a player makes at a certain stage. In the wireless scenario, actions can be transmission rates, modulation schemes, backoff time or transmit power level [89]. Meanwhile, a strategy specifies the action a player will take at every stage of the game, given what he or she knows about the actions of other players and any other information that a player may learn over the course of the game [37]. In a static game, the strategy and action are the same. However, in a dynamic game, the strategy and action should be differentiated. When each

player chooses a strategy, the resulting "strategy profile" determines the "outcome" of the game. (A strategy profile is a vector containing the strategy of every player, one for each player.)

Note that there are two types of strategies defined in game theory: a pure strategy and a mixed strategy. A pure strategy selects unambiguously some specific course of action. A mixed strategy corresponds to a player choosing a probability distribution over his set of pure strategies [101].

There are two important concepts which are usually mentioned in game theory: "best response" and "dominant strategy". Best response is a strategy which maximizes a player's payoff, given a particular strategy choice of other players. Then, dominant strategy is a strategy which maximizes a player's payoff, regardless of the strategy choice of other players [37, 89, 101].

Payoffs For each strategy profile, each player receives a "payoff", which represents the value of the outcome to the user. In particular, a payoff is a number assigned to each possible outcome through a utility function. A higher payoff represents a more desirable outcome [89]. In the wireless scenario, energy saving and throughput are some examples of players' payoff.

Nash equilibrium One of the goals of game theory is to predict what will happen when a game is played. The most common prediction of what will happen is called an "equilibrium", a stable solution. The most well-known equilibrium concept in game theory is the "Nash equilibrium" [89]. A Nash equilibrium is an strategy profile at which no player has any incentive for unilateral deviation [89, 147]. Despite its shortcomings such as being not unique and optimal for players, Nash equilibrium has emerged as the central solution concept in game theory, with extremely diverse applications. A Nash equilibrium is stable; hence, once proposed, the players do not want to individually deviate [101].

An alternative interpretation of the definition of Nash equilibrium is that it is a

mutual best response from each player to other players' strategies. Then, if every player has a dominant strategy, the strategy profile including the dominant strategy of every player is a Nash equilibrium. [37] provides a good discussion of how to realize a Nash equilibrium if there is any.

Mechanism design

Interactive decision-making processes may lead to an inefficient outcome (i.e. a game with an inefficient Nash equilibrium). Then, mechanism design can be used to achieve a desired outcome such as a social optimum.

Mechanism design is a subfield of economic theory that attempts to implement optimal system allocation with "rational" individuals who aim to maximize their own payoffs [29, 101]. A mechanism is a pair of a cross product of the strategy spaces of every individuals and an outcome function which maps a strategy profile to a decision and transfers [63]. In general, mechanisms can be classified into mechanisms with money (or pricing mechanism) and mechanisms without money, the detailed description of which can be found in [101].

An important concept in mechanism design is "incentive compatibility". A mechanism is called "incentive-compatible" if every individual prefers a certain strategy which reflects his truthful information because this gives him higher utility [101]. An example of "incentive-compatible" mechanisms widely applied in many fields are Vickrey-Clarke-Groves (VCG) mechanisms, the details of which are presented in the following.

VCG mechanisms are among the most efficient mechanisms which not only tackle the dishonesty of individuals in choosing a strategy not reflecting his truthful private information, but also guarantee achieving the maximum social welfare (i.e. the optimum network utility) [29]. The mechanisms incur payments to individuals, which encourages them to declare their private information truthfully. The payment in these mechanisms represents the loss in value that is imposed on the other individuals due to the change in decision that results from the presence of an individual in

the society [63].

Application of game theory and mechanism design

Although game theory was first developed for use in economics, it has been applied in many other fields such as political science, psychology, biology, communication and networking [147].

In communication networks, many problems such as flow and congestion control, network routing, trust management, resource allocation, and QoS provision have been modeled and analyzed using game theory [16, 48]. Especially, game theory has been widely used to study a variety of issues in wireless networks such as random access control, power control, rate control, power allocation of MIMO channels, and packet forwarding [48, 81, 89, 128]. In the following, I will discuss in more detail its application in random access control, power control, rate control and QoS provision in wireless networks, especially in WLANs.

There has been much work on games of random access control. The earliest random access games analyze ALOHA with selfish users [15, 64, 90]. In 802.11 WLANs, the backoff misbehavior (nodes deliberately fail to follow the IEEE 802.11 MAC protocol) has been studied extensively because of their easy operation and potential catastrophic impact on network performance [86]. Lu et al. [86] studied different kinds of backoff misbehaviors and showed that the fixed window backoff misbehavior (keeping contention window constant regardless of the backoff stage) is much more harmful than than others because it has scalable gain, which means its throughput gain ratio goes to infinity as the number of legitimate nodes increases to infinity. In general, the existing work in the 802.11 WLANs analyzes static and/or dynamic games where each stage of the game is one slot. [73, 147] consider one-shot CSMA/CA game and find that a noncooperative CSMA/CA game then arises with a payoff structure characteristic of a Prisoners' Dilemma. For dynamic games, [26] considers fixed window backoff misbehaviors, which shows that the existence of a small population of selfish and noncooperative nodes leads to a network collapse and

there exists an infinite number of Nash equilibria in the network of a small population of selfish nodes. Besides, it also shows that if selfish users cooperate by using a Pareto-optimal strategy and penalizing those deviating from that strategy, the system can stabilize at a Nash equilibrium which is also Pareto optimal. Similarly, [73, 74] also consider the CSMA/CA game as a dynamic and repeated game and propose a strategy to converge to a symmetric Nash equilibrium which maximizes the long-term utility (SPELL) [73] or Pareto efficient (CRISP) [74] . [153] extends the approach in [90] to find a symmetric Nash equilibrium which maximizes each player's payoff in a CSMA/CA 802.11 dynamic and repeated game, without assuming that players know the number nodes in the network. However, similar to [90], it does not prove the uniqueness of that solution or propose how the stable solution can be reached.

Beside its popular application in random access control, game theory has also been widely used in rate control and power control in wireless networks. In particular, rate and power control game in cellular networks have been much investigated [18, 50, 122, 156]. In WLANs, rate games are studied in [27, 28, 131] while [28] considers power control game. Moreover, the joint power and rate control game has also been considered in [17, 28], where players want to maximize their throughput with minimum energy consumption.

Furthermore, QoS provisioning with selfish users has also been a subject to study of game theory. In particular, game theory has been used in WLANs to analyze the interaction between the service provider (e.g. AP) and new users with QoS constraints in admission control, and study the incentive of users when classifying their traffic class. In particular, the former issue has been studied in [53, 79], in which the AP wants to increase its utility by improving the channel utilization and accommodating more new users while new users want to maximize its own utility by achieving the highest QoS if possible. Both show that there exists a Nash equilibrium in the game. However, [79] does not provide an 802.11e model to calculate the performance metrics required to determine the utility of players in the

game model while [53] does. The application of game theory and mechanism design in the latter issue of selfish users in choosing a traffic class, which is the focus of the thesis, will be presented in detail as follows.

Application for service differentiation provision in WLANs

Recall that the default EDCA parameter setting in 802.11e and most of the proposed schemes of service differentiation mentioned above provide higher priority for realtime traffic (e.g. better service in all respects). This creates an incentive for applications of "lower priority" traffic to use the traffic class of the highest priority [29, 102], which can lead to no service differentiation and worsen the overall network performance. To cope with this issue, approaches in prior work can be classified into rewarding schemes [102, 109] and pricing schemes [29, 45, 110]. The first approach uses 802.11e's contention-free period (CFP) to provide extra throughput to the data class, which is problematic because current wireless NICs do not implement the CFP. While [102] assumes one service class per station and does not consider admission control, [109] assumes two service classes per station and captures admission control. The pricing approach either requires micropayments of monetary prices, which makes implementation difficult, or must impose prices through some other form of service degradation such as packet drops, which seems counter-productive. In particular, the pricing scheme in [29] uses Vickrey-Clarke-Groves mechanism to achieve the implementation of the socially optimal allocation in dominant strategies. This method requires users to report their entire utility function and usually requires a centralized allocation of resources [110]. The scheme in [110] proposes a two-dimensional bid mechanism to ensure socially optimal operation as a Nash equilibrium strategy among users whose utility functions are not known and who attempt to access the channel in a decentralized manner. In [45], untruthful users are detected and then penalized by jamming their transmission, which is not efficient and requires cumbersome implementation.

Therefore, I propose not to use prioritization as a means of service differentiation

provision. Instead I use the concept of "fair differentiated service" which has already been proposed in [42, 44, 60, 65] for wired networks. Recall that the idea of "fair differentiated services" is to differentiate a number of traffic classes by trading loss for delay without giving an absolute better service to any class. This has not been widely deployed, because it requires complex scheduling and queueing management in the core network. In contrast, for wireless links connected directly to the host running the application, I find that there exists a fair service differentiation scheme where no protocol changes are needed.

In brief, my scheme of service differentiation in Chapter 4 is different from prior work in the way that it not only provides service differentiation for users of different traffic types but also guarantees the right incentive to users. This is done by proposing a fixed set of values for MAC parameters (TXOP limit and CW_{min}), which is simpler than the previous work and complies with the standard. To analyze and prove the properties of my proposed scheme, I need a tractable and reasonably accurate model of 802.11e EDCA which considers heterogeneous traffic (e.g. both saturated and unsaturated) and captures both MAC parameters: TXOP limit and CW_{min}. The model should give the prediction of the performance measures of users such as mean delay and throughput. In the next section, I will summarize the existing models of 802.11e EDCA and point out the gaps in the literature which motivate us to develop another model.

2.3 Modeling IEEE 802.11e EDCA WLANs

The existing models of 802.11 WLANs can generally be classified into two types: (1) modeling the legacy DCF without service differentiation and (2) modeling the 802.11e with service differentiation. Most of the existing 802.11e models extend the approach used in 802.11 DCF models to capture different ACs with different MAC parameters: AIFS, CW_{min}, CW_{max} and TXOP limit. Hence, despite my interest in 802.11e EDCA models, I will first summarize the approaches used to model an

802.11 DCF and then describe how these approaches have been extended to model 802.11e EDCA.

2.3.1 Models of IEEE 802.11 DCF

There has been much work modeling the DCF mechanism in WLANs [14, 21, 34, 41, 77, 82, 85, 91, 121, 134, 135, 140, 143, 155]. In general, most models require to solve a fixed point system between the probability τ a station transmits a packet in a given slot and the probability p a station collides when transmitting, which was originally proposed by Bianchi [21]. A key approximation in [21] is that a packet collides with the same probability on each of its attempts, regardless of the backoff stage. Then, the fixed point system is given by

$$\tau = f(p) \tag{2.2}$$

$$p = 1 - g(\tau) \tag{2.3}$$

where $g(\tau)$ is the probability that no other stations transmit in a given slot. By solving this system, the probabilities τ and p are determined, which are then used to calculate the performance measures of the network such as throughput or delay. A common method to find the fixed point is using the approach of fixed point iteration. It is possible that the fixed point system has multiple solutions. Hence, it is important to investigate the uniqueness of the fixed point solution and the convergence of the fixed point iteration. In the literature, there have been only a few works studying these issues while most of the 802.11 WLAN models ignore them. In particular, [77, 113] studied the uniqueness of fixed point under saturated condition. They showed that with IEEE 802.11 DCF parameters, the fixed point system has a unique solution. Moreover, [77] provided a relaxed fixed point iteration for computing the fixed point and the conditions under which the sequence of relaxed iteration converges to the fixed point. Under unsaturated conditions, [155] studied the uniqueness of the fixed point for the small buffer and infinite buffer models.

It proved that for small buffer model, the fixed point system always has a unique solution while for the infinite buffer model, it may have multiple solutions in a transition regime from light to heavy traffic loads. It also provides the conditions under which the sequence of relaxed iteration converges to a general fixed point.

Generally, models are different in the way the attempt probability is determined. In particular, there are three main approaches widely used to calculate the attempt probability in the existing models of 802.11 DCF WLANs as presented below. For each approach, models can be further classified by the traffic type: saturated vs. unsaturated. Note that most work modeling networks of unsaturated sources claim that their models can cover saturated sources by simply setting the queue utilization [85] or probability that there is at least another packet waiting to transmit after a successful transmission [91] to 1.

The first approach is developed in [21], which explicitly models the backoff process as a two-dimensional Markov chain where the vertical dimension represents different backoff stages and the horizontal dimension shows the decrementing process of the backoff counter and obtain the attempt probability by explicitly solving Markov chain (e.g. determining the entire stationary distribution of the Markov chain and summing the probability of states corresponding to attempts). The model in [21] assumes that stations are saturated and can retransmit packets for an infinite number of times. By solving the fixed point system to get the attempt probability and the collision probability, the throughput of a saturated user is calculated as the average payload successfully transmitted per slot duration. This approach is extended in [140, 143] to capture the retransmission limit for saturated networks. Specifically, a recent work [41] claims to improve the accuracy of the model by including the channel status into the transition probability between two states of continuous backoff value at the same backoff stage in the Markov chain.

To model a network of unsaturated users, [34, 82, 91] extend the Markov chain in [21] by adding additional stages to capture the process when a new packet arrives at an empty queue and whether the post-backoff has expired or not at that time. [41]

extends their Markov chain proposed for saturated sources to capture unsaturated sources by replacing the number of saturated sources with the number of active unsaturated sources (e.g. unsaturated sources with a packet waiting to transmit) where the distribution of this is given as a function of the queue utilization. Among the above unsaturated models, [34, 41, 82] consider network of homogeneous unsaturated users. In contrast, [91] investigates networks of heterogeneous unsaturated users (e.g. users with different arrival rate and packet size).

All of the unsaturated models calculate an important measure of unsaturated sources, delay. By considering a small buffer model, the delay of each packet in [91] only consists of access delay (also known as service time), where the access delay of a packet is defined as the duration between the instant when the packet reaches the head of the transmission queue and the time when it is successfully received. Besides, [34, 41, 82] use the M/G/1 queue model for unsaturated sources. However, [34, 41] calculate only the mean access delay while [82] calculates the average total delay of an unsaturated user which is the sum of the mean access delay and the mean queueing delay. Note that the method to calculate the mean access delay in [34] is based on the counting the total number of slots during the service time of a source multiplied by the mean slot time. In [41, 91], the mean access delay of a source is the sum of the mean backoff time, the collision time and the successful transmission time. Meanwhile, the method in [82] uses the mean value of components contributing to the access delay during the service time of a tagged unsaturated source developed in [135], the detail of which will be discussed below.

The second approach is to use the mean-value technique to calculate the probability each station attempts to transmit per slot, proposed in [134] for saturated sources. In particular, the attempt probability of a saturated source is given by the inverse of the average number of backoff slots per stage, which represents one attempt per backoff stage. The model also involves solving a fixed point system to get the value of collision probability, from which the saturation throughput is determined as the fraction of channel bandwidth used to transmit payload success-

fully. Also using this approach, [121] proposes a unified delay model for saturated sources, which allows to obtain the explicit moments of different order as well as generating function. Using the proposed model, [121] proves that the binary exponential backoff mechanism induces a heavy-tailed delay distribution for the case of unlimited transmission and shows through numerical examples that the distribution has a truncated power-law tail with limited retransmission limit.

Besides, this approach is extended in [85, 135] to capture networks of heterogeneous unsaturated sources (e.g. users of different arrival rate and/or packet size). In particular, to model an unsaturated source, the attempt probability of the source is determined by the attempt probability on condition that the source has a packet waiting to be transmitted multiplied by the probability the source has a packet at a given time. The conditional attempt probability can be determined similar to the calculation of the attempt probability of a saturated source using mean-value technique. The probability the source has a packet waiting to be transmitted at a given time depends on the queueing model considered; however, in most cases, it involves the mean access delay of an unsaturated source. In [85, 135], the mean access delay of a tagged unsaturated source is calculated as the sum of mean value of contributing components during its service time such as the average number of idle backoff slots, the number of successful transmission from the tagged unsaturated source and every other stations, and the average number of collisions of the tagged sources and other sources). The model in [85] is similar to the one in [135] with a slight modification to the conditional attempt probability. The calculation of the mean access delay in [85] uses a similar approach to [135], with a difference in the calculation of the number of collisions during the service time of the tagged unsaturated node. In particular, [135] calculates this as the sum of the number of collisions suffered by all stations, which is overestimated because collision involves at least two stations. In contrast, [85] fixes this by taking only half of that, based on the approximation that a collision only occurs between two stations.

The third approach to determine the attempt probability uses the renewal reward

theorem proposed in [77] for saturated sources. This approach comes from the fact that the total backoff period of each packet is a renewal cycle with the number of attempts a packet makes during that period being the "reward". Then, the attempt probability is given by the average number of attempts of a packet divided by the average number of slots collapsing from the time the packet reaches head of the queue until it is successfully transmitted or dropped due to exceeding retransmission limit. Although this approach also implicitly assumes a Markovian structure to the back-off process and provides the same result as explicitly solving Markov chain, it is a much more convenient way to determine the attempt probability and directly reveals that back-off distributions only enter the attempt probability formula through their mean values. Therefore, for clarity, I present this approach as a separate category. [155] extends this approach to model unsaturated sources. In particular, similar to the mean-value approach, the attempt probability of an unsaturated source is equal to the attempt probability of the source on condition that it has a packet waiting to transmit multiplied by the probability the source has a packet at a given time. The conditional attempt probability of an unsaturated source is determined in the same way as the attempt probability of a saturated source using renewal reward theorem. Besides, the access delay in [155] is calculated in a similar way to [34]. Based on the proposed model, [155] also studies the uniqueness of the fixed point as mentioned above.

Note that although most of the DCF models for unsaturated networks based on Markov chains take into account the probability that a packet arrives at an empty queue in the Markov chain to calculate the attempt probability, they do not consider this probability when determining the access delay. This may affect the accuracy of the access delay calculation because the access delay in that case does not include backoff time.

2.3.2 Models of IEEE 802.11e EDCA

Recall that EDCA is mainly different from DCF in that it allows service differentiation by defining several ACs with different MAC parameters: AIFS, CW_{min}, CW_{max}, and TXOP limit. There is also a minor difference in the way the backoff counter decrements. In EDCA, decrementing is resumed one slot time before the expiration of AIFS after a channel activity period ends while in DCF, the decrementing is not resumed until after the expiration of DIFS [23]. These main differences between EDCA and DCF, and the property of AIFS and CW in providing service differentiation have been thoroughly investigated in [23] and the performance of TXOP differentiation has been investigated in [55, 107].

In general, when service differentiation is captured in the models, stations of different traffic types (e.g saturated or unsaturated) or different MAC parameters will have different attempt probability and hence different collision probability. Then, unlike (2.2), the fixed point system will consist of more than two equations, which reflects this heterogeneity.

Note that before the official release of the IEEE 802.11e standard in 2005, there have been several works which seek to provide service differentiation in 802.11 WLAN by differentiating CW [51, 112, 144]. Those have proposed models to investigate how it can be used to provide service differentiation, which are based on Markov chain and for saturated networks. In particular, [51] studied the effect of service differentiation caused by CW_{min} and CW_{max} with infinite retransmission by extending the model in [21] for multiple classes. [144] proposed a Markov-based model for a simple priority scheme by differentiating CW_{min}, the exponential backoff factor and the maximum backoff stage with no constraint on CW_{max}. Instead of using the exponential backoff proposed in the standard, the model in [112] considers different sources with different fixed contention windows which are optimized as a function of the number of stations in the network as well as maintain a weighted fairness in terms of throughput among stations. A recent work [148] developed a

model to capture CW_{min} and CW_{max} differentiation with finite retransmission under unsaturated condition. This model modifies the Markov chain of [21] with an extra state added to capture the probability that the queue is empty when a packet is successfully transmitted. Obviously, these models only cover CW differentiation, which showed the need for the following 802.11e EDCA models with more differentiation parameters such as AIFS and TXOP limit.

Most of the current EDCA models extend DCF ones to capture the differences mentioned above. Therefore, most EDCA models are based on a fixed point system of attempt probabilities and collision probabilities. Correspondingly, the EDCA models can be classified by the approach of calculating attempt probability: explicitly solving Markov chain [39, 40, 52, 55, 56, 62, 68, 118, 133, 146, 154, 157], mean-value analysis [20, 84, 107, 149] or renewal reward theorem [113]. These models can also be classified into modeling saturated sources [52, 68, 72, 78, 84, 107, 113, 118, 133, 146, 149, 154, 157], unsaturated sources [20, 55, 56, 62] or both [39, 75]. Among those, some models only capture one differentiation parameter such as CW differentiation [146] or TXOP differentiation [55, 107]. Whereas, there has been much work concentrating on modeling both AIFS and CW differentiation [39, 68, 72, 75, 78, 84, 113, 118, 133, 154, 157]. Only a few models [20, 40, 52, 56, 62, 149] cover all of these three parameters.

Like DCF models, most EDCA models are based on a fixed point system, which is solved using fixed point iteration technique. Therefore, the importance of studying the uniqueness of the fixed point and convergence of the fixed point iteration still remains. However, very few works such as [113] investigate these issues while most EDCA models do not consider those due to their complication. In [113], the uniqueness of fixed point is analyzed in the case of CW differentiation and AIFS differentiation in saturated networks, which uses the proposed model based on the renewal reward theorem. In particular, for the case of CW differentiation, [113] provides conditions on the retransmission limit, exponential backoff factor, and the mean backoff slots at the first attempt so that the fixed point model has a unique

solution. For AIFS differentiation, [113] comes up with the same condition on the retransmission limit, exponential backoff factor and the mean backoff slot of each station as for CW differentiation so that the fixed point is unique. Note that my model in Chapter 3 capturing CW and TXOP differentiation is also based on a fixed point system. Like most of the existing EDCA models, my model assumes that the solution found by the fixed point iteration is unique.

In the following, I provide the detailed discussion of modeling each of MAC parameters in the previous work mentioned above. Note that I am interested in a model of 802.11e EDCA with CW differentiation and TXOP differentiation. However, to be complete, I also present how to model AIFS differentiation.

CW differentiation

CW differentiation provides service differentiation in that users of different CW_{min} or CW_{max} have different probability to transmit a packet in a given slot. Recall that CW determines the number of backoff slots a station has to wait before being allowed to transmit a packet/burst. Stations with smaller CW_{min} and CW_{max} will have higher probability to transmit a packet, which implies that they can gain higher share of link capacity. This parameter has been shown in [23, 146] to be able to provide service differentiation under different traffic load. However, [23] shows that at high traffic load, the tradeoff of the aggregate performance for service differentiation may become high due to excessive collision in any slots.

Modeling CW differentiation can be straightforwardly incorporated with the existing DCF models, regardless of the modeling approach. In particular, although different sources may have different CW_{min} and CW_{max}, which lead to different attempt probability, the methods to calculate the attempt probability and collision probability of each source are the same as those developed for DCF models. For Markov-based models, the attempt probability can be determined by solving the Markov chain, which is the total probability that a station is at the state with the backoff counter's value of 0. Models using mean-value analysis calculate the attempt

probability by taking the inverse of the mean number of backoff slots per stage while the attempt probability in those using renewal reward theorem is given as the ratio of the mean number of slots during which a given packet/burst is transmitted and the mean number of slots during its service time. Moreover, the collision of a station is also given by the complement of the probability that no other stations transmit in a given slot as calculated in the DCF models.

Due to its simplicity, most of EDCA models [20, 39, 40, 52, 56, 62, 68, 84, 113, 118, 133, 146, 149, 154, 157] capture CW differentiation.

AIFS differentiation

Recall that AIFS is the time that a station has to defer transmission after a busy period. The idea of using AIFS to provide service differentiation is to reserve some slots for higher priority sources. This means that there are some slots after a busy period where only sources of a higher priority source can access, which creates different contention zones. Then, stations of different AIFSs are different in the number of slots where they are allowed to transmit a packet right after a busy period. A station with smaller AIFS means that it is able to transmit in more slots right after a busy period. As a result, the benefit of AIFS will increase with traffic load due to the increase of busy period which makes the ratio of the number of slots where only stations of higher AIFS can transmit (these slots have smaller collision probability) and the number of slots between two consecutive busy periods increase.

To model AIFS differentiation, models have to capture different contention zones due to the fact that some slots after a busy period can only be accessed by users of certain classes. Although the models incorporating AIFS differentiation are still based on solving fixed point system of the collision probability and attempt probability of different sources of different ACs, the calculation of the collision probability should be different to capture the fact the collision probability of a particular source can be different in different slots after a busy period. There has been much work modeling AFIS differentiation [39, 40, 52, 56, 62, 68, 72, 75, 78, 84, 113,

118, 133, 149, 154, 157]. Among those, a few models actually capture the fact that collision probability may be different at different slots by calculating collision probability as the sum of the collision probability at different slots between two busy periods weighted by the probability of that slot which is obtained by solving a Markov chain which shows the transition of slots between two busy periods [40, 62, 68, 113, 118, 149], or using the collision probability at each slot directly [133, 154].

Work capturing AIFS differentiation in the calculation of the collision probability may capture that in the calculation of attempt probability. In particular, [133, 154] proposes three dimension Markov chain with an additional dimension of the physical time slot in an operation period (defined as the period between two busy periods). Then, attempt probability of a particular AC is given as a function of the time slot and similarly for collision probability. The others [40, 62, 68, 113, 118, 149] do not consider this because their approach only requires the conditional attempt probability (e.g. the probability that an AC transmits in a contention-allowed slot), which can be calculated using the same method as in DCF models. For example, [40, 68, 118] propose two dimension Markov chain of backoff counter and backoff stage used to determine the attempt probability. The model in [40] is an extension of [118], which considers more ACs (more than 2) and TXOP differentiation. [149] is based on mean-value analysis while Ramaiyan et al. [113] extends their DCF model [77], which is based on renewal reward theorem.

In contrast, many models [39, 56, 72, 75, 84, 157] do not capture AIFS differentiation in the calculation of collision probability but consider it in the calculation of the attempt probability instead. In particular, [56, 72] propose to add another dimension to the two dimension Markov chain, where the additional dimension denotes the remaining time during either the frozen, transmission, or collision period [72] or the remaining number of time slots during the deferring period between the minimum AIFS and the AIFS of a considered AC [56]. Differently, [39, 75] still keep two-dimension Markov chain where AIFS differentiation is captured in the probabil-

ity that the channel is busy in a given slot which decides whether state changes to lower backoff counter or remains the same. [39] and [75] are different in the way to calculate this probability. The model in [84] is based on mean-value analysis which calculates the attempt probability in any slot in which AIFS differentiation is captured by assuming slots during the AIFS of an AC is considered its additional backoff slots. Then, it calculates the additional backoff slots caused by the interruption of any station in a given slot.

TXOP differentiation

TXOP differentiation supports service differentiation by allowing users of different TXOP limit to transmit for different duration without contending again once they gain channel access. This means that stations with larger TXOP limit will transmit for a longer period of time and hence their throughput/delay can be improved. During a TXOP duration, multiple packets may be transmitted, which is called a "burst".

Unlike AFIS, TXOP differentiation does not receive much attention in prior work. In particular, [39] stated that TXOP can be easily captured by simply inflating the packet size (e.g. summing up subsequent packets per burst and the SIFS between them). However, it has been shown in [149] that creating an accurate model of TXOP differentiation requires more than simply inflating the packet length and is a nontrivial extension that requires careful consideration; i.e., the duration of a collision only involves the first packet in a burst.

Among those models which explicitly consider TXOP differentiation, [40, 52, 107, 149] model saturated traffic while [20, 55, 56, 62] consider unsaturated traffic. Note that modeling TXOP for saturated sources is much easier than that for unsaturated sources. This is because saturated sources always have packets waiting to transmit; hence, it always uses up the whole TXOP duration, which makes it easy to calculate number of packets sent per TXOP (hereafter called "burst size"). In contrast, the number of packets sent per TXOP of unsaturated sources depend on the number of

packets present in the queue which vary with time.

For saturated networks, the models in [40, 52, 107] are based on Markov chain developed in [21] to calculate the attempt probability of each source. While [107] only considers the immediate ACK bursting scheme, [52] considers three different bursting schemes including immediate ACK scheme, two block ACK schemes where multiple frames are either separated by SIFS or not. Also modeling saturated sources, the model in [149] is based on mean-value analysis, which considers immediate ACK bursting scheme.

To capture TXOP differentiation under unsaturated condition, [20] assumes that all stations send the same number of packets per successful transmission, which does not reflect the real behavior of unsaturated sources as mentioned above. In contrast, the models in [55, 56, 62] capture the fact that the number of packets sent per channel access vary with the queue occupancy of each unsaturated source. Although those are based on Markov chain, their approach to capture TXOP in the model is different. The model in [62] adds an additional dimension in the Markov chain to reflect the number of packets buffered for transmission at the MAC layer, which takes into account the loss due to exceeding retransmission limit and buffer limit. However, adding an extra dimension into the Markov chain makes it hard to solve and to obtain an explicit solution. Differently, [55, 56] propose a separate Markov chain for the queue of an unsaturated source, from which the distribution of burst size is calculated from the distribution of queue size. The method to determine the distribution of queue size in those works requires a burdensome matrix calculation on each iteration when solving the fixed point system. Besides, it does not capture the effect on the distribution of the loss probability due to exceeding retransmission limit. Moreover, the models in [55, 56, 62] miss an important aspect in the network of large TXOP limit, which is the residual time of an ongoing transmission from other stations seen by a burst of an unsaturated source arriving during that transmission as a component of the burst's delay. The importance of capturing this aspect will be shown in Chapter 3.

From the above discussion, there are only a few existing models which capture both CW and TXOP differentiation with heterogeneous traffic (e.g. both saturated and unsaturated). Among those, all work which properly models TXOP by taking into account the dependence of the actual number of packets per burst on the queue occupancy is Markov chain based, which is not tractable because it can be hard to obtain explicit solution to gain some analytical insight. Besides, they all do not consider an important aspect in the network of large TXOP limit, the residual time of an ongoing transmission from other stations seen by a burst of an unsaturated source arriving during that transmission as a component of the burst's delay. The work in this thesis will fill these gaps in the literature by proposing a tractable model of heterogeneous traffic (e.g saturated and unsaturated), based on the renewal reward theory in [77]. My model captures both CW and TXOP differentiation in which TXOP is properly taken into account through obtaining a closed form distribution of the burst size and considering the effect of the residual time of an ongoing transmission on the delay.

Moreover, the performance metrics investigated in the existing EDCA models are throughput and/or access delay for saturated sources, and delay (e.g. access delay or total delay) and/or packet loss probability for unsaturated sources. Only a few investigate the delay distribution [38, 117, 133, 149]. The delay distribution can be obtained using a computational approach based on the transient analysis of a Markov chain [133]. In [117], the cumulative distribution function (CDF) of access delay is achieved by expanding the probability generating function into a power series where the generating function is obtained by combining the signal flow graphs for backoff and transmission states and analyzing the path from the start to the end points of the signal transfer function of delay. [38, 149] use a more direct method to determine the delay distribution by inverting the generating function of the delay distribution, where the generating function in [149] is more detailed and accurate than that of [38]. To calculate the mean delay, [38] uses the first order derivative of the generating function while [149] derives it via direct probabilistic arguments.

My model in Chapter 3 calculates the throughput of saturated sources. It also determines the access delay of unsaturated sources by extending the method in [149]. However, instead of using the traditional approach of inverting the generating function to obtain the delay distribution, I propose a simple method to approximate the distribution of access delay in Chapter 3.

Also like DCF, most of EDCA models under unsaturated condition do not consider the probability that a packet arrives at an empty queue in the calculation of the access delay. In fact, my model presented in Chapter 3 considers this probability and I find that taking this into account can improve the accuracy of delay up to 25%.

2.4 Conclusion

In this chapter, the literature review of QoS provision in WLANs has been provided. I have found that most of the existing proposals to provide service differentiation are based on prioritization, which creates an incentive for selfish users to use the access class of the highest priority to gain a higher share of the channel. This can degrade the overall performance of the network and result in no service differentiation. I also found that the existing solutions to this issue are either complicated or impractical to implement, which shows the need for a scheme to provide QoS which is easy to implement, compatible with the 802.11e standard and robust against selfish users. This gap is filled by the proposed QoS scheme in this thesis, the analysis of which requires an 802.11e EDCA model. Therefore, the thesis also proposes a model of 802.11e EDCA WLANs, which will be presented in Chapter 3. The proposed model addresses the gaps in the literature which have been identified in this chapter.

Chapter 3

Model of IEEE 802.11e EDCA WLANs

3.1 Introduction

The purpose of the IEEE 802.11e EDCA mechanism is to provide service differentiation in WLANs. The "qualitative" implication of differentiation of MAC parameters may be clear (i.e., users with higher TXOP limit value can have higher throughput). However, it is also important to quantify how much service differentiation is provided with different parameters, which leads to the need for an 802.11e EDCA model.

The contribution in this chapter is a novel model of 802.11 EDCA WLANs with a mixture of saturated non-realtime sources which seek high throughput and unsaturated (Poisson) real-time sources which demand low delay, assuming no buffer overflow. The motivation is to enable the study of MAC mechanisms that improve service for both types of users by means of EDCA parameters: TXOP limit, CW_{min} and CW_{max}. I do not model the parameter AIFS because it provides load-dependent prioritization, which does not help to achieve the "fair" service differentiation.

Recall from Chapter 2 that there are only a few existing models which capture both CW and TXOP limit differentiation with heterogeneous traffic (e.g. both saturated and unsaturated) as the proposed model. I will show later in the chapter that there are two important aspects of large TXOP limit which can affect the accuracy of a model. Those are the dependence of the actual number of packets sent per channel access (called "burst size") on the queue occupancy, and the residual time of an ongoing transmission from other stations seen by a burst (e.g. a sequence of packets sent by a source per channel access) of an unsaturated source arriving

during that transmission as a component of the burst's delay.

The former aspect (e.g. the dependence of burst size on queue occupancy) is taken into account in the proposed model and a few prior models. The novelty of the proposed model is that it uses the renewal reward theory, which is tractable. In contrast, the prior models capturing this aspect are Markov chain based [56, 62], which is not tractable because it can be hard to obtain explicit solution to gain some analytical insight. Besides, the proposed model is the first to provide a closed form expression of the distribution of the burst size, which shows that when the TXOP limit for unsaturated sources is greater than one packet, bursts are approximately distributed as a geometric random variable clipped to TXOP limit. Another novel feature of the proposed model is that it captures the latter aspect (e.g. the residual time as a component of the delay), which is not considered in the previous models.

Moreover, the proposed model is also novel in that it considers the case when a burst from an unsaturated source arrives at idle channel (asynchronously) in the calculation of the access delay (e.g. duration between the instant when the burst reaches the head of the queue and begins contending for the channel, and the time when it is successfully received), which is not taken into account in the previous models. I find that this can have an effect of up to 25% on the accuracy of delay estimates when load is light.

Another contribution in this chapter is that based on the proposed model, asymptotic results for the distribution of access delay are provided. In particular, a simple method to approximate the distribution of the access delay with infinite retransmission limit is proposed. This allows us to approximately derive the slope of distribution's tail. Then, the lower bound on the number of saturated sources at which excessive queueing delay will be seen by unsaturated sources of arbitrary load is determined, on condition that all sources use the same MAC parameters.

The rest of the chapter is organized as follows. I first introduce notation and modeling assumptions in Section 3.2. Then, I present a model of EDCA WLANs in Section 3.3, which is validated in Section 3.4. The asymptotic results for the access

delay distribution are presented in Section 3.5.

3.2 Notation and modeling assumptions

I model an 802.11 EDCA WLAN with a set $\mathbb{U}$ of $N_u \geq 0$ unsaturated Poisson sources and a set $\mathbb{S}$ of $N_s \geq 1$ saturated, bulk data sources, which always have packets to transmit.

The model assumes an ideal channel so that packets are received correctly unless multiple sources transmit at the start of the same slot (a "collision"). Sources do not use RTS/CTS. All packets from a given source have equal size, and unsaturated sources can accommodate an arbitrary number of packets. All stations use the same AIFS.

In the following description of notation, $s \in SS$, $u \in \mathbb{U}$ and $x, y \in SS \cup \mathbb{U}$ denote arbitrary sources, $U[a, b]$ denotes an integer uniformly distributed on $[a, b]$, $A \sim B$ denotes that A and B are equal in distribution, and $\mathbb{E}[\cdot]$ is ensemble average.

Source x emits packets of constant size l_x in bursts of a (possibly random) number of packets η_x, bounded above by the constant r_x.

Packets arrive at a source u as a Poisson process of rate λ_u and are queued. Source u has a packet to transmit a fraction ρ_u of the time. If a packet arrives when u has no packets to transmit, then with probability denoted $1 - b_u$ it observes the channel idle and transmits immediately. Such arrivals (termed "asynchronous") do not experience collisions, due to carrier sensing by the other stations at the start of the next slot.

The backoff mechanism imposes a slotted structure on time, with slot sizes independently distributed as a random variable Y, which is σ if the slot is idle or longer if a transmission is attempted. In each slot, x attempts to transmit with an "attempt probability" denoted by τ_x and, conditional on making an attempt, collides with a "collision probability" denoted by p_x. Following [21], these are assumed independent of the number of previous attempts of this packet, or packets from other

stations. If the first packet in the burst collides, the remainder are not transmitted. Transmissions of subsequent packets in a burst, not subject to contention, are not considered "attempts".

Each burst is attempted up to K times, with the jth attempt occurring after a backoff of $U_{xj} \sim U[0, 2^{\min(j,m)}W_x - 1]$ slots, where W_x is called the minimum contention window. I assume U_{xj} is independent of random variables mentioned above. The size of a slot *conditioned* on source u performing a backoff is distributed as Y_u.

With probability L_x, all attempts of a burst suffer collisions, in which case the first packet is discarded.

Slots that are idle, collisions and successful transmissions are denoted by superscripts i, c, and s.

Note that the thesis considers EDCA with the immediate ACK bursting scheme. Then, the (random) time that a burst sent by a source x occupies the channel if it is successfully transmitted is given by

$$T_x^s = AIFS + \eta_x(T_{\mathrm{p}x} + T_{\mathrm{ACK}}) + (2\eta_x - 1)SIFS \tag{3.1}$$

where T_{ACK} is the duration of an ACK packet, and $T_{\mathrm{p}x}$ is the transmission time of a packet from the source x. The deterministic value of T_x^s conditioned on $\eta_x = 1$ is denoted T_x.

The duration of a collision slot is the maximum of T_x over all sources x involved in the collision.[1]

3.3 Model

I now present a model that takes the system parameters W_x, r_x, $T_{\mathrm{p}x}$, and λ_u, as input, and predicts the throughput of a source $s \in SS$ and the access delay of a

[1]This is because stations involved in the collision wait for the ACK as usual, and other stations wait for an EIFS [7].

source $u \in \mathbb{U}$.

Without loss of generality, sources are indexed in non-increasing order of their packet duration, regardless of whether they are saturated or unsaturated. That is, $T_x \geq T_y$ for $x < y$.

3.3.1 Fixed point model

The model is a set of fixed-point equations, where the collision probabilities are expressed in terms of the attempt probabilities, and vice versa. I will now derive the fixed point equations which will be presented in (3.9) below.

First, to determine the collision probability, denote the probability that no sources transmit in a given slot by

$$G = \prod_{x \in \mathbb{S} \cup \mathbb{U}} (1 - \tau_x). \tag{3.2}$$

The collision probability of a given source $x \in \mathbb{S} \cup \mathbb{U}$ is

$$p_x = 1 - \frac{G}{1 - \tau_x}, \tag{3.3}$$

which is based on the common approximation [20, 56, 85, 155] that all bursts from a source x have the same collision probability at each attempt. This includes unsaturated sources, regardless of whether they arrive asynchronously or not.

Second, the attempt probability of a saturated source s is the mean number of attempts per burst divided by the mean number of slots per burst

$$\tau_s = \frac{\sum_{k=0}^{K} p_s^k}{\sum_{k=0}^{K} (\mathbb{E}[U_{sk}] + 1) p_s^k} \tag{3.4}$$

where at each stage, one slot is used for transmission and the mean number of backoff slots is

$$\mathbb{E}[U_{sk}] = 2^{\min(k,m)-1} W_s - 1/2 \tag{3.5}$$

for U_{sk} of uniform distribution mentioned in Section 3.2.

Next, I determine τ_u, the attempt probability of an unsaturated source u. First, consider the number of packets u "serves" for each burst formed. With probability $L_u = p_u^{K+1}$, the first packet in the burst is discarded. Otherwise, u successfully sends on average $\mathbb{E}[\eta_u]$ packets. (The latter depends on the queue size distribution at the node; for light load, $\mathbb{E}[\eta_u] = 1$, and in general it is given by (3.35) in Section 3.3.4.) Thus bursts are formed at rate

$$\frac{\lambda_u}{L_u + (1 - L_u)\mathbb{E}[\eta_u]}. \tag{3.6}$$

Next, determine the mean number of attempts per burst from u under the usual approximation [20, 56, 85, 155] that all bursts contend for the channel, even if they arrive asynchronously. The mean number of attempts is then approximated by

$$1 + \sum_{j=1}^{K} p_u^j = \frac{1 - p_u^{K+1}}{1 - p_u}. \tag{3.7}$$

Simulations suggest this is reasonably accurate, which appears to be due to the presence of saturated sources. This approximation is not required in the delay model of Section 3.3.3.

From (3.6), (3.7) and the fact that there are $1/\mathbb{E}[Y]$ slots per second, the attempt probability of the source u is the total number attempts per second divided by the number of slots per second:

$$\tau_u = \frac{\lambda_u}{L_u + (1 - L_u)\mathbb{E}[\eta_u]} \frac{1 - p_u^{K+1}}{1 - p_u} \mathbb{E}[Y] \tag{3.8}$$

A special case of Eq. (3.8) in 802.11 DCF WLANs without saturated sources coincides with the model of [155].

The fixed point is between the collision probabilities in (3.3) and the attempt probabilities derived from (3.4) and (3.8):

$$\tau_s = 2(1 - p_s^{K+1}) / \Big(W_s(1 - (2p_s)^{m+1}) \frac{1 - p_s}{1 - 2p_s}$$

$$+ (2^m W_s + 1)(1 - p_s^{K+1}) - 2^m W_s (1 - p_s^{m+1})\Big) \tag{3.9a}$$

$$\tau_u = \frac{\lambda_u}{L_u + (1 - L_u)\mathbb{E}[\eta_u]} \mathbb{E}[Y] \frac{1 - p_u^{K+1}}{1 - p_u} \tag{3.9b}$$

$$p_x = 1 - \frac{G}{1 - \tau_x}. \tag{3.9c}$$

It remains to determine the mean slot time $\mathbb{E}[Y]$. This can be expressed in terms of the probabilities P^i, P_x^s and P_x^c that a given slot contains (a) no transmissions, (b) a successful burst transmission from source x, or (c) a collision involving the source x and only sources $y > x$ with packets no larger than T_x. Specifically,

$$\mathbb{E}[Y] = P^i \sigma + \sum_{x \in \mathbb{S} \cup \mathbb{U}} P_x^s \mathbb{E}[T_x^s] + \sum_{x \in \mathbb{S} \cup \mathbb{U}} P_x^c T_x \tag{3.10a}$$

$$P^i = G \tag{3.10b}$$

$$P_x^s = \frac{\tau_x}{1 - \tau_x} G \tag{3.10c}$$

$$P_x^c = \frac{\tau_x}{1 - \tau_x} \left(\prod_{y \le x} (1 - \tau_y) - G \right) \tag{3.10d}$$

$$\mathbb{E}[T_x^s] = AIFS + \mathbb{E}[\eta_x](T_{\mathrm{p}x} + T_{\mathrm{ACK}}) + (2\mathbb{E}[\eta_x] - 1) SIFS. \tag{3.10e}$$

Note that all $N_s + N_u$ values of P_x^c can be calculated in $O(N_s + N_u)$ time, by the nested structure of the products in (3.10d).

The fixed point (3.9) involves $\mathbb{E}[\eta_u]$ and $\mathbb{E}[Y]$. For light load, $\mathbb{E}[\eta_u] = 1$; hence, solving (3.9) requires only (3.10). In general, $\mathbb{E}[\eta_u]$ is given by (3.35) derived from the delay model; hence, the delay model in Section 3.3.3 forms part of the fixed point.

Simpler form for $K = m = \infty$ Although the retry limit K for non-RTS/CTS mode is 7 in IEEE 802.11 standard [10], in many settings a source rarely uses all seven retransmissions. In that case, it is reasonable to reduce the complexity of the model by approximating K and m as infinite. Then, the fixed point (3.9) simplifies

to

$$\tau_s = \frac{2}{W_s \frac{1-p_s}{1-2p_s} + 1}, \qquad s \in \mathbb{S} \tag{3.11a}$$

$$\tau_u = \frac{\lambda_u}{\mathbb{E}[\eta_u]} \mathbb{E}[Y] \frac{1}{1-p_u}, \qquad u \in \mathbb{U} \tag{3.11b}$$

$$p_x = 1 - \frac{G}{1-\tau_x}, \qquad x \in \mathbb{S} \cup \mathbb{U}. \tag{3.11c}$$

3.3.2 Throughput of saturated sources

The throughput in packets/s of a saturated source $s \in \mathbb{S}$ is the average number of packets successfully transmitted per slot divided by the average slot length [21]

$$S_s = \frac{\mathbb{E}[\eta_s]\tau_s(1-p_s)}{\mathbb{E}[Y]}. \tag{3.12}$$

where $\mathbb{E}[\eta_s]$ is the average number of packets per burst given by (3.28), and the rest of the numerator is the probability the source s successfully transmits a burst in a given slot.

3.3.3 Delay model

I now calculate the access delay of bursts from an unsaturated source. This is not only an important performance metric, but also used to determine $\mathbb{E}[\eta_x]$ in (3.9).

I first propose an access delay model for a burst that arrives at an empty transmission queue. Recall that my delay model captures two important features in this case: the case when the burst arrives at idle channel with the probability b_u, and the residual time $T_{\text{res},u}$ of the busy period during which the burst arrives.

Let D_u be the random access delay of a burst from an unsaturated source $u \in \mathbb{U}$. Also let F_u be the random total backoff and collision time of the burst before it is successfully transmitted. Then,

$$D_u = T_u^s + F_u \tag{3.13}$$

where T_u^s, given by (3.1), is random since η_u is random. F_u has the distribution

$$F_u = \begin{cases} 0 & \text{w.p. } \dfrac{1-b_u}{1-b_u+b_u(1-p_u^{K+1})} \\ F_{uk} & \text{w.p. } \dfrac{b_u p_u^k(1-p_u)}{1-b_u+b_u(1-p_u^{K+1})},\ K \geq k \geq 0 \end{cases} \tag{3.14}$$

in which F_{uk} is the random total backoff and collision time of the burst provided that it is successfully transmitted in the kth backoff stage. The remainder of the complexity of the delay model comes from estimating the duration of the backoff slots which comprise F_{uk}. Write

$$F_{uk} = \sum_{j=0}^{k} B_{uj} + \sum_{j=1}^{k} T_u^c + T_{\text{res},u} \tag{3.15}$$

where T_u^c is the random duration of a collision involving u, and the random backoff time in the jth stage is

$$B_{uj} = \sum_{k=1}^{U_{uj}} Y_{u,k}. \tag{3.16}$$

Here U_{uj} is the number of backoff slots in the jth backoff stage, and the $Y_{u,k} \sim Y_u$ are the independent, identically distributed (i.i.d.) durations of a slot conditional on source u not transmitting, namely

$$Y_u = \begin{cases} \sigma & \text{w.p. } P_u^i \\ T_x & \text{w.p. } P_{xu}^c, \quad x \in \mathbb{S} \cup \mathbb{U} \setminus \{u\} \\ T_x^s & \text{w.p. } P_{xu}^s, \quad x \in \mathbb{S} \cup \mathbb{U} \setminus \{u\} \end{cases} \tag{3.17}$$

where P_u^i, P_{xu}^c and P_{xu}^s are the probabilities, conditional on u not transmitting, of an idle slot, a collision between a source x and sources $y > x$ with packets no larger than T_x, and a success of a burst from a source x. P_u^i and P_{xu}^s are obtained by dividing the analogous quantities in (3.10b)–(3.10c) by $1-\tau_u$ while P_{xu}^c is given by

$$P_{xu}^c = \frac{\tau_x}{1-\tau_x}\left(\prod_{y \leq x, y \neq u}(1-\tau_y) - \frac{G}{1-\tau_u}\right). \tag{3.18}$$

Note that this is not $P_x^c/(1-\tau_u)$ because from Bayes' theorem, the conditional prob-

ability P_{xu}^c is given by the probability that a collision slot is equal to T_x and source u does not transmit divided by the probability that source u does not transmit.

The random collision time T_u^c is the duration of the longest packet involved in a collision involving source u,

$$T_u^c = \max(T_u, T_x) \quad \text{w.p.} \quad P_{xu}^{cu}, \quad x \in \mathbb{S} \cup \mathbb{U} \setminus \{u\} \tag{3.19}$$

where P_{xu}^{cu} is the probability that the source u collides with the source x and possibly sources $y > x$ with packets no larger than T_x, given by

$$P_{xu}^{cu} = \frac{\tau_x}{1 - P_u^i} \prod_{\substack{y<x \\ y \neq u}} (1 - \tau_y). \tag{3.20}$$

Finally, the probability b_u can be estimated as the fraction of busy slots as follows.

$$b_u = 1 - \frac{P_u^i \sigma}{\mathbb{E}[Y_u]} \tag{3.21}$$

Mean access delay

From (3.13), the mean access delay is

$$\mathbb{E}[D_u] = \mathbb{E}[T_u^s] + \mathbb{E}[F_u]. \tag{3.22}$$

where $\mathbb{E}[T_u^s]$ is given by (3.10e) and $\mathbb{E}[F_u]$ is

$$\begin{aligned}
\mathbb{E}[F_u] \approx & \frac{b_u}{1 - b_u + b_u(1 - p_u^{K+1})} \Bigg(\frac{W_u}{2} \Big(\frac{2(1 - (2p_u)^{m+1})(1 - p_u)}{1 - 2p_u} - 1 + p_u^{m+1} \\
& + (-1 + 2^{m+1} - m2^m)(p_u^{m+1} - p_u^{K+1}) \\
& + 2^m \Big(\frac{p_u^{m+1} - p_u^{K+1}}{1 - p_u} + m p_u^{m+1} - K p_u^{K+1} \Big) \Big) \mathbb{E}[Y_u] \\
& + \mathbb{E}[T_u^c] \Big(\frac{1 - p_u^K}{1 - p_u} p_u - K p_u^{K+1} \Big) + \mathbb{E}[T_{\text{res},u}](1 - p_u^{K+1}) \Bigg).
\end{aligned} \tag{3.23}$$

The detailed derivation of (3.23) is provided in Appendix A.1. Note that the approximation in (3.23) comes from the approximation in (A.5).

The mean slot duration $\mathbb{E}[Y_u]$ observed by the source u and the mean collision delay $\mathbb{E}[T_u^c]$ can be found from (3.17) and (3.19), respectively. The mean residual time $\mathbb{E}[T_{\text{res},u}]$ is given by [71]

$$\mathbb{E}[T_{\text{res},u}] = \frac{\mathbb{E}[Y_u^b]}{2} + \frac{\text{Var}[Y_u^b]}{2\mathbb{E}[Y_u^b]}, \tag{3.24}$$

where Y_u^b is the duration of a busy period caused by transmissions of other sources. Its distribution is similar to that of Y_u of (3.17), conditioned on the slot not being idle.

Simpler form for $K = m = \infty$ The mean access delay again simplifies when K and m are infinite, becoming

$$\mathbb{E}[F_u] \approx b_u \Bigg(\left(\frac{1}{2(1-2p_u)} \right) W_u \mathbb{E}[Y_u] + \frac{\mathbb{E}[Y_u]}{2(1-p_u)} + \frac{p_u}{1-p_u} \mathbb{E}[T_u^c] + \mathbb{E}[T_{\text{res},u}] \Bigg). \tag{3.25}$$

Remark 1 *Although $\mathbb{E}[Y_u]$ and $\mathbb{E}[Y_u^b]$ can be calculated using (3.17), it is simpler to use*

$$\mathbb{E}[Y_u] = \frac{\mathbb{E}[Y] - P_u^s \mathbb{E}[T_u^s] - \mathbb{E}[T_u^c]\tau_u p_u}{1-\tau_u}, \tag{3.26}$$

which comes from the fact that Y_u is Y excluding components involving the source u which are successful transmission of u or collision involving u and the fact that the probabilities a slot is idle, contains a successful transmission, or contains a collision among an arbitrary number of sources of Y_u are similar to those of Y scaled by $1-\tau_u$.

Then, $\mathbb{E}[Y_u^b]$ is given from $\mathbb{E}[Y_u]$ as

$$\mathbb{E}[Y_u^b] = \frac{\mathbb{E}[Y_u] - \sigma P_u^i}{1 - P_u^i}. \tag{3.27}$$

However, the form (3.17) is needed to calculate $Var[Y_u^b]$, and the distribution of delay as done in Appendix A.2.

Under high load, a burst of an unsaturated source is likely to see a non-empty queue when arriving. Hence, it will have queueing delay in addition to access delay.

One of the methods to calculate the mean queueing delay is to use the P-K formula [71] for an M/G/1 queue with the mean and variance of the service time. Using the proposed access delay model above for service time, some preliminary numerical results of queueing delay obtained from this method are not very accurate, but investigating this is out of scope of the thesis.

Recall that the above delay model is for the case a burst arrives mostly at an empty transmission queue. In practice, the network can be of any load; hence, it is possible that a burst often arrives at non-empty queue. To see how the access delay model above is applicable in the presence of queueing, consider the following three possibilities that a burst of an unsaturated source can observe when arriving at the transmission queue.

- Empty queue and channel idle for AIFS. For this case, $F_u = 0$ as in the first case of (3.14).
- Empty queue but channel not idle for AIFS. For this case, $F_u = F_{uk}$ with F_{uk} given in (3.15).
- Non-empty queue. For this case, $F_u = F_{uk}$ with F_{uk} given in (3.15) but without $\mathbb{E}[T_{\text{res},u}]$.

The last two cases can be approximated by the second term of (3.14) when $\mathbb{E}[T_{\text{res},u}]$ is small. The probability of $F_u = 0$ is slightly over-estimated by (3.14), but this effect is small at high load, since $b_u \to 1$ as load increases. It is confirmed by simulation in Section 3.4 that (3.14) is often a good approximation for delay at high load.

Note that the above delay model becomes inaccurate in the uncommon case that $\mathbb{E}[T_{\text{res},u}]$ is significant compared with the access delay, which occurs when the arrival rate from source u is high while the arrival rate from other stations is light and other stations use very large TXOP limit. A more accurate but less tractable model considers all of three possibilities of a burst of an unsaturated source when it arrives at the transmission queue as mentioned above, and hence is obtained by replacing

(3.15) and (3.14) respectively by

$$F'_{uk} = \sum_{j=0}^{k} B_{uj} + \sum_{j=1}^{k} T_u^c$$

$$F'_u = \begin{cases} 0 & \text{w.p.} \quad (1-b_u)(1-\rho_u)/\Theta \\ F'_{uk} + \mathbb{E}[T_{\text{res},u}] & \text{w.p.} \quad b_u(1-\rho_u)/\Theta \\ F'_{uk} & \text{w.p.} \quad p_u^k(1-p_u)\rho_u/\Theta \end{cases}$$

where $\Theta = (1-b_u)(1-\rho_u) + (1-(1-b_u)(1-\rho_u))(1-p_u^{K+1})$.

3.3.4 Distribution of burst size

Recall that the fixed point (3.9) involves the mean burst size $E[\eta_u]$. To solve the fixed point, we need to find the expression of the mean burst size as a function of other variables in the fixed point. In this section, I will first determine the distribution of burst size and then calculate the mean burst size from the distribution.

Saturated sources

The burst size η_s of a saturate source s is a constant and equal to r_s, the maximum number of packets that fit in TXOP limit of the source s. This is because a saturated source always has a packet waiting to transmit.

In particular, by (3.1),

$$\eta_s = r_s = \left\lfloor \frac{\text{TXOP limit} - AIFS + SIFS}{T_{px} + T_{\text{ACK}} + 2SIFS} \right\rfloor. \tag{3.28}$$

Non-saturated sources

A non-saturated source u will send in bursts up to r_u or the number of packets in the queue, whichever is less. To estimate the distribution of these burst sizes, I first model the queue size process. Note that in this model, packets arrive separately. In practice, packets may arrive in bursts. The model could be extended to one such as [108], but that is out of the scope of this thesis.

Distribution of queue size Model the queue size process as the Markov chain in Fig. 3.1, with state $k = 0, 1, 2 \ldots$ corresponding to having k packets in the queue. From state k, there are transitions at rate λ_u to state $k+1$ corresponding to packet arrivals. From state $k \geq 1$, there are transitions to state $k-1$ at rate $\mu_u L_u$, corresponding to the loss of a single packet due to excess collisions. In states $k = 1, \ldots, r_u$, all packets can form a single batch, and so there are transitions to state 0 at rate $\mu_u(1 - L_u)$ due to the successful transmission of this batch. In states $k > r_u$, each batch consists of r_u packets and so there are transitions to state $k - r_u$ at rate $\mu_u(1 - L_u)$. Note that this Markov approximation is only useful for estimating the queue distribution for low occupancies. I will show in Section 3.5 that the tail of the service time distribution can be heavy, which means this Markov approximation does not capture the tail properties of the queue size. However, the burst size distribution does not depend on the tail; therefore, the results provided here to estimate the burst size are still useful.

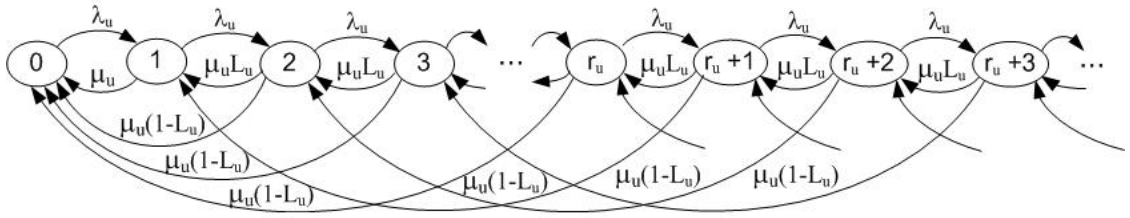


Figure 3.1: The transition diagram of queue size of an unsaturated source u.

In the above Markov chain, the total service rate at each state is the same and determined by

$$\mu_k = \mu_u = 1/\mathbb{E}[D_u], \quad \forall k \geq 1 \tag{3.29}$$

where μ_k is the total service rate at state k; μ_u is the mean service rate of source u; $\mathbb{E}[D_u]$ is given by (3.22).

As noted in [54], the service rate may actually differ between states. However, as will be shown by simulation below, the approximation of constant service rate is

actually more accurate than the approximation in [54] under the considered circumstances, as well as being more tractable.

Let Q_u be a random variable representing the queue size of an unsaturated source u in this Markov model.

Observe that Fig. 3.1 is similar to that of bulk service systems in [71], except there is an additional transition from every state k to the previous state $k-1$ which represents the case when the head of queue packet is dropped due to exceeding the retry limit. This suggests the following result.

Theorem 3.1 *If* $0 < \lambda_u < \mu_u(L_u + r_u(1 - L_u))$ *then the above Markov chain has a geometric steady state distribution,*

$$P[Q_u = k] = \Big(1 - \frac{1}{z_0}\Big)\Big(\frac{1}{z_0}\Big)^k, \quad k = 0, 1, 2, \cdots \tag{3.30}$$

where $z_0 > 1$ *is a solution of*

$$\rho_u z^{r_u+1} - (1 + \rho_u)z^{r_u} + L_u z^{r_u-1} + 1 - L_u = 0 \tag{3.31}$$

where $\rho_u = \lambda_u/\mu_u$.

Proof: The proof decomposes the transition matrix A of the Markov chain as the sum of those of an M/M/1 queue and a bulk service queue, with equal steady state distributions.

Let A'_x be the transition matrix of an M/M/1 queue with service rate $L_u\mu_u$ and arrival rate $x\lambda_u$, and A''_x be the transition matrix of a bulk service queue [71] with service rate $(1 - L_u)\mu_u$ and arrival rate $(1 - x)\lambda_u$. For $x \in (0, L_u\mu_u/\lambda_u)$, the M/M/1 queue has geometric steady state probabilities Q'_x whose mean q'_x increases continuously from 0 to ∞. For $x \in (1 - (1 - L_u)\mu_u/\lambda_u, 1)$, the bulk service queue has geometric steady state probabilities Q''_x whose mean q''_x decreases continuously from ∞ to 0. Let (a, b) be the intersection of those intervals. This is non-empty by the upper bound on λ_u. Then $q'_x - q''_x$ increases continuously on (a, b). It is negative as $x \to a$, as either $q'_a = 0$ if $a = 0$ or $q''_x \to \infty$ as $x \to \infty$ if $a > 0$. Similarly, it is

positive as $x \to b$. Hence there is an $\tilde{x} \in (a, b) \subseteq (0, 1)$ such that $Q'_{\tilde{x}} = Q''_{\tilde{x}}$. Then $0 = Q'_{\tilde{x}}(A' + A'') = Q'_{\tilde{x}}A$, and so the geometric distribution $Q'_{\tilde{x}}$ is the steady state distribution of the original Markov chain.

Substituting(3.30) into the following balance equation of the Markov chain

$$\begin{aligned}&(\lambda_u + \mu_u L_u + \mu_u(1 - L_u))P[Q_u = k] = \\ &\lambda_u P[Q_u = k - 1] + \mu_u L_u P[Q_u = k + 1] + \mu_u(1 - L_u)P[Q_u = k + r_u]\end{aligned} \tag{3.32}$$

gives

$$(1 - \frac{1}{z_0})(\frac{1}{z_0})^{k-1}(\mu_u(1 - L_u)(\frac{1}{z_0})^{r+1} + \mu_u L_u(\frac{1}{z_0})^2 - (\lambda_u + \mu_u)\frac{1}{z_0} + \lambda) = 0 \tag{3.33}$$

Dividing (3.33) by $\mu_u(1-\frac{1}{z_0})(\frac{1}{z_0})^{k-1}$ and then multiplying by $(1/z_0)^{r_u+1}$ gives (3.31). Then, z_0 in (3.30) is the solution greater than 1 of (3.31). ■

Distribution of burst size Here I determine the distribution of burst size η_u of an unsaturated source u, which is a function of the queue size. Since the transmission rate is equal (μ_u) in each state, the distribution of burst size η_u is equal to that of $\min(Q_u, r_u)$ conditioned on $Q_u \geq 1$, which has complementary cumulative distribution function (ccdf)

$$P[\eta_u > k] = \begin{cases} (1/z_0)^k & 0 \leq k < r_u \\ 0 & k \geq r_u. \end{cases} \tag{3.34}$$

Then, the mean burst size is the sum of its ccdf as follows.

$$\mathbb{E}[\eta_u] = \sum_{k=0}^{\infty} P[\eta_u > k] = \sum_{k=0}^{r_u-1}(1/z_0)^k = \frac{1 - (1/z_0)^{r_u}}{1 - 1/z_0} \tag{3.35}$$

To justify the need for my method of burst size calculation, I will next discuss the common method used in [54, 55, 56] and compare that with mine.

Comparison with other work [54] proposed a Markov chain of the queue size

similar to the above except that it (a) assumes different service rates for different states, (b) ignores the transition when the retry limit is exceeded, and (c) has a finite buffer. Then, the distribution of queue size Q_u is determined by numerically solving balance equations and the distribution of burst size is approximated by the (time average) distribution of $\min(Q_u, r_u)$ conditioned on $Q_u > 0$. One drawback of that approach is that it does not admit a closed-form solution for the distribution. Hence, it is computationally costly due to matrix calculation on each iteration when solving the fixed point, especially when the buffer size is large.

Using the fixed-point model (3.9)–(3.10), I investigate the mean burst size $\mathbb{E}[\eta_u]$ determined from two Markov chains of queue size distribution: mine in Fig. 3.1 and the one in [54]. In particular, I compare the performance of two approaches in the scenarios where both approaches hold: L_u is assumed to be 0 and the buffer capacity is set to be large (100 packets). The highest difference in $\mathbb{E}[\eta_u]$ between two Markov chains occurs when the network load is light and the arrival rate of source u is reasonably high. I simulate such a scenario, specifically one with one saturated source and one unsaturated source with the arrival rate changing from small to large.

It is not explicitly stated in [54] how the service rate in each state is determined. Since it is constant for states greater than r_u, I assume that the service rate at state k satisfies

$$1/\mu_k = \mathbb{E}[F_u] + T_u^s|_{\eta_u=k}, \quad \forall k \geq 1 \tag{3.36}$$

where $T_u^s|_{\eta_u=k}$ is the duration of a successful transmission of a burst of k packets, given by (3.1) with $\eta_u = k$.

The results in Fig. 3.2 shows that $\mathbb{E}[\eta_u]$ from my Markov chain is closer to the simulation than that from the Markov chain of [54]. At this light load, the truncation to an occupancy of 100 packets is insignificant, and $L_u = 0$; hence, the two Markov chains only differ in whether the service rate μ_k is constant or given by (3.36). I believe the inaccuracy of [54] is because (3.36) neglects the fact that some fraction

of the access delay $\mathbb{E}[F_u]$ has already elapsed by the time state k is reached, and so should not be reflected in (the reciprocal of) the transition rate. Since the true mean transmission time is the sum of an increasing term and a decreasing term, it is not clear *a prior* whether the constant rate μ_u or the increasing rate (3.36) would be a better model.

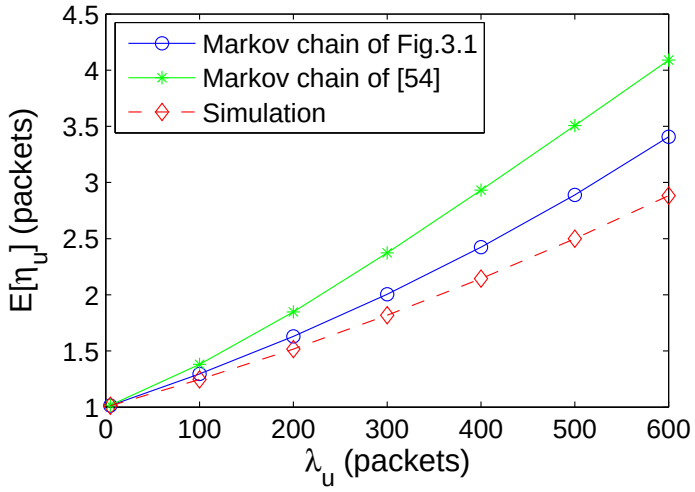


Figure 3.2: The average burst size $\mathbb{E}[\eta_u]$ of an unsaturated source u as a function of its arrival rate λ_u. (Unsaturated stations: Poisson arrivals with rate λ_u, $N_u = 1$, $l_u = 100$ Bytes, $W_u = 32$, $r_u = 7$; Saturated stations: $N_s = 1$, $l_s = 1040$ Bytes, $W_s = 32$, $\eta_s = 1$.)

Another possible source of error is in obtaining the burst size distribution from the queue occupancy distribution. In [54] the burst size distribution was approximated by the *time average* distribution of $\min(Q_u, r_u)$ conditioned on $Q_u > 0$. However, the burst size depends on the queue size not at a typical point in time, but at a service instant. Thus, the weights given to different queue occupancies should be proportional to $\mu_k P[Q_u = k]$, rather than $P[Q_u = k]$. In my model, μ_k is independent of k and so these become equivalent.

3.3.5 Model summary

My model from previous sections is summarized as follows.

At low load, $\mathbb{E}[\eta_u] = 1$ for $u \in \mathbb{U}$; hence, the fixed point consists of (3.9), (3.10) and (3.28).

At high load, $\mathbb{E}[\eta_u]$ ($u \in \mathbb{U}$) depends on the distribution of queue size which involves the access delay; hence, the fixed point includes not only (3.9), (3.10) and (3.28) but also the delay model (3.13)–(3.24) and the burst size model (3.29)–(3.35).

The outputs p_x, τ_x, S_s and $\mathbb{E}[D_u]$ can be determined by iteratively solving the fixed point numerically and applying (3.12).

Consistency of the model

For my model to be physically meaningful, the rate of successful channel accesses per second of source u should be less than that of a saturated source with the same CW_{min}, m, and K.[2] When all sources have equal CW_{min}, m, and K, this implies that for all $s \in \mathbb{S}$ and $u \in \mathbb{U}$,

$$\frac{\lambda_u}{\mathbb{E}[\eta_u]} < \frac{S_s}{\mathbb{E}[\eta_s]}. \tag{3.37}$$

For situations where the burst arrival rate $\lambda_u/\mathbb{E}[\eta_u]$ does not satisfy (3.37), an alternate instance of model (3.9)–(3.37) should be used, in which source u is replaced by a saturated source.

3.4 Numerical Evaluation and Discussion

To validate the model (3.9)–(3.10),(3.13)–(3.24),(3.28)–(3.35), and (3.12), it was compared with simulations (using *ns-2.33* [1] and [139]) and, where possible, two existing models [85], [91]. Note that EDCA implemented in [139] uses the immediate ACK bursting scheme.

I simulated networks of unsaturated and saturated sources sending packets to an access point using DCF and EDCA. All sources use UDP. Saturated sources receive

[2]It is not trivial that a saturated source achieves higher throughput than an unsaturated one; a network of only unsaturated sources can obtain a higher throughput than one of saturated sources [21, Fig. 3] because of the lower collision rate. However, within a given network, a saturated source gets a higher throughput than an unsaturated one with the same parameters.

Table 3.1: MAC and PHY parameters for 802.11b systems

Parameter	Symbol	Value
Data bit rate	r_{data}	11 Mbps
Control bit rate	r_{ctrl}	1 Mbps
PHYS header	T_{phys}	192 μs
MAC header	l_{mac}	288 bits
UDP/IP header	l_{udpip}	160 bits
ACK packet	l_{ACK}	112 bits
Slot time	σ	20 μs
SIFS		10 μs
AIFS, DIFS		50 μs
Retry limit	K	7
Doubling limit	m	5
Buffer capacity		50 packets

constant bit rate (CBR) traffic faster than they can transmit. Unsaturated sources use either Poisson or "quasi-periodic" traffic. By "quasi-periodic", I mean unsaturated sources with the packet inter-arrival times set to be uniformly distributed in the range $1/\lambda_u \pm 1\%$. This quasi-periodic model represents voice traffic (which is often treated as periodic CBR traffic [93]), subject to jitter such as that caused by the operating system. Explicitly including this jitter is necessary to avoid "phase effect" artifacts in the results. I use the 802.11b parameters in Table 3.1. The $T_{\text{p}x}$ and T_{ACK} in (3.1) are

$$T_{\text{p}x} = T_{phys} + \frac{l_{mac} + l_{udpip} + l_x}{r_{data}}, \quad x \in \mathbb{S} \cup \mathbb{U}$$
$$T_{\text{ACK}} = T_{phys} + l_{\text{ACK}}/r_{ctrl}.$$

Simulation results are shown with 95% Student-t confidence intervals [119]. In some figures, the confidence intervals are too small to be seen.

3.4.1 Validation and comparison with existing DCF models

As summarized in Section 2.3.1, there have been quite a few 802.11 DCF models which capture heterogenous traffic. Among those, two models [85] and [91] are

chosen as representatives for comparison with the proposed model because they use different modeling approaches and explicitly state how to modify an unsaturated model to also capture saturated sources. To apply my model to DCF, I adjusted the backoff decrement rule by replacing T_x^s and T_x in (3.10a) and (3.17) by $(T_x^s + \sigma)$ and $(T_x + \sigma)$.

Summary of two benchmark models

I first recall the models in [85] and [91].

Markov chain The model in [91] is based on a Markov chain similar to that of [21], with additional states for unsaturated sources. It assumes that unsaturated sources have minimal buffers; therefore, when a packet arrives at a busy source, it will be dropped. This causes the collision probability computed from this model to be smaller than that of models with non-zero buffers, such as my model.

Mean-based In [85] the mean-based approach is used for heterogeneous traffic where the attempt probability of an unsaturated source is multiplied by the probability ρ that the source has a packet to send. For saturated sources, $\rho = 1$. Unsaturated sources are assumed to have infinite buffers.

It will be shown later in Figs. 3.3 and 3.4 that the results of this model are not very accurate in settings I consider. In [85], the queue utilization involves the service time of an unsaturated source. Therefore, the fixed point involves the calculation of service time and hence may amplify the error over fixed point iteration. I propose a modification to the model [85] which replaces ρ by

$$\rho_{slot} = \frac{\lambda_u(\bar{w}_u + \mathbb{E}[R_u])}{S_s(\bar{w}_s + \mathbb{E}[R_s])}, \tag{3.38}$$

where the numerator is the mean number of slots per second in which an unsaturated source has a packet, and the denominator is the mean total number of system slots per second; S_s and λ_u are the throughput of a saturated source s and the arrival

rate of an unsaturated source u; $\bar{w}_u$ and $\mathbb{E}[R_u]$ are the mean number of backoff slots and attempts that a packet from source u encounters before being successfully sent; and $\bar{w}_s$ and $\mathbb{E}[R_s]$ are the corresponding values for source s. In (3.38), the service time of source u is not used and hence not involved in the fixed point equations as it is in [85]. The proposed modification improves the match between the model of [85] and simulated values of the collision probabilities and throughput, but the match to mean access delay remains poor.

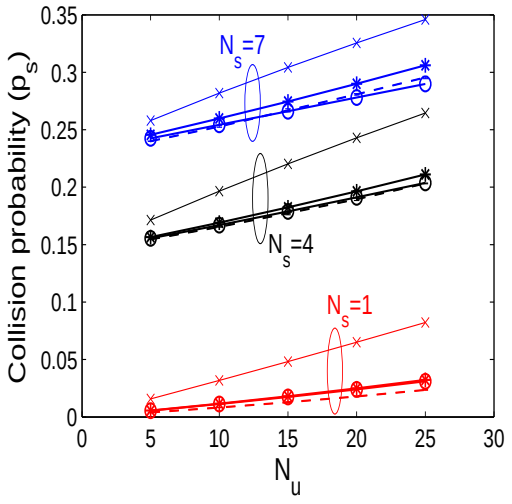
Validation

I simulated networks of N_u identical unsaturated sources sending packets of size l_u with Poisson arrival of rate λ_u, and N_s identical saturated sources sending packets of size l_s. The values of N_u, N_s, λ_u and l_u are varied. All sources have the same MAC parameters $\langle CW_{\min} = 32, \eta = 1 \rangle$. Note that the values of parameters used in each scenario are shown in the caption of figures which show the corresponding results of that scenario.

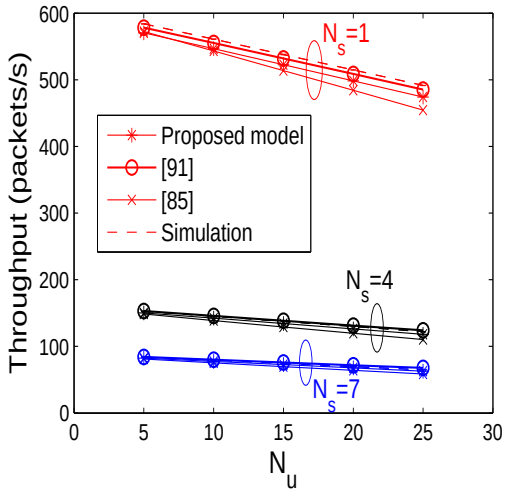
Scenario 1 In this scenario, the number of unsaturated and saturated sources, N_u and N_s, are varied. Then, the collision probability and throughput of a saturated source, and the collision probability and mean access delay of an unsaturated source are shown in Fig. 3.3 as functions of N_u at different N_s. These figures show results from my model as well as from [85], [91] and simulation.

My model and the model [91] accurately capture the increase in collision probabilities when N_s and N_u increases, and the resulting decrease in throughput and increase in mean access delay. However, collision probabilities and mean access delay from [85] are much higher than those of the simulation.

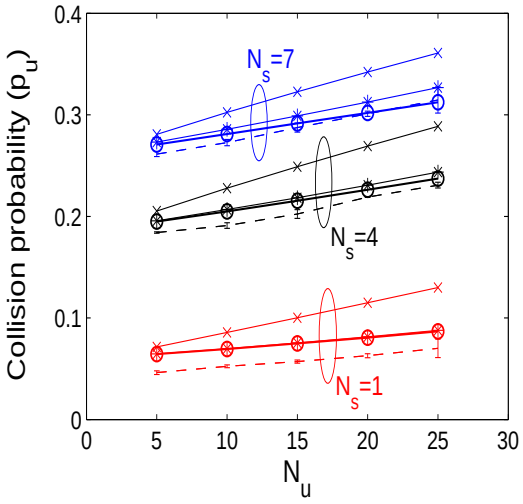
Scenario 2 In Scenario 2, the packet size of an unsaturated sources l_u and its arrival rate λ_u are varied. The collision probability and throughput of each saturated source, and the collision probability and mean access delay of an unsaturated source



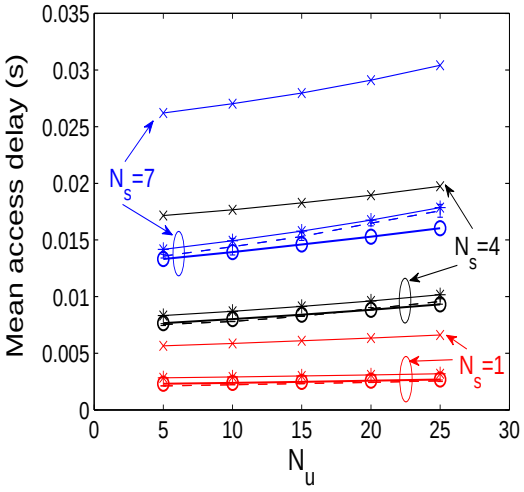
(a) Collision probability of a saturated source



(b) Throughput of a saturated source



(c) Collision probability of an unsaturated source



(d) Mean access delay of an unsaturated source

Figure 3.3: Collision probabilities, throughput, and mean access delay for Scenario 1. Figs. 3.3(a), 3.3(c) and 3.3(d) clearly show that my model is much more accurate than the model in [85]. (Unsaturated stations: Poisson arrivals with rate $\lambda_u = 10$ packets/s, $l_u = 100$ Bytes, $W_u = 32$, $\eta_u = 1$; Saturated stations: $l_s = 1040$ Bytes, $W_s = 32$, $\eta_s = 1$; Buffer size: 50 packets.)

are shown in Fig. 3.4 as functions of l_u at different λ_u. Results are obtained from my model, [85], [91] and simulation.

Figure 3.4 shows that results from my model correctly capture the increase in collision probability with increasing l_u and λ_u, and the resulting decrease in throughput and increase in mean access delay. As for Scenario 1, the model in [85] overestimates the collision probabilities and mean access delay.

This scenario violates the zero-buffer assumption of [91], which hence becomes inaccurate when the packet arrival rate of unsaturated sources is 50 packets/s. That model predicts a high packet drop rate at high traffic load, which causes the collision probabilities to be underestimated.

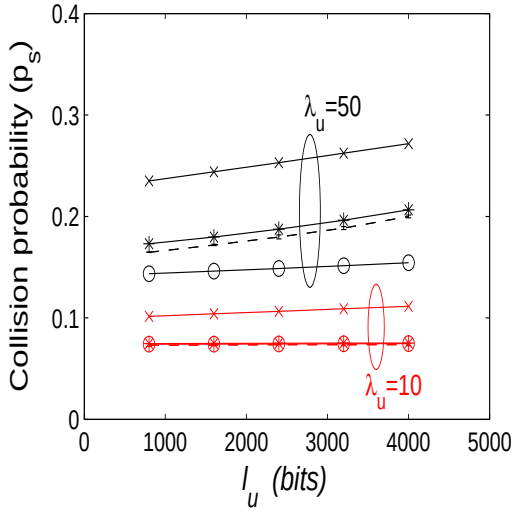
In summary, my model for a network with both unsaturated and saturated sources developed in Section 3.3 is simple and versatile, and provides results more accurate than two existing models when buffers are large.

3.4.2 Validation in 802.11e EDCA

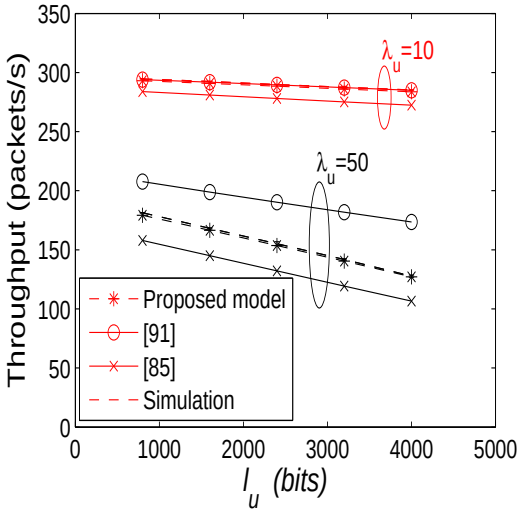
Having validated the proposed scheme in 802.11 DCF in the previous section, I now validate it in 802.11e EDCA WLANs. Different from 802.11 DCF where all stations have the same MAC parameters, I will consider 802.11e EDCA scenarios with different number of traffic types using different MAC parameters. Note that the number of traffic types and the values of the MAC parameters of each type will be defined in each scenario. Also the values of other parameters will be provided in the caption of figures related to that scenario.

Scenario 3

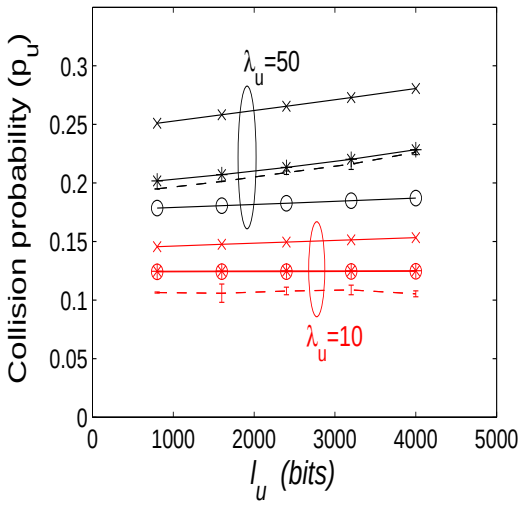
I simulated networks with 4 traffic types, denoted $u1$, $u2$, $s1$ and $s2$, of which the first two are unsaturated. The number of sources N, burst size η and packet size l are distinguished by subscripts $u1$ to $s2$. Unsaturated sources of types $u1$ and $u2$ have arrival rates λ_{u1} and λ_{u2}.



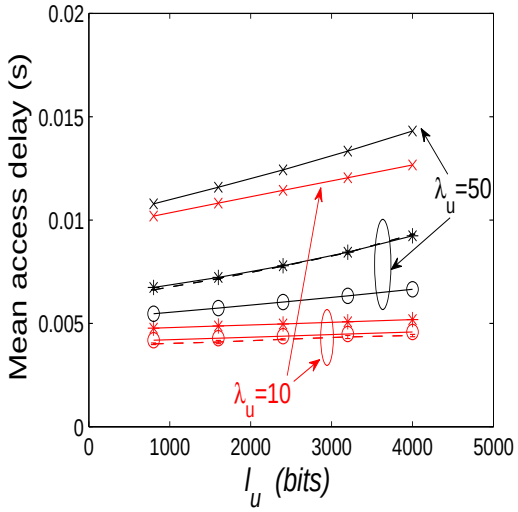
(a) Collision probability of a saturated source



(b) Throughput of a saturated source



(c) Collision probability of an unsaturated source



(d) Mean access delay of an unsaturated source

Figure 3.4: Collision probabilities, throughput, and mean access delay for Scenario 2. Figs. 3.4(b) and 3.4(d), respectively, show clearly that my model is much more accurate than the models in [85] and [91]. (Unsaturated stations: Poisson arrivals with rate λ_u, $N_u = 10$, $W_u = 32$, $\eta_u = 1$; Saturated stations: $N_s = 2$, $l_s = 1040$ Bytes, $W_s = 32$, $\eta_s = 1$; Buffer size: 50 packets.)

QoS parameters $\langle CW_{\min}, \eta \rangle$ of sources of types $u1$, $u2$, $s1$ and $s2$, respectively, are $\langle 32, 2 \rangle$, $\langle 32, 5 \rangle$, $\langle 96, 1 \rangle$ and $\langle 96, 2 \rangle$.

The throughput of a source of type $s1$ and $s2$, and the mean access delay of a source of type $u1$ are shown in Figs. 3.5(a) and 3.5(b) as functions of the number of sources per type.

From Fig. 3.5(a), the throughput of a saturated source of type $s1$ is less than that of type $s2$. This is because types $s1$ and $s2$ have the same CW_{min} but type $s1$ has smaller TXOP limit and larger packet size. My model provides a surprisingly accurate estimate of the throughput.

Fig. 3.5(b) shows that my model provides a reasonably accurate estimate of the mean access delay despite its simplicity compared with Markov chain based models. The model also predicts the access delay of sources of type $u2$ with accuracy similar to that of type $u1$.

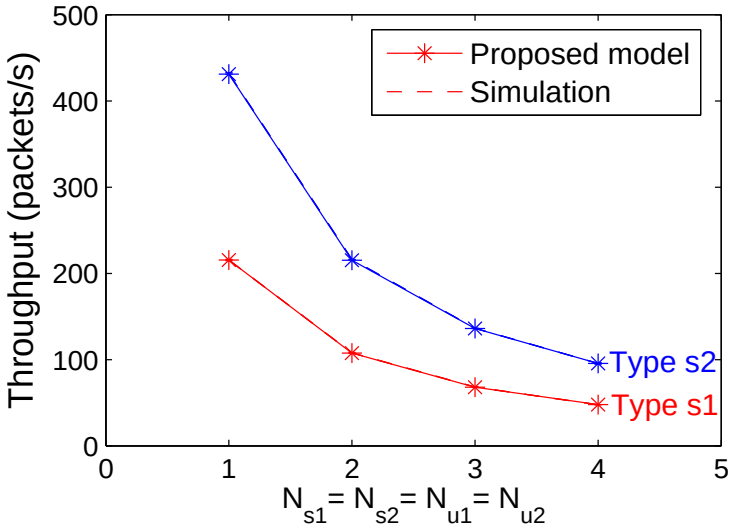
Scenario 4

I simulated networks of N_u identical unsaturated sources sending bursts of η_u packets of size l_u with the packet arrival rate λ_u, and N_s identical saturated sources sending fixed bursts of η_s packets of size l_s.

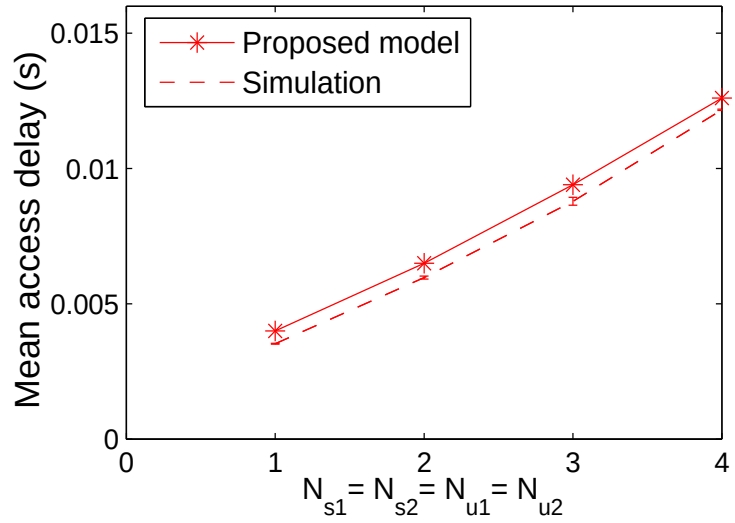
QoS parameters $\langle CW_{\min}, \eta \rangle$ of unsaturated and saturated sources, respectively, are $\langle 32, 1 \rangle$ and $\langle 32\eta_s, \eta_s \rangle$.

The packet inter-arrival times of unsaturated sources are set to be uniformly distributed in the range $1/\lambda_u \pm 1\%$. This quasi-periodic model represents voice traffic (which is often treated as periodic CBR traffic [93]), subject to jitter such as that caused by the operating system. Explicitly including this jitter is necessary to avoid "phase effect" artifacts in the results.

The throughput in packets/s of a saturated source is shown in Fig. 3.6(a) as a function of η_s, parameterized by N_s. When η_s increases, there are fewer bursts from saturated sources contending for the channel, which decreases their collision probability. As a result, the throughput increases.

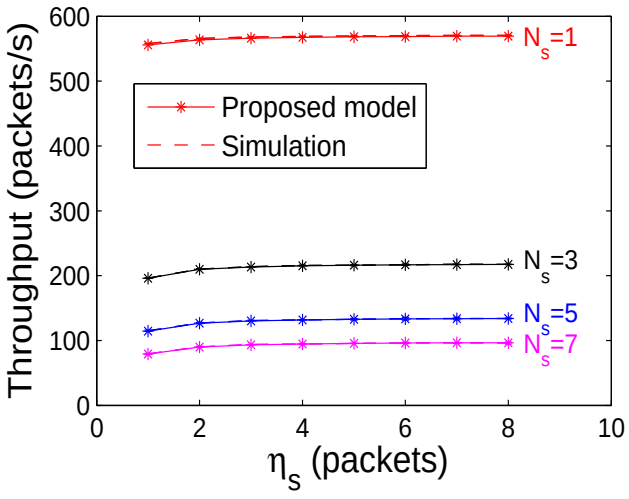


(a) Throughput of a saturated source

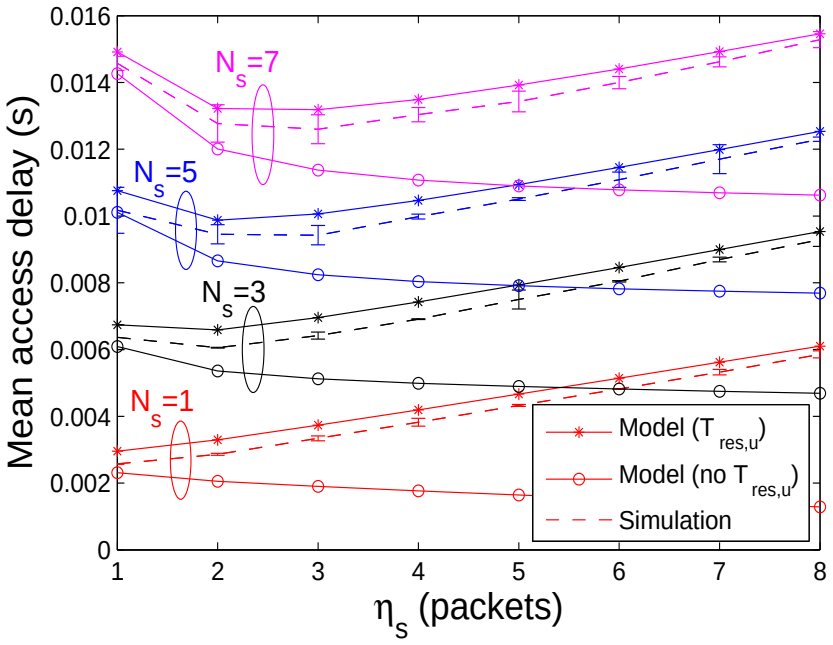


(b) Mean access delay of an unsaturated source of type $u1$

Figure 3.5: Throughput of a source of type $s1$ and $s2$ and mean access delay of a source of type $u1$, Scenario 3. (Unsaturated stations of type $u1$: Poisson arrivals with $\lambda_{u1} = 10$ packets/s, $l_{u1} = 500$ Bytes, $W_{u1} = 32$, $\eta_{u1} = 2$; Unsaturated stations of type $u2$: Poisson arrivals with $\lambda_{u2} = 45$ packets/s, $l_{u2} = 100$ Bytes, $W_{u2} = 32$, $\eta_{u2} = 5$; Saturated stations of type $s1$: $l_{s1} = 1200$ Bytes, $W_{s1} = 96$, $\eta_{s1} = 1$; Saturated stations of type $s2$: $l_{s2} = 800$ Bytes, $W_{s2} = 96$, $\eta_{s2} = 2$.)



(a) Throughput of a saturated source.



(b) Mean access delay of an unsaturated source

Figure 3.6: Mean access delay and throughput when W_s and η_s are scaled together, Scenario 4. (Unsaturated stations: "quasi-periodic" traffic with rate $\lambda_u = 10$ packets/s, $N_u = 10$, $l_u = 200$ Bytes, $W_u = 32$, $\eta_u = 1$; Saturated stations: $N_s = \{1, 3, 5, 7\}$, $l_s = 1040$ Bytes, $W_s = \eta_s W_u$.)

One of my model's contributions is to capture the residual time of busy period during which a burst arrived $T_{\text{res},u}$, which was not important in DCF and has often been overlooked in EDCA models. Fig. 3.6(b) shows the mean access delay of a burst from unsaturated sources with and without $T_{\text{res},u}$ in the access delay models under the same scenario. As seen, when η_s is large, $T_{\text{res},u}$ has significant effect on delay estimation.

Also from Fig. 3.6(b), when η_s increases, for $N_s > 1$, there is a local minimum access delay. For small η_s, the dominant effect is the decrease in collisions due to the larger backoff window W_s of saturated sources. For larger η_s, the increase in residual time $T_{\text{res},u}$ dominates this. This suggests there is an optimal value for η_s where the access delay of unsaturated sources is minimum. This qualitative effect is not captured by models that neglect $T_{\text{res},u}$. More importantly, Fig. 3.6 shows that increasing W_s and η_s together can benefit both unsaturated and saturated sources. Although the optimal value of η_s may vary in different scenarios, in most cases, η_s of 2 provides an improvement in the throughput of a saturated source and a reduction in mean access delay of unsaturated sources. My model can be used to estimate the optimal η_s in this scenario.

3.5 Application of the model

To demonstrate the usefulness of my model, I will use it to determine the distribution of access delay experienced by a burst from an unsaturated source. This is useful for tasks such as determining the appropriate size for jitter buffers.

For tractability, here I approximate K and m to be infinite in the whole model and $b_u = 1$ in the delay model. Simulation results show that this gives accurate estimates of delay in the typical range of interest, from 10 ms to 1 s.

3.5.1 Analysis of access delay distribution

Note that access delay distribution can be calculated using transform methods. The generating function of ccdf of access delay can be derived from its probability mass function (pmf). The distribution can then be obtained by numerical inversion of the z-transform, using the Lattice-Poisson algorithm [12]. The details are not illuminating and hence referred to Appendix A.2.

Approximation method

It is more informative to consider a simple approximate model of the access delay. The total burst access delay is the sum of many random variables: the backoff delays at each stage. However, at particular points, the ccdf of the access delay can be estimated accurately, from which the remainder can be estimated by interpolation. I will now derive such an approximation.

Let $W_{\text{med}}(k)$ be the median number of backoff slots used by bursts which succeed at the kth backoff stage (starting from $k = 0$). Since the number of slots at each stage j, U_{uj}, is symmetric about its median $M[U_{uj}] = (2^j W_u - 1)/2$, the median of their sum is

$$\begin{aligned} W_{\text{med}}(k) &= \sum_{j=0}^{k} M[U_{uj}] = \sum_{j=0}^{k} (2^j W_u - 1)/2 \\ &= \frac{W_u}{2} \sum_{j=0}^{k} 2^j - \frac{k+1}{2} = \frac{W_u}{2} \frac{1 - 2^{k+1}}{1 - 2} - \frac{k+1}{2} \\ &= \left(2^k - \frac{1}{2}\right) W_u - \frac{k+1}{2}. \end{aligned} \tag{3.39}$$

Note that $W_{\text{med}}(k)$ is larger than $(2^k - 1)W_u - k$, the maximum number of backoff slots that could be experienced by a burst that succeeds at stage $k - 1$ or earlier. It is possible for a burst which succeeds at stage $k + 1$ or later also to experience $W_{\text{med}}(k)$ backoff slots but the probability of that is small, especially if p_u is small. Thus the unconditional ccdf of experiencing $W_{\text{med}}(k)$ backoff slots is slightly below

the following upper bound

$$\begin{aligned} ccdf_W(W_{\text{med}}(k)) &\leq 1 - \left(\sum_{j=0}^{k-1} (1-p_u) p_u^j + \frac{1}{2}(1-p_u)p_u^k \right) \\ &= 1 - \left((1-p_u)\frac{1-p_u^k}{1-p_u} + \frac{1}{2}(1-p_u)p_u^k \right) \\ &= p_u^k \left(\frac{1+p_u}{2} \right), \end{aligned} \tag{3.40}$$

which becomes tight for $p_u \ll 1$.

So far, this gives a good approximation for the ccdf of the number of backoff slots experienced. This can be related to the actual delay distribution by approximating the duration of each backoff slot by its mean, and adding the additional overhead of each stage. Thus, the delay associated with $W_{\text{med}}(k)$ backoff slots is approximately

$$\begin{aligned} D(W_{\text{med}}(k)) &\approx W_{\text{med}}(k)\mathbb{E}[Y_u] + k\mathbb{E}[T_u^c] + \mathbb{E}[T_{\text{res},u}] + \mathbb{E}[T_u^s] \\ &= 2^k W_u \mathbb{E}[Y_u] + k(\mathbb{E}[T_u^c] - \mathbb{E}[Y_u]/2) + M_1 \\ &\equiv f(k). \end{aligned} \tag{3.41}$$

where M_1 is a constant representing the remaining components. The approximation becomes tight for large k by the law of large numbers. This implies $k \approx f^{-1}(D(W_{\text{med}}(k))$, and so when $D = D(W_{\text{med}}(k))$ for some k,

$$ccdf_D(D) \approx \left(\frac{1+p_u}{2} \right) p_u^{f^{-1}(D)}. \tag{3.42}$$

It turns out that (3.42) is a good approximation for any delay $D \geq D(W_{\text{med}}(0))$.

However, for delay $D < D(W_{\text{med}}(0))$, which corresponds to the total number of backoff slots from 0 to $W_u/2-1$, a much better approximation is possible. Note that the most likely way to back off for a small number of slots is to back off once, which gives a uniform distribution of the number of slots. Thus for $j = 0, 1, \ldots, W_u/2-1$, the ccdf of a delay

$$D(j) = j\mathbb{E}[Y_u] + \mathbb{E}[T_{\text{res},u}] + \mathbb{E}[T_u^s]$$

is approximately

$$\begin{aligned} ccdf_D(D(j)) &\approx 1-(1-p_u)\frac{j+1}{W_u} \\ &= 1-\frac{1-p_u}{W_u}\Big(1+\frac{D(j)-\mathbb{E}[T_{\text{res},u}]-\mathbb{E}[T_u^s]}{\mathbb{E}[Y_u]}\Big). \end{aligned} \tag{3.43}$$

Thus, I propose the approximation that finds the ccdf from (3.43) for delays less than $D((W_u-1)/2)$, and from (3.42) for larger delays.

Power law delay distribution

In the proposed model, with unlimited retransmissions, the distribution of burst access delays has a power law tail ($Bt^kP(D>t)\to 1$ as $t\to\infty$ for some B, k). Although the true delay cannot be strictly heavy tailed when retry limit is finite, the approximation holds for delays in the typical range of interest, from 10 ms to 1 s [132].

This power law arises since the duration and probability of occurrence of the kth backoff stage increase geometrically in k. This is distinct from the heavy tailed delays in ALOHA, which are caused by heavy-tailed numbers of identically distributed backoffs. Although the latter effect is very sensitive to the assumption of infinite retransmissions and the lack of burst fragmentation, 802.11 can be usefully modeled as heavy tailed even with typical limits of 6 to 8 retransmissions.

Note from (3.41) that $f(k)=2^kW_u\mathbb{E}[Y_u]+O(k)$, where $h(m)=O(g(m))$ means that there exists a C such that for all sufficiently large m, $|h(m)|<Cg(m)$. Thus, by (3.42), the complementary CDF of a large delay D is approximately

$$\begin{aligned} ccdf_D(D) \approx& \frac{1+p_u}{2}p_u^{\log_2\left(\frac{D}{W_u\mathbb{E}[Y_u]}\right)} = \frac{1+p_u}{2}2^{\log_2\left(p_u^{\log_2\left(\frac{D}{W_u\mathbb{E}[Y_u]}\right)}\right)} \\ =& \frac{1+p_u}{2}2^{\log_2\left(\frac{D}{W_u\mathbb{E}[Y_u]}\right)^{\log_2(p_u)}} = \frac{1+p_u}{2}\left(2^{\log_2\left(\frac{D}{W_u\mathbb{E}[Y_u]}\right)}\right)^{\log_2(p_u)} \\ =& \frac{1+p_u}{2}\left(\frac{D}{W_u\mathbb{E}[Y_u]}\right)^{\log_2(p_u)}. \end{aligned} \tag{3.44}$$

That is, the distribution has power law tail with slope $\log_2(p_u)$, which increases (becomes heavier) with increasing congestion, as measured by the collision probability p_u. This is consistent with the more detailed calculations of [31]. This insight would not be obtained by the direct use of the z-transform.

Excessive queueing delay

One application of the preceding result is to determine the congestion level at which the expected queueing delay for unsaturated sources becomes excessive. Although "excessive" will depend on the specific application, I will use the criterion that the expected queueing delay is infinite in my model with no limit on the BEB. If each source is assumed to implement an M/G/1 queue, then this corresponds to the service time having infinite variance.

Consider a log-log plot of the ccdf of a random variable D whose ccdf is the right hand side of (3.44). The minimum (steepest) slope for which the variance of D becomes infinite is -2 [31]. The right hand side of (3.44) suggests that this slope is $\log p_u / \log 2$. Thus the variance of D is infinite when $p_u \geq 2^{-2} = 1/4$. Under the model (3.11) and (3.12)–(3.37), I will now derive the minimum number of saturated sources N_s for which this occurs; that is, the N_s such that, for any number of unsaturated source N_u with arbitrary arrival rate, unsaturated sources using the same backoff parameters as saturated sources will have $p_u \geq 1/4$. Let us start with the following lemma, proved in Appendix A.3.

Lemma 3.2 *Let s and u denote an arbitrary saturated and unsaturated source. Under the model (3.11) and (3.12) with all sources using the same CW_{min},*

$$\frac{\tau_s}{\tau_u} = \frac{S_s \mathbb{E}[\eta_u]}{\lambda_u \mathbb{E}[\eta_s]} \frac{1-\tau_s}{1-\tau_u}.$$

If, in addition, (3.37) holds then $p_u > p_s$.

Theorem 3.3 *Consider the model (3.11) and (3.12)–(3.37), with all sources using*

the same CW_{min} *(*$W_x = W, \forall x \in \mathbb{S} \cup \mathbb{U}$*). If*

$$N_s \geq 1 + \frac{\log(3/4)}{\log(1 - \frac{4}{3W+2})} \tag{3.45}$$

then for any $N_u \geq 1$ *and* $\lambda_u > 0$*, the variance of the random variable whose ccdf is the right hand side of (3.44) is infinite.*

The proof is in Appendix A.4. Surprisingly, the sufficient condition for infeasibility (3.45) depends only on W, the minimum contention window, and not settings such as channel data rate, traffic of real-time source, or the *TXOP limit*.

From (3.44), the distribution of an unsaturated source's access delay D_u under the model (3.11)–(3.37) has a tail which is approximately power law, given by the right hand side of (3.44). Hence, under the condition (3.45), the variance of the access delay D_u is predicted to be infinite.

Note that the variance of the delays in the real system will not be infinite, due to the truncation of the backoff process. However, the high variability is enough to cause significant degradation of the user experience.

3.5.2 Numerical validation and discussion

This section is to validate: (i) approximation method of determining access delay distribution; (ii) the slope of the distribution curve's tail; (iii) the condition (3.45) for the infinite variance of unsaturated sources' access delay.

The simulated network is the same as that in Section 3.4. In the simulation, all sources have the retry limit of 7 and the doubling limit of 5.

Validation of the distribution of access delay

The distribution of unsaturated sources' access delay determined from approximation and z-transform methods and simulation in different scenarios are shown in Fig. 3.7 and 3.8. Although assuming infinite retransmission, both the approximation and z-transform methods provide accurate estimates in the typical range of

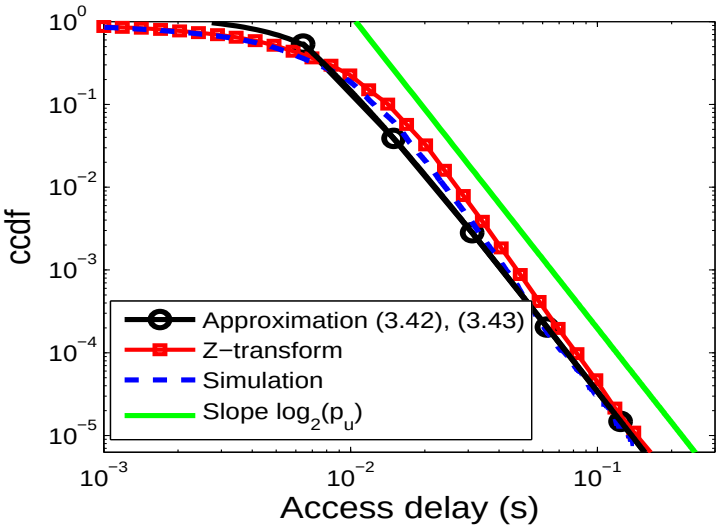


Figure 3.7: Distribution of access delay. (Unsaturated stations: Poisson arrivals with rate $\lambda_u = 10$ packets/s, $N_u = 15$, $l_u = 100$ Bytes, $W_u = 32$, $\eta_u = 1$; Saturated stations: $N_s = 2$, $l_s = 1040$ Bytes, $W_s = 3W_u$, $\eta_s = 4$.)

interest, from 10 ms to hundreds of ms. The approximation is of comparable accuracy to the z-transform method.

Slope of distribution curve's tail

The straight line in Fig. 3.8 shows the slope $\log_2(p_u)$. It captures the trend of the distribution curve reasonably well in the typical delay range from tens to hundreds of ms.

Validation of Theorem 3.3

From (3.45), when W is 32 as in 802.11 DCF, the minimum number of saturated sources required for infinite variance of unsaturated sources' access delay is 8. This is validated in Fig. 3.9 which shows the access delay distribution of unsaturated sources from NS-2 simulation. As seen, the slope of distribution curve's tail is slightly greater than -2 in the typical range of interest, from tens to hundreds of ms. This implies that these delays will occur as often as if the system had a power law tail with infinite variance.

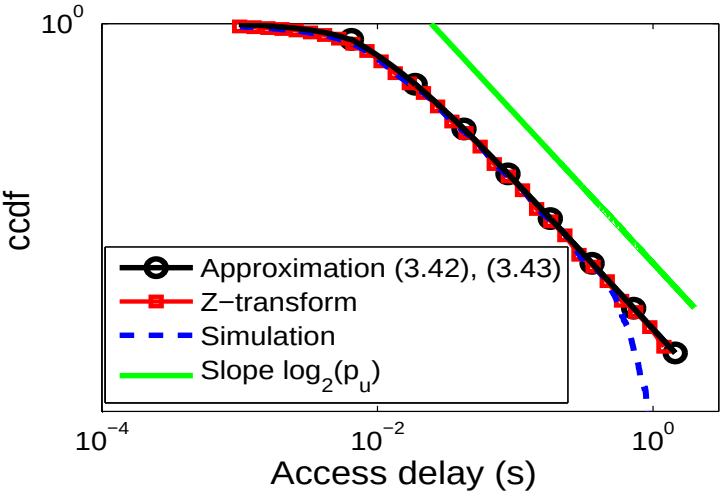


Figure 3.8: Distribution of access delay. (Unsaturated stations: Poisson arrivals with rate $\lambda_u = 10$ packets/s, $N_u = 20$, $l_u = 100$ Bytes, $W_u = 32$, $\eta_u = 1$; Saturated stations: $N_s = 6$, $l_s = 1040$ Bytes, $W_s = 32$, $\eta_s = 1$.)

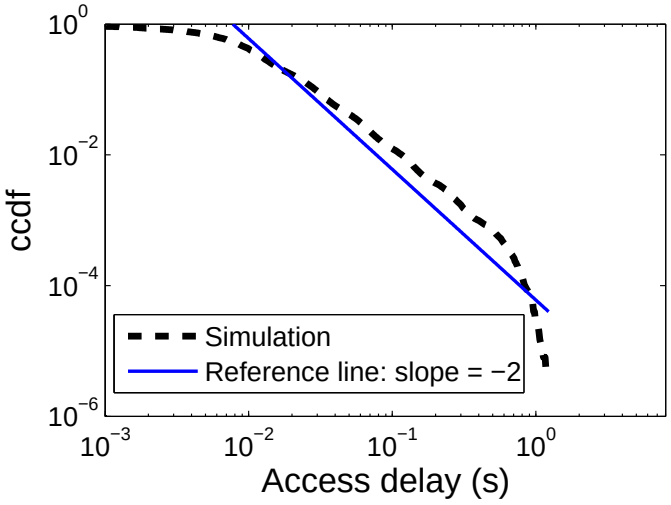


Figure 3.9: Access delay distribution of an unsaturated source. (Unsaturated stations: Poisson arrivals with rate $\lambda_u = 10$ packets/s, $N_u = 1$, $l_u = 100$ Bytes, $W_u = 32$, $\eta_u = 1$; Saturated stations: $N_s = 8$, $l_s = 1040$ Bytes, $W_s = 32$, $\eta_s = 1$.)

3.6 Conclusion

In this chapter, I have provided a comprehensive but tractable fixed point model of 802.11 WLANs with both unsaturated (Poisson) real-time and saturated non-realtime sources, assuming no buffer overflow. The proposed model has been shown to provide accurate estimates of delay, throughput and collision probability in comparison with two of the existing models when buffers are large. Using the model to investigate the interaction between these two traffic types, I have briefly shown that "fair" service differentiation can be achieved based on two QoS parameters, TXOP limit and CW_{min}.

One of the contributions of my model is that I have proposed a closed form approximation for the distribution of the queue size of unsaturated sources, which is sufficiently accurate at low queue occupancies to predict the burst size distribution. Besides, I have also modeled the residual time of an ongoing transmission in unsaturated sources' delay and shown the importance of modeling it.

Moreover, I have proposed a simple method to approximate access delay distribution, which has been shown to be reasonably accurate in the typical delay range of interest. Based on this method, the slope $\log_2(p_u)$ of distribution curve's tail has been easily obtained and then used to determine the lower bound on the number of saturated sources at which excessive queueing delay will be seen by unsaturated sources of arbitrary load, when all sources use the same MAC parameters. This information can be taken in account in network design.

Based on the analysis in this chapter, I will propose a scheme to provide service differentiation in 802.11e EDCA without prioritizing one type of traffic over another in the next chapter. The model proposed here will be used to analyze the performance of the proposed scheme.

Chapter 4

Service differentiation without priority

4.1 Introduction

With the rapid growth of WLANs and new applications, WLANs now carry a diverse mix of traffic, from voice with tight delay constraints to bulk file downloads with only long-term throughput requirements. Efficient use of the network requires services tailored to each of these traffic types. This leads to the need for service differentiation.

Recall from Chapter 2 that most of the previous work and the default EDCA parameters of the 802.11e standard [7] to provide service differentiation in WLANs are based on prioritization, which provides better performance in all respects for a "higher priority" class. This creates an incentive for rational users, who try to optimize their performance without changing the network stack (e.g. application writers who optimize their code based on measured performance using all the available services), to use the access class of the highest priority to gain a higher share of the channel. This can degrade network performance drastically and lead to no service differentiation. To cope with that issue, approaches in prior work [29, 45, 102, 109, 110] such as rewarding schemes or pricing schemes are either complicated or impractical to implement.

The novelty of the work in this chapter is that I seek to provide service differentiation without prioritizing one class over another, that is, there is no ordering of the classes such that one gets better performance in all respects than the later ones. My aim is to provide "different but fair" services for different traffic types, by allowing users to choose different points on a throughput-delay tradeoff curve.

I do this by choosing ACs such that some parameters are less aggressive whenever others are more aggressive, which is motivated by the observations in the previous chapter.

In particular, this chapter contributes a scheme to provide service differentiation which is easy to implement, compatible with the 802.11e standard and robust against rational users, by scaling two MAC parameters: CW_{min} and TXOP limit. The proposed scheme does not require any additional mechanisms such as fair queueing or traffic policing.

In this chapter, the proposed scheme will be constructed in the following steps. I first propose in Section 4.2 the "proportional" scheme which improves service for both throughput- and delay-sensitive types of traffic by scaling CW_{min} and TXOP limit in equal proportion. Then, to analyze its properties, I propose a general game-theoretic framework in which users choose whichever traffic class maximizing their desired performance. Their performance metrics are determined based on the model proposed in Chapter 3. Then, I apply the game framework to analyze the properties of this "proportional" scheme in Section 4.4. Also in this section, I use simulation to validate these theoretical properties of the "proportional" scheme. The results show that the "proportional scheme" can provide better service for both types of traffic; however, there is still a slight incentive for data users to use the real-time class. Then, in Section 4.5, a simple change to the proportional scheme (called the "proportional incentive adjusted" scheme or "PIA") is suggested to give throughput-sensitive applications the incentive to use the bulk-data service class while giving improved performance to both classes.

4.2 Proposed proportional tradeoff scheme

I propose a mechanism which improves service for both data and real-time traffic by increasing CW_{min} and TXOP limit. I do not use the AIFS parameter because it provides load-dependent prioritization which makes it difficult to achieve a "fair"

service differentiation.

In particular, I define $n > 1$ service classes, denoted by B_k ($k \in \{1, \cdots, n\}$). These classes can cover different types of users with different requirements of delay and throughput. Users which demand higher throughput and can tolerate higher delay can transmit more packets per channel access but less often. To achieve this, class B_k with higher k has a higher TXOP limit but commensurately higher CW_{min}. This is similar to the method in [138] to ensure fairness.

Let $\mathcal{T}$ be the TXOP limit of class B_1, which is chosen to fit one packet at the lowest data rate supported by the standard. Then,

$$\text{TXOP limit of class } B_k = \eta_k \mathcal{T}, \tag{4.1a}$$

where η_k ($k = 1, \cdots, n$) satisfies $\eta_k < \eta_{k+1}$ and $\eta_1 = 1$.

Let W_{B_k} be the value of CW_{min} used by class B_k. Then

$$W_{B_{k+1}} = \frac{\eta_{k+1}}{\eta_k} W_{B_k}. \tag{4.1b}$$

My scheme provides several classes for different types of traffic; however, to retain simplicity, I only consider in this chapter two extreme types of traffic: delay-sensitive and throughput-sensitive traffic. Note that class B_1 is designed for delay-sensitive traffic while class B_n is suitable for throughput-sensitive traffic. The logic is that real-time traffic requires low delay and often has only one packet to send at a time but the packet needs to be sent as soon as possible; hence, it always uses class B_1. In contrast, a data source requires high throughput; hence, it may be willing to wait a little longer, if an increase in the amount it can transmit per channel access makes its overall throughput higher.

I will show below when all data users use class B_k, their throughput improves when k increases. When η_k is appropriately chosen, this scheme improves service for both traffic types. This benefit comes from the reduction of collision probability in the network due to the lower attempt probability of data sources.

4.3 Model

Here I present a model of 802.11e EDCA WLANs with rational data and real-time users. Consider an infrastructure network with a set $\mathbb{S}$ of $N_s \geq 1$ saturated sources and a set $\mathbb{U}$ of $N_u \geq 0$ unsaturated Poisson sources with negligible queueing. (For a discussion of unsaturated sources with non-negligible queueing, see Section 4.5.3.) The non-saturated sources represent the real-time users while data users are modeled as saturated sources. (For a discussion of data sources using TCP, see Section 5.2 of Chapter 5.) For simplicity, I make the standard assumption that each station transmits packets of only one source, although this is not required by the scheme itself. (For a discussion of multiple sources per station, see Section 4.5.4.)

The natural framework for considering incentive issues is game theory. WLANs with rational users can be modeled as a game in which users are players. A player i chooses an action which is to use any of classes B_k-s. Based on other players' actions and its action, the player i will get a payoff, which is the throughput for a saturated user or the reciprocal of delay for an unsaturated user.

Using class B_1 is a dominant strategy for unsaturated stations, since it reduces their delay regardless of what other stations do. For this reason, I will not treat unsaturated stations as players, but simply model their effect on the throughput obtained by the saturated users.

In the following description of notation, s, s_k and u denote any saturated user, a saturated user using class B_k and an unsaturated user.

Let N_x ($x \in \{s, s_k, u\}$) denote the number of users of type x. Note that $N_s = \sum_{k=1}^{n} N_{s_k}$ where n is the number of classes. Besides, let W_x ($x \in \{s_k, u\}$) be the minimum contention window of users of type x. Note that my model considers $W_x > 11$ to guarantee system stability as explained later in the wireless model of Section 4.3.1.

Different nodes may use different physical layer bit rates. To avoid inefficiencies [131], I aim at time fairness among saturated users, and so measure throughput

as the amount of time each can transmit. In addition to the natural measure of the *fraction* of time (S_x, called "dimensionless" throughput), some of my results apply to the more tractable measure of throughput in seconds/slot, denoted C_x. By *slot*, I mean MAC slot.

My model makes the standard assumption that the network is in equilibrium. It also assumes that a saturated source sends data for the whole duration of TXOP limit. This is because a saturated source is defined as always having packets waiting to transmit.

4.3.1 Game Framework

A game of the wireless network described above is denoted by a quadruple $\langle \mathcal{P}, A, (u_i)_{i\in\mathcal{P}}, N_u \rangle$ where

- $\mathcal{P} = \{1, \ldots, N_s\}$, the set of players, contains the saturated users.
- For every $i \in \mathcal{P}$, $A_i = \{B_k : k \in [1, n]\}$ is the set of actions available to player i, where action B_k is to use MAC parameters $(CW_{\min}, TXOP) = (W_{B_k}, \eta_k \mathcal{T})$, with $W_{B_k} > W_{B_{k-1}}$ and $\eta_k > \eta_{k-1}$. Note that all the players have the same action space; hence, A is used to denote a general action space of each player. However, the game in Section 4.4 has a different action space from that in Section 4.5.
- For every $i \in \mathcal{P}$, the payoff $u_i(a)$ is the throughput of player i under the action profile a which is a vector containing the action of every player, $(a_1, \cdots, a_{N_s})$. There are two forms of the game, corresponding to the two types of throughput which are determined using the wireless model below.
  - Game 1: $u_i(a)$ is given by throughput in seconds/slot. Then, it is denoted by $C_i(a)$, given by (4.3);
  - Game 2: $u_i(a)$ is given by the dimensionless throughput. Then, it is denoted by $S_i(a)$, given by (4.4).

My results use action profiles defined as follows

$$a_{(X;\cdot)} \in \{a \in A^{N_s} : a_1 = X\}, \quad \forall X \in A$$
$$a_{(X;\cdot;Z;\cdot)} \in \{a \in A^{N_s} : a_1 = X \text{ and } a_j = Z\}, \quad \forall X, Z \in A$$

Wireless model

I now summarize the wireless model to determine the throughput of a saturated station as payoff of a player in the game framework, which is derived, justified and validated in Chapter 3.

The model assumes that sources have no limit on the number of retransmission and CW_{max}. This is made for notational and computational simplicity; however, simulations show that qualitative results from this model still hold when these two backoff parameters are truncated as in the standard.

Central to the model is a set of fixed point equations. I only consider balanced fixed points, i.e., ones in which all the nodes of the same type have same value of collision probability, based on the following observations. The minimum contention window I consider is $W_x > 11$ ($x \in \{s_k, u\}$), for which binary backoff satisfies the condition of Theorem 5.4 in [113]; hence, the system has a unique fixed point which is balanced when $N_u = 0$. For $N_u > 0$, I assume that the load of unsaturated users is light enough that there again exists a unique and balanced fixed point as most analyses assume.

Fixed point model The attempt probability τ_s of a saturated source $s \in SS$ is from (3.11a)

$$\frac{1}{\tau_s} = \frac{W_s}{2}\frac{1-p_s}{1-2p_s} + \frac{1}{2}. \tag{4.2a}$$

Note that all saturated users using class B_k have the same CW_{min}, $W_s = W_{B_k}$ and hence, the same attempt probability and collision probability, denoted by τ_{s_k} and p_{s_k}, respectively.

Next, the attempt probability of an unsaturated source $u \in \mathbb{U}$ with the arrival rate λ_u is from (3.11b)

$$\tau_u = \frac{\lambda_u \sum_{j=0}^{\infty} p_u^j}{(1/\mathbb{E}[Y])} = \lambda_u \mathbb{E}[Y] \frac{1}{1-p_u}. \tag{4.2b}$$

Finally, the collision probability of source $x \in SS \cup \mathbb{U}$ is from (3.11c)

$$p_x = 1 - \frac{G}{1-\tau_x}. \tag{4.2c}$$

where G is given by (3.2).

Throughput of data users The throughput in seconds/slot C_{s_k} of a saturated source of class B_k is given by the probability the source transmits successfully a burst in a slot multiplied by the duration it can transmit.

$$C_{s_k} = \tau_{s_k}(1-p_{s_k})\eta_k \mathcal{T}. \tag{4.3}$$

The dimensionless throughput S_{s_k} of a saturated source of class B_k is given by the throughput in seconds/slot divided by the average duration of a slot.

$$S_{s_k} = \frac{C_{s_k}}{\mathbb{E}[Y]}. \tag{4.4}$$

Another measure called "relative throughput" is also used. This is the throughput of a saturated source under the given scheme divided by that under the scheme with no service differentiation ($\eta_k = 1$, $\forall k$).

4.4 Properties of the proportional tradeoff scheme

I now consider the first specific game in the foregoing framework, which is based on the proportional scheme to provide service differentiation. An alternative based on the PIA scheme will be considered in Section 4.5.

Under the proportional scheme given by (4.1), the action space of the game is

$$A0 = \Big\{ \big(\eta_k W_{B_1}, \eta_k \mathcal{T}\big) : k \in [1, n] \Big\}, \quad \text{where } \eta_1 = 1$$

and $(\eta_k W_{B_1}, \eta_k \mathcal{T})$ are the MAC parameters of class B_k.

4.4.1 Theoretical results

The following results will be proved for unbounded retransmission and CW_{max} and some results are for networks with only data users. However, I will show by simulation they apply when these assumptions are relaxed.

Service differentiation property

I first show that the proportional scheme improves service for both data and realtime traffic by considering the network in which all users use the class designed for them in Theorems 4.1 and 4.3. In particular, all saturated sources use class $\mathrm{B}_{k>1}$ and all unsaturated sources use class B_1.

I start with Theorem 4.1, proven in Appendix B.8, which states that, in a network without real-time users, when all data users uses class B_k with $\eta_k > 1$ under the proportional scheme, they will receive higher throughput than when there is no service differentiation ($\eta_k = 1$).

Theorem 4.1 *Consider the wireless model (4.2)–(4.4), in the game $\langle \mathcal{P}, A0, (S_i)_{i\in\mathcal{P}}, 0\rangle$ with all data users using the same class B_k. The dimensionless throughput of data users increase when they use class with a higher η_k.*

The above theorem is based on the following lemma proven in Appendix B.7.

Lemma 4.2 *Under the wireless model (4.2), in the game $\langle \mathcal{P}, A, (S_i)_{i\in\mathcal{P}}, 0\rangle$ with all data users using class B_k, the collision probability and attempt probability of all data users decrease with the increase of their CW_{min}.*

This lemma suggests that under the proportional scheme, when data users use higher class (higher η_k), their CW_{min} increases. Therefore, their collision probability reduces, which explains for their throughput increase as stated in Theorem 4.1.

The above result show the benefit of proportional scheme for data users only. It would be more interesting and complete to show the benefit of the proportional scheme for both data and real-time users. Because the proof for the network of only data users is quite complicated, I study analytically a simple two-user network of mixed traffic in Theorem 4.3 and use simulation to study networks with higher load. Theorem 4.3 is proved in Appendix B.1 using the wireless model with (4.2a) simplified to

$$\tau_{s_k} = \frac{2}{W_{s_k}} \frac{1 - 2p_{s_k}}{1 - p_{s_k}}. \tag{4.5}$$

to keep the algebra tractable, assuming that $W_{s_k} \gg 1$.

Theorem 4.3 *Consider the wireless model (4.2)–(4.3) with (4.2a) replaced by (4.5), in game* $\langle \mathcal{P}, A0, (S_i)_{i\in\mathcal{P}}, N_u \rangle$ *with* $N_u = N_s = N_{s_k} = 1$, $\max(T_u, T_s) < 2\mathcal{T}$, *and* $\lambda_u T_u \leq 1$.

(T4.3-1) *The throughput in seconds/slot of the saturated station increases when* $\eta_k \geq 1$ *increases.*

(T4.3-2) *The collision probability of the unsaturated station decreases when* $\eta_k \geq 1$ *increases.*

Although the result in Theorem 4.1 is for scenarios with only data users and that in Theorem 4.3 is for a simple mixed-traffic scenario, I will show by simulation that they hold for more general scenarios. In particular, simulation shows that the reduction in collision probability of unsaturated sources is accompanied by a reduction in the mean delay, except at light load.

Incentive property

Here I will investigate the incentive of bulk-data users under the proportional scheme by examining different actions of theirs in Theorems 4.4 and 4.6. In particular, I am

interested in the Nash equilibrium of the game where the action space is A0 because system with selfish users often operates at the Nash Equilibrium. Recall that an action profile is a *Nash equilibrium* if no player gets higher payoff by changing its action while others keep theirs unchanged [43].

Theorem 4.4 *Under the wireless model (4.2)–(4.4), in the game* $\langle \mathcal{P}, A0, (S_i)_{i\in\mathcal{P}}, N_u \rangle$ *with* $W_i > 11$*, a data user using class* B_1 *has higher throughput than any other data user using any class* $B_{k>1}$ *in the same network. Specifically,* $S_1(a_{(B_1;\cdot;B_{k>1};\cdot)}) \geq S_j(a_{(B_1;\cdot;B_{k>1};\cdot)})$.

This theorem is proved in Appendix B.3 and based on the following lemma which is proved in Appendix B.2.

Lemma 4.5 *Consider the wireless model (4.2), in the game* $\langle \mathcal{P}, A, (S_i)_{i\in\mathcal{P}}, N_u \rangle$ *with* $N_{s_j} \geq 1$ *and* $N_{s_{j+i}} \geq 1$ $(i, j > 0)$*. If* $W_{j+i} \geq W_j > 11$ *then data users using class* B_j *have an attempt probability equal to or higher than those using class* B_{j+i}, $\tau_{s_j} \geq \tau_{s_{j+i}}$. *Moreover, if* $W_{j+i} > W_j > 11$ *then* $\tau_{s_j} > \tau_{s_{j+i}}$.

The following theorem proven in Appendix B.5 states that, regardless of the actions of other data users, the remaining user is better off by using class B_1.

Theorem 4.6 *Consider the wireless model based on (4.2)–(4.3) with (4.2a) replaced by (4.5), in the game* $\langle \mathcal{P}, A0, (C_i)_{i\in\mathcal{P}}, 0 \rangle$ *with* $W_i > 11$*. We have*

$$C_1(a_{(B_{k>1};\cdot)}) < C_1(a_{(B_1;\cdot)}). \tag{4.6}$$

Although the throughput in Theorem 4.6 is in seconds/slot, simulation demonstrates this result still holds for the dimensionless throughput. The proof of Theorem 4.6 is based on the following lemma.

Lemma 4.7 *Consider the wireless model based on (4.2)–(4.3) with (4.2a) replaced by (4.5), in the game* $\langle \mathcal{P}, A, (u_i)_{i\in\mathcal{P}}, 0 \rangle$ *with* $W_{s_k} > W_u > 11$*. Data user 1 has a higher attempt probability and other data users has lower attempt probability when data user 1 uses class* B_1 *than when it uses any class* $B_{k>1}$.

This lemma explains for the increase of the throughput of data user 1 when it uses class B_1 as stated in Theorem 4.6.

From Theorem 4.6, the action profile with all data users using class B_1 is a unique Nash equilibrium. Then, according to Theorem 4.1, the throughput of a data user at Nash equilibrium is less than that when all data users use class $\mathrm{B}_{k>1}$. Section 4.5 will consider an improved scheme that avoids that issue.

From Theorem 4.6, using class B_1 is a dominant strategy, which means that regardless of actions of other users, a given user always get the highest throughput by using class B_1. Hence, even if the action space consists of mixed strategies [43] (i.e., randomly selecting a class from a given probability distribution), the action profile with all data users always using class B_1 is still a unique Nash equilibrium.

4.4.2 Simulation results and discussion

Recall that the properties of the proportional scheme in Sec. 4.4.1 are proved for unbounded retransmission and CW_{max}, and some of them are for a network with only data users. Herein I will use simulation (*ns-2.33* [1] [139]) to validate those in more general scenarios with both data and real-time users, and a limited number of retransmissions.

In the simulated networks, unsaturated and saturated sources send packets to an access point, using UDP. Unsaturated sources have the same packet size and produce Poisson traffic of the same arrival rate. Saturated sources have the same packet size and receive CBR traffic faster than they can transmit. I use the 802.11g parameters in Table 4.1. Note that similar results were obtained if not all users use the same data bit rate, or the network is based on the 802.11b MAC.

For tractability, I only consider two classes ($k \in \{1, 2\}$). The MAC parameters specific to classes B_1 and B_2 in the proportional scheme are given in Table 4.2 with $\mathcal{T} = 0.72$ms. Note that the value of $\mathcal{T}$ is chosen such that for the rates given in Table 4.1, it is at least the duration of the largest packet over all sources, which is 1400 bytes in this section.

Table 4.1: 802.11g MAC and PHY parameters

Parameter	Symbol	Value
Data bit rate	r_{data}	54 $Mbps$
Control bit rate	r_{ctrl}	1 $Mbps$
Basic rate		2 $Mbps$
PHYS header	T_{phys}	192 μs
MAC header	l_{mac}	288 $bits$
ACK packet	l_{ACK}	112 $bits$
Slot time	σ	20 μs
SIFS		10 μs
DIFS		50 μs
Retry limit	K	7
Doubling limit	m	5
Buffer capacity		50 packets

Table 4.2: MAC parameters of classes B_1 and B_2 used in Section 4.4.2.

Class	CW_{min}	CW_{max}	AIFSN	TXOP limit
B_2	$W_{B_2} = \eta W_{B_1}$	$2^5 W_{B_2}$	2	$\eta\mathcal{T}$
B_1	W_{B_1}	$2^5 W_{B_1}$	2	$\mathcal{T}$

In this section, I also compare the performance of throughput-sensitive bulk data and delay-sensitive voice under the proposed scheme with that using the default EDCA parameters shown in Table 4.3 taken from Table 7-37 of [8].

Table 4.3: Default EDCA parameters (DSSS-OFDM 54Mbps) [8].

AC	Traffic	CW_{min}	CW_{max}	AIFSN	TXOP limit
AC_BE	Data	15	1023	3	1 packet
AC_VO	Real-time	3	7	2	1.504 ms

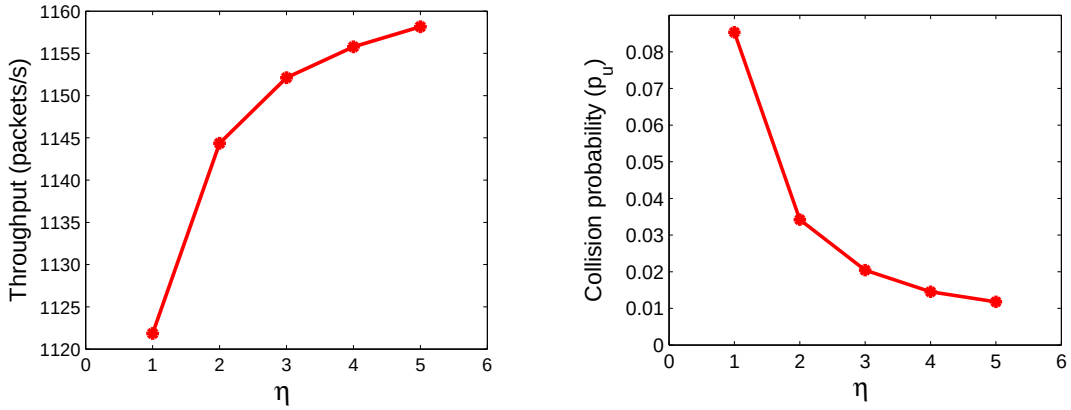
Note that the throughput in simulation results are measured in packets/s, which can be converted to the dimensionless throughput by multiplying by the packet duration. At 54 Mbps, this is 345 us for 1000 bytes, 375 us for 1200 bytes and 405 us for 1400 bytes.

Service differentiation property

To validate service differentiation property, I consider the network with all users using the class designed for them ($N_{s_1} = 0$). Realtime users use class B_1 and data

users use class B_2.

Scenario 1 ($N_{s_2} = N_u = 1$) Fig. 4.1 shows the throughput of a data user and the collision probability of an unsaturated station. When η increases, the throughput increases and the collision probability decreases, which shows the benefit of the proportional scheme. This confirms the result of Theorem 4.3.



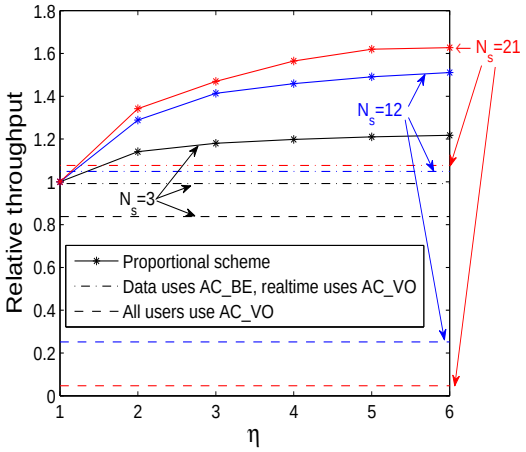
(a) (b)

Figure 4.1: Throughput of a data user and collision probability of a real-time user as a function of class B_2's TXOP limit in units of $\mathcal{T}$ (η). ($\lambda_u = 50$ packets/s, $l_s = 1400$ bytes, $l_u = 400$ bytes, $N_{s_2} = N_u = 1$, $N_{s_1} = 0$, $W_{B_1} = 16$, $W_{B_2} = \eta W_{B_1}$.)

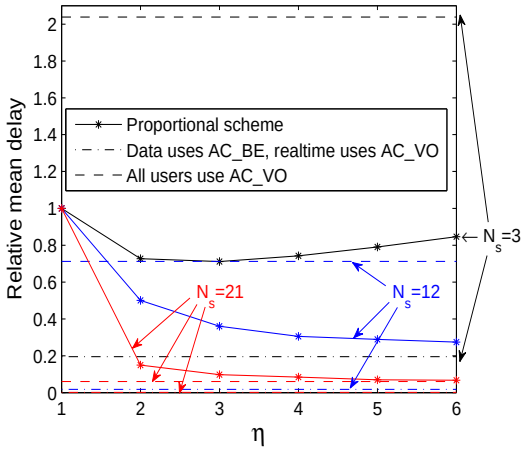
Scenario 2 ($N_{s_2} > 1$, $N_u > 1$) To investigate the ability of my proportional scheme to give benefits to both classes of traffic in larger systems, I compare it with the default EDCA parameters (Table 4.3) within the same scenarios.

The throughput of a data user, and the mean delay of a real-time user under the proportional scheme are shown in Fig. 4.2, as functions of η for different N_s. Moreover, the performance metrics under the default EDCA setting (Table 4.3) with all data users using class AC_BE and real-time users using class AC_VO, and under the default EDCA setting with all users using class AC_VO are also shown for comparison. In Fig. 4.2, the performance metric of the proportional scheme at each η and of the default EDCA setting are normalized by that of the proportional

scheme at $\eta = 1$. Note that the actual throughput and mean delay degrade as N_s increases; however, the relative performances improve as N_s increases.



(a) Data throughput. Proportional allocation gives higher throughput than the default EDCA setting with data users using AC_BE class. It also gives better throughput than the default with data users using AC_VO, especially at heavy load.



(b) Real-time delay. Proportional allocation gives lower delay than the default EDCA setting with data users using AC_VO except at heavy load ($N_s = 21$) due to high loss probability under the default EDCA setting (see Table 4.4), though higher delay than the default with data users using AC_BE.

Figure 4.2: Performance of proportional allocation as a function of class B_2's TXOP limit in units of $\mathcal{T}$ (η). ($\lambda_u = 35$ packets/s, $l_s = 1000$ bytes, $l_u = 200$ bytes, $N_u = 6$, $W_{B_1} = 16$, $W_{B_2} = \eta W_{B_1}$.)

Since the relative throughput is greater than 1 for $\eta > 1$ in Fig. 4.2(a), the proportional scheme with $\eta > 1$ always provides better service for data users compared to no service differentiation ($\eta = 1$). This corroborates the result of Theorem 4.1. Note that the benefit of the proposed scheme increases with contention level in the network. Fig. 4.2(a) also shows that the throughput of a data user in the proportional scheme is always higher than that in the default EDCA scheme with all data users using class AC_BE. Moreover, when traffic load is high enough, my scheme significantly improves the throughput of data users compared to the default EDCA setting with all data users using class AC_VO.

In Fig. 4.2(b), when the load is high enough, my scheme with $\eta > 1$ provides

Table 4.4: Loss probability of a real-time user (%)

	η	$N_s = 3$	$N_s = 12$	$N_s = 21$
Proportional tradeoff	1	–	–	0.85
	2	–	–	–
	3	–	–	–
	4	–	–	–
	5	–	–	–
	6	–	–	–
Data uses AC_BE, real-time uses AC_VO		–	–	–
All users use AC_VO		7.315	79.35	97.85

("–" denotes no loss found during simulation time)

significant improvement in mean delay of real-time users compared to the case of no service differentiation ($\eta = 1$). Additional simulation shows that at light load, the improvement is negligible. This is acceptable because delay only becomes a problem at high load. Fig. 4.2(b) also suggests that at each network load, there exists an optimal value of η at which mean delay is minimum (e.g. $\eta = 2$ at $N_s = 3$ and $\eta > 5$ at $N_s = 12$). I find that this optimal η increases with the network load. Besides, Table 4.4 shows that the loss probability of real-time traffic decreases with the increase of η.

Compared with the default EDCA setting with all data users using class AC_VO, my proportional scheme provides much better service for real-time users in terms of both mean delay and loss, as shown in Fig. 4.2(b) and Table 4.4. The apparent exception for $N_s = 21$ is due to the much higher loss rate under the default EDCA setting with all data users using AC_VO.

Compared with the default parameter setting which prioritizes real-time traffic with all data users using class AC_BE, we expect the performance will be worse for real-time users under the proportional scheme. This is seen in Fig. 4.2(b).

Although the optimal η in my proportional scheme depends on traffic load, the majority of the benefit for both data and realtime users is obtained at $\eta = 2$. Fig. 4.2 suggests that increasing η beyond 6 does not improve performance significantly,

which is because the contention level does not decrease much further then.

Incentive property

Here I will investigate the incentive of data users in choosing a class under the proportional tradeoff scheme, by comparing the payoff of a particular data user in different action profiles. I assume realtime users always choose class B_1.

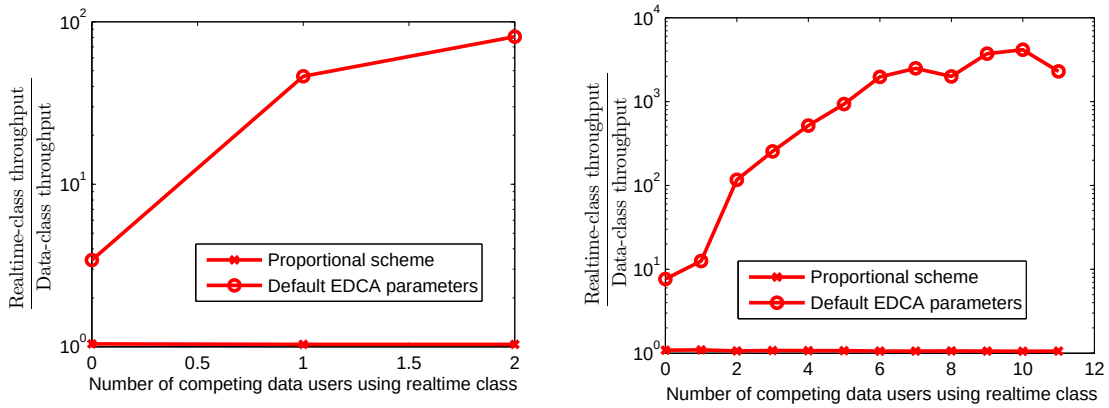
A class-B_1 user has higher throughput than a class-B_2 user I simulated the network scenario with $\lambda_u = 35$ packets/s, $l_s = 1000$ bytes, $l_u = 200$ bytes, $N_u = 6$, $W_{B_1} = 16$, $W_{B_2} = \eta W_{B_1}$, η varied from 2 to 5, and $N_s = \{3, 12, 21\}$. The obtained results show that a data user using class B_1 gains higher throughput than another data user using class B_2, which confirms the result of Theorem 4.4.

Nash equilibrium The results in Fig. 4.3 show that a data user achieves higher throughput by using class B_1 regardless of the other data users' choice under the proportional scheme. However, a data user has less incentive to use B_1 in this case than it does to use AC_VO under the default EDCA scheme, because the latter provides a larger increase in throughput relative to AC_BE.

This implies that the action profile in which all data users use realtime class B_1 is the only Nash equilibrium, which confirms the result of Theorem 4.6. However, this equilibrium gives a lower throughput than could be obtained when all data users use class B_2, as shown by the increase in relative throughput with η in Fig. 4.2(a). I next investigate a way to avoid this undesirable equilibrium.

4.5 Incentive adjusted scheme, PIA

Section 4.4.2 showed that for networks with both data and realtime users, my proportional scheme can improve service for both traffic types relative to the scheme with no service differentiation, especially at high load. However, when a small fraction of data users uses class B_1, their throughput can be slightly improved. Although



(a) $N_s = 3$ (b) $N_s = 12$

Figure 4.3: Ratio of throughput of a data user when it uses "real-time" class to that when it uses "bulk data" class, as a function of the number of competing data users using realtime class. The figures show there is a big incentive for data users to use realtime class under the default EDCA parameters while this incentive seems negligible under the proportional scheme. ($\lambda_u = 35$ pkts/s, $l_s = 1000$B, $l_u = 200$B, $N_u = 6$, $\eta = 2$, $W_{B_1} = 16$, $W_{B_2} = \eta W_{B_1}$.)

the improvement is small, measurement-driven application design will still result in class B_1 being chosen by throughput-sensitive applications. However, I will now show a slight modification to the proposed scheme can eliminate this incentive issue. This is in contrast to priority-based schemes, which require explicit policing or pricing mechanisms [29, 45, 102, 109, 110].

4.5.1 Description of the PIA scheme

I modify the proportional scheme by reducing CW_{min} of class $\mathrm{B}_{k>1}$ by an amount $\epsilon_k > 0$, which provides higher benefit for users of class $\mathrm{B}_{k>j}$ than users of class B_j. The reduction in CW_{min} for class $\mathrm{B}_{k>1}$ results in more throughput for a data user when it uses $\mathrm{B}_{k>1}$ compared to when it uses class B_1, and thus data users have no incentive to use the real-time class B_1 but have incentive to use the class providing the highest throughput B_n. Recall that users can only select one of the access classes determined by the access point, and cannot choose arbitrary combinations of parameters.

Note that the performance of delay-sensitive users degrades as ϵ_k increases, and so I would like to use the smallest ϵ_k such that bulk data users using class B_k get a higher throughput than those using class B_{k-1}, regardless of the network load; any larger value of ϵ_k will increase that benefit but degrade realtime sources' performance. The absolute smallest such ϵ_k is given in the following theorem. Importantly, it depends only on η_k and η_{k-1}, and not the number of users of each type in the network.

Theorem 4.8 *Under the wireless model based on (4.2)–(4.4), in the game* $\langle \mathcal{P}, A, (S_i)_{i\in\mathcal{P}}, N_u \rangle$ *with* $W_{s_k} = \frac{\eta_k}{\eta_{k-1}} W_{s_{k-1}} - \epsilon_k > W_{s_{k-1}} > 11$*, when*

$$\epsilon_k \geq \epsilon_k^0 = 4(\frac{\eta_k}{\eta_{k-1}} - 1), \tag{4.7}$$

data users using B_k *get higher throughput than those using* B_{k-1}*. That is,* $S_1(a_{(B_k;\cdot;B_{k-1};\cdot)}) > S_j(a_{(B_k;\cdot;B_{k-1};\cdot)})$.

The above result is proved in Appendix B.6.

Specifically, under the PIA scheme, the action space in the game framework has the form

$$A1 = \left\{ (W_{B_1}, \mathcal{T}), \left\{ \left(\frac{\eta_k}{\eta_{k-1}} W_{B_{k-1}} - \epsilon_k^0, \eta_k \mathcal{T}\right) \right\}_{k\in[2,n]} \right\}, \tag{4.8}$$

where $(W_{B_1}, \mathcal{T})$ and $\left(\frac{\eta_k}{\eta_{k-1}} W_{B_{k-1}} - \epsilon_k^0, \eta_k \mathcal{T}\right)$, respectively, are MAC parameters of class B_1 and $\mathrm{B}_{k>1}$.

4.5.2 Properties of the PIA scheme

In this section, I first use the game framework above to derive some properties of the PIA scheme. Then, I validate these results using ns-2 simulation.

Theoretical results

Here the results will be proved for unbounded retransmission CW_{max} and some results are for networks with only data users. However, simulation shows they still apply when these assumptions are relaxed.

Service differentiation property To show that the PIA scheme improves service for data traffic, I consider the network in which all users use the class designed for them in Theorem 4.9. It states that the PIA scheme provides better service for data users using a class with higher η_k, which is proved in Appendix B.8.

Theorem 4.9 *Consider the wireless model (4.2)–(4.4), in the game* $\langle \mathcal{P}, A1, (S_i)_{i\in\mathcal{P}}, 0\rangle$ *with all data users using the same class* B_k*. The dimensionless throughput of data users increase when they use a class with higher* η_k*.*

The following corollary comes from the above theorem.

Corollary 4.10 *Consider the wireless model based on (4.2)–(4.4), in the game* $\langle \mathcal{P}, A1, (S_i)_{i\in\mathcal{P}}, 0\rangle$ *with* $W_i > 11$ *and all data users using class* B_k*. The dimensionless throughput of each data user using class* $B_{k>1}$ *under the PIA scheme is higher than that under no service differentiation (all use class* B_1*).*

Incentive property To see if the PIA scheme eliminates incentive for data users to use realtime class B_1, we look at their performance under different actions.

The following theorem, proven in Appendix B.10, implies that the action profile with all data users using the highest class B_n is the unique Nash equilibrium.

Theorem 4.11 *Consider the wireless model based on (4.2)–(4.3) with (4.2a) replaced by (4.5), in the game* $\langle \mathcal{P}, A1, (C_i)_{i\in\mathcal{P}}, 0\rangle$ *with* $W_i > 11$*. For any action profiles in which not all data users use class* B_n*, a data user using a class other than* B_n *can improve its throughput by using* B_n*.*

Note that this theorem is a natural consequence of Lemma 4.12. This lemma, proved in Appendix B.9, states that, if there exists at least another data user using the class with the index equal to or higher than the class used by a given user, the given user can get a higher throughput per slot by using the highest class, B_n.

Lemma 4.12 *Consider the wireless model based on (4.2)–(4.3) with (4.2a) replaced by (4.5), in the game* $\langle \mathcal{P}, A1, (C_i)_{i\in\mathcal{P}}, 0\rangle$ *with* $W_i > 11$*. For all* $i \geq 0$*,*

$$C_1(a_{(B_{k<n};\cdot;B_{k+i};\cdot)}) < C_1(a_{(B_n;\cdot;B_{k+i};\cdot)}). \tag{4.9}$$

Although the throughput in Lemma 4.12 is in seconds/slot, simulation shows that this result still holds for the dimensionless throughput.

Under the PIA scheme, the action profile with all data users using the highest class B_n is the unique Nash equilibrium. Then, according to Corollary 4.10, the throughput of each data user at this Nash equilibrium is greater than that when all data users use class B_1. This suggests that the PIA scheme achieves the desired goal of providing a scheme in which rational users will all gain improved performance.

Note that, when mixed strategies are allowed, it remains an open question whether the equilibrium in which all users use class B_n is the unique Nash equilibrium.

Simulation results and discussion

Recall that the properties of the PIA scheme are proved for networks with only data users and for unbounded retransmission and CW_{max}. Here I will use simulation (*ns-2.33* [1] [139]) to validate those in more general scenarios with both data and real-time users and a limited number of retransmissions. Noticeably, simulation results show that the PIA scheme is actually incentive-compatible, which means that using class B_n is the best strategy regardless of what other data users choose.

The simulated network in this section is the same as one in Sec. 4.4.2. Note that all the results are still valid when 802.11b parameters are used.

Incentive compatibility I verify here that throughput-sensitive users have an incentive to choose class with the highest TXOP limit. I consider the case of three ACs per station: B_1, B_2 and B_3. The MAC parameters specific to these ACs are given in Table 4.5 with $\mathcal{T} = 0.72$ms.

Table 4.5: MAC parameters of classes B_1, B_2 and B_3 used in Fig. 4.4.

Class (B_k)	TXOP limit	ϵ_k^0	$CW_{\min}$	CW_{max}	AIFSN
B_1	$\mathcal{T}$	0	W_{B_1}	$2^5 W_{B_1}$	2
B_2	$2\mathcal{T}$	4	$W_{B_2} = 2W_{B_1} - 4$	$2^5 W_{B_2}$	2
B_3	$3\mathcal{T}$	2	$W_{B_3} = \frac{3}{2}W_{B_2} - 2$	$2^5 W_{B_3}$	2

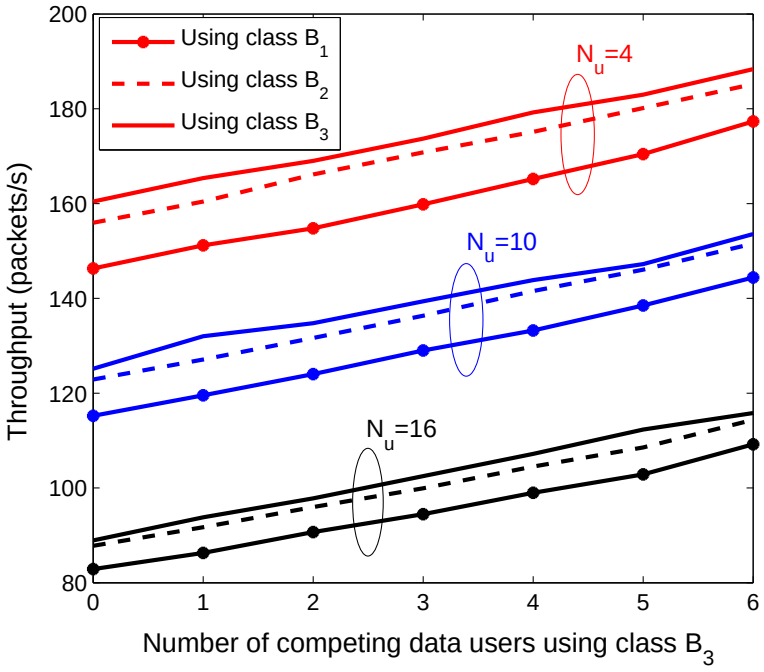


Figure 4.4: Throughput in packets/s of a data user when it uses classes B_1, B_2 and B_3 as a function of the number of competing data users using data class B_3. The throughput improvement of a data source under the PIA scheme at a given N_u is the ratio of the throughput when all data users use class B_3 to the throughput when all data users use class B_1, which is about 22% at $N_u = 4$ and larger when N_u increases. ($\lambda_u = 35$ pkts/s, $l_s = 1000$B, $l_u = 200$B, $N_s = 7$, $W_{B_1} = 16$, $\mathcal{T} = 0.72$ms.)

Figure 4.4 displays the throughput of a data user at different choices of a class when other data users arbitrarily choose class B_1 or B_3. The results show that a data user obtains higher throughput by using class B_3 than by using B_1 or B_2, regardless of the choice between B_1 and B_3 of other users. This validates the result of Theorem 4.11. The total throughput is maximum when all data users choose class B_3, which is about 22% higher than the case when they all choose class B_1 at $N_u = 4$. This ratio becomes larger when N_u increases.

Note that the results for the case of two ACs per station (e.g. only B_1 and B_3 classes are available) can also be inferred from Figure 4.4. In particular, a data user always gets higher throughput by using class B_3, regardless of the choice of other users. This suggests that the PIA scheme is incentive compatible, resulting in all

data users to choose class B_3. This property of the PIA scheme is even stronger than the one proven in Theorem 4.11. Note that the throughput improvements reported here are at the MAC layer only and without considering any congestion control mechanism of the higher layers.

Comparison with the default EDCA parameters We can now compare the performance of the proposed PIA scheme with that of the default EDCA parameter setting in Table 4.3, under the assumption that all users will use whatever class gives them the best performance.

I consider the case of two ACs per station: B_1 and B_2. The MAC parameters specific to classes B_1 and B_2 are given in Table 4.6 with $\mathcal{T} = 0.72$ms. In the scenarios considered here, the value of η is varied from 1 to 5. The case of no service differentiation ($\eta = 1$) is included for comparison.

Table 4.6: MAC parameters of classes B_1 and B_2 used in Figs. 4.5, 4.6, 4.7 and 4.8.

Class	CW_{min}	CW_{max}	AIFSN	TXOP limit
B_2	$W_{B_2} = \eta W_{B_1} - \epsilon^0$ $(\epsilon^0 = 4(\eta - 1))$	$2^5 W_{B_2}$	2	$\eta\mathcal{T}$
B_1	W_{B_1}	$2^5 W_{B_1}$	2	$\mathcal{T}$

Under the default EDCA parameters, all users will use AC_VO, and under the PIA scheme, bulk data users will use class B_2 and real-time users will use class B_1.

The relative throughput of a saturated user under the PIA scheme is shown in Fig. 4.5 as functions of η for different numbers of saturated users, N_s. For comparison, the throughput under the default EDCA parameter setting (Table 4.3) is also shown. The throughput is again normalized by that of the PIA scheme at $\eta = 1$.

It can be seen from Figs. 4.5 and 4.2(a) that the throughput increases faster with η under the PIA scheme than it did under the proportional scheme, which reflects the reduction in CW_{min}. Fig. 4.5 shows that the PIA scheme provides better service for data users than the case of no service differentiation ($\eta = 1$), especially at high load. This is in contrast to the default EDCA parameter setting with all data users

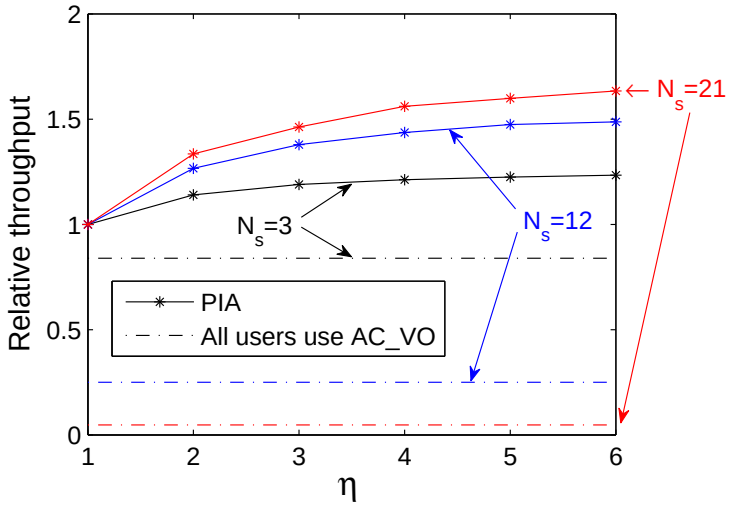


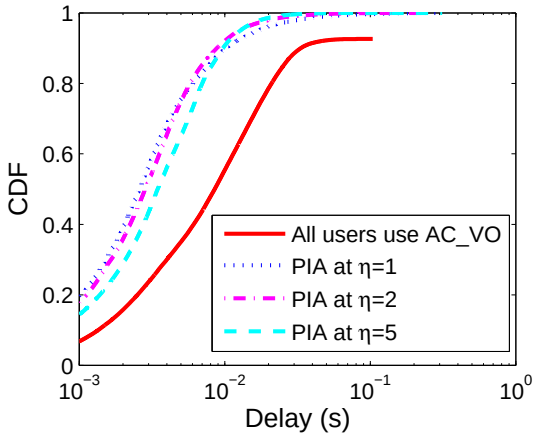
Figure 4.5: Throughput of a data user under the PIA scheme as a function of class B_2's TXOP limit in units of $\mathcal{T}$ (η), scaled by that of the PIA scheme at $\eta = 1$. The PIA scheme gives better throughput than the default EDCA setting with data users using AC_VO class. ($\lambda_u = 35$ packets/s, $l_s = 1000$ bytes, $l_u = 200$ bytes, $N_u = 6$, $W_{B_1} = 16$, $W_{B_2} = \eta W_{B_1} - \epsilon^0$, $\epsilon^0 = 4(\eta - 1)$.)

using real-time class (AC_VO), for which the performance degrades rapidly at high load.

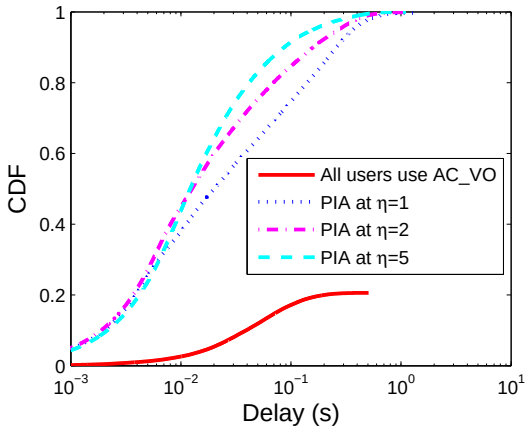
This improvement in throughput of the PIA scheme compared with that of the proportional scheme comes at the expense of increased delay for real-time users. Fig. 4.6 shows the probability that a packet of a real-time user is successfully transmitted before a given delay, for different η and loads $N_s = 3$, $N_s = 12$, and $N_s = 21$.

Fig. 4.6(a) shows that the PIA scheme at both $\eta = 2$ and $\eta = 5$ gives a higher probability that a packet is successfully delivered at a given delay than the default EDCA setting with all data users using the real-time class AC_VO. This means that the average packet delay under the PIA scheme is smaller. The cumulative distribution of delay for the default EDCA setting never reaches 1, which indicates loss rate. In this lightly loaded case, $\eta = 2$ provides comparable service to $\eta = 1$ (no service differentiation), and $\eta = 5$ provides slight degradation compared to $\eta = 1$, but less than that caused by the default prioritization setting.

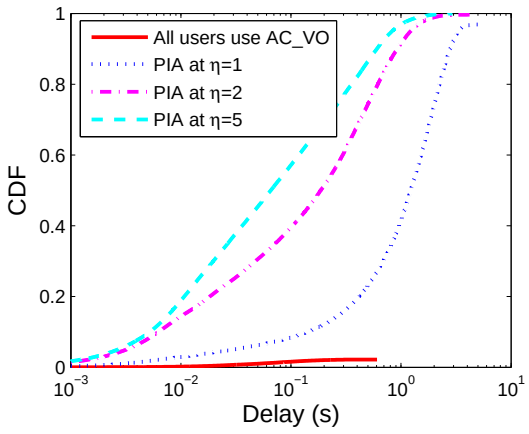
In the heavily loaded case of Figs. 4.6(b) and 4.6(c), the cumulative distribution



(a) $N_s = 3$



(b) $N_s = 11$



(c) $N_s = 21$

Figure 4.6: Probability a packet of real-time users is successfully delivered as a function of delay. ($\lambda_u = 35$ pkts/s, $l_s = 1000$B, $l_u = 200$B, $N_u = 6$, $W_{B_1} = 16$, $W_{B_2} = \eta W_{B_1} - \epsilon^0$, $\epsilon^0 = 4(\eta - 1)$.)

of delay for the default EDCA setting is much lower than 1, which indicates a high loss rate. In contrast, the PIA scheme has a low loss rate for all values of η tested, although some packets have very high delays. In this case, the benefit increases as η increases. Together with the result in Fig. 4.6(a), this implies that under the PIA scheme, the optimal η for real-time users increases with traffic load, as was observed for the proportional scheme. However, even using $\eta = 2$ for all loads appears to provide improvement over the default EDCA parameters.

In brief, although the optimal η in the PIA scheme depends on traffic load, it is clear that when $\eta = 2$, the PIA scheme provides better service for both traffic types under typical scenarios considered in this section and Section 4.5.3. This implies that when designing a network with an unknown number of users, the PIA scheme can be implemented by simply setting $\eta = 2$ and $\epsilon^0 = 4$. Adaptive schemes that set η dependent on the estimated load are possible, but out of the scope of this thesis.

4.5.3 Additional simulation results

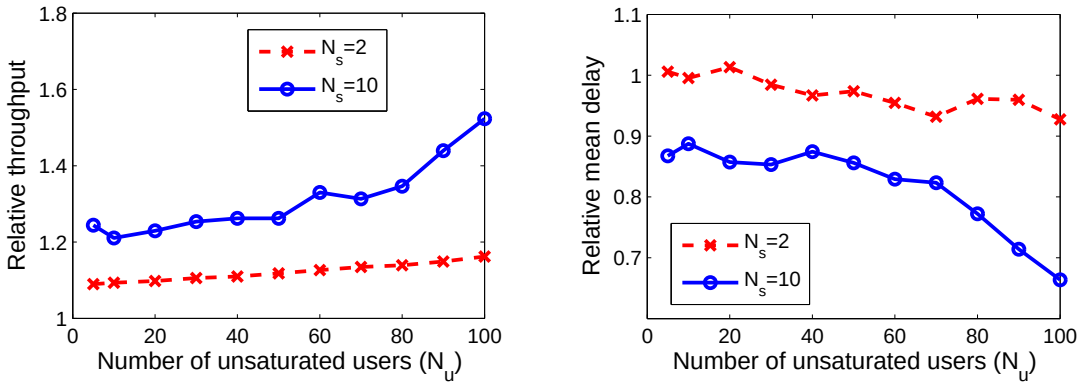
This section extends Section 4.5.2, by continuing the numerical study of the PIA scheme with two ACs, B_1 and B_2, per station under 802.11g for a wider range of traffic loads. The MAC parameters specific to these ACs are given in Table 4.6 with $\mathcal{T} = 0.72$ms.

Scenario 1

The following simulation results show that when $\eta = 2$, the PIA scheme provides better service for both real-time and data traffic than the case of no service differentiation ($\eta = 1$).

The ratio of the throughput of a data user under the PIA scheme at $\eta = 2$ to that under the case of no service differentiation ($\eta = 1$) is shown in Fig. 4.7(a) and the corresponding ratio of the mean delay of an unsaturated user is shown in Fig. 4.7(b), as functions of the number of unsaturated users N_u for different N_s.

Fig. 4.7(a) shows that when traffic load increases (N_s and/or N_u increases), the



(a) (b)

Figure 4.7: Ratio of throughput of a data user under the PIA scheme at $\eta = 2$ to that under no service differentiation ($\eta = 1$) and the corresponding ratio of mean delay of a real-time user, as a function of the number of realtime users. ($\lambda_u = 7$ packets/s, $l_s = 1200$ bytes, $l_u = 100$ bytes, $W_{B_1} = 16$.)

improvement in throughput under the PIA scheme in comparison with the case of no service differentiation increases. A similar trend is visible in the delay performance of an unsaturated user shown in Fig. 4.7(b). When traffic load increases, the delay performance under the PIA scheme becomes better than that under no service differentiation.

Note that under light load, the original 802.11 without service differentiation is good enough for both data users and realtime users; although the PIA scheme does not help much, help is not necessary, and the PIA scheme does not hurt. However, under heavy load where the original 802.11 needs help, my proposed scheme provide significant improvement for both types of traffic.

Scenario 2

So far, I have only considered the incentives facing saturated data users. Now I consider the incentive of realtime users in choosing a class.

I consider a network with 5 saturated users and 10 unsaturated users. Among the 10 unsaturated users, I tag a particular user and change its arrival rate in a wide range from 10 packets/s to 210 packets/s, while keeping the arrival rate of the other

9 unsaturated users at 40 packets/s. Then, I investigate how the mean delay of the tagged user changes with its arrival rate when it uses classes B_1 and B_2, respectively.

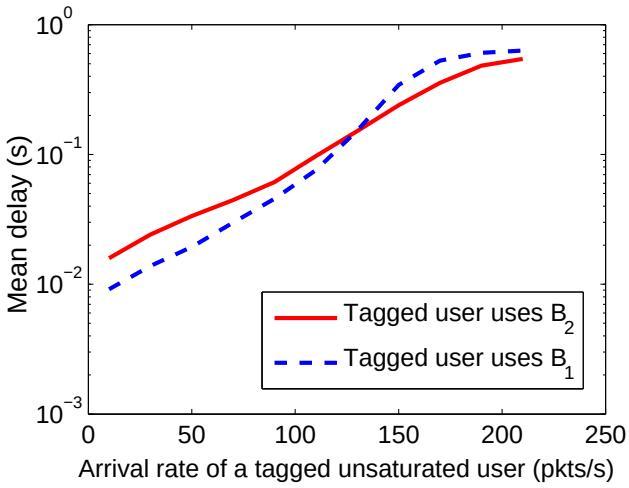


Figure 4.8: Mean delay of of the tagged "unsaturated user" as a function of its arrival rate. ($\lambda_{u\neq 1} = 40$ pkts/s, $l_s = 1200$B, $l_u = 100$B, $N_u = 10$, $N_s = 5$, $W_{B_1} = 16$, $\eta = 2$.)

Fig. 4.8 shows that when traffic load (e.g., the arrival rate of the tagged user increases) is not very high, the mean delay of the tagged "unsaturated" user is lower if it uses realtime class B_1, as expected. This is because the tagged user has negligible queueing at that traffic load. However, when traffic load reaches a certain threshold (e.g., the arrival rate of the tagged user is around 130 pkts/s), using the "bulk data" class B_2 gives the tagged "unsaturated" user lower delay. This is because the station incurs non-negligible queuing, and is no longer always unsaturated. This means that my previous assessments of the throughput of class $B_{k>1}$ assuming that realtime users use realtime class B_1 may not reflect the realtime users' actual choice. However, it is conservative because class B_1 causes the most collisions; if a realtime users change to using the "bulk data" class at high traffic load, this not only improves its delay but also helps data users due to the reduction in collisions.

4.5.4 Implication of multiple sources per station

The MAC model can be used to describe many situations in which a single station has multiple sources of data. For example, if the station has two saturated sources using the same access class, then the aggregate throughput is the same as if it had a single saturated source; the fraction of capacity going to each source is beyond the control of the MAC. If the station has two sources using different ACs, then it can be modelled as two separate stations, since each AC operates independently, with its own backoff counter.

Similarly, the game model remains appropriate in the (typical) case that each source does not base its choice on the class chosen by other sources on the same station. That occurs, for example, when an application has the choice of class hard-coded based on performance measurements made by the application writer. Since the fraction of time that typical wireless node is actively transmitting is small, the application should be designed on the assumption that it is the only active source.

If applications dynamically choose the class based on the choices of other sources on the same station, the situation becomes more complicated. Consider a station in which one saturated source has chosen to use the class n with the largest TXOP limit, giving throughput S. If another saturated source also chooses class n then both will get throughput $S/2$. However, if it chooses class $n-1$, which uses a different AC, then the independence between ACs means that it will achieve the same throughput as if it was the only source at that station, which is only slightly below S. That means that it is no longer a Nash equilibrium for all saturated sources to choose class n. Studying this situation is an interesting direction for future work, although its practical relevance is reduced by the fact that most applications will choose a suitable class at design time rather than run time.

4.6 Conclusion

In this chapter, I have shown that the priority-based service differentiation implied by the default 802.11e EDCA parameters creates an incentive for rational users to use the highest priority class. This can lead to the performance degradation of the whole network. Therefore, it is important to provide different but fair services, without giving all users the incentive to use the "highest priority" class.

I have also shown through both analysis and simulation that allowing users to adjust CW_{min} and TXOP limit in the same proportion provides service differentiation in WLANs. This scheme improves service for both data and real-time traffic, especially at high load. However, it still provides a slight incentive for data users to use real-time class's parameters although this incentive is much smaller than that caused by the default EDCA parameters.

This misalignment of incentives can be removed by increasing CW_{min} by a slightly smaller factor than the TXOP limit. My incentive adjusted scheme has many advantages over prior proposals: it improves service for both data and real-time traffic and provides the correct incentives for application optimizers, while allowing easy implementation: a single set of 802.11e MAC parameters provides tradeoff between throughput and delay over the range of load studied.

Note that in this chapter, the analysis is based on the assumption that data traffic is saturated and does not use congestion control. Moreover, it is also based on the modeling approximation that each source has the same collision probability at every attempt. In the next chapter, I will consider scenarios where those no longer hold.

Chapter 5

Extended validation

The models in the previous chapters made some assumptions about the collision probability of sources and the property of data sources. In particular, the general model in Chapter 3 uses a common mean-field approximation that, at every transmission attempt, and regardless of the number of retransmissions suffered, each packet of a source collides with constant and independent probability. I have shown that it is reasonably accurate under typical scenarios. Besides, the proposed model also assumes that data sources are saturated.

In this chapter, I will consider specific scenarios where those assumptions no longer hold. It is important to investigate the accuracy of the model under these scenarios and the implication on the performance of the PIA scheme proposed in Chapter 4. (Note that an extensive validation of the common hypotheses used in the modeling of saturated and unsaturated 802.11 infrastructure mode networks and saturated 802.11e networks with AIFS differentiation is provided in [58].)

Firstly, I find that under specific scenarios when there is big variability of packet sizes in the network, the collision probability of small packets from unsaturated sources is no longer the same on its first and subsequent attempts. This violates the mean field approximation of constant collision probability used in most previous models of 802.11 WLANs and the proposed model in Chapter 3, which may affect the accuracy of those models. Therefore, it is important to find out its cause and conditions when it is observed. Besides, it is also important to quantify how much the collision probabilities on the first and subsequent attempts are different, and how much improvement in model accuracy can be obtained by modeling this correctly. In this chapter, I will explain this phenomenon in Section 5.1.2, state the conditions

to observe this in Section 5.1.3, and extend the proposed model in Chapter 3 to capture that effect in Section 5.1.4. I find that correctly modeling this can improve the accuracy of collision probability of unsaturated sources up to 30%, which implies its importance for work requiring the accurate capture of collision probability. An example of such work will be provided in Section 5.1.5. Moreover, I also find that the effect of accurately modeling this phenomenon on the performance measures of users such as throughput or mean access delay is not significant (e.g. observed accuracy improvement is at most 8%), which explains the consistency of the properties of the PIA between the analytical model in Chapter 4 and simulation.

Secondly, I relax the assumption of saturated data users by considering data users governed by TCP. TCP implements congestion control which adapts the sending rate of TCP sources to the network condition, based on the use of acknowledgements (hereafter called "TCP ACKs") for every packet sent successfully. It has been shown in [25] that in 802.11 DCF WLAN, a TCP source may not be saturated due to the congestion of TCP ACK at the receiver (i.e. the AP for uplink traffic) and that this congestion can be reduced with the use of prioritization of access parameters defined in the 802.11e standard. In Section 5.2, I will discuss the implication of replacing saturated data sources by TCP sources on the modeling of 802.11 WLAN and the performance of the PIA scheme.

5.1 Packet size variability affects collisions

5.1.1 Introduction

The majority of existing analytical models to evaluate the performance of MAC protocol in WLANs including my model in Chapter 3 have been based on a mean-field approximation introduced in a seminal paper of Bianchi [21] which stated that, at each transmission attempt, and regardless of the number of retransmissions suffered, each packet of a source collides with constant and independent probability p given

by

$$p = 1 - (1 - \tau)^{n-1} \tag{5.1}$$

My main contribution here is to show that the existence of big packets in WLANs such as what occurs under my proposed scheme in Chapter 4 can make the above assumption inappropriate in estimating the collision probability of sources sending small packets. The inaccuracy stems from the fact that packets may experience different collision probabilities at different times, i.e., the collision probability is not homogeneous across time slots in the system. In particular, I will investigate the conditions where this effect is significant in general CSMA networks and show example scenarios of IEEE 802.11e EDCA WLANs in which significant difference in the packet size among different types of traffic is allowed by varying TXOP limit. The collision probability is of importance because the energy consumption of the battery powered mobile devices depends on the number of packet transmissions, which is directly related to the collision probability. Moreover, I also extend the proposed model in Chapter 3 to capture this effect and use it to optimize IEEE 802.11e EDCA parameters to minimize collision probability.

The effect described here applies to all networks based on CSMA which have significant persistence, such as the dominant IEEE 802.11 standards.

5.1.2 Main finding: Impact of big packets

Consider a WLAN with a large number N_u of unsaturated sources sending small packets, each with rate λ_u, and one source sending big packets of size l_b and transmission duration T_b. In this scenario, it is possible for sufficient small packets to accumulate during the transmission of a large packet, that the collision probability of small packets is significantly under-estimated by the mean field approximation (5.1).

While a large packet is being sent, on average $N_u \lambda_u T_b$ new small packets will arrive at the system. These will all attempt to transmit within the short persistence time, which in 802.11 is uniformly distributed up to 32 slots. As a result, the longer

the big packet is, the more small packets will attempt to transmit soon afterwards, and the higher the collision probability during that period.

This issue is illustrated in Figure 5.1. The curve U-U shows the probability that a small packet will collide with another small packet as a function of the number of slots since the most recent big packet. This is clearly elevated in the 32 slots corresponding to packets which arrived during the busy period. The scenario is for an 802.11e EDCA network, in which a station which wins a contention can send a burst of packets, whose length is TXOP limit. This has an effect analogous to a single long packet. From the results in Chapter 4, to balance the traffic load in the network, when a station sends bigger packets, its initial persistence period (CW_{min}) is also increased. In the example shown in Figure 5.1, one greedy source has an "effective" packet size of $l_b = 6000$ bytes by using a large TXOP limit, and its CW_{min} is increased in proportion to 192 slots. The small packets were from 10 sources sending 100 byte packets with a rate of 30 packets/s. These sources were quasi-periodic (see page 71 for the definition), in that the inter-packet time varied slightly around 1/30 second to eliminate phase effects.

Due to the effect of large packets, there exist high-contention and low-contention periods, which makes the contention level of slots not homogeneous. However, the mean-field approximation (5.1) assumes that the contention level is the same for every slot, which does not take into account the effect of large packets.

In systems such as 802.11, in which backoff intervals are measured in slots rather than absolute time, this effect primarily affects the first transmission attempt. On retransmission attempts, the sources are synchronized to the slot times, and are no more likely to transmit after a large (busy) slot than an idle slot. As a result, the collision probability of the first attempt is significantly larger than that of retransmissions. This is in contrast to the effect identified in [59] which occurs with unsaturated sources with large buffers. In that case, the collision probability of the first attempt can be significantly lower than retransmission attempts, because the first attempt may occur when few stations have packets to transmit, whereas

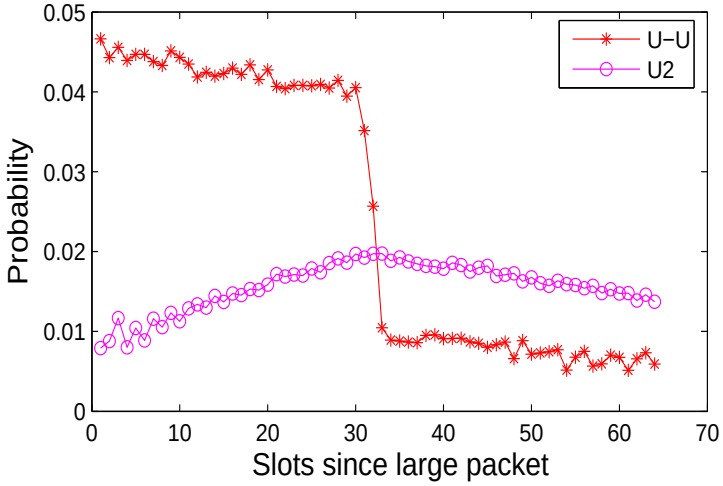


Figure 5.1: Collision probability between small packets in each slot of 64 slots right after a large packet (U-U) and the number of retransmission attempts of small packets in each slot normalized by dividing by the total number of retransmissions of small packets in those 64 slots (U2). (Scenario: an 802.11e EDCA WLAN with one greedy source sending large packets of 6000 bytes, 10 quasi-periodic sources sending small packets of 100 bytes with rate of 30 packets/s.)

retransmissions only occur during times of congestion.

The impact of large packets on the collision probability of a small packet on its first attempt and retransmission attempts is illustrated in Figure 5.2, which shows the probability that a small packet will collide with (a) another small packet, on its first attempt (U1-U), (b) another small packet on a retransmission attempt (U2-U), (c) another small packet, as determined from (5.1) with τ being the attempt probability of a small packet measured from simulation (U-U), or (d) a large packet (U-T). The same scenario is applied here as in Figure 5.1, with one greedy source sending large packets and 10 quasi-periodic sources sending small packets of 100 bytes with a rate of 30 packets/s. In Figure 5.2, the "effective" packet size l_b of the greedy source is varied by adjusting TXOP limit, and its CW_{min} is increased in proportion.

As can be seen from Figure 5.2, when the size of big packets increases, U1-U increases significantly while U2-U increases slowly. This means that, as big packets' size increases, the collision probability of a small packet on its first attempt becomes

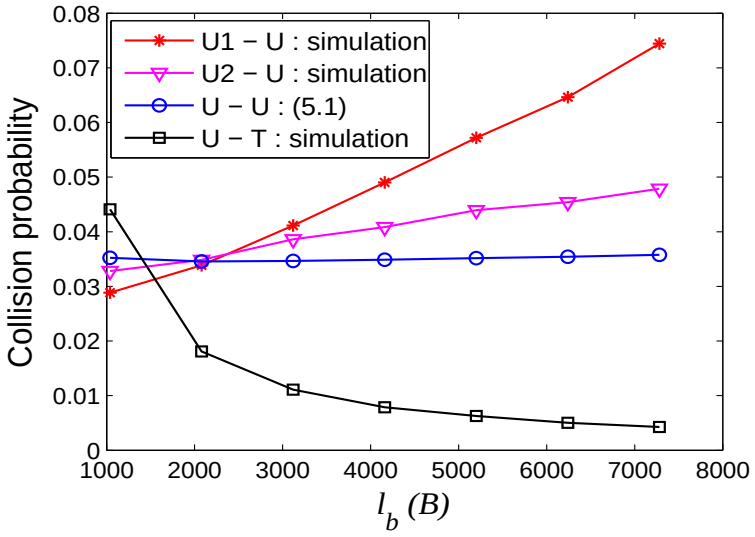


Figure 5.2: The classification of the collision probability of a small packet. (Scenario: an 802.11e EDCA WLAN with one greedy source sending large packets of l_b bytes, 10 quasi-periodic sources sending small packets of 100 bytes with rate of 30 packets/s.)

higher than that on retransmission attempts and the gap between those becomes bigger.

The estimated collision probability based on (5.1) is even lower than U2-U, which can be understood as follows. Without considering the effect of large packets, the collision probability of a small packet estimated by the mean-field approximation (5.1) will be similar to the 160-slot low-contention period (e.g. 160 is the difference in CW_{min} between a small-packet source and a big-packet source), the start of which is after the slot 32 as shown in the curve U-U of Figure 5.1. However, with the existence of big packets, retransmissions are much more likely to occur in the high-contention period than at other times as shown through the curve U2 in Figure 5.1 as well as in [59]. (Note that the curve U2 in Figure 5.1 displays the number of retransmission attempts of small packets in each slot of 64 slots after a large packet, normalized by dividing by the total number of retransmissions of small packets in those 64 slots.) This means that the presence of large packets also increases the collision probability of retransmissions beyond that predicted by (5.1), although less than for initial transmissions.

The significant decrease in U-T in Figure 5.2 occurs because of the increase in CW_{min} of the greedy source. This demonstrates that increasing TXOP limit and CW_{min} can make collisions with large packets negligible. The rest of Section 5.1 will focus on collisions between the unsynchronized small packets. Note that in this chapter, the term "unsynchronized" refers to the fact that a packet arrives at an empty transmission queue.

5.1.3 When does this effect occur?

Given the marked discrepancy between these simulation results and Bianchi's successful model, it is fair to ask why this effect has not been described before. Let us now consider the conditions under which this effect occurs.

Variable packet size

This effect will only occur when the expected number of arrivals, $N_u\lambda_u T_b$, is large (at least comparable to 1). If all packets are of equal duration T_b, then this corresponds to a heavily overloaded system. The fraction of time spent sending first attempts of unsynchronized packets, not counting retransmissions, is $N_u\lambda_u T_u$ where T_u is the size of such packets. In order for this to consume less than 100% of the resources and still to have $N_u\lambda_u T_b$ large, it is necessary that T_b be much larger than T_u. When this ratio is 1500/64, the maximum possible ratio under standard 802.11, the phenomenon only has a small effect on the performance metrics usually studied, namely packet delay and throughput. However, the introduction of 802.11e will make the effect more important.

Unsaturated sources

The arrival rate of small-packet sources should be small enough so that the queue of small-packet sources rarely builds up. The first attempt of a packet which arrives at a non-empty queue will be synchronized with the slot structure induced by 802.11.

Such packets do not contribute to the inspection paradox [120] of large numbers of packets arriving while the big packets are being transmitted.

Moderate spacing between big packets

According to the inspection paradox [120], a small packet which arrives at an empty transmission queue is more likely to see a long busy period than short busy period or idle slot. However, if the spacing between big packets is too large compared with the time to clear the backlog of unsaturated sources, then many small packets will still observe the "background noise" of independent small transmissions, which is captured well by the mean-field model. Conversely, if the spacing between big packets is small compared to the time to clear the backlog, then congestion periods will overlap each other, and the collision probability is again fairly constant in time.

The impact of three factors mentioned above are illustrated in Figure 5.3. Let p_{u1} and p_{u2} be the collision probability of a small packet on its first attempt and retransmission attempts respectively. The figure shows the ratio of p_{u1} to p_{u2} in two scenarios of an 802.11e EDCA WLAN with N_u quasi-periodic sources (see page for their definition) sending 100 byte packets with the rate of λ_u and one greedy source sending big packets with size l_b. The ratio of p_{u1} to p_{u2} is shown as a function of the time interval between big packets which is varied by changing CW_{min} of the greedy source denoted by W_s.

When W_s is small, the impact of big packets is small. This is partly because the contention does not have time to abate between transmissions, and partly because the probability that the queue of small-packet sources builds up is higher. When W_s is very high, the impact of big packet is again small because the probability that a small packet comes and senses channel busy due to the transmission of big packets is small.

The impact of big packets is most pronounced for W_s near 200 or 300, when W_s is high enough for the backlog of unsaturated sources to clear between big packets, but small enough that a large proportion of the unsynchronized packets arrive during

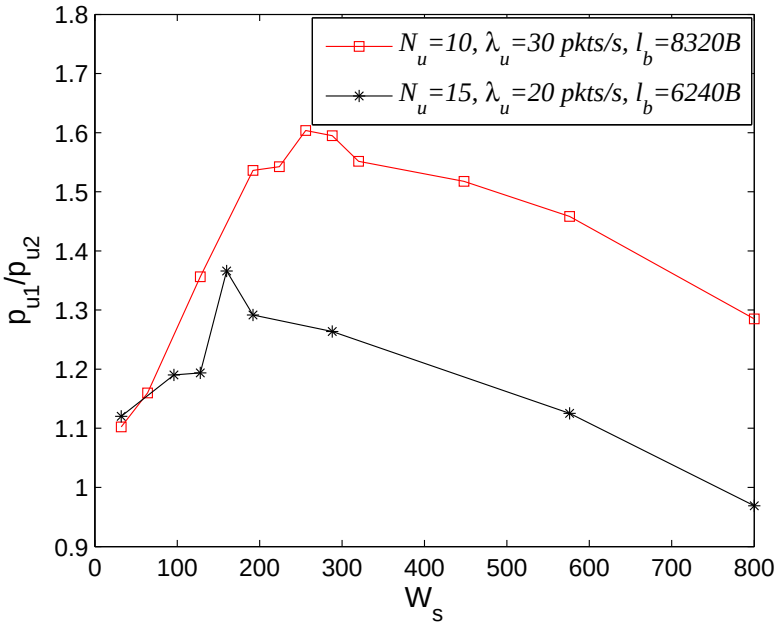


Figure 5.3: The ratio of the collision probability of a small packet on its first attempt to that on retransmission attempts, as a function of CW_{min} of saturated sources (W_s). (Scenario: an 802.11e EDCA WLAN with one saturated source sending large packets of l_b bytes, N_u quasi-periodic sources sending small packets of 100 bytes with rate of λ_u.)

the transmission of big packets.

Summary

The effect described above is largest when the following conditions occur:

- $N_u \lambda_u T_b$ is large (at least comparable to 1), which implies that
- the ratio of big packets' size to small packets' size is reasonably large;
- The time interval between big packets is of the same order as the time to clear the backlog of unsaturated sources caused by a busy period;
- Stations sending small packets are very unsaturated, so that minimal queue builds up even during the big packet transmissions, which requires that
- The number of unsaturated stations is large.

Moreover, the impact is clearer when the arrival process of small packets at a source is quasi-periodic than when it is Poisson, because this maximizes the number of unsynchronized arrivals.

5.1.4 Case study: 802.11e

In this section, I look at a particular scenario that covers the impact of big packets, propose an extended model which is based on the one in Chapter 3 to capture that impact and then evaluate the extended model.

Consider an 802.11e EDCA WLAN with N_u identical low-rate realtime sources with Poisson arrival of rate λ_u and several (N_s) identical greedy data sources sending data to an AP. Greedy data sources send big packets by using large TXOP limit ($\eta_s = const >> 1$) while realtime sources send small packets ($\eta_u = 1$). The model is for the scheme proposed in Chapter 4 in which, to balance the traffic load in the network, when a greedy data source sends bigger packets, it must increase the time interval between its attempts by using higher CW_{min}. The traffic load in the network is kept at a level which rarely makes the queue of realtime traffic build up.

According to the analysis in Section 5.1.2, the existence of big packets from greedy data sources can make the collision probability of a small packet from realtime sources on its first attempt much different from retransmission attempts.

Extended model

Motivated by the above observations, I now relax assumption that sources collide with a constant probability as in Chapter 3 and allow a packet from a non-saturated station to have a different collision probability on its first attempt p_{u1} from that on retransmission attempts p_{u2}.

To define the extended model, consider a tagged non-saturated station with a tagged packet arriving during a busy slot (so that it is synchronized with the global slots). Let N_{u1} be the (random) number of non-saturated stations with a packet arriving during a busy slot within W_u slots *before* the first attempt of the tagged

packet. Let I be the condition that a given non-saturated station is neither one of these N_{u1} nor the tagged node, and let R_u be the event that a given non-saturated station makes a *re*transmission attempt in a given slot.

I assume that the retransmission limit is infinite. Then, let $g(p_{u1}, p_{u2})$ be the average number of attempts per packet from an unsaturated source, given by

$$\begin{aligned} g(p_{u1}, p_{u2}) &= 1 + p_{u1} + p_{u1}p_{u2} + p_{u1}p_{u2}^2 + \cdots \\ &= 1 + p_{u1}(1 + p_{u2} + p_{u2}^2 + \cdots) \\ &= 1 + \frac{p_{u1}}{1 - p_{u2}}. \end{aligned} \tag{5.2}$$

The extended model is then based on (3.11) with the attempt probability of unsaturated stations in (3.11b) governed by

$$\begin{aligned} \tau_u &= \lambda \mathbb{E}[Y] g(p_{u1}, p_{u2}) \\ &= \lambda \mathbb{E}[Y](1 + \frac{p_{u1}}{1 - p_{u2}}), \end{aligned} \tag{5.3a}$$

where the collision probability on the first attempt is

$$p_{u1} = h(\tau_s, \tau_u) = b_u \left(1 - (1 - \tau_s)^{N_s} \left(1 - \frac{1}{W_u} \right)^{\mathbb{E}[N_{u1}]} (1 - P(R_u|I))^{N_u - (\mathbb{E}[N_{u1}]+1)} \right). \tag{5.3b}$$

and the collision probability on the subsequent attempts is given by (3.11c).

Equation (5.3b) reflects the fact that the tagged packet will collide with an attempt from any of the saturated stations (which attempt with probability τ_s), or of the N_{u1} non-saturated stations in their first attempt (which attempt with probability $1/W_u$), or of the $N_u - (N_{u1} + 1)$ non-saturated stations in their subsequent attempts (which attempt with probability $P(R_u|I)$). Note that a more accurate model would use the expectation of the exponentiation (e.g. $\mathbb{E}[(1 - \frac{1}{W_u})^{N_{u1}}$) but for tractability, I use the expectation of N_{u1} by applying the binomial approximation. This is justified because $1/W_u$ is close to 0 for large enough W_u, which is usually the case where the

effect is observed.

The mean of N_{u1} is given by

$$\mathbb{E}[N_{u1}] = (N_u - 1)\lambda_u(\mathbb{E}[2T_{res,u}] + b_u(W_u - 1)\mathbb{E}[Y_u]), \tag{5.3c}$$

as the arrival process is the superposition of rate-λ_u Poisson arrival processes from $(N_u - 1)$ non-saturated stations, observed during both the busy slot in which the tagged packet arrived (of mean duration $\mathbb{E}[2T_{res,u}]$) and the remaining $W_u - 1$ slots of mean duration $\mathbb{E}[Y_u]$. The busy probability b_u arises because packets arriving when the channel is idle are transmitted immediately, and carrier sensing inhibits collisions with the packets synchronized to slot boundaries.

To calculate $P(R_u|I)$, note first that the $N_{u1} + 1$ nodes ready for their first transmissions cannot retransmit, whence $P(R_u|\overline{I}) = 0$. Hence, the law of total probability gives

$$P(R_u|I) = \frac{N_u}{N_u - (\mathbb{E}[N_{u1}] + 1)} P(R_u). \tag{5.3d}$$

where

$$\begin{aligned} P(R_u) &= \mathbb{E}[\text{Retransmission attempts per source per slot}] \\ &= \left(\frac{g(p_{u1}, p_{u2}) - 1}{g(p_{u1}, p_{u2})}\right) \tau_u \end{aligned} \tag{5.3e}$$

which completes the extended model.

Note that the average collision probability p_u is the weighted sum

$$\begin{aligned} p_u &= \frac{1}{g(p_{u1}, p_{u2})} p_{u1} + \left(1 - \frac{1}{g(p_{u1}, p_{u2})}\right) p_{u2} \\ &= \frac{p_{u1}}{1 + p_{u1} - p_{u2}} \end{aligned} \tag{5.4}$$

which is approximately the first attempt collision probability when collisions are rare.

Evaluation

In this section, I demonstrate that, in appropriate circumstances, the extended model captures important qualitative properties of the collision probabilities which are not captured by the mean-field based model (hereafter called the "traditional model"). This is done by comparing these models with simulations performed using the *ns-2* simulator (version 2.33) [1], combined with an EDCA module [139].

Consider a network which consists of N_u non-saturated sources sending small packets and N_s saturated sources sending bursts of η_s packets. These stations will send packets to an access point in ideal channel conditions. As for Figures 5.1 and 5.2, saturated sources increase the spacing between their packets (W_s) in proportion to their burst size (η_s) to balance the throughput. As opposed to Poisson unsaturated sources assumed in the model, the packet inter-arrival times of unsaturated sources are uniformly distributed in the range $1/\lambda_u \pm 10\%$, to model voice traffic with enough jitter to avoid phase effects. The rate λ_u was sufficiently low that queues rarely built up.

Both saturated stations and non-saturated stations use UDP. The MAC and physical layer parameters were the default values in IEEE 802.11b, as shown in Table 3.1.

The collision probability of a small packet from an unsaturated station is shown in Figure 5.4 as a function of burst size of saturated stations (η_s). This figure shows the collision probability determined from the traditional model and simulation, and the collision probability on the first attempt and retransmission attempts determined from the extended model and simulation. The traditional model incorrectly predicts the collision probability to decrease monotonically, while the extended model can capture the right trend of the collision probability on both the first and retransmission attempts.

The behavior can be understood by comparing with Fig. 5.2. When η_s increases, the first attempt collision probability initially decreases because the increase in W_s

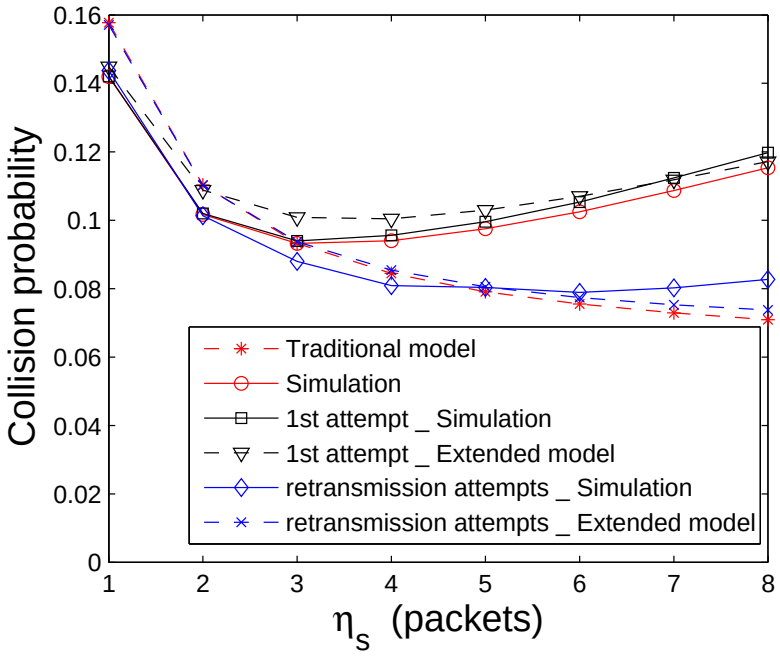


Figure 5.4: Collision probability of a small packet from an unsaturated source, as a function of burst size of saturated sources (η_s). ($N_u = 10$, $N_s = 2$, $\lambda_u = 30$ packets/s, $l_s = 1040$ bytes, $l_u = 100$ bytes, $W_u = 32$, $W_s = \eta_s W_u$.)

decreases U-T. When U-U begins to dominate, the collision probability increases, with U1-U increasing more markedly. The extended model does not capture the eventual increase in the retransmission collision probability, which occurs when the increase in U2-U exceeds the decrease in U-T.

After considering many scenarios where this phenomenon is observed, I find that the effect of modeling the phenomenon accurately on the performance measures of users such as throughput or mean access delay is not significant (e.g. observed accuracy improvement is at most 8%). This explains the fact that the theoretical properties of the PIA scheme proven from the model in Chapter 4 are consistent with those from simulation. However, modeling this accurately is useful for work requiring the accurate description of collision probability. In the next section, I will provide such an example of the application of the extended model.

5.1.5 Application to energy efficiency

The energy consumption of a wireless transmitter is proportional to the number of packets transmitted. Since collisions are wasted transmissions, an energy-saving design will seek system parameters which reduce collisions, possibly at the expense of higher delay.

A natural tradeoff in an 802.11e network is to encourage delay-insensitive stations to transmit seldom (large CW_{min}), and to achieve fair throughput by sending very large bursts when they do send (large TXOP limit). Delay-sensitive stations will still send frequent small packets, leading to the burstiness effect studied above.

If the tradeoff between TXOP limit, represented by burst size η_s, and CW_{min} is chosen using the standard model (5.1), then Fig. 5.4 suggests that an infinite TXOP limit may minimize the predicted collision probability of a small packet due to the trend of collision probability decreasing with burst size η_s.

In contrast, the extended model allows the optimal TXOP limit, represented by the optimal burst size (η_s), to be determined quite accurately. Note that the collision probability of a small packet from a non-saturated source is given by (5.4).

Fig. 5.5 shows the optimal burst size determined from the extended model for a typical situation, which is quite close to that from simulation.

5.2 TCP data sources

Let us now consider the other important modeling assumption of Chapter 3.

The model proposed in Chapter 3 assumes that data sources are saturated, which means that it always has at least one packet waiting to transmit at the MAC sublayer. This implies data sources are inelastic, and do not adapt their sending rate to the available channel capacity. However, in practice, Internet traffic is dominated by the TCP-based applications [92, 114, 151]; hence, most data traffic in WLANs is carried via TCP.

Existing research on TCP in WLANs focuses on the following categories: (1)

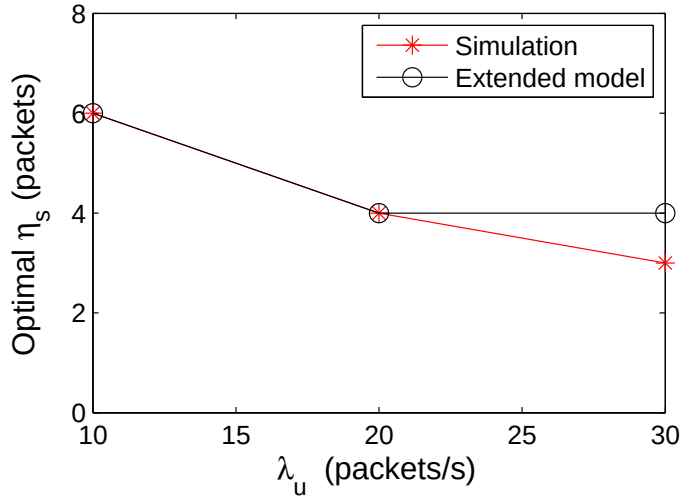


Figure 5.5: Value of burst size of greedy sources (η_s) which minimizes collision probability of small packets, as determined by (a) the extended model which considers bursty collisions, and (b) simulation. The optimal η_s predicted from the traditional model is infinite, and off the scale of the graph. ($N_u = 10$, $N_s = 2$, $l_s = 1040$ bytes, $l_u = 100$ bytes.)

modeling the interaction between TCP flow control and medium access control mechanism [24, 25, 32, 76, 94, 150, 151]; (2) enhancing TCP performance over WLANs by prioritizing TCP ACKs [83, 92, 114] or reducing resource inefficiency due to spurious timeouts and retransmissions [30, 70, 87, 92, 123, 152]; (3) improving TCP fairness in WLANs [47, 57, 66, 83, 125].

Recall that TCP is different from the inelastic sources assumed in the model in the way that it uses a congestion control algorithm to adapt the sending rate to network condition inferred from the reception status of TCP ACKs [33]. In the following, I will discuss the implication of TCP data sources on WLANs modeling and on the PIA scheme.

5.2.1 Implication on modeling

In my proposed model as well as other models assuming saturated data sources, the complex interaction between the collision avoidance mechanisms employed by the 802.11 MAC protocol and the closed-loop behavior of TCP is not captured.

First of all, the flow control algorithm employed by the TCP will limit the number of packets sent to the MAC sublayer. Wireless stations which are TCP senders will have TCP data packets to transmit depending on the pace of TCP ACKs received from TCP receivers; hence, the throughput of a TCP source depends much on how fast TCP ACKs are received. In an infrastructure WLAN, the AP is the receiver of upload TCP flows. Hence, the AP is usually the bottleneck of upload TCP ACKs. This can cause the wireless stations or TCP senders to be in idle state, without data packets to transmit at the MAC. Hence, the assumption of saturated data sources can be violated with data sources using TCP.

There have been several works modeling the TCP performance over IEEE 802.11 DCF WLANs [24, 25, 32, 94, 150]. Among those, [24, 25, 32, 150] use a Markov chain to model the number of TCP packets stored in the stations' queues, based on which TCP throughput performance can be calculated. [94] uses a simpler analysis where the statistics of the number of stations with non-empty queue (also called "active" stations) are obtained by assuming that it follows a Bernoulli distribution law. Noticeably, [25] proved that, in 802.11 DCF WLANs, TCP stations are sporadically active, whereas the AP stores most of the traffic generated by the TCP connections. This is because the AP has to contend for channel access with several uplink CSMA instances with the same MAC parameters. It also shows that without UDP traffic, the total TCP throughput is basically independent of the number of open TCP connections and the aggregate TCP traffic can be equivalently modeled as two saturated flows.

However, in 802.11e, MAC parameters can be used to prioritize TCP ACKs, which helps to reduce the congestion at the TCP receiver (i.e. the AP for uploading). This makes the packets queue up at the stations instead, which means stations are saturated [151]. This helps to increase TCP throughput. There have been several proposals using MAC parameters to prioritize TCP ACKs. In particular, [83] suggests to use smaller AIFS and CW_{min} to send upload TCP ACKs, which has been shown to increase TCP throughput and fairness among TCP flows. Similarly,

[92] proposes to create a separate AC for TCP ACKs, the AIFS and CW_{min} of which are smaller than those of the class AC_VO but the TXOP limit value is as large as that of the class AC_VO. [114] also suggests to use the highest priority class for TCP ACKs, together with a dynamic adaptation of MAC parameters to improve TCP throughput with minimal negative impact on high priority delay-sensitive UDP-based traffic.

5.2.2 Implication on the PIA scheme

The properties of the PIA proven in Chapter 4 is based on the model which assumes saturated data sources, inherited from the one in Chapter 3. As mentioned above, without the prioritization of TCP ACKs at the AP, the congestion of TCP ACKs will cause TCP sources to be unsaturated. In contrast, with the prioritization of TCP ACKs using access parameters in 802.11e WLANs, this assumption is more realistic. However, these access parameters have to be carefully chosen so that it improves the TCP throughput without harming delay-sensitive traffic too much.

I have collected some preliminary simulation results to evaluate the performance of the PIA scheme where data sources use TCP. The access parameters of the AP were set to be the same as those of the class AC_VO. These results suggest that much further work is required to investigate PIA into an environment dominated by TCP. However, I leave further investigation for future work.

5.3 Conclusion

In this chapter, I have investigated the case where the standard mean-field approximation of constant collision probability used in the proposed model in Chapter 3 no longer holds and then briefly discussed how the analysis in Chapter 4 may be affected. Besides, I have also discussed the implications of replacing saturated data sources assumed in the model by TCP data sources on the 802.11 WLAN modeling and the performance of the PIA scheme.

I have first considered wireless networks with heterogeneous packet sizes, in which some sources are unsaturated. It has shown that the accumulation of small packets from unsaturated sources during the transmission of a large packet can cause the collision probability of small packets to be much larger than predicted by previous models. This effect is particularly marked on the packet's first transmission attempt. When this occurs, it invalidates the common assumption that collision probabilities are independent and identically distributed. I have also proposed a model capturing this effect which can be used to optimize energy consumption of a station by minimizing its collision probability. This shows that the observed effect may have important implications, which should be considered in future models of CSMA-based networks with high heterogeneity of packet sizes and unsaturated sources. I have also found that the influence of this on performance measures such as throughput or mean delay is not significant, which explains why the proven properties of the PIA scheme is consistent with the simulation results.

Besides, when TCP data sources are considered instead of saturated ones, my preliminary results of the performance of the PIA scheme differs significantly, which suggests that much further work remains to be done. This may be related to the congestion control algorithm, which controls the sending rate of a TCP source in accordance with the return of TCP ACKs. However, further investigation is out of the scope of the thesis.

Chapter 6

Conclusion

Service differentiation has become more and more important in WLANs due to the increasing diversity of applications with different QoS requirements. Most of the existing proposals to provide service differentiation are based on prioritization, which provides better service in all aspects for a service class with higher priority. These proposals create an incentive for selfish users to use the access class of the highest priority to gain a higher share of the channel. This can degrade the overall performance of the network and result in no service differentiation, which shows the importance of QoS provision for WLANs with selfish users. The existing solutions to this issue in the literature are either complicated or impractical to implement. This raises a research question about the existence of a scheme to provide service differentiation which is easy to implement, compatible with the 802.11e standard and robust against selfish users, which have been addressed in the thesis.

6.1 Contributions

The thesis has contributed a scheme to provide service differentiation which is easy to implement and compatible with the 802.11e EDCA standard by scaling two parameters of MAC layer, TXOP limit and CW_{min}, in nearly the same ratio. In the proposed scheme, the access class for delay-sensitive traffic has the smallest CW_{min} and TXOP limit because sources of this type require their packets to be transmitted as soon as possible but usually has only one packet waiting to transmit. In contrast, throughput-intensive traffic has the largest CW_{min} and TXOP limit because it may be willing to wait a little longer, if an increase in the amount it can transmit per

channel access makes its overall throughput higher.

The proposed scheme has been shown to be robust against selfish users who try to maximize their performance at the cost of others, by using a game framework based on a novel model of 802.11e EDCA proposed in the thesis. The proposed model of 802.11e EDCA WLANs is more tractable and more accurate by capturing several features introduced by TXOP limit differentiation, which have not been taken into account in the previous models in the literature. Those include the closed form distribution of the number of packets sent by an unsaturated source per channel access, the residual time of an ongoing transmission from other stations seen by a burst of an unsaturated source arriving during that transmission, and the probability that a packet arrives at an empty buffer in the delay model. The model is not only useful for analyzing the proposed scheme of service differentiation but also helps to gain some insight into the properties of 802.11e EDCA WLANs such as the asymptotic analysis of delay distribution.

Moreover, an extension of the proposed model has also been proposed to model scenarios with the big variability of packet sizes in the network. In these scenarios, the collision probability of small packets from unsaturated sources is no longer the same on its first and subsequent attempts, which violates the mean field approximation of constant collision probability used in most previous models of 802.11 WLANs.

6.2 Implications of the work in the thesis

Through the work in this thesis, there are several implications which are generally noteworthy, especially for researchers in the same field.

- With its nice properties, the proposed scheme of QoS provision is promising to be implemented in practice, although its performance with TCP traffic needs further investigation.

- The proposed scheme is based on an idea of providing "different but fair"

services, which is originally developed for wired networks. This implies that an idea developed for one context may not work well in that context, but can be suitable for another.

- The fact that I find some scenarios which significantly violate the standard assumption of constant collision probability implies that when extending a model to capture new features, special attention needs to be paid to assumptions.

- My work is another example which shows that game theory is a very useful tool to analyze the network with selfish users in wireless networks.

6.3 Future work

Note that for the proposed scheme to be implemented, there requires the support of a signaling mechanism between layers of the network stack to use services supported at the MAC layer, just the same as what required for 802.11e EDCA. I leave this for others to work on. This research has opened up new avenues of investigation, with the following possible directions.

- Consider the case where applications dynamically choose the class based on the choices of other sources on the same station. Recall from the discussion in Chapter 4 that the choice of an access class of a source may be dependent on the choices of other sources in the same station, which is because different access classes maintain different queues with independent backoff procedure. I think this will be interesting to study.

- Conduct more validation of the proposed scheme with TCP traffic and investigate how to improve the proposed scheme in this case. Recall from Chapter 5 that I obtained some simulation results with TCP traffic and the results are not impressive. Because TCP traffic is widely used in practice, improving the proposed scheme with TCP traffic is worth as future work.

6.4 Final remarks

To conclude, the thesis has proposed a potential scheme to provide service differentiation which is easy to implement, compatible with the 802.11e EDCA standard of WLANs, and robust against selfish users who choose any access class which maximizes their own performance at the cost of others. The properties of this scheme have been verified using the proposed analytical model as well as ns-2 simulation. Due to its nice properties, the proposed scheme can be promising to be implemented in practice once its performance with TCP traffic is improved.

Bibliography

[1] The network simulator ns-2. `http://www.isi.edu/nsnam/ns/`.

[2] Qbone home page. http://qbone.internet2.edu/.

[3] IEEE standard for information technology - telecommunications and information exchange between systems - local and metropolitan area networks - specific requirements - part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications, 1999.

[4] Supplement to IEEE standard for information technology telecommunications and information exchange between systems local and metropolitan area networks specific requirements part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications: Higher-Speed physical layer extension in the 2.4 GHz band, 1999.

[5] Supplement to IEEE standard for information technology telecommunications and information exchange between systems local and metropolitan area networks specific requirements part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications: High-speed physical layer in the 5 GHz band, 1999.

[6] IEEE standard for information technology - telecommunications and information exchange between systems - local and metropolitan area networks - specific requirements - part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications - amendment 4: Further higher data rate extension in the 2.4 GHz band, 2003.

[7] IEEE standard for information technology - telecommunications and information exchange between systems - local and metropolitan area networks - specific requirements - part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications - amendment 8: Medium access control (MAC) quality of service enhancements, 2005.

[8] IEEE standard for information technology - telecommunications and information exchange between systems - local and metropolitan area networks - specific requirements - part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications, 2007.

[9] IEEE standard for information technology - telecommunications and information exchange between systems - local and metropolitan area networks - specific requirements - part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications - amendment 5: Enhancements for higher throughput, 2009.

[10] IEEE standard for information technology - telecommunications and information exchange between systems - local and metropolitan area networks - specific requirements - part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications, 2012.

[11] I. Aad and C. Castelluccia. Differentiation mechanisms for IEEE 802.11. In *Proceedings of IEEE International Conference on Computer Communications (INFOCOM'01)*, volume 1, pages 209–218, 2001.

[12] J. Abbate and W. Whitt. Numerical inversion of probability generating functions. In *Operations Research Letters 12*, pages 245–251, 1992.

[13] O. Abu-sharkh and A. Tewfik. Toward accurate modeling of the IEEE 802.11e EDCA under finite load and error-prone channel. *IEEE Transactions on Wireless Communications*, 7(7):2560 –2570, july 2008.

[14] H. M. K. Alazemi, A. Margolis, J. Choi, R. Vijaykumar, and S. Roy. Stochastic modelling and analysis of 802.11 DCF with heterogeneous non-saturated nodes. *Computer Communications*, 30:3652–3661, 2007.

[15] E. Altman, R. E. Azouzi, and T. Jimenez. Slotted aloha as a stochastic game with partial information. In *Proceedings of WiOpt*, 2002.

[16] E. Altman, T. Boulogne, R. El-Azouzi, T. Jiménez, and L. Wynter. A survey on networking games in telecommunications. *Computers & Operations Research*, 33(2):286–311, February 2006.

[17] E. Altman, A. Kumar, D. Kumar, and R. Venkatesh. Cooperative and Non-Cooperative control in IEEE 802.11, August 2005.

[18] G. Bacci and M. Luise. A noncooperative approach to joint rate and power control for infrastructure wireless networks. In *Proceedings of the 2009 International Conference on Game Theory for Networks (GameNets'09)*, pages 33–42, 2009.

[19] A. Banchs and X. Perez. Distributed weighted fair queuing in 802.11 wireless LAN. In *Proceedings of IEEE International Conference on Communications (ICC'02)*, volume 5, pages 3121 – 3127, 2002.

[20] B. Bellalta, C. Cano, M. Oliver, and M. Meo. Modeling the IEEE 802.11e EDCA for MAC parameter optimization. In *Proceedings of IEEE Consumer Communications and Networking Conference (CCNC'06)*, volume 1, pages 390–394, 2006.

[21] G. Bianchi. Performance analysis of the IEEE 802.11 distributed coordination function. *IEEE Journal on Selected Areas in Communications*, 18(3):535–547, 2000.

[22] G. Bianchi, I. Tinnirello, and L. Scalia. Understanding 802.11e contention-based prioritization mechanisms and their coexistence with legacy 802.11 stations. *IEEE Network*, 19(4):28 – 34, july-aug. 2005.

[23] G. Bianchi, I. Tinnirello, and L. Scalia. Understanding 802.11e contention-based prioritization mechanisms and their coexistence with legacy 802.11 stations. *IEEE Network*, 19(4):28–34, 2005.

[24] R. Bruno, M. Conti, and E. Gregori. Performance modelling and measurements of TCP transfer throughput in 802.11-based WLANs. In *Proceedings of the 9th ACM Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems (ACM MSWiM'06)*, volume 2006, pages 4–11, 2006.

[25] R. Bruno, M. Conti, and E. Gregori. Throughput analysis and measurements in IEEE 802.11 WLANs with TCP and UDP traffic flows. *IEEE Transactions on Mobile Computing*, 7(2):171 – 186, Feb. 2008.

[26] M. Cagalj, S. Ganeriwal, and J.-P. Hubaux. On selfish behavior in CSMA/CA networks. In *Proceedings of IEEE International Conference on Computer Communications (INFOCOM'05)*, 2005.

[27] P. Chaporkar, A. Proutiere, and B. Radunovi. Rate adaptation games in wireless LANs: Nash equilibrium and price of anarchy. In *Proceedings of IEEE International Conference on Computer Communications (INFOCOM'10)*, 2010.

[28] L. Chen and J. Leneutre. A game theoretic framework of distributed power and rate control in IEEE 802.11 WLANs. *IEEE Journal on Selected Areas in Communications*, 26(7):1128–1137, 2008.

[29] M. H. Cheung, A. H. Mohsenian-Rad, V. W. S. Wong, and R. Schober. Random access protocols for WLANs based on mechanism design. In *Proceedings of IEEE International Conference on Communications (ICC'09)*, pages 4892–4897, 2009.

[30] I. Cho, J. Han, and J. Lee. Enhanced response algorithm for spurious TCP timeout (ER-SRTO). In *International Conference on Information Networking (ICOIN'08)*, 2008.

[31] J. Cho and Y. Jiang. Basic theorems on the backoff process in 802.11. *ACM SIGMETRICS Performance Evaluation Review*, 37(2):18–20, 2009.

[32] S. Choi, K. Park, and C.-K. Kim. Performance impact of interlayer dependence in infrastructure WLANs. *IEEE Transactions on Mobile Computing*, 5(7):829–845, 2006.

[33] D. E. Comer. *Internetworking with TCP/IP: Principles, protocols, and architecture*. Upper Saddle River, N.J. : Pearson Prentice Hall, 5th edition, 2006.

[34] N. Dao and R. Malaney. A new markov model for Non-Saturated 802.11 networks. In *IEEE Consumer Communications and Networking Conference (CCNC'08)*, pages 420–424, 2008.

[35] A. Demers, S. Keshav, and S. Shenker. Analysis and simulation of a fair queueing algorithm. *ACM SIGCOMM Computer Communication Review*, 19(4):1–12, August 1989.

[36] J. Deng and R.-S. Chang. A priority scheme for IEEE 802.11 DCF access method. *IEICE Transactions on Communications*, E82-B(1):96–102, 1999.

[37] P. K. Dutta. *Strategies and Games: Theory and Practice*. MIT Press, 2nd edition, 1999.

[38] P. E. Engelstad and O. N. sterb. Analysis of the total delay of IEEE 802.11e EDCA and 802.11 DCF. In *IEEE International Conference on Communications (ICC'06)*, volume 2, pages 552–559, 2006.

[39] P. E. Engelstad and O. N. Østerbø. Non-saturation and saturation analysis of IEEE 802.11e EDCA with starvation prediction. In *Proceedings of ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM)*, 2005.

[40] Y. Pourmohammadi Fallah, S. El-Housseini, and H. Alnuweiri. A generalized saturation throughput analysis for IEEE 802.11e contention-based MAC. *Wireless Personal Communications*, 47(2):235–245, 2008.

[41] E. Felemban and E. Ekici. Single hop IEEE 802.11 DCF analysis revisited: Accurate modeling of channel access delay and throughput for saturated and unsaturated traffic cases. *IEEE Transactions on Wireless Communications*, 10(10):3256–3266, 2011.

[42] V. Firoiu, X. Zhang, and Y. Guo. Best effort differentiated services: Trade-off service differentiation for elastic applications. In *Proceedings of IEEE International Conference on Telecommunications (ICT)*, June 2001.

[43] D. Fudenberg and J. Tirole. *Game Theory*. The MIT Press, 1991.

[44] B. Gaidioz and P. Primet. EDS: A new scalable service differentiation architecture for internet. In *Proceedings of IEEE Symposium on Computers and Communications*, pages 777–782, July 2002.

[45] L. Galluccio. A Game-Theoretic approach to prioritized transmission in wireless CSMA/CA networks. In *IEEE Vehicular Technology Conference*, April 2009.

[46] S. Jamaloddin Golestani. Self-clocked fair queueing scheme for broadband applications. In *Proceedings of IEEE International Conference on Computer Communications (INFOCOM'94)*, volume 2, pages 636–646, 1994.

[47] J. Ha, E.-C. Park, K.-J. Park, and C.-H. Choi. A cross-layer dual queue

approach for improving TCP fairness in infrastructure WLANs. *Wireless Personal Communications*, 51(3):499–516, 2009.

[48] Z. Han, D. Niyato, W. Saad, T. Basar, and A. Hjrungnes. *Game Theory in Wireless and Communication Networks: Theory, Models, and Applications*. Cambridge University Press, 2012.

[49] L. Harte. *Introduction to 802.11 Wireless LAN (WLAN): Technology, Market, Operation, Profiles, Services*. Fuquay Varina, N.C. : ALTHOS Publishing Inc, 2004.

[50] M. Hayajneh and C. T. Abdallah. Distributed joint rate and power control game-theoretic algorithms for wireless data. *IEEE Communications Letters*, 8(8):511–513, 2004.

[51] J. He, L. Zheng, Z. Yang, and C. T. Chou. Performance analysis and service differentiation in IEEE 802.11 WLAN. In *Proceedings of IEEE Conference on Local Computer Networks (LCN)*, page 691697, 2003.

[52] R. P. F. Hoefel. Frame aggregation and concatenation schemes for IEEE EDCF 802.11e: A first order MAC and PHY cross-layer model to estimate the throughput. In *IEEE Vehicular Technology Conference*, 2009.

[53] J. Hu, G. Min, W. Jia, and M. E. Woodward. Admission control in the IEEE 802.11e WLANs based on analytical modelling and game theory. In *IEEE Global Telecommunications Conference*, 2009.

[54] J. Hu, G. Min, and M. E. Woodward. Analysis and comparison of burst transmission schemes in unsaturated 802.11e WLANs. In *Proceedings of IEEE Global Telecommunications Conference*, 2007.

[55] J. Hu, G. Min, and M. E. Woodward. Modelling of IEEE 802.11e contention free bursting scheme with non-identical stations. In *IEEE International Work-*

shop on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, pages 88–94, 2007.

[56] J. Hu, G. Min, M. E. Woodward, and W. Jia. A comprehensive analytical model for IEEE 802.11e QoS differentiation schemes under non-saturated traffic loads. In *Proceedings of IEEE International Conference on Communications (ICC'08)*, 2008.

[57] J. Huang, J. Wang, and J. Ye. Buffer allocation management for improving TCP fairness in IEEE 802.11 WLANs. In *The 6th International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM'10)*, 2010.

[58] K. Huang, K. R. Duffy, and D. Malone. On the validity of IEEE 802.11 MAC modeling hypotheses. *IEEE/ACM Transactions on Networking*, 18(6):1935–1948, 2010.

[59] K. D. Huang, K. R. Duffy, D. Malone, and D. J. Leith. Investigating the validity of IEEE 802.11 MAC modeling hypotheses. In *Proceedings of IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2008.

[60] P. Hurley and J.Y. Le Boudec. ABE: Providing a low-delay service within best effort. *IEEE Network*, 15(3):60–69, 2001.

[61] O. C. Ibe. *Fundamentals of applied probability and random processes*. Elsevier Inc., United Kingdom, 2005.

[62] I. Inan, F. Keceli, and E. Ayanoglu. Modeling the 802.11e enhanced distributed channel access function. In *Proceedings of IEEE Global Telecommunications Conference*, 2007.

[63] M. O. Jackson. Mechanism theory, 2003.

[64] Y. Jin and G. Kesidis. Equilibria of a noncooperative game for heterogeneous users of an ALOHA network. *IEEE Communications Letters*, 6(7):282–284, 2002.

[65] M. Karsten, Y. Lin, and K. Larson. Incentive-Compatible differentiated scheduling. In *HotNets IV*, 2005.

[66] N. Khademi and M. Othman. Guaranteeing per station and per flow fairness of downstream and upstream flows over IEEE 802.11 WLAN. In *International Conference on Information and Multimedia Technology (ICIMT'09)*, pages 431–435, 2009.

[67] J.-D. Kim and C.-K. Kim. Performance analysis and evaluation of IEEE 802.11e EDCF. *Wireless Communications and Mobile Computing*, 4(1):55–74, 2004.

[68] J.-D. Kim and C.-K. Kim. Performance analysis and evaluation of IEEE 802.11e EDCF. *Wireless Communications and Mobile Computing*, 4(1):55–74, 2004.

[69] T. King and D. Newson. *Data Network Engineering*. Kluwer Academic Publishers, 1999.

[70] T.E. Klein, K.K. Leung, R. Parkinson, and L.G. Samuel. Avoiding spurious TCP timeouts in wireless networks by delay injection. In *IEEE Global Telecommunications Conference (GLOBECOM'04)*, volume 5, pages 2754–2759, 2004.

[71] L. Kleinrock. *Queueing systems*, volume 1. John Wiley & Sons, Inc., NewYork, 1975.

[72] Z.-N. Kong, D. H. K. Tsang, B. Bensaou, and D. Gao. Performance analysis of IEEE 802.11e contention-based channel access. *IEEE Journal on Selected Areas in Communications*, 22(10):2095–2106, 2004.

[73] J. Konorski. Playing CSMA/CA game to deter backoff attacks in ad hoc wireless LANs. In *Proceedings of the 4th international conference on Ad-Hoc, Mobile and Wireless Networks (ADHOC-NOW'05)*, pages 127–140, Berlin, Heidelberg, 2005. Springer-Verlag.

[74] J. Konorski. A Game-Theoretic study of CSMA/CA under a backoff attack. *IEEE/ACM Transactions on Networking*, 14(6):1167 –1178, dec. 2006.

[75] K. Kosek-Szott, M. Natkaniec, and A. R. Pach. A simple but accurate throughput model for IEEE 802.11 EDCA in saturation and non-saturation conditions. *Computer Networks*, 55(3):622 – 635, 2011.

[76] S. Krishnasamy and A. Kumar. Modeling the effect of transmission errors on TCP controlled transfers over infrastructure 802.11 wireless LANs. In *Proceedings of the 14th ACM International Conference on Modeling, Analysis, and Simulation of Wireless and Mobile Systems (MSWiM'11)*, pages 275–284, 2011.

[77] A. Kumar, E. Altman, D. Miorandi, and M. Goyal. New insights from a fixed point analysis of single cell IEEE 802.11 WLANs. In *Proceedings of IEEE International Conference on Computer Communications (INFOCOM'05)*, volume 3, pages 1550 – 1561, march 2005.

[78] Y.-L. Kuo, C.-H. Lu, E. H.-K. Wu, G.-H. Chen, and Y.-H. Tseng. Performance analysis of the enhanced distributed coordination function in the IEEE 802.11e. In *IEEE Vehicular Technology Conference*, volume 58, pages 3488–3492, 2003.

[79] Y.-L. Kuo, E.H.-K. Wu, and G.-H. Chen. Noncooperative admission control for differentiated services in IEEE 802.11 WLANs. In *IEEE Global Telecommunications Conference*, volume 5, pages 2981–2986, 2004.

[80] Y.-C. Lai, Y.-H. Yeh, and C.-L. Wang. Dynamic backoff time adjustment with considering channel condition for IEEE 802.11e EDCA. *Lecture Notes in Computer Science*, 5200:445–454, 2008.

[81] S. Lasaulce and H. Tembine. *Game Theory and Learning for Wireless Networks: Fundamentals and Applications*. Academic Press, 2011.

[82] W. Lee, C. Wang, and K. Sohraby. On use of traditional M/G/1 model for IEEE 802.11 DCF in unsaturated traffic conditions. In *IEEE Wireless Communications and Networking Conference (WCNC'06)*, volume 4, pages 1933 –1937, april 2006.

[83] D. J. Leith, P. Clifford, D. Malone, and A. Ng. TCP fairness in 802.11e WLANs. *IEEE Communications Letters*, 9(11):964–966, 2005.

[84] Y. Lin and V. W. S. Wong. Saturation throughput of IEEE 802.11e EDCA based on mean value analysis. In *Proceedings of IEEE Wireless Communications and Networking Conference (WCNC)*, 2006.

[85] X. Ling, L. X. Cai, J. W. Mark, and X. Shen. Performance analysis of IEEE 802.11 DCF with heterogeneous traffic. In *Proceedings of IEEE Consumer Communications and Networking Conference (CCNC'07)*, pages 49–53, 2007.

[86] Z. Lu, W. Wang, and C. Wang. Modeling and performance evaluation of backoff misbehaving nodes in CSMA/CA networks. *IEEE Transactions on Mobile Computing*, 11(8):1331 – 1344, 2011.

[87] R. Ludwig and R.H. Katz. The Eifel algorithm: Making TCP robust against spurious retransmissions. In *Computer Communication Review*, volume 30, pages 30–36, 2000.

[88] J. Lv, X. Zhang, and X. Han. A novel dynamic tuning of the contention window (CW) for IEEE 802.11e enhanced distributed control function. In *Pro-*

ceedings of International Conference on Networked Computing and Advanced Information Management (NCM), 2008.

[89] A. MacKenzie and L. A. DaSilva. *Game Theory for Wireless Engineers*. Morgan & Claypool Publishers, 2006.

[90] A. B. MacKenzie and S. B. Wicker. Selfish users in Aloha: A game-theoretic approach. In *IEEE Vehicular Technology Conference*, volume 3, pages 1354–1357, 2001.

[91] D. Malone, K. Duffy, and D. Leith. Modeling the 802.11 distributed coordination function in nonsaturated heterogeneous conditions. *IEEE/ACM Transactions on Networking*, 15(1):159–172, 2007.

[92] C. Mbarushimana and A. Shahrabi. TCP enhancement in IEEE 802.11e wireless networks. In *Proceedings of the International Conference on Parallel and Distributed Systems (ICPADS)*, pages 407–414, 2008.

[93] M. Menth, A. Binzenhfer, and S. Mhleck. Source models for speech traffic revisited. *IEEE/ACM Transactions on Networking*, 17(4):1042–1051, 2009.

[94] D. Miorandi, A. A. Kherani, and E. Altman. A queueing model for HTTP traffic over IEEE 802.11 WLANs. *Computer Networks*, 50(1):63–79, 2006.

[95] S. H. Nguyen, L. L. H. Andrew, and H. L. Vu. Service differentiation without prioritization in IEEE 802.11 WLANs. In *Proceedings of IEEE Local Computer Networks (LCN'11)*, pages 109–116, 2011.

[96] S. H. Nguyen, H. L. Vu, and L. L. H. Andrew. Packet size variability affects collisions and energy efficiency in WLANs. In *Proceedings of IEEE Wireless Communications and Networking Conference (WCNC'10)*, pages 1–6, 2010.

[97] S. H. Nguyen, H. L. Vu, and L. L. H. Andrew. Performance analysis of IEEE 802.11 WLANs with saturated and unsaturated sources. *IEEE Transactions on Vehicular Technology*, 61(1):333–345, 2012.

[98] S. H. Nguyen, H. L. Vu, and L. L. H. Andrew. Service differentiation without prioritization in IEEE 802.11 WLANs. *IEEE Transactions on Mobile Computing*, 2012.

[99] Q. Ni, L. Romdhani, and T. Turletti. A survey of QoS enhancements for IEEE 802.11 wireless LAN. *Wireless Communications and Mobile Computing*, 4(5):547–566, 2004.

[100] T. Nilsson and J. Farooq. A novel MAC scheme for solving the QoS parameter adjustment problem in IEEE 802.11e EDCA. In *IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks (WoWMoM)*, 2008.

[101] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani. *Algorithmic Game Theory*. Cambridge University Press, 2007.

[102] P. Nuggehalli, M. Sarkar, K. Kulkarni, and R. R. Rao. Game-theoretic analysis of QoS in wireless MAC. In *Proceedings of IEEE International Conference on Computer Communications (INFOCOM'08)*, 2008.

[103] C. N. Ojeda-Guerra and I. Alonso-Gonzalez. Adaptive tuning mechanism for EDCA in IEEE 802.11e wireless LANs. In *The 14th European Wireless Conference (EW'08)*, 2008.

[104] S.-W. Pan and J.-S. Wu. Throughput analysis of IEEE 802.11e EDCA under heterogeneous traffic. *Computer Communications*, 32(5):935 – 942, 2009.

[105] A. Parekh. *A Generalized Processor Sharing Approach to Flow Control In Integrated Services Networks*. PhD thesis, Massachusetts Institute of Technology, Feb 1992.

[106] W. Pattara-Atikom, S. Banerjee, and P. Krishnamurthy. Starvation prevention and quality of service in wireless LANs. In *Proceedings of International Sym-*

posium on Wireless Personal Multimedia Communications (WPMC), pages 1078–1082, 2002.

[107] F. Peng, H. M. Alnuweiri, and V. C. M. Leung. Analysis of burst transmission in IEEE 802.11e wireless LANs. In *Proceedings of IEEE International Conference on Communications (ICC'05)*, pages 535–539, 2005.

[108] W. B. Powell. Iterative algorithms for bulk arrival, bulk service queues with poisson and Non-Poisson arrivals. *Transportation Science*, 20(2), 1986.

[109] J. Price, P. Nuggehalli, and T. Javidi. Incentive compatible MAC-Layer QoS design. In *Proceedings of IEEE Consumer Communications and Networking Conference (CCNC'08)*, 2008.

[110] J. Price, P. Nuggehalli, and T. Javidi. Pricing and QoS in wireless random access networks. In *Proceedings of IEEE Global Telecommunications Conference*, 2008.

[111] I. A. Qazi. *An efficient framework of congestion control for next-generation networks*. PhD thesis, University of Pittsburgh, 2010.

[112] D. Qiao and K. G. Shin. Achieving efficient channel utilization and weighted fairness for data communications in IEEE 802.11 WLAN under the DCF. In *Proceedings of IEEE International Workshop on Quality of Service (IWQoS)*, pages 227–236, 2002.

[113] V. Ramaiyan, A. Kumar, and E. Altman. Fixed point analysis of single cell IEEE 802.11e WLANs: Uniqueness and multistability. *IEEE/ACM Transactions on Networking*, 16(5):1080 – 1093, 2008.

[114] D. A. Rambim, M. Mzyece, and K. Djouani. Enhancement of TCP in 802.11e wireless local area networks. In *IEEE Vehicular Technology Conference*, 2011.

[115] N. Ramos, D. Panigrahi, and S. Dey. ChaPLeT: Channel-dependent packet level tuning for service differentiation in IEEE 802.11e. In *Proceedings of International Symposium on Wireless Personal Multimedia Communications (WPMC)*, 2003.

[116] Wi-Fi Alliance Press Releases. Wi-Fi innovations and user enthusiasm propel continued sales growth. `http://www.wi-fi.org/media/press-releases/wi-fi\ discretionary{-}{}{}innovations-and-user-enthusiasm-propel-continued-sale`

[117] J. W. Robinson. An analytical model for the service delay distribution of IEEE 802.11e enhanced distributed channel access. Master's thesis, Simon Franser University, Vancouver, Canada, 2005.

[118] J. W. Robinson and T. S. Randhawa. Saturation throughput analysis of IEEE 802.11e enhanced distributed coordination function. *IEEE Journal on Selected Areas in Communications*, 22(5):917–928, 2004.

[119] W. A. Rosenkrantz. *Introduction to Probability and Statistics for Science, Engineering, and Finance*. CRC Press, 2008.

[120] S. M. Ross. *Introduction to Probability Models*. Academic Press, 2006.

[121] T. Sakurai and H. L. Vu. MAC access delay of IEEE 802.11 DCF. *IEEE Transactions on Wireless Communications*, 6(5):1702 –1710, may 2007.

[122] C. U. Saraydar, N. B. Mandayam, and D. J. Goodman. Efficient power control via pricing in wireless data networks. *IEEE Transactions on Communications*, 50(2):291–303, 2002.

[123] P. Sarolahti, M. Kojo, and K. Raatikainen. F-RTO: An enhanced recovery algorithm for TCP retransmission timeouts. In *Computer Communication Review*, volume 33, pages 51–63, 2003.

[124] P. Serrano, A. Banchs, P. Patras, and A. Azcorra. Optimal configuration of 802.11e EDCA for real-time and data traffic. *IEEE Transactions on Vehicular Technology*, 59(5):2511–2528, 2010.

[125] M. Seyedzadegan and M. Othman. Weighted window and class-based weighted window methods for per-station TCP fairness in IEEE 802.11 WLANs. *Eurasip Journal on Wireless Communications and Networking*, 2010, 2010.

[126] M. Shreedhar and G. Varghese. Efficient fair queuing using deficit round-robin. *IEEE/ACM Transactions on Networking*, 4(3):375–385, 1996.

[127] J. L. Sobrinho and A. S. Krishnakumar. Real-time traffic over the IEEE 802.11 medium access control layer. *Bell Labs Technical Journal*, 1(2):172–187, 1996.

[128] V. Srivastava, J. Neel, A. B. Mackenzie, R. Menon, L. A. Dasilva, J. E. Hicks, J. H. Reed, and R. P. Gilles. Using game theory to analyze wireless ad hoc networks. *IEEE Communications Surveys Tutorials*, 7(4):46 – 56, quarter 2005.

[129] T. Szigeti and C. Hattingh. *End-to-End QoS Network Design: Quality of Service in LANs, WANs, and VPNs*. Indianapolis, Ind. : Cisco Press, 2005.

[130] N. Tadayon, S. Zokaei, and E. Askari. A novel prioritization scheme to improve QoS in IEEE 802.11e networks. *Journal of Computer Systems, Networks, and Communications*, 2010.

[131] G. Tan and J. Guttag. The 802.11 MAC protocol leads to inefficient equilibria. In *Proceedings of IEEE International Conference on Computer Communications (INFOCOM'05)*, volume 1, pages 1–11, 2005.

[132] J. Tan and N. B. Shroff. Transition from heavy to light tails in retransmission durations. In *Proceedings of IEEE International Conference on Computer Communications (INFOCOM'10)*, 2010.

[133] Z. Tao and S. Panwar. An analytical model for the IEEE 802.11e enhanced distributed coordination function. In *IEEE International Conference on Communications (ICC'04)*, volume 7, pages 4111–4117, 2004.

[134] Y. C. Tay and K. C. Chua. A capacity analysis for the IEEE 802.11 MAC protocol. *Wireless Networks*, 7(2):159–171, March 2001.

[135] O. Tickoo and B. Sikdar. A queueing model for finite load IEEE 802.11 random access MAC. In *Proceedings of IEEE International Conference on Communications (ICC'04)*, volume 1, pages 175–179, 2004.

[136] I. Tinnirello and S. Choi. Efficiency analysis of burst transmissions with block ACK in contention-based 802.11e WLANs. In *IEEE International Conference on Communications (ICC'05)*, volume 5, pages 3455 – 3460, May 2005.

[137] N. H. Vaidya, P. Bahl, and S. Gupta. Distributed fair scheduling in a wireless LAN. In *Proceedings of the 6th annual international conference on Mobile computing and networking (MobiCom)*, pages 167–178, 2000.

[138] N. H. Vaidya, A. Dugar, S. Gupta, and P. Bahl. Distributed fair scheduling in a wireless LAN. *IEEE Transactions on Mobile Computing*, 4(6):616–629, Nov/Dec 2005.

[139] S. Wietholter and C. Hoene. An IEEE 802.11e EDCF and CFB simulation model for ns-2. `http://www.tkn.tu-berlin.de/research/802.11e_ns2/`.

[140] H. Wu, Y. Peng, K. Long, S. Cheng, and J. Ma. Performance of reliable transport protocol over IEEE 802.11 wireless LAN: analysis and enhancement. In *Proceedings of IEEE International Conference on Computer Communications (INFOCOM'02)*, volume 2, pages 599 – 607, 2002.

[141] X. Xiao. *Technical, Commercial and Regulatory Challenges of QoS : An Internet Service Model Perspective*. Morgan Kaufmann, 2008.

[142] X. Xiao and L. M. Ni. Internet QoS: A big picture. *IEEE Network*, 13(2):8 – 18, 1999.

[143] Y. Xiao. Saturation performance metrics of the IEEE 802.11 MAC. In *IEEE Vehicular Technology Conference*, volume 58, pages 1453–1457, 2003.

[144] Y. Xiao. A simple and effective priority scheme for IEEE 802.11. *IEEE Communications Letters*, 7(2):70–72, 2003.

[145] Y. Xiao. IEEE 802.11E: QoS provisioning at the MAC layer. *IEEE Wireless Communications*, 11(3):72–79, 2004.

[146] Y. Xiao. Performance analysis of IEEE 802.11e EDCF under saturation condition. In *IEEE International Conference on Communications (ICC'04)*, volume 1, pages 170–174, 2004.

[147] Y. Xiao, X. Shan, and Y. Ren. Game theory models for IEEE 802.11 DCF in wireless ad hoc networks. *IEEE Communications Magazine*, 43(3):S22–S26, 2005.

[148] C. Xu, K. Liu, and J. Zhou. Performance analysis of service differentiation schemes for IEEE 802.11 wireless LANs in non-saturated conditions. In *Proceedings of the third International Conference on Communications and Networking*, pages 12 – 16, 2008.

[149] D. Xu, T. Sakurai, and H. L. Vu. An access delay model for IEEE 802.11e EDCA. *IEEE Transactions on Mobile computing*, 8(2):261–275, 2009.

[150] J. Yu and S. Choi. Modeling and analysis of tcp dynamics over ieee 802.11 wlan. In *The fourth Annual Conference on Wireless on Demand Network Systems and Services (WONS'07)*, pages 154–161, 2007.

[151] J. Yu, S. Choi, and D. Qiao. Analytical study of TCP performance over IEEE 802.11e WLANs. *Mobile Networks and Applications*, 14(4):470–485, 2009.

[152] S.M.S. Zabir, A. Ashir, and N. Shiratori. An efficient flow control approach for TCP over wireless networks. *Journal of Circuits, Systems and Computers*, 13(2):341–360, 2004.

[153] G. Zhang and H. Zhang. Modelling IEEE 802.11 DCF in wireless LANs as a dynamic game with incompletely information. In *IET Conference Publications*, number 535 CP, pages 215–218, 2008.

[154] J. Zhao, Z. Guo, Q. Zhang, and W. Zhu. Performance study of MAC for service differentiation in IEEE 802.11. In *IEEE Global Telecommunications Conference (GLOBECOM'02)*, volume 1, pages 778–782, 2002.

[155] Q. Zhao, D. H. K. Tsang, and T. Sakurai. A simple and approximate model for nonsaturated IEEE 802.11 DCF. *IEEE Transactions on Mobile Computing*, 8, 2009.

[156] P. Zhou, W. Liu, W. Yuan, and W. Cheng. Energy-efficient joint power and rate control via pricing in wireless data networks. In *IEEE Wireless Communications and Networking Conference (WCNC'08)*, pages 1091–1096, 2008.

[157] H. Zhu and I. Chlamtac. Performance analysis for IEEE 802.11e EDCF service differentiation. *IEEE Transactions on Wireless Communications*, 4(4):1779–1788, 2005.

[158] H. Zhu, M. Li, I. Chlamtac, and B. Prabhakaran. A survey of quality of service in IEEE 802.11 networks. *IEEE Wireless Communications*, 11(4):6 – 14, 2004.

Appendix A

Derivation and proofs in Chapter 3

A.1 Derivation of (3.23)

The detailed derivation of $\mathbb{E}[F_u]$ in (3.23) is presented as follows.

From (3.14) and (3.15),

$$\begin{aligned}\mathbb{E}[F_u] &= b_u \frac{1-p_u}{1-b_u+b_u(1-p_u^{K+1})} \sum_{k=0}^{K} F_{uk} \\ &= b_u \frac{1-p_u}{1-b_u+b_u(1-p_u^{K+1})} \Big(\sum_{k=0}^{K} p_u^i (\sum_{j=0}^{k} \mathbb{E}[B_{uj}] + k\mathbb{E}[T_u^c] + \mathbb{E}[T_{\text{res},u}]) \Big) \end{aligned} \quad \text{(A.1)}$$

By Wald's theorem for sums of i.i.d. random variables [120] and the fact that $Y_{u,k} \sim Y_u$, we have from (3.16)

$$\mathbb{E}[B_{uj}] = E[Y_u]\mathbb{E}[U_{uj}] \quad \text{(A.2)}$$

Then, (A.1) becomes

$$\begin{aligned}\mathbb{E}[F_u] &= b_u \frac{1-p_u}{1-b_u+b_u(1-p_u^{K+1})} (\mathbb{E}[Y_u] \sum_{k=0}^{K} p_u^k (\sum_{j=0}^{k} \mathbb{E}[U_{uj}]) + \sum_{k=0}^{K} p_u^k (k\mathbb{E}[T_u^c] + \mathbb{E}[T_{\text{res},u}])) \\ &= b_u \frac{1-p_u}{1-b_u+b_u(1-p_u^{K+1})} (L_1 \mathbb{E}[Y_u] + L_2) \end{aligned} \quad \text{(A.3)}$$

where

$$\begin{aligned} L_1 &= \sum_{k=0}^{K} p_u^k (\sum_{j=0}^{k} \mathbb{E}[U_{uj}]) = \sum_{k=0}^{K} p_u^k (\sum_{j=0}^{k} (\frac{2^{\min(j,m)} W_u + 1}{2})) \\ &\approx \frac{W_u}{2} (\sum_{k=0}^{K} p_u^k (\sum_{j=0}^{k} 2^{\min(j,m)})) \end{aligned} \quad \text{(A.4)}$$

Note that the approximation in (A.4) comes from approximating

$$\mathbb{E}[U_{uj}] = \frac{2^{\min(j,m)}W_u - 1}{2} \approx 2^{\min(j,m)-1}W_u. \tag{A.5}$$

Then, (A.4) becomes

$$\begin{aligned}
L_1 \approx & \frac{W_u}{2}\Big(\sum_{k=0}^{m} p_u^k(\frac{1-2^{k+1}}{1-2}) + \sum_{k=m+1}^{K} p_u^i(\frac{1-2^{m+1}}{1-2} + (k-m)2^m)\Big) \\
= & \frac{W_s}{2}\Big(2\sum_{k=0}^{m}(2p_u)^k - \sum_{k=0}^{m} p_u^k + (-1+2^{m+1}-m2^m)\sum_{k=m+1}^{K} p_u^k + 2^m\sum_{m+1}^{K} kp_u^k\Big) \\
= & \frac{W_s}{2}\Big(2\frac{1-(2p_u)^{m+1}}{1-2p_u} - \frac{1-p_u^{m+1}}{1-p_u} + (-1+2^{m+1}-m2^m)(\frac{1-p_u^{K+1}}{1-p_u} - \frac{1-p_u^{m+1}}{1-p_u}) \\
& + 2^m(\frac{1-p_u^K}{(1-p_u)^2}p_u - \frac{Kp_u^{K+1}}{1-p_u} - \frac{1-p_u^m}{(1-p_u)^2}p_u + \frac{mp_u^{m+1}}{1-p_u})\Big) \\
= & \frac{W_s}{2(1-p_u)}\Big(2\frac{(1-(2p_u)^{m+1})(1-p_u)}{1-2p_u} - (1-p_u^{m+1}) \\
& + (-1+2^{m+1}-m2^m)(p_u^{m+1}-p_u^{K+1}) + 2^m(\frac{p_u^{m+1}-p_u^{K+1}}{1-p_u} + mp_u^{m+1} - Kp_u^{K+1})\Big)
\end{aligned} \tag{A.6}$$

Moreover,

$$\begin{aligned}
L_2 &= \sum_{k=0}^{K} p_u^k(k\mathbb{E}[T_u^c] + \mathbb{E}[T_{\text{res},u}]) \\
&= \mathbb{E}[T_u^c](\frac{1-p_u^K}{(1-p_u)^2}p_u - K\frac{p_u^{K+1}}{1-p_u}) + \mathbb{E}[T_{\text{res},u}]\frac{1-p_u^{K+1}}{1-p_u}
\end{aligned} \tag{A.7}$$

Substituting (A.6) and (A.7) into (A.3) gives (3.23).

A.2 The z-transform of access delay

The generating function of the pmf of a non-negative integer-valued random variable X is defined as

$$\hat{X}(z) = \sum_{k=0}^{\infty} P(X=k)z^k, \quad \text{for } z \in \mathbb{C} \tag{A.8}$$

To apply the z-transform, the continuous r.v.s D_u, F_u, F_{ui}, and $T_{\text{res},u}$ were quantized in steps of δ. Other random variables in the delay model in Chapter 3 are non-negative and discrete, but some are not integer-valued. However, they can be transformed to integer-valued random variables, using the scale factor δ. Similarly, positive real variables such as σ, T_x, and T_x^s ($x \in \mathbb{S} \cup \mathbb{U}$) are also transformed to integers using δ.

By (3.13), the generating function of the access delay is

$$\hat{D}_u(z) = \hat{T}_u^s(z)\hat{F}_u(z) \tag{A.9}$$

Note that $\hat{T}_x^s(z)$ can be calculated from the distribution of η_x given by (3.34) for $x \in \mathbb{U}$ or (3.28) for $x \in \mathbb{S}$.

From (3.14), $\hat{F}_u(z)$ is given by

$$\hat{F}_u(z) = \frac{1}{1 - b_u + b_u(1 - p_u^{K+1})}\Big(b_u(1-p_u)\sum_{k=0}^{K} p_u^k \hat{F}_{uk}(z) + (1 - b_u)\Big) \tag{A.10}$$

where, by (3.15), $\hat{F}_{uk}(z)$ is

$$\hat{F}_{uk}(z) = \hat{T}_u^c(z)^k \hat{T}_{\text{res},u}(z) \prod_{j=0}^{k} \hat{B}_{uj}(z). \tag{A.11}$$

$\hat{T}_{\text{res},u}(z)$ in (A.11) is given by

$$\hat{T}_{\text{res},u}(z) = \frac{z}{(1-z)\mathbb{E}[Y_u^b]}\left(1 - \frac{1}{1 - P_u^i}\left(\sum_{x \in \mathbb{S}\cup\mathbb{U}\setminus\{u\}} P_{xu}^s \hat{T}_x^s(z) + \sum_{x \in \mathbb{S}\cup\mathbb{U}\setminus\{u\}} P_{xu}^c \hat{T}_x(z)\right)\right) \tag{A.12}$$

with the detailed derivation presented in Section A.2.1 below. Besides, $\hat{T}_u^c(z)$ is determined from (3.19) as follows.

$$\hat{T}_u^c(z) = \sum_{x \in \mathbb{S}\cup\mathbb{U}\setminus\{u\}} P_{xu}^{cu} \widehat{\max(T_u, T_x)}(z) \tag{A.13}$$

Moreover, from (3.16), $\hat{B}_{uj}(z)$ is the z-transform of random sum of random variables [61]

$$\hat{B}_{uj}(z) = \hat{U}_{uj}(\hat{Y}_u(z)), \tag{A.14}$$

where $\hat{U}_{uj}(z)$ is given by

$$\hat{U}_{uj}(z) = \frac{1 - z^{2^{\min(j,m)}W_u}}{2^{\min(j,m)}W_u(1-z)} \tag{A.15}$$

and $\hat{Y}_u(z)$ is determined from (3.17) as

$$\hat{Y}_u(z) = \hat{\sigma}(z)P_u^i + \sum_{x\in\mathbb{S}\cup\mathbb{U}\setminus\{u\}} P_{xu}^s \hat{T}_x^s(z) + \sum_{x\in\mathbb{S}\cup\mathbb{U}\setminus\{u\}} P_{xu}^c \hat{T}_x(z) \tag{A.16}$$

In summary, $\hat{D}_u(z)$ is given by

$$\begin{aligned}\hat{D}_u(z) =& \frac{\hat{T}_u^s(z)}{1 - b_u + b_u(1-p_u^{K+1})} \\ & . \left(b_u(1-p_u)\sum_{k=0}^{K} p_u^k \hat{T}_u^c(z)^k \hat{T}_{\text{res},u}(z) \prod_{j=0}^{k} \hat{U}_{uj}(\hat{Y}_u(z)) + (1-b_u) \right)\end{aligned} \tag{A.17}$$

Then, the generating function of the ccdf, $\hat{D}_u^c(z)$ can be obtained from $\hat{D}_u(z)$ via the identity

$$\hat{D}_u^c(z) = \frac{1 - \hat{D}_u(z)}{1-z}. \tag{A.18}$$

The access delay ccdf is the inverse z-transform of $\hat{D}_u^c(z)$.

A.2.1 Derivation of (A.12)

Recall that $T_{\text{res},u}$ is quantized in steps of δ to make it an integer and it depends on the distribution of observed busy slots Y_u^b. Then, its distribution is

$$P[T_{\text{res},u} = t, Y_u^b = y_u] = \frac{1}{y_u} P[Y_u^b = y_u] \frac{y_u}{\mathbb{E}[Y_u^b]} = \frac{1}{\mathbb{E}[Y_u^b]} P[Y_u^b = y_u], \quad \text{for } t \le y_u \tag{A.19}$$

and

$$P[T_{\mathrm{res},u}=t]=\sum_{y_u\geq t}P[T_{\mathrm{res},u}=t,Y_u^b=y_u]=\frac{1}{\mathbb{E}[Y_u^b]}\sum_{y_u\geq t}P[Y_u^b=y_u] \tag{A.20}$$

Then, the z-transform of $T_{\mathrm{res},u}$ is as follows. First, let T_x^b denote a mixed sequence of T_y and T_y^s ($y\in\mathbb{S}\cup\mathbb{U}\setminus\{u\}$) which is indexed in non-increasing order of duration. Let N be the maximum value of x.

$$\begin{aligned}
\hat{T}_{\mathrm{res},u}(z)=&\sum_{t=0}^{T_N^b}P[T_{\mathrm{res},u}=t]z^t\\
=&\frac{1}{\mathbb{E}[Y_u^b]}\sum_{t=0}^{T_N^b}\Big(\sum_{y_u\geq t}P[Y_u^b=y_u]\Big)z^t\\
=&\frac{1}{\mathbb{E}[Y_u^b]}\Bigg(\sum_{t=T_2^b+1}^{T_1^b}P[Y_u^b=T_1^b]z^t+\sum_{t=T_3^b+1}^{T_2^b}(P[Y_u^b=T_1^b]+P[Y_u^b=T_2^b])z^t\\
&+\cdots+\sum_{t=T_N^b+1}^{T_{N-1}^b}\sum_{j=1}^{N-1}P[Y_u^b=T_j^b]z^t+\sum_{t=0}^{T_N^b}z^t\Bigg)\\
=&\frac{1}{\mathbb{E}[Y_u^b]}\Bigg(P[Y_u^b=T_1^b]\big(\sum_{t=0}^{T_1^b}z^t-\sum_{t=0}^{T_2^b}z^t\big)+(P[Y_u^b=T_1^b]+P[Y_u^b=T_2^b])\\
&.\big(\sum_{t=0}^{T_2^b}z^t-\sum_{t=0}^{T_3^b}z^t\big)+\cdots+\sum_{j=1}^{N-1}P[Y_u^b=T_j^b]\big(\sum_{t=0}^{T_{N-1}^b}z^t-\sum_{t=0}^{T_N^b}z^t\big)+\sum_{t=0}^{T_N^b}z^t\Bigg)\\
=&\frac{1}{\mathbb{E}[Y_u^b]}\Bigg(\sum_{j=1}^{N}P[Y_u^b=T_j^b]\sum_{t=0}^{T_j^b}z^t\Bigg)\\
=&\frac{1}{\mathbb{E}[Y_u^b]}\Bigg(\sum_{x\in\mathbb{S}\cup\mathbb{U}\setminus\{u\}}P[Y_u^b=T_x]\sum_{t=0}^{T_x}z^t+\sum_{x\in\mathbb{S}\cup\mathbb{U}\setminus\{u\}}P[Y_u^b=T_x^s]\sum_{t=0}^{T_x^s}z^t\Bigg)\\
=&\frac{1}{\mathbb{E}[Y_u^b]}\Bigg(\sum_{x\in\mathbb{S}\cup\mathbb{U}\setminus\{u\}}P[Y_u^b=T_x]\frac{1-z^{T_x+1}}{1-z}+\sum_{x\in\mathbb{S}\cup\mathbb{U}\setminus\{u\}}P[Y_u^b=T_x^s]\frac{1-z^{T_x^s+1}}{1-z}\Bigg)\\
=&\frac{z}{(1-z)\mathbb{E}[Y_u^b]}\Bigg(1-\Big(\sum_{x\in\mathbb{S}\cup\mathbb{U}\setminus\{u\}}P[Y_u^b=T_x]z^{T_x}+\sum_{x\in\mathbb{S}\cup\mathbb{U}\setminus\{u\}}P[Y_u^b=T_x^s]z^{T_x^s}\Big)\Bigg)
\end{aligned}$$

$$=\frac{z}{(1-z)\mathbb{E}[Y_u^b]}\left(1-\frac{1}{1-P_u^i}\left(\sum_{x\in\mathbb{S}\cup\mathbb{U}\setminus\{u\}}P_{xu}^s\hat{T}_x^s(z)+\sum_{x\in\mathbb{S}\cup\mathbb{U}\setminus\{u\}}P_{xu}^c\hat{T}_x(z)\right)\right) \tag{A.21}$$

A.3 Lemma 3.2

Proof: Dividing p_s from (3.11c) by p_u from (3.11c), we have

$$\frac{1-p_u}{1-p_s}=\frac{1-\tau_s}{1-\tau_u} \tag{A.22}$$

Moreover, by (3.12),

$$\tau_s=\frac{S_s}{\mathbb{E}[\eta_s]}\frac{\mathbb{E}[Y]}{1-p_s} \tag{A.23}$$

Dividing (A.23) by τ_u from (3.11b), and applying (A.22) gives

$$\frac{\tau_s}{\tau_u}=\frac{S_s\mathbb{E}[\eta_u]}{\lambda_u\mathbb{E}[\eta_s]}\frac{1-p_u}{1-p_s}=\frac{S_s\mathbb{E}[\eta_u]}{\lambda_u\mathbb{E}[\eta_s]}\frac{1-\tau_s}{1-\tau_u} \tag{A.24}$$

which establishes the first claim.

By (3.37), this implies $\tau_s>\tau_u$, whence $p_u>p_s$ by (A.22). ■

A.4 Theorem 3.3

Proof: The result is a consequence of Lemma 3.2 and the following observations, which will be established below.

1. When a new unsaturated user is added to the existing network (equivalent to N_u increasing), p_s is increasing.

2. If there are $N_u=0$ unsaturated sources and

$$N_s\geq 1+\frac{\log(3/4)}{\log(1-\frac{4}{3W+2})} \tag{A.25}$$

then $p_s\geq 1/4$.

3. If $p_u \geq 1/4$ then the variance of the random variable whose ccdf is the right hand side of (3.44) is infinite.

These can be shown as follows:

1. This follows from (3.11c) since $\tau_u \in [0, 1]$, and τ_s is decreasing in p_s.

2. When $N_u = 0$, (3.11c) becomes $p_s = 1 - (1 - \tau_s)^{N_s - 1}$. Thus $p_s \geq 1/4$ if

$$\tau_s \geq 1 - \left(\frac{3}{4}\right)^{1/(N_s-1)}. \tag{A.26}$$

   Conversely, (3.11a) is decreasing in p_s, and so $p_s \geq 1/4$ if

$$\tau_s \leq \frac{4}{3W + 2} \tag{A.27}$$

   Combining (A.26) and (A.27), $p_s \geq 1/4$ if

$$1 - \left(\frac{3}{4}\right)^{1/(N_s-1)} \leq \tau_s \leq \frac{4}{3W + 2}$$

   which upon rearrangement gives (A.25).

3. If $p_u \geq 1/4$, then the random variable D whose ccdf is the right hand side of (3.44) has a tail heavier than kD^{-2} for some k, and hence infinite variance.

■

Appendix B

Proofs in Chapter 4

In this appendix, the proofs of all lemmas and theorems in Sections 4.4.1 and 4.5.2 are shown.

Note that the mean slot time $\mathbb{E}[Y]$ in (4.4) is given by (3.10a)-(3.10d) where durations T_x^s and T_x of source x using class B_k are

$$T_x^s = E + \eta_k \mathcal{T}, \qquad \text{with} \quad E > \sigma \tag{B.1}$$

and

$$T_x = E + T_{\text{p}x} + SIFS + T_{\text{ACK}} \tag{B.2}$$

where E is the interval during which a station needs to sense channel free before transmitting (e.g. AIFS or DIFS). T_{ACK} is the duration of an ACK packet and $T_{\text{p}x}$ is the transmission time of a packet from the source x.

B.1 Theorem 4.3

Proof: I first prove Claim (T4.3-1) that under the proportional scheme (4.1), the throughput per slot of a saturated source using class B_k increases when η_k increases. Then, I prove Claim (T4.3-2) that the collision probability of unsaturated sources decreases with the increase of $\eta_k \geq 1$.

Although the scenario is simple, the proof is complicated by the mutual independence between attempt probabilities and collision probabilities in the fixed point, their dependence on the scale η_k, and the dependence of the slot time $\mathbb{E}[Y]$ on η_k.

B.1.1 Proof of Claim (T4.3-1)

Substituting (4.1), (4.5) into (4.3) gives

$$C_{s_k} = \frac{2}{W_{B_1}}(1 - 2p_{s_k})\mathcal{T}$$

which is decreasing in p_{s_k}. Thus, to prove Claim (T4.3-1), it is sufficient to show that $dp_{s_k}/d\eta < 0$. To prove $dp_{s_k}/d\eta < 0$, I first find a closed-form expression of $\frac{dp_{s_k}}{d\eta_k}$ and then prove it is less than 0. This closed-form expression can be achieved by solving a system of two linear equations with two variables: $dp_{s_k}/d\eta$ and $d\tau_{s_k}/d\eta$. Those are done as follows.

Firstly, I find $dp_{s_k}/d\eta$ as a linear function of $\frac{d\tau_{s_k}}{d\eta_k}$. Recall that $N_u = N_{s_k} = N_s = 1$, whence by (4.2c), $p_{s_k} = \tau_u$ and $p_u = \tau_{s_k}$. Hence by (4.2b),

$$p_{s_k} = \lambda_u \mathbb{E}[Y] \frac{1}{1 - \tau_{s_k}}. \tag{B.3}$$

Taking the derivative of (B.3) with respect to η_k gives

$$\begin{aligned}
\frac{dp_{s_k}}{d\eta_k} &= \lambda_u \Big(\frac{d\mathbb{E}[Y]}{d\eta_k} \frac{1}{1 - \tau_{s_k}} + \mathbb{E}[Y] \frac{d(\frac{1}{1-\tau_{s_k}})}{d\eta_k} \Big) \\
&= \lambda_u \Big(\frac{1}{1 - \tau_{s_k}} \frac{d\mathbb{E}[Y]}{d\eta_k} + \mathbb{E}[Y] \frac{1}{(1 - \tau_{s_k})^2} \frac{d\tau_{s_k}}{d\eta_k} \Big) \\
&= \frac{\lambda_u}{(1 - \tau_{s_k})^2} \left(\frac{d\mathbb{E}[Y]}{d\eta_k} (1 - \tau_{s_k}) + \mathbb{E}[Y] \frac{d\tau_{s_k}}{d\eta_k} \right).
\end{aligned} \tag{B.4}$$

To have $dp_{s_k}/d\eta$ as a linear function of $\frac{d\tau_{s_k}}{d\eta_k}$ from (B.4), I now express $\frac{d\mathbb{E}[Y]}{d\eta_k}$ in terms of $d\tau_{s_k}/d\eta_k$ as follows.

Since $N_u = N_s = N_{s_k} = 1$ by hypothesis, substituting (4.2c) into (3.10a) gives

$$\mathbb{E}[Y] = \sigma(1 - \tau_{s_k})(1 - p_{s_k}) + T_u \tau_u (1 - p_u) + T^s_{s_k} \tau_{s_k} (1 - p_{s_k}) + T_c \tau_{s_k} \tau_u \tag{B.5}$$

where $T_c = \max(T_u, T_s)$.

Then, substituting $p_{s_k} = \tau_u$ and $p_u = \tau_{s_k}$ into (B.5) gives

$$\begin{aligned}\mathbb{E}[Y] &= \sigma(1-\tau_{s_k})(1-p_{s_k}) + T_u p_{s_k}(1-\tau_{s_k}) + T^s_{s_k}\tau_{s_k}(1-p_{s_k}) + T_c\tau_{s_k}p_{s_k} \\ &= (\sigma - T_u)(1-\tau_{s_k})(1-p_{s_k}) + T_c + (T_u - T_c)(1-\tau_{s_k}) + (T^s_{s_k} - T_c)\tau_{s_k}(1-p_{s_k})\end{aligned} \tag{B.6}$$

From (4.5), $W_{s_k} = \eta_k W_{B_1}$ implies

$$p_{s_k} = 1 - \frac{2/W_{B_1}}{4/W_{B_1} - \eta_k\tau_{s_k}}. \tag{B.7}$$

Substituting (B.7) into (B.6) gives

$$\begin{aligned}\mathbb{E}[Y] =&(\sigma - T_u)(1-\tau_{s_k})\frac{2/W_{B_1}}{4/W_{B_1} - \eta_k\tau_{s_k}} + T_c \\ &+ (T_u - T_c)(1-\tau_{s_k}) + (T^s_{s_k} - T_c)\tau_{s_k}\frac{2/W_{B_1}}{4/W_{B_1} - \eta_k\tau_{s_k}}.\end{aligned} \tag{B.8}$$

This shows that $\mathbb{E}[Y]$ is a function of τ_{s_k} and $T^s_{s_k}$, both of which depend on η_k.

By (B.2), T_u and T_c are independent of η_k and $dT^s_{s_k}/d\eta_k = \mathcal{T}$ from (B.1). Besides, τ_{s_k} is also a function of η_k. Then, taking derivative of (B.8) gives

$$\begin{aligned}\frac{d\mathbb{E}[Y]}{d\eta_k} =&(\sigma - T_u)(2/W_{B_1})\frac{d(\frac{1-\tau_{s_k}}{4/W_{B_1}-\eta_k\tau_{s_k}})}{d\eta_k} + (T_u - T_c)\frac{d(1-\tau_{s_k})}{d\eta_k} \\ &+ (2/W_{B_1})\frac{d(\frac{(T^s_{s_k}-T_c)\tau_{s_k}}{4/W_{B_1}-\eta_k\tau_{s_k}})}{d\eta_k} \\ =&(\sigma - T_u)(2/W_{B_1})\frac{1}{(4/W_{B_1} - \eta_k\tau_{s_k})^2} \\ &\cdot\Big(- \frac{d\tau_{s_k}}{d\eta_k}(4/W_{B_1} - \eta_k\tau_{s_k}) - (1-\tau_{s_k})\frac{d(4/W_{B_1} - \eta_k\tau_{s_k})}{d\eta_k}\Big) \\ &- (T_u - T_c)\frac{d\tau_{s_k}}{d\eta_k} + (2/W_{B_1})(\frac{1}{(4/W_{B_1} - \eta_k\tau_{s_k})^2}) \\ &\cdot\Big(\frac{d((T^s_{s_k} - T_c)\tau_{s_k})}{d\eta_k}(4/W_{B_1} - \eta_k\tau_{s_k}) - (T^s_{s_k} - T_c)\tau_{s_k}\frac{d(4/W_{B_1} - \eta_k\tau_{s_k})}{d\eta_k}\Big) \\ =&\frac{(\sigma - T_u)(2/W_{B_1})}{(4/W_{B_1} - \eta_k\tau_{s_k})^2}\Big(- (4/W_{B_1} - \eta_k\tau_{s_k})\frac{d\tau_{s_k}}{d\eta_k} + (1-\tau_{s_k})\frac{d(\eta_k\tau_{s_k})}{d\eta_k}\Big)\end{aligned}$$

Table B.1: Math expression of symbols in Theorem 4.3.

Symbol	Expression
K_1	$(\sigma - T_u)(2/W_{B_1})(\eta_k - 4/W_{B_1})$ $-(T_u - T_c)(4/W_{B_1} - \eta_k\tau_{s_k})^2 + (T^s_{s_k} - T_c)(8/W^2_{B_1})$
K_2	$(2/W_{B_1})\tau_{s_k}\Big((\sigma - T_u)(1-\tau_{s_k}) + \mathcal{T}(4/W_{B_1}) + (E - T_c)\tau_{s_k}\Big)$
L_1	$1 + \frac{\lambda_u}{(1-\tau_{s_k})^2}\Big(\frac{K_1(1-\tau_{s_k})}{(4/W_{B_1} - \eta_k\tau_{s_k})^2} + \mathbb{E}[Y]\Big)\frac{(2/W_{B_1})}{\eta_k}\frac{1}{(1-p_{s_k})^2}$
L_2	$\frac{K_2(1-\tau_{s_k})}{(4/W_{B_1} - \eta_k\tau_{s_k})^2} - \Big(\frac{K_1(1-\tau_{s_k})}{(4/W_{B_1} - \eta_k\tau_{s_k})^2} + \mathbb{E}[Y]\Big)\frac{\tau_{s_k}}{\eta_k}$
H_1	$\frac{1}{(4/W_{B_1} - \eta_k\tau_{s_k})^2}\Big(-(2/W_{B_1})\tau_{s_k} - \frac{\lambda_u K_2}{1-\tau_{s_k}}\Big)$
H_2	$\frac{\eta_k(2/W_{B_1})}{(4/W_{B_1} - \eta_k\tau_{s_k})^2} + \frac{\lambda_u}{1-\tau_{s_k}}\Big(\frac{\mathbb{E}[Y]}{1-\tau_{s_k}} + \frac{K_1}{(4/W_{B_1} - \eta_k\tau_{s_k})^2}\Big)$

$$
\begin{aligned}
&- (T_u - T_c)\frac{d\tau_{s_k}}{d\eta_k} + \frac{2/W_{B_1}}{(4/W_{B_1} - \eta_k\tau_{s_k})^2} \\
&.\Big(\big(\frac{d\tau_{s_k}}{d\eta_k}(T^s_{s_k} - T_c) + \tau_{s_k}\frac{d(T^s_{s_k} - T_c)}{d\eta_k}\big)(4/W_{B_1} - \eta_k\tau_{s_k}) + (T^s_{s_k} - T_c)\tau_{s_k}\frac{d(\eta_k\tau_{s_k})}{d\eta_k}\Big) \\
=&\frac{(\sigma - T_u)(2/W_{B_1})}{(4/W_{B_1} - \eta_k\tau_{s_k})^2}\Big(-(4/W_{B_1} - \eta_k\tau_{s_k})\frac{d\tau_{s_k}}{d\eta_k} + (1-\tau_{s_k})(\tau_{s_k} + \eta_k\frac{d\tau_{s_k}}{d\eta_k})\Big) \\
&- (T_u - T_c)\frac{d\tau_{s_k}}{d\eta_k} + \frac{2/W_{B_1}}{(4/W_{B_1} - \eta_k\tau_{s_k})^2} \\
&.\Big(\big((T^s_{s_k} - T_c)\frac{d\tau_{s_k}}{d\eta_k} + \tau_{s_k}\mathcal{T}\big)(4/W_{B_1} - \eta_k\tau_{s_k}) + (T^s_{s_k} - T_c)\tau_{s_k}(\tau_{s_k} + \eta_k\frac{d\tau_{s_k}}{d\eta_k})\Big)
\end{aligned}
\tag{B.9}
$$

Substituting (B.1) into (B.9) gives

$$
\begin{aligned}
\frac{d\mathbb{E}[Y]}{d\eta_k} =&\frac{(\sigma - T_u)(2/W_{B_1})}{(4/W_{B_1} - \eta_k\tau_{s_k})^2}\Big(\big(\eta_k - \frac{4}{W_{B_1}}\big)\frac{d\tau_{s_k}}{d\eta_k} + \tau_{s_k} - \tau^2_{s_k}\Big) - (T_u - T_c)\frac{d\tau_{s_k}}{d\eta_k} \\
&+ \frac{(2/W_{B_1})}{(4/W_{B_1} - \eta_k\tau_{s_k})^2}\Big((T^s_{s_k} - T_c)(4/W_{B_1})\frac{d\tau_{s_k}}{d\eta_k} + \mathcal{T}(4/W_{B_1})\tau_{s_k} + (E - T_c)\tau^2_{s_k}\Big) \\
=&\frac{1}{(4/W_{B_1} - \eta_k\tau_{s_k})^2}\Big(K_1\frac{d\tau_{s_k}}{d\eta_k} + K_2\Big)
\end{aligned}
\tag{B.10}
$$

where K_1 and K_2 are given in Table B.1.

Then, substituting (B.10) into (B.4) gives

$$\begin{aligned}\frac{dp_{s_k}}{d\eta_k} =& \frac{\lambda_u}{(1-\tau_{s_k})^2}\Big(\frac{1-\tau_{s_k}}{(4/W_{B_1}-\eta_k\tau_{s_k})^2}(K_1\frac{d\tau_{s_k}}{d\eta_k}+K_2)+\mathbb{E}[Y]\frac{d\tau_{s_k}}{d\eta_k}\Big)\\ =& \frac{\lambda_u}{(1-\tau_{s_k})^2}\bigg(\frac{K_2(1-\tau_{s_k})}{(4/W_{B_1}-\eta_k\tau_{s_k})^2}+\Big(\frac{K_1(1-\tau_{s_k})}{(4/W_{B_1}-\eta_k\tau_{s_k})^2}+\mathbb{E}[Y]\Big)\frac{d\tau_{s_k}}{d\eta_k}\bigg). \quad \text{(B.11)}\end{aligned}$$

Secondly, I find $\frac{d\tau_{s_k}}{d\eta_k}$ as a linear function of $dp_{s_k}/d\eta$. Substituting $W_{s_k}=\eta_k W_{B_1}$ into (4.5) gives

$$\tau_{s_k} = \frac{(2/W_{B_1})}{\eta_k}\Big(2-\frac{1}{1-p_{s_k}}\Big) \quad \text{(B.12)}$$

Then, differentiating (B.12) with respect to η_k gives

$$\begin{aligned}\frac{d\tau_{s_k}}{d\eta_k} &= (2/W_{B_1})\Big(\frac{d(\frac{1}{\eta_k})}{d\eta_k}\Big(2-\frac{1}{1-p_{s_k}}\Big)+\frac{1}{\eta_k}\frac{d(2-\frac{1}{1-p_{s_k}})}{d\eta_k}\Big)\\ &= (2/W_{B_1})\bigg(-\frac{1}{\eta_k^2}\Big(2-\frac{1}{1-p_{s_k}}\Big)+\frac{1}{\eta_k}\Big(-\frac{1}{(1-p_{s_k})^2}\frac{dp_{s_k}}{d\eta_k}\Big)\bigg)\\ &= -\frac{(2/W_{B_1})}{\eta_k^2}\Big(2-\frac{1}{1-p_{s_k}}\Big)+\frac{(2/W_{B_1})}{\eta_k}\Big(-\frac{1}{(1-p_{s_k})^2}\frac{dp_{s_k}}{d\eta_k}\Big)\\ &= -\frac{1}{\eta_k}\frac{(2/W_{B_1})}{\eta_k}\Big(2-\frac{1}{1-p_{s_k}}\Big)-\frac{(2/W_{B_1})}{\eta_k(1-p_{s_k})^2}\frac{dp_{s_k}}{d\eta_k}\\ &= -\frac{1}{\eta_k}\tau_{s_k}-\frac{2/W_{B_1}}{\eta_k}\frac{1}{(1-p_{s_k})^2}\frac{dp_{s_k}}{d\eta_k}. \quad \text{(B.13)}\end{aligned}$$

Note that the last expression uses (B.12).

Solving two linear equations (B.11) and (B.13) gives

$$\frac{dp_{s_k}}{d\eta_k}L_1 = \frac{\lambda_u}{(1-\tau_{s_k})^2}L_2 \quad \text{(B.14)}$$

where L_1 and L_2 are given in Table B.1. From (B.14), to show $dp_{s_k}/d\eta_k<0$, it is sufficient to show $L_1>0$ and $L_2<0$ as follows.

First, I show $L_1>0$. I start with determining the term $\frac{K_1(1-\tau_{s_k})}{(4/W_{B_1}-\eta_k\tau_{s_k})^2}+\mathbb{E}[Y]$ which appears in both L_1 and L_2. Let

$$J = (4/W_{B_1}-\eta_k\tau_{s_k})^2 \quad \text{(B.15)}$$

Substituting (B.8) and K_1 from Table B.1 into this term gives

$$\begin{aligned}
&\frac{K_1(1-\tau_{s_k})}{J} + \mathbb{E}[Y] \\
=&\Bigg((\sigma - T_u)\frac{(2/W_{B_1})(\eta_k - 4/W_{B_1})}{J} - (T_u - T_c) + (T^s_{s_k} - T_c)\frac{8/W_u^2}{J}\Bigg)(1-\tau_{s_k}) \\
&+ (\sigma - T_u)(1-\tau_{s_k})\frac{2/W_{B_1}}{\sqrt{J}} + T_c + (T_u - T_c)(1-\tau_{s_k}) + (T^s_{s_k} - T_c)\tau_{s_k}\frac{2/W_{B_1}}{\sqrt{J}} \\
=&\frac{1}{J}\Bigg((\sigma - T_u)(1-\tau_{s_k})(2/W_{B_1})\Big((\eta_k - 4/W_{B_1}) + (4/W_{B_1} - \eta_k\tau_{s_k})\Big) + T_c J \\
&+ (T^s_{s_k} - T_c)(2/W_{B_1})\Big((4/W_{B_1})(1-\tau_{s_k}) + \tau_{s_k}(4/W_{B_1} - \eta_k\tau_{s_k})\Big)\Bigg) \\
=&\frac{1}{J}\Big((\sigma - T_u)(1-\tau_{s_k})^2(2/W_{B_1})\eta_k + T_c J + (T^s_{s_k} - T_c)(2/W_{B_1})(4/W_{B_1} - \eta_k\tau^2_{s_k})\Big).
\end{aligned} \tag{B.16}$$

From (B.7),

$$\frac{1}{(1-p_{s_k})^2} = \frac{(4/W_{B_1} - \eta_k\tau_{s_k})^2}{(2/W_{B_1})^2} = \frac{J}{(2/W_{B_1})^2}. \tag{B.17}$$

Substituting (B.16) and (B.17) into L_1 from Table B.1 gives

$$L_1 = 1 + \lambda_u\Big((\sigma - T_u) + T_c\frac{(4/W_{B_1} - \eta_k\tau_{s_k})^2}{(1-\tau_{s_k})^2}\frac{1}{(2/W_{B_1})\eta_k} + (T^s_{s_k} - T_c)\frac{(4/W_{B_1} - \eta_k\tau^2_{s_k})}{(1-\tau_{s_k})^2\eta_k}\Big) \tag{B.18}$$

Since $\lambda_u T_u \leq 1$ by hypothesis, $T^s_{s_k} \geq T_c$ by (B.1) and (B.2), and

$$4/W_{B_1} - \eta_k\tau^2_{s_k} > 4/W_{B_1} - \eta_k\tau_{s_k} > 0 \tag{B.19}$$

, we have $L_1 > 0$ from (B.7).

Next, I show $L_2 < 0$. Substituting K_2 from Table B.1, (B.16), and (B.1) into L_2 from Table B.1 gives

$$\begin{aligned}
L_2 =&\frac{(1-\tau_{s_k})(2/W_{B_1})\tau_{s_k}}{J}\Big((\sigma - T_u)(1-\tau_{s_k}) + \mathcal{T}(4/W_{B_1}) + (E - T_c)\tau_{s_k}\Big) \\
&- \frac{\tau_{s_k}}{\eta_k J}\Big((\sigma - T_u)(1-\tau_{s_k})^2(2/W_{B_1})\eta_k + T_c J
\end{aligned}$$

$$
\begin{aligned}
&+ (E + \eta_k \mathcal{T} - T_c)(2/W_{B_1})(4/W_{B_1} - \eta_k \tau_{s_k}^2)\Big) \\
=&\frac{(1-\tau_{s_k})\tau_{s_k}(2/W_{B_1})}{J}\Big(\mathcal{T}\frac{4}{W_{B_1}} + (E - T_c)\tau_{s_k}\Big) \\
&- \frac{\tau_{s_k}}{\eta_k J}\Big(T_c J + (E + \eta_k \mathcal{T} - T_c)(2/W_{B_1})(4/W_{B_1} - \eta_k \tau_{s_k}^2)\Big) \\
=&\mathcal{T}\frac{(2/W_{B_1})\tau_{s_k}}{J}\Big(\eta_k \tau_{s_k}^2 - (4/W_{B_1})\tau_{s_k}\Big) + (E - T_c)\frac{2/W_{B_1}}{J\eta_k}\tau_{s_k}(\eta_k \tau_{s_k} - 4/W_{B_1}) - T_c\frac{\tau_{s_k}}{\eta_k} \\
=&\frac{(2/W_{B_1})\tau_{s_k}}{J}(\eta_k \tau_{s_k} - 4/W_{B_1})\Big(\tau_{s_k}\mathcal{T} + \frac{E - T_c}{\eta_k}\Big) - \frac{\tau_{s_k}}{\eta_k}T_c \\
=&-\frac{(2/W_{B_1})\tau_{s_k}}{4/W_{B_1} - \eta_k \tau_{s_k}}\Big(\tau_{s_k}\mathcal{T} + \frac{E}{\eta_k}\Big) + \Big(\frac{\tau_{s_k}}{\eta_k}\Big)\Big(\frac{-2/W_{B_1} + \eta_k \tau_{s_k}}{4/W_{B_1} - \eta_k \tau_{s_k}}\Big)T_c
\end{aligned} \tag{B.20}
$$

From (4.5), we have $\tau_{s_k} < \frac{2}{\eta_k W_{B_1}}$ due to $p_{s_k} \in (0, 1)$. Then,

$$
-2/W_{B_1} + \eta_k \tau_{s_k} < -2/W_{B_1} + \eta_k \frac{2}{\eta_k W_{B_1}} = 0 \tag{B.21}
$$

This, together with $\mathcal{T} > 0$, $E > 0$ and (B.19), implies that $L_2 < 0$.

B.1.2 Proof of Claim (T4.3-2)

By (4.2c), $p_u = \tau_{s_k}$, it is sufficient to show that τ_{s_k} decreases when η_k increases. I first find a closed form expression of $d\tau_{s_k}/d\eta_k$ and then prove it to be less than 0. Recall from Appendix B.1.1 that the closed form of $d\tau_{s_k}/d\eta_k$ can be found by solving two linear equations (B.11) and (B.13), which gives

$$
H_1 = \frac{d\tau_{s_k}}{d\eta_k} H_2 \tag{B.22}
$$

where H_1 and H_2 are given in Table B.1. From (B.22), to show $d\tau_{s_k}/d\eta_k < 0$, it is sufficient to prove that $H_1 < 0$ and $H_2 > 0$ as follows.

First, I show $H_1 < 0$. Substituting K_2 from Table B.1 and $J = (4/W_{B_1} - \eta_k \tau_{s_k})^2$ into H_1 from Table B.1 gives

$$
H_1 = -\frac{(2/W_{B_1})\tau_{s_k}}{J} - \frac{\lambda_u}{1 - \tau_{s_k}}\frac{(2/W_{B_1})\tau_{s_k}}{J}\Big((\sigma - T_u)(1 - \tau_{s_k}) + \mathcal{T}(4/W_{B_1}) + (E - T_c)\tau_{s_k}\Big)
$$

$$= - \frac{(2/W_{B_1})\tau_{s_k}}{J}\Big(1 + \lambda_u(\sigma - T_u) + \frac{\lambda_u}{1-\tau_{s_k}}\Big((4/W_{B_1})\mathcal{T} + (E - T_c)\tau_{s_k}\Big)\Big) \tag{B.23}$$

Since $\lambda_u T_u < 1$ and $\tau_{s_k} \in (0,1)$ by hypothesis, to show $H_1 < 0$ it is sufficient to show that $(4/W_{B_1})\mathcal{T} + (E - T_c)\tau_{s_k} > 0$ as follows. From (4.5), we have $\tau_{s_k} < 2/(\eta_k W_{B_1})$ due to $p_{s_k} \in (0,1)$ and $T_c - E > 0$ by (B.2). Then,

$$\begin{aligned}(4/W_{B_1})\mathcal{T} + (E - T_c)\tau_{s_k} &> (4/W_{B_1})\mathcal{T} - (T_c - E)\frac{(2/W_{B_1})}{\eta_k} \\ &= (2/W_{B_1})\Big(2\mathcal{T} - \frac{T_c}{\eta_k} + \frac{E}{\eta_k}\Big)\end{aligned} \tag{B.24}$$

Since $T_c = \max(T_u, T_s) < 2\mathcal{T}$ by hypothesis, the left hand side of the inequality (B.24) is greater than 0, which proves that $H_1 < 0$.

Second, I show $H_2 > 0$. Substituting (B.8) and K_1 from Table B.1 into H_2 from Table B.1 gives

$$\begin{aligned}H_2 =& \frac{\eta_k(2/W_{B_1})}{(4/W_{B_1} - \eta_k\tau_{s_k})^2} + \frac{\lambda_u}{1-\tau_{s_k}}\Big((\sigma - T_u)\frac{(2/W_{B_1})}{4/W_{B_1} - \eta_k\tau_{s_k}}\big(1 + \frac{\eta_k - 4/W_{B_1}}{4/W_{B_1} - \eta_k\tau_{s_k}}\big) \\ &+ \frac{T_c}{1-\tau_{s_k}} + (T^s_{s_k} - T_c)\frac{(2/W_{B_1})}{4/W_{B_1} - \eta_k\tau_{s_k}}\big(\frac{\tau_{s_k}}{1-\tau_{s_k}} + \frac{4/W_{B_1}}{4/W_{B_1} - \eta_k\tau_{s_k}}\big)\Big) \\ =& \frac{\eta_k(2/W_{B_1})}{(4/W_{B_1} - \eta_k\tau_{s_k})^2} + \frac{\lambda_u}{1-\tau_{s_k}}\Big((\sigma - T_u)\frac{(2/W_{B_1})}{(4/W_{B_1} - \eta_k\tau_{s_k})^2}\eta_k(1-\tau_{s_k}) + \frac{T_c}{1-\tau_{s_k}} \\ &+ (T^s_{s_k} - T_c)\frac{(2/W_{B_1})(4/W_{B_1} - \eta_k\tau^2_{s_k})}{(1-\tau_{s_k})(4/W_{B_1} - \eta_k\tau_{s_k})^2}\Big) \\ =& \frac{(2/W_{B_1})\eta_k}{(4/W_{B_1} - \eta_k\tau_{s_k})^2}\big(1 + \lambda_u(\sigma - T_u)\big) \\ &+ \frac{\lambda_u}{(1-\tau_{s_k})^2}\Big(T_c + (T^s_{s_k} - T_c)\frac{(2/W_{B_1})(4/W_{B_1} - \eta_k\tau^2_{s_k})}{(4/W_{B_1} - \eta_k\tau_{s_k})^2}\Big)\end{aligned} \tag{B.25}$$

Moreover, since $\lambda_u T_u \leq 1$ by hypothesis, $T^s_{s_k} \geq T_c$ by (B.1) and (B.2), and (B.19), it follows from (B.25) that $H_2 > 0$. ■

B.2 Lemma 4.5

Proof: From (4.2a), we have

$$\frac{1}{1-p_{s_k}} = 2 - \frac{W_{B_k}}{2/\tau_{s_k} - 1}. \tag{B.26}$$

Moreover, dividing $1-p_{s_{j+i}}$ from (4.2c) by $1-p_{s_j}$ from (4.2c) gives

$$\frac{1-p_{s_{j+i}}}{1-p_{s_j}} = \frac{1-\tau_{s_j}}{1-\tau_{s_{j+i}}}. \tag{B.27}$$

To simplify notation, define

$$g(\tau, W) = \frac{1-\tau}{2-\frac{W}{2/\tau-1}} = \frac{1}{2}\frac{1-\tau}{1-\frac{W}{2}\frac{1}{2-\tau}\tau}. \tag{B.28}$$

Substituting $1-p_{s_{j+i}}$ and $1-p_{s_j}$ from (B.26) into (B.27) gives

$$g(\tau_{s_{j+i}}, W_{j+i}) = g(\tau_{s_j}, W_j). \tag{B.29}$$

Since $W_{s_k} > 4$ by hypothesis and $\tau \leq 1$, the coefficient $\frac{W}{2}\frac{1}{2-\tau}$ of τ in the denominator of (B.28) is greater than 1 and increasing in τ. Hence, $g(\tau, W)$ is increasing in τ. Moreover, $g(\tau, W)$ is increasing in W. Therefore, from (B.29), $W_{j+i} > W_j$ implies $\tau_{s_{j+i}} < \tau_{s_j}$ and $W_{j+i} \geq W_j$ implies $\tau_{s_{j+i}} \leq \tau_{s_j}$. ■

B.3 Theorem 4.4

Proof: Under the action profile $a_{(B_1;\cdot;B_{k\geq 1};\cdot)}$, we have $N_{s_k} \geq 1$, $N_{s_1} \geq 1$, $S_1(a_{(B_1;\cdot;B_{k\geq 1};\cdot)}) = S_{s_1}$ and $S_j(a_{(B_1;\cdot;B_{k\geq 1};\cdot)}) = S_{s_{k\geq 1}}$. Thus it is required to show $S_{s_1} \geq S_{s_k}$ under (4.2)–(4.4), with strict inequality if $\eta_k > 1$.

Dividing S_{s_k} by S_{s_1} from (4.4) and substituting (B.27) gives

$$\frac{S_{s_k}}{S_{s_1}} = \frac{\tau_{s_k}(1-p_{s_k})\eta_k}{\tau_{s_1}(1-p_{s_1})} = \frac{\tau_{s_k}(1-\tau_{s_1})\eta_k}{\tau_{s_1}(1-\tau_{s_k})}. \tag{B.30}$$

To show $S_{s_1} \geq S_{s_k}$ it is sufficient to show that the denominator of (B.30) is at least

as large as its numerator.

First, by (4.2a) and the fact that $W_{B_k} = \eta_k W_{B_1}$,

$$\tau_{s_1}(1-\tau_{s_k}) - \tau_{s_k}(1-\tau_{s_1})\eta_k \tag{B.31}$$

$$=\tau_{s_k}\tau_{s_1}\left(\frac{1}{\tau_{s_k}} - \frac{\eta_k}{\tau_{s_1}} + \eta_k - 1\right)$$

$$=\tau_{s_k}\tau_{s_1}\left(\frac{\eta_k W_{B_1}}{2}\left(\frac{1-p_{s_k}}{1-2p_{s_k}} - \frac{1-p_{s_1}}{1-2p_{s_1}}\right) + \frac{\eta_k - 1}{2}\right). \tag{B.32}$$

To show that (B.32) is non-negative, it is sufficient to show that $\frac{1-p_{s_k}}{1-2p_{s_k}} \geq \frac{1-p_{s_1}}{1-2p_{s_1}}$, or equivalently that $p_{s_k} \geq p_{s_1}$, since $p_{s_1} \geq 0$.

Under the action space $A0$ and by hypothesis, $W_{B_k} = \eta_k W_{B_1} \geq W_{B_1} > 4$, which satisfies the conditions of Lemma 4.5. Hence $\tau_{s_1} \geq \tau_{s_k}$, and by (B.27), $p_{s_k} \geq p_{s_1}$. If $\eta_k > 1$, these inequalities are all strict. ■

B.4 Lemma 4.7

Proof: To see how the attempt probability of the user 1 changes when its action changes from B_k to B_{k+i} ($i > 0$), consider an arbitrary action profile of the form $a_{(X;)}$ for some $X \in A$. Then there are a $j \neq 1$ and a variable c which depends on X and a_j, such that

$$W_1 = cW_j \tag{B.33}$$

By hypothesis, $W_1 > 11$, whence $cW_j > 11$. Note that subscripts 1, i and i in this proof are to denote the quantities for user 1, i and j.

I first prove that there exists a unique solution of the fixed point model and find that solution. I then show how the solution changes with the action choice of user 1.

Since $N_u = 0$ by hypothesis, (4.2c) implies

$$p_i = 1 - \frac{\prod_{k=1}^{N_s}(1-\tau_k)}{1-\tau_i}, \quad \forall i \in \mathcal{P}. \tag{B.34}$$

whence

$$(1-p_i)(1-\tau_i) = (1-p_j)(1-\tau_j), \quad \forall i \neq j \tag{B.35}$$

From (4.5),

$$p_i = 1 - \frac{2}{4 - W_i\tau_i}, \quad \forall i \in \mathcal{P}. \tag{B.36}$$

Replacing $1-p_j$ and $1-p_i$ from (B.36) into (B.35) gives

$$\frac{1-\tau_i}{4-W_i\tau_i} = \frac{1-\tau_j}{4-W_j\tau_j}. \tag{B.37}$$

This is equivalent to

$$\tau_i = \frac{(4-W_j)\tau_j}{4 - W_i + (W_i - W_j)\tau_j}, \quad \forall i \in \mathcal{P} \setminus \{j\}. \tag{B.38}$$

Substituting (B.33) and p_1 from (B.34) into τ_1 from (4.5) gives

$$\tau_1 = \frac{2}{cW_j}\left(2 - \frac{1}{\prod_{i=2}^{N_s}(1-\tau_i)}\right) \equiv f_1(\tau_j, c). \tag{B.39}$$

Note that f_1 is a function of τ_j due to the relation between τ_i-s in the denominator and τ_j given in (B.38).

Substituting $1-p_j$ from (B.36) into $1-p_j$ from (B.34) gives

$$\tau_1 = 1 - \frac{2(1-\tau_j)}{(4-W_j\tau_j)\prod_{i=2}^{N_s}(1-\tau_i)} \equiv f_2(\tau_j). \tag{B.40}$$

Then, a solution of the fixed point model is any solution to $f_1(\tau_j, c) = f_2(\tau_j)$ with $\tau_j \in [0,1]$. I first prove there exists such a solution and then prove its uniqueness.

Now $f_1(\tau_j, c)$ and $f_2(\tau_j)$ are decreasing functions of τ_j on $[0, 2/W_j]$, as illustrated in Fig. B.1.

Moreover, at $\tau_j = 0$, we have $f_2(0) > f_1(0,c) > 0$. Besides, let τ_j^* be the solution to $f_1(\tau_j) = 0$. Then, (B.39) implies $\prod_{i=2}^{N_s}(1-\tau_i) = 1/2$. Substituting this into (B.40)

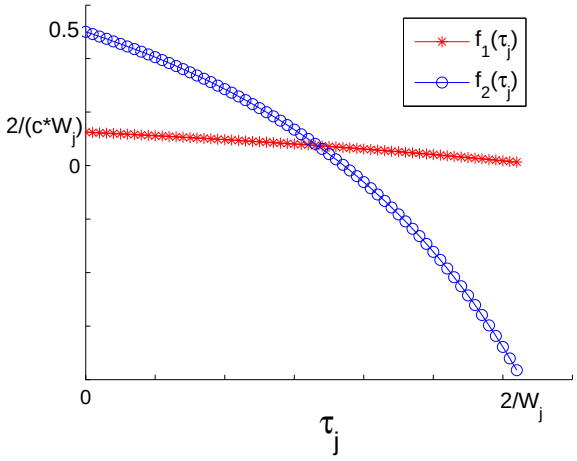


Figure B.1: $f_1(\tau_j, c)$ and $f_2(\tau_j)$

gives

$$f_2(\tau_j) = 1 - \frac{2(1-\tau_j)}{(4 - W_j\tau_j)(1/2)} = \frac{(4-W_j)\tau_j}{4 - W_j\tau_j} < 0 \tag{B.41}$$

If τ_j^* is in (0,1) and unique, then these, together with the continuity of $f_1(\tau_j, c)$ and $f_2(\tau_j)$, imply that there exists a solution to $f_1(\tau_j, c) = f_2(\tau_j)$ with $\tau_j \in (0, \tau_j^*)$. The following proves that τ_j^* is unique solution in (0,1) of $\prod_{i=2}^{N_s}(1-\tau_i) = 1/2$.

Let $g(\tau_j) = \prod_{i=2}^{N_s}(1-\tau_i)$. At $\tau_j = 0$, we have $\tau_{k\neq 1} = 0$ from (B.38); then, $g(0) = 1 > 1/2$. Moreover, at $\tau_j = 1$, we have $\tau_{k\neq 1} = 1$ from (B.38); then, $g(1) = 0 < 1/2$. These, together with the fact that $g(\tau_j)$ is a decreasing function of τ_j (due to $\tau_{k\neq 1}$ increasing with τ_j from (B.38)), imply that $f_1(\tau_j) = 0$ has unique solution τ_j^* in (0,1).

Next, to see that the solution to $f_1(\tau_j, c) = f_2(\tau_j)$ is unique, let $f(\tau_j, c) = f_1(\tau_j, c) - f_2(\tau_j)$, which is given by

$$\frac{2}{\prod_{i\in\mathcal{P}\setminus\{1,j\}}(1-\tau_i)}\left(\frac{1}{cW_j(1-\tau_j)} - \frac{1}{4-W_j\tau_j}\right) + 1 - \frac{4}{cW_j} \equiv g_1(\tau_j)g_2(\tau_j) + 1 - \frac{4}{cW_j}$$

Clearly $g_1(\tau_j)$ is increasing and positive for $\tau_j \in [0, 1)$. Moreover, $g_2(\tau_j)$ is negative since (4.5) implies the second term is negative, and the hypothesis $W_i > 11$ for all i implies that $cW_j(1-\tau_j) > 4(1-\tau_j) > 4 - W_j\tau_j$. Similarly, g_2 is decreasing because

its derivative

$$g_2'(\tau_j) = \frac{1}{cW_j}\frac{1}{(1-\tau_j)^2} - \frac{W_j}{(4-W_j\tau_j)^2}$$
$$< \frac{1}{4(1-\tau_j)^2} - \frac{W_j}{(4-W_j\tau_j)^2} = \frac{(4-W_j)(4-W_j\tau_j^2)}{4(1-\tau_j)^2(4-W_j\tau_j)^2} < 0$$

which uses the fact that $4 - W_j\tau_j^2 \geq 4 - W_j\tau_j > 0$ by (B.36) and $4 - W_j < 0$. Thus $f(\tau_j, c)$ is decreasing in τ_j. This implies that the solution to $f_1(\tau_j, c) = f_2(\tau_j)$ is unique.

I will now investigate how this unique solution changes with the action of user 1. When user 1 changes its action, its CW_{min} (W_1) changes, causing the coefficient c in (B.33) to change. Let τ_{j1} and τ_{j2} be the solutions to $f(\tau_j, c) = 0$ for $c = c_1$ and $c = c_2 > c_1$, respectively.

It is clear that $f(\tau_j, c)$ is also increasing in c; hence, $f(\tau_{j1}, c_2) > f(\tau_{j1}, c_1) = f(\tau_{j2}, c_2) = 0$. This, together with the fact that $f(\tau_j, c)$ is a decreasing function of τ_j, implies that $\tau_{j1} < \tau_{j2}$. Therefore, when c increases or W_1 increases, τ_j increases and τ_1 decreases.

In particular, c decreases when a_1 changes from B_k to B_{k-1} while a_j remain unchanged; hence, this change decreases τ_j and increases τ_1. ■

B.5 Theorem 4.6

Let $\tau_i(a)$, $p_i(a)$ and $W_i(a)$ denote the attempt probability, collision probability and minimum contention window of a player $i \in \mathcal{P}$ under the action profile a. Let j denote any player in $\mathcal{P} \setminus \{1\}$.

Proof: The successful transmission rate per slot of the data user in accordance with each action profiles $a_{(B_1;\cdot)}$ and $a_{(B_{k>1};\cdot)}$, respectively, are given from (4.3) as follows

$$C_1(a_{(B_{k>1};\cdot)}) = \eta_k\tau_1(a_{(B_{k>1};\cdot)})\big(1 - p_1(a_{(B_{k>1};\cdot)})\big)\mathcal{T} \tag{B.42a}$$

$$C_1(a_{(B_1;\cdot)}) = \tau_1(a_{(B_1;\cdot)})\big(1 - p_1(a_{(B_1;\cdot)})\big)\mathcal{T}. \tag{B.42b}$$

To show $C_1(a_{(B_{k>1};\cdot)}) < C_1(a_{(B_1;\cdot)})$, it's sufficient to show

$$\eta_k \tau_1(a_{(B_{k>1};\cdot)}) > \tau_1(a_{(B_1;\cdot)}) \tag{B.43a}$$

$$p_1(a_{(B_{k>1};\cdot)}) > p_1(a_{(B_1;\cdot)}). \tag{B.43b}$$

Those will be proven as follows.

The conditions of this theorem satisfy those of Lemma 4.7. In the action space $A0$, I partition the cases by the action a_1 of user 1.

Consider $a_1 = B_1$. From (B.39),

$$\tau_1(a_{(B_1;\cdot)}) = \frac{2}{W_{B_1}}\left(2 - \frac{1}{\prod_{i=2}^{N_s}(1 - \tau_i(a_{(B_1;\cdot)}))}\right). \tag{B.44a}$$

Otherwise, $a_1 = B_k$. From (B.39),

$$\tau_1(a_{(B_{k>1};\cdot)}) = \frac{2}{\eta_k W_{B_1}}\left(2 - \frac{1}{\prod_{i=2}^{N_s}(1 - \tau_i(a_{(B_{k>1};\cdot)}))}\right). \tag{B.44b}$$

When a_1 changes from $B_{k>1}$ to B_1, c in (B.33) decreases because class $B_{k>1}$ has higher CW_{min} than class B_1. Then, from Lemma 4.7, for any player $j \neq 1$, we have

$$\tau_j(a_{(B_{k>1};\cdot)}) > \tau_j(a_{(B_1;\cdot)}). \tag{B.45}$$

From (B.44) and (B.45), I obtain (B.43a) as follows

$$\begin{aligned}\tau_1(a_{(B_1;\cdot)}) &= \eta_k \frac{2}{\eta_k W_{B_1}}\left(2 - \frac{1}{\prod_{i=2}^{N_s}(1 - \tau_i(a_{(B_1;\cdot)}))}\right) \\ &> \eta_k \frac{2}{\eta_k W_{B_1}}\left(2 - \frac{1}{\prod_{i=2}^{N_s}(1 - \tau_i(a_{(B_{k>1};\cdot)}))}\right) = \eta_k \tau_1(a_{(B_{k>1};\cdot)}).\end{aligned}$$

Applying (B.45) to (B.34) gives (B.43b). ■

B.6 Theorem 4.8

Proof: First, note that $S_1(a_{(B_k;\cdot;B_{k-1};\cdot)}) = S_{s_k}$ and $S_j(a_{(B_k;\cdot;B_{k-1};\cdot)}) = S_{s_{k-1}}$ under the wireless model (4.2)–(4.4). Therefore, it is sufficient to show that all ϵ_k satisfying (4.7) will satisfy $S_{s_k}/S_{s_{k-1}} > 1$, as follows.

Let $\phi(W,p) = W(1-p)/(1-2p)$. With this notation, dividing τ_{s_k} from (4.2a) by $\tau_{s_{k-1}}$ from (4.2a), gives

$$\frac{\tau_{s_k}}{\tau_{s_{k-1}}} = \frac{\phi(W_{B_{k-1}}, p_{s_{k-1}}) + 1}{\phi(W_{B_k}, p_{s_k}) + 1}. \tag{B.46}$$

Moreover, we can apply Lemma 4.5 since, by hypothesis,

$$W_{B_k} = \frac{\eta_k}{\eta_{k-1}} W_{B_{k-1}} - \epsilon_k > W_{B_{k-1}} > 11$$

Hence $p_{s_k} > p_{s_{k-1}}$ by (4.2c). Since $W_{B_k} > W_{B_{k-1}}$, this implies $\phi(W_{B_k}, p_{s_k}) > \phi(W_{B_{k-1}}, p_{s_{k-1}})$, whence (B.46) gives

$$\frac{\tau_{s_k}}{\tau_{s_{k-1}}} > \frac{\phi(W_{B_{k-1}}, p_{s_{k-1}})}{\phi(W_{B_k}, p_{s_k})}. \tag{B.47}$$

By (4.3), dividing S_{s_k} from (4.4) by $S_{s_{k-1}}$ from (4.4), and then substituting (B.47) and the definition of ϕ gives

$$\frac{S_{s_k}}{S_{s_{k-1}}} = \frac{\eta_k \tau_{s_k}(1-p_{s_k})}{\eta_{k-1}\tau_{s_{k-1}}(1-p_{s_{k-1}})} > \frac{\eta_k W_{B_{k-1}}(1-2p_{s_k})}{\eta_{k-1} W_{B_k}(1-2p_{s_{k-1}})}. \tag{B.48}$$

It remains to show that the right hand side exceeds 1.

Dividing $1-\tau_{s_k}$ by $1-\tau_{s_{k-1}}$ with τ_k ($k \in \{t,d\}$) from (4.2a) gives

$$\begin{aligned}\frac{1-\tau_{s_k}}{1-\tau_{s_{k-1}}} &= \frac{1 - \dfrac{2}{\phi(W_{B_k}, p_{s_k}) + 1}}{1 - \dfrac{2}{\phi(W_{B_{k-1}}, p_{s_{k-1}}) + 1}} \\ &< \frac{1 - 2/\phi(W_{B_k}, p_{s_k})}{1 - 2/\phi(W_{B_{k-1}}, p_{s_{k-1}})} = \frac{W_{B_k} - 2 - p_{s_k}(W_{B_k} - 4)}{W_{B_{k-1}} - 2 - p_{s_{k-1}}(W_{B_{k-1}} - 4)} \frac{W_{B_{k-1}}}{W_{B_k}} \frac{1-p_{s_{k-1}}}{1-p_{s_k}}.\end{aligned} \tag{B.49}$$

since $\phi(W_{B_k}, p_{s_k}) > \phi(W_{B_{k-1}}, p_{s_{k-1}}) > 1$.

The final factor of (B.49) cancels with the left hand side by (B.27), and so the hypothesis $W_{B_k} > 4$ implies

$$\begin{aligned} 1 - 2p_{s_k} >& 1 - 2\frac{W_{B_k} - 2 - \left(W_{B_{k-1}} - 2 - (W_{B_{k-1}} - 4)p_{s_{k-1}}\right)\frac{W_{B_k}}{W_{B_{k-1}}}}{W_{B_k} - 4} \\ &= \frac{W_{B_k}}{W_{B_{k-1}}}\frac{W_{B_{k-1}} - 4}{W_{B_k} - 4}(1 - 2p_{s_{k-1}}). \end{aligned}$$

Substituting this into (B.48) and using the fact that $1 - 2p_{s_{k-1}} > 0$ we obtain

$$\frac{S_{s_k}}{S_{s_{k-1}}} > \frac{\eta_k W_{B_{k-1}}(1 - 2p_{s_k})}{\eta_{k-1} W_{B_k}(1 - 2p_{s_{k-1}})} > \frac{\eta_k}{\eta_{k-1}}\frac{W_{B_{k-1}} - 4}{W_{B_k} - 4}. \tag{B.50}$$

For $W_{B_k} = \frac{\eta_k}{\eta_{k-1}} W_{B_{k-1}} - \epsilon_k$ with $\epsilon_k \geq 4(\frac{\eta_k}{\eta_{k-1}} - 1)$, the most right hand side of (B.50) is at least 1, which implies that $S_{s_k} > S_{s_{k-1}}$. ■

B.7 Lemma 4.2

Proof: To prove that the attempt probability of a data user reduces when its CW_{min} increases, I first find the solution of the fixed point and then prove its property when CW_{min} changes.

By hypothesis, I will consider the network with $N_u = 0$ and $N_s = N_{s_k}$. Then, (4.2) becomes

$$\tau_{s_k} = \frac{2}{W_{B_k}\frac{1-p_{s_k}}{1-2p_{s_k}} + 1} \equiv g_1(p_{s_k}) \tag{B.51a}$$

$$p_{s_k} = 1 - (1 - \tau_{s_k})^{N_s - 1}. \tag{B.51b}$$

From (B.51b),

$$\tau_{s_k} = 1 - (1 - p_{s_k})^{1/(N_s - 1)} \equiv g_2(p_{s_k}). \tag{B.52}$$

The solution of (B.51) is the solution to $g_1(p_{s_k}) = g_2(p_{s_k})$. Next, I will prove that there exists a solution to $g_1(p_{s_k}) = g_2(p_{s_k})$ and the solution is unique.

First, for finite N_s,

$$g_1(0) = \frac{2}{W_{B_k}+1} > g_2(0) = 0$$
$$g_1(1/2) = 0 < g_2(1/2) = 1 - (1/2)^{1/(N_s-1)}.$$

This, together with the fact that $g_1(p_{s_k})$ and $g_2(p_{s_k})$ are continuous functions over $[0, \frac{1}{2}]$, implies that there exists solution to $g_1(p_{s_k}) = g_2(p_{s_k})$.

Second, $g_2(p_{s_k})$ is an increasing function of p_{s_k} and $g_1(p_{s_k})$ is a decreasing function of p_{s_k}. Hence, it can be concluded that the solution to $g_1(p_{s_k}) = g_2(p_{s_k})$ is unique.

Now I will show how the solution of the fixed point changes with CW_{min} as follows.

Define $g(p_{s_k}, W_{B_k})$ by

$$g(p_{s_k}, W_{B_k}) = g_1(p_{s_k}) - g_2(p_{s_k}).$$

Let $p_{s_{k1}}$ and $p_{s_{k2}}$, respectively, be the solution to $g(p_{s_k}, W_{B_k}) = 0$ at $W_{B_k} = W_{B_{k1}}$ and $W_{B_k} = W_{B_{k2}} > W_{B_{k1}}$.

It is clear that $g(p_{s_k}, W_{B_k})$ is a decreasing function of W_{B_k}; hence, $g(p_{s_{k2}}, W_{B_{k1}}) > g(p_{s_{k2}}, W_{B_{k2}}) = g(p_{s_{k1}}, W_{B_{k1}}) = 0$. This, together with the fact that $g(p_{s_k}, W_{B_k})$ is a decreasing function of p_{s_k}, implies that $p_{s_{k2}} < p_{s_{k1}}$.

From (B.51b), $p_{s_{k2}} < p_{s_{k1}}$ implies $\tau_{s_{k2}} < \tau_{s_{k1}}$. This is illustrated in Fig. B.2. ■

B.8 Theorems 4.1 and 4.9

Theorems 4.1 and 4.9 are immediate corollaries of the following result, with $(M_1, M_2) = (0, 0)$ and $(4, 4)$ respectively.

Lemma B.1 *Consider the wireless model (4.2)–(4.4) with $N_u = 0$, when all data users use class B_k with $W_{B_k} = \frac{\eta_k}{\eta_{k-1}}(W_{B_{k-1}} - M_1) + M_2$ for constants $M_1 < W_{B_{k-1}}$, $M_2 \geq 0$ and $M_2 \leq M_1$, their dimensionless throughput increases in comparison with using class B_{k-1} .*

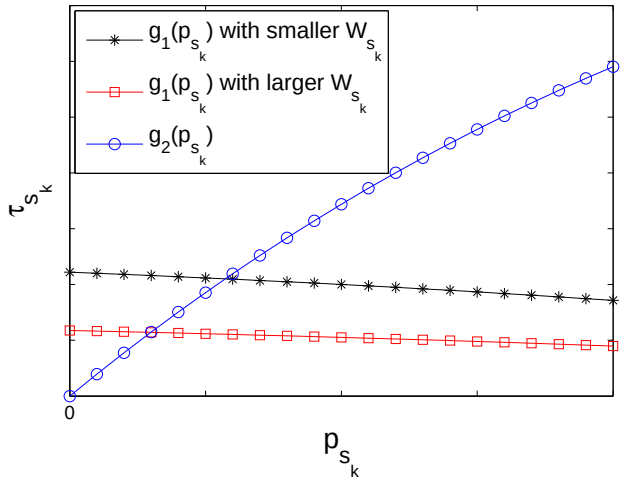


Figure B.2: Graphs of (B.51a) and (B.52) at different W_{B_k}.

Proof: Consider two networks with $N_s > 0$, identical except that one has all data users using class B_{k-1} and the other has all data users using class B_k. Quantities pertaining to the two networks will be designated by subscripts $i \in \{k-1, k\}$.

Substituting (4.2c), $N_u = 0$ by hypothesis, and the fact that all data users are of the same type into (3.10a) gives

$$\begin{aligned}\mathbb{E}[Y_i] =& \sigma(1-\tau_{s_i})(1-p_{s_i}) + N_s T^s_{s_i}\tau_{s_i}(1-p_{s_i}) \\ &+ \sum_{x\in S} T_x \tau_{s_i}\big((1-\tau_{s_i})^{N_{<x}} - (1-\tau_{s_i})^{N_s-1}\big)\end{aligned} \tag{B.53}$$

where $N_{<x}$ is the number of saturated sources with packets no larger than T_x.

Substituting (B.53) and (4.3) into (4.4) and then dividing numerator and denominator by $\tau_{s_i}(1-p_{s_i})\eta_i$ gives

$$\frac{\mathcal{T}}{S_{s_i}} = \sigma\big(\frac{1-\tau_{s_i}}{\eta_i\tau_{s_i}}\big) + T^s_{s_i}\frac{N_s}{\eta_i} + \sum_{x\in S}\frac{T_x}{\eta_i}\big(\frac{1}{(1-\tau_{s_i})^{N_s-N_{<x}-1}} - 1\big) \tag{B.54}$$

To show $S_{s_{k-1}} < S_{s_k}$, it's sufficient to show that the right hand side of (B.54) is higher for $S_{s_{k-1}}$ than for S_{s_k}. Since $\eta_k > \eta_{k-1}$, it is sufficient that both

$$\frac{1}{(1-\tau_{s_k})^{N_s-N_{<x}-1}} \le \frac{1}{(1-\tau_{s_{k-1}})^{N_s-N_{<x}-1}}, \tag{B.55a}$$

$$\sigma\Big(\frac{1-\tau_{s_k}}{\eta_k\tau_{s_k}}\Big)+T^s_{s_k}\frac{N_s}{\eta_k}<\sigma\Big(\frac{1-\tau_{s_{k-1}}}{\eta_{k-1}\tau_{s_{k-1}}}\Big)+T^s_{s_{k-1}}\frac{N_s}{\eta_{k-1}}. \tag{B.55b}$$

B.8.1 Proof of (B.55a)

Because the conditions of Lemma B.1 satisfy those of Lemma 4.2, we have

$$\tau_{s_{k-1}}>\tau_{s_k} \qquad\qquad p_{s_{k-1}}>p_{s_k}. \tag{B.56}$$

Since $\tau_{s_{k-1}}>\tau_{s_k}$ by (B.56), the fact that $(1-\tau_{s_i})^{N_s-N_{<x}-1}$ is non-increasing with the increase of τ_{s_i} establishes (B.55a).

B.8.2 Proof of (B.55b)

Showing (B.55b) is equivalent to showing the right hand side of (B.55b) subtracted by the left hand side is greater than 0.

From (B.1),

$$\begin{aligned}&\Big(\sigma\Big(\frac{1-\tau_{s_{k-1}}}{\eta_{k-1}\tau_{s_{k-1}}}\Big)+T^s_{s_{k-1}}\frac{N_s}{\eta_{k-1}}\Big)-\Big(\sigma\Big(\frac{1-\tau_{s_k}}{\eta_k\tau_{s_k}}\Big)+T^s_{s_k}\frac{N_s}{\eta_k}\Big)\\&=\sigma\Big(\frac{1}{\eta_{k-1}\tau_{s_{k-1}}}-\frac{1}{\eta_k\tau_{s_k}}\Big)+\Big(\frac{1}{\eta_{k-1}}-\frac{1}{\eta_k}\Big)(EN_s-\sigma).\end{aligned} \tag{B.57}$$

Since $E>\sigma$ by (B.1), to show that (B.57) is greater than 0, it suffices to show

$$\eta_{k-1}\tau_{s_{k-1}}<\eta_k\tau_{s_k} \tag{B.58}$$

as below.

Multiplying $\tau_{s_{k-1}}$ and τ_{s_k} from (4.2a) by η_{k-1} and η_k, respectively, gives

$$\frac{2}{\tau_{s_{k-1}}\eta_{k-1}}=\frac{W_{B_{k-1}}}{\eta_{k-1}}\frac{1-p_{s_{k-1}}}{1-2p_{s_{k-1}}}+\frac{1}{\eta_{k-1}} \tag{B.59}$$

$$\frac{2}{\tau_{s_k}\eta_k}=(\frac{W_{B_{k-1}}-M_1}{\eta_{k-1}}+\frac{M_2}{\eta_k})\frac{1-p_{s_k}}{1-2p_{s_k}}+\frac{1}{\eta_k}. \tag{B.60}$$

Applying $\eta_k > \eta_{k-1}$ and $M_2 \leq M_1$ to (B.59) and (B.60),

$$\frac{1}{\eta_k} < \frac{1}{\eta_{k-1}}, \tag{B.61}$$

$$\frac{W_{B_{k-1}} - M_1}{\eta_{k-1}} + \frac{M_2}{\eta_k} < \frac{W_{B_{k-1}}}{\eta_{k-1}}. \tag{B.62}$$

By (B.56),

$$\frac{1 - p_{s_{k-1}}}{1 - 2p_{s_{k-1}}} < \frac{1 - p_{s_k}}{1 - 2p_{s_k}}. \tag{B.63}$$

Substituting (B.61) and (B.63) into (B.59) and (B.60) implies (B.58). ■

B.9 Lemma 4.12

Proof: Consider action profiles $a_{(B_k;\cdot;B_{k+i};\cdot)}$ ($k < n$, $i \geq 0$ and $k + i \leq n$) and $a_{(B_n;\cdot;B_{k+i};\cdot)}$.

To show (4.9), I first show that

$$C_j(a_{(B_n;\cdot;B_{k+i};\cdot)}) > C_j(a_{(B_k;\cdot;B_{k+i};\cdot)}) \tag{B.64}$$

as follows.

When a_1 changes from using class B_k to B_n, we have the following from Lemma 4.7 due to $W_{B_k} < W_{B_n}$

$$\tau_j(a_{(B_k;\cdot;B_{k+i};\cdot)}) < \tau_j(a_{(B_n;\cdot;B_{k+i};\cdot)}). \tag{B.65}$$

We know from (B.36) that p_i is decreasing in τ_i, and so by (B.65) we have

$$p_j(a_{(B_k;\cdot;B_{k+i};\cdot)}) > p_j(a_{(B_n;\cdot;B_{k+i};\cdot)}). \tag{B.66}$$

From (4.3), the successful transmission rates per slot of the data user j under

the action profile $a_{(B_h;\cdot;B_{k+i};\cdot)}$ $(h \leq n)$ is

$$C_j(a_{(B_h;\cdot;B_{k+i};\cdot)}) = \eta_{k+i}\tau_j(a_{(B_h;\cdot;B_{k+i};\cdot)})\big(1 - p_j(a_{(B_h;\cdot;B_{k+i};\cdot)})\big)\mathcal{T} \tag{B.67}$$

Substituting (B.65) and (B.66) into $C_j(a_{(B_k;\cdot;B_{k+i};\cdot)})$ from (B.67) and $C_j(a_{(B_n;\cdot;B_{k+i};\cdot)})$ from (B.67) gives (B.64).

Then, applying Theorem 4.8 in the action profile $a_{(B_k;\cdot;B_{k+i};\cdot)}$ and $a_{(B_n;\cdot;B_{k+i};\cdot)}$ gives

$$C_1(a_{(B_k;\cdot;B_{k+i};\cdot)}) \leq C_j(a_{(B_k;\cdot;B_{k+i};\cdot)}) \tag{B.68}$$

$$C_1(a_{(B_n;\cdot;B_{k+i};\cdot)}) \geq C_j(a_{(B_n;\cdot;B_{k+i};\cdot)}). \tag{B.69}$$

From (B.64), (B.68), and (B.69), we have (4.9). ■

B.10 Theorem 4.11

Proof: Note that the conditions of this theorem satisfy those of Lemma 4.12.

Consider an action profile with at least one data user using a class other than B_n. Choose the data user using the lowest class among all users under this action profile. Then, according to Lemma 4.12, that user has incentive to change its action to using class B_n to improve its throughput. Therefore, it can be concluded that no action profile in which at least one data user using lower class than B_n is a Nash equilibrium. ■