I-WLAN: Intelligent Wireless Local Area Networking

by

Mustafa Ergen

B.S. (Middle East Technical University, Ankara) 2000

M.S. (University of California, Berkeley) 2002

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering-Electrical Engineering

and Computer Sciences

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Pravin Varaiya, Chair

Professor Jean Walrand

Professor John Chuang

Professor Ahmad Bahai

Fall 2004

The dissertation of Mustafa Ergen is approved:

Professor Pravin Varaiya, Chair Date

Professor Jean Walrand Date

Professor John Chuang Date

Professor Ahmad Bahai Date

University of California, Berkeley

Fall 2004

I-WLAN: Intelligent Wireless Local Area Networking

Abstract

I-WLAN: Intelligent Wireless Local Area Networking

by

Mustafa Ergen

Doctor of Philosophy in Engineering-Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Pravin Varaiya, Chair

Future wireless networks require intelligent components that can automate, scale and manage the network in order to handle the demand for ubiquitous access. In this dissertation, we first evaluate existing problems, and then we introduce components that improve the performance of existing systems.

We introduce a new Markov model for Distributed Coordination of Function of IEEE 802.11 with which we formulate the throughput and delay for saturated and non-saturated traffic. We introduce a novel formulation for individual throughput when stations operate with mixed data rates. We introduce an admission control mechanism to maintain the highest achievable throughput by controlling the access. After that we use our throughput formulation to obtain the performance of an indoor network. We introduce a packet size adjustment scheme with respect to data rate so that a slow station sends small packets to prevent throttling of the network. We introduce a frame aggregation scheme for wireless

voice over IP. We introduce a fast and fair sub-optimal algorithm to allocate sub-carriers and bits adaptively in an OFDMA system for point-to-multipoint communication, and investigate MAC performance of the wireless LAN with adaptive antennas for point-to-point communication.

We find that ignoring the consecutive transmission probability in the previous Markov models is incomplete. Consequently, our model which takes this into account is closer to the standard. We find that the individual throughput of a station is the same for fast and slow stations, and slow stations throttle the performance, since the CSMA/CA scheme gives equal access but not equal time of channel usage. Our packet size optimization scheme increases the throughput of the total network and the fast station but not that of the slow station. The admission control mechanism can tune the network from random access to controlled access. Network management can monitor the whole network and do real-time adjustments for optimum performance. Our frame aggregation scheme for wireless voice over IP reduces the number of access by concatenating the packets in the accesses point in order to be sent at one time. The results of our resource allocation scheme for adaptive sub-carrier and bit allocation is appealing and can be close to optimal. We find that a directional antenna increases the performance significantly and can provide enhancements for wireless LANs.

Prof. Pravin Varaiya
Dissertation Committee Chair

To my wife Sinem and my family

Contents

List of Figures

List of Tables

Preface

Wireless networks are among society's most important infrastructure technologies and are vital to the operation of many systems. The extensive deployment of wireless LAN systems has brought wireless networks to the forefront of human communication. Wireless network technology has seen a rapid advancement over last decade; however, existing network architectures are reaching capacity limits. Substantial modifications and innovations are required to fulfill future demands.

The proliferation of wireless users necessitates the extension of coverage and requires "any-time", "any-where" connectivity. The wireless network architecture of tomorrow must be endowed with intelligence to provide the best performance, and be able to scale and support evolving capability requirements.

It is hoped that this work will provide a valuable summary of the wireless local area networks and introduce possible new components for the next generation wireless networks.

Chapter 1 is an introduction to the subject, including a historical perspective. The core of the chapter provides a description of the areas where wireless networking must change to support future demands. The chapter concludes with a summary of contribu-

tions of this dissertation.

Chapter 2 is devoted to a review of Wireless Local Area Networking. It gives an extensive analysis of MAC and PHY layers of IEEE 802.11 standard. Finite state machine representation is presented for access functions and the OFDM PHY layer is analyzed in detail. A reader familiar with the subject may wish to skip this chapter.

Chapter 3 describes the access mechanism for Distributed Coordination Function (DCF). Emphasis is placed on a Markov model analysis of the Carrier Sense Multiple Access/ Collision Avoidance (CSMA/CA) scheme and a comparison is made with the existing model, [5].

Chapter 4 considers the throughput formulation of a network when stations access the channel with different data rates.

Chapter 5 introduces delay analysis of the DCF mechanism. This chapter investigates the statistics of the time interval between two successful transmission per station.

Chapter 6 is focused on an admission control mechanism which could introduce adjustable control over DCF.

Chapter 7 provides an algorithm to characterize and manage a network with more than one access point. The emphasis is to investigate throughput with interference of the access points.

Chapter 8 introduces a cross layer optimization for the degradation of mixed data rates introduced in Chapter 4. It describes a fair scheduling scheme based on the length of channel usage.

Chapter 9 provides a description of Wireless Voice over IP and introduces a frame aggregation algorithm to decrease the delay between inter-arrival times. The chapter first

focuses on the components of the VoIP system and then describes the algorithm fortified with the performance results.

Chapter 10 focuses on a sub-optimal algorithm for an OFDMA system. It describes an adaptive algorithm for sub-carrier allocation and bit loading. This chapter can be read independently of the other chapters.

Chapter 11 focuses on adaptive antenna systems. The performance of an adaptive antenna in an access point network and in an ad-hoc network is given with throughput analysis.

Chapter 12 concludes the work.

Acknowledgements

I have had the opportunity to work with and learn from, my advisor, Prof. Pravin Varaiya whom I wish to express my sincere gratitude. He did his best to ensure me the perfect conditions, encouraged my ambitions and promoted my work. I am particularly indebted to Prof. Suha Sevuk and Prof. Kemal Inan for their help in this great opportunity of working with Prof. Varaiya. I also thank Prof. Jean Walrand, Prof. Ahmad Bahai, and Dr. Burton Saltzberg. I would like to acknowledge Farhana Sheikh for editing. I would like to salute my wife Sinem for her constant care and support. Last, but by no means least, I am thankful to my family for their support and patience.

Chapter 1

Introduction

Information technology has evolved remarkably over the past years. After the advent of silicon technology, technological and economical forces guided advances in the computer industry that aroused the vigorous need for networking technology. This technology is currently forcing the economy and society to adapt to the big invention of this century:the Internet. Internet, the largest of all data networks, is now affecting virtually every aspect of the modern society, just as electricity once did. The world has absorbed this innovation rapidly and has quickly demonstrated that legacy networks are inadequate to meet the vigorous expectations of a modern lifestyle. The next wave in information technology stems from these continually increasing expectations: ubiquity, autonomicity, evolvability and scalability.

Ubiquitous access is the ability to support seamless access to information and services by anyone, any time, at high speed. Autonomicity is the ability of the network to autonomously manage, protect and heal itself with minimal human intervention. Evolvability refers to the ability of the network to evolve over time to incorporate new tech-

nologies and meet new requirements while supporting new applications. Scalability is the ability of the network to accommodate unforeseen growth across many dimensions without significant performance degradation.

Wi-Fi is used for Wireless Local Area Networking (WLAN), which brings a mechanism to connect wireless users to backbone in high speed. The best-known WLAN standard is IEEE 802.11, which has several supplementary standards. The legacy IEEE 802.11 was introduced in 1997 with carrier sense multiple access/collision avoidance (CSMA/CA) MAC protocol and three different physical layer mechanisms: direct sequence spread spectrum (DSSS), frequency hopping spread spectrum (FHSS), and infrared (IR). Since then, the standard has been enhanced with two physical layer standards: IEEE 802.11b and IEEE 802.11a. IEEE 802.11b uses High Rate DSSS (HR/DSSS) and IEEE 802.11a uses orthogonal frequency division multiplexing (OFDM). The IEEE 802.11e MAC protocol is expected to be ratified for providing quality of access in 2006.

Over the last few years, Wi-Fi is usually deployed as the last hop of the Internet or wireline telephone network, thereby working in conjunction with the wireline networks. The biggest advantages that Wi-Fi provides are mobility and coverage. Together, these characteristics provide “any-time”, “any-where” connectivity. Meanwhile, a new standard for broadband wireless connectivity, known as WiMAX, is emerging. WiMAX (which stands for worldwide interoperability for microwave access), a non-line-of-sight, point-to-multipoint broadband wireless access (BWA) technology, is emerging as a potential competitor to wireline DSL and cable for last-mile access and a strong backhaul option for Wireless-Fidelity (Wi-Fi). Wi-Fi (up to 100 meters) and WiMAX (up to 50 kilometers) are expected to provide new broadband wireless connectivity. Wi-Fi provides service for

client-to-access point (AP) communications. WiMAX is for implementations of AP-to-AP and AP-to-service providers that are typically needed for wireless last-mile [1, 2].

Recent developments in this area have given WiMAX-based BWA a new hope for growth in client-to-service providers with IEEE 802.16e standard, which is based on Orthogonal Frequency Division Multiple Access (OFDMA) system, expected to be ratified in early 2005. Wi-Fi is moving into the last-mile access with WLAN with directional antennas, or WLAN with a mesh-network topology [2, 3]. Next-generation wireless systems are expected to be a mix of many standards with intelligent network management in order to overcome the limitations of existing networks, and add new capabilities and services. In this dissertation our goals are to sustain the science and technology needed to fulfill our vision for next-generation networks, as well as address the limitations of existing networks. We present a convergent network architecture, which addresses many of the problems underlying the existing technology.

1.1 Motivation

The wireless medium is a complex system that comes with many challenges. These challenges arise due to the fact that the wireless medium is a limited resource in terms of spectrum, and bandwidth is therefore regulated. This limits the performance of the system because the speed and the bandwidth are strongly coupled. Radio propagation is another challenge since now the waves traverse through walls, moving objects, and so on, creating multi-path interference and deep fades in the signal. These propagation issues result in loss of data. The lack of physical connectivity between the stations causes

unpredicted transmissions and results in collision of packets. Furthermore, non-uniform connection of various nodes requires management to mitigate overall performance degradation. As a result, intelligent protocols are required to cope with the growing need of networking.

Distributed Coordination Function (DCF), which implements CSMA/CA, is one of the widely used access mechanism of IEEE 802.11. The increase in wireless users results in a high demand for a wireless network that can provide high speed and seamless connectivity. This increases the importance of CSMA/CA, since random access scheme is a good choice for ad-hoc schemes, which will be present in the next-generation wireless networks. Access point networks will connect wireless to other networks, or a station could function as an access point from time to time, resulting in an ad-hoc configuration of centralized networks. It is essential to formulate the DCF thoroughly and characterize the performance.

The legacy systems use each communication layer independently, which means each layer is pursuing its own task independently. They interact only with the upper or lower layer, transporting the data from one end to the other. Although it provides a clean interface between layers, a better way of structuring is needed to feed the desire for high-speed networking. The lack of interaction between layers results in an uncoordinated architecture, ultimately affecting the performance. Cross layer optimization, or the use of information from a layer by another layer, can be applied to legacy systems in order to provide optimal performance. For example, the physical layer is capable of giving rich information to upper layers; in the same way, upper layers can transmit requests to lower layers to adjust their parameter set.

It is possible to use a WLAN system for voice communication. A significant benefit of carrying voice communications is to eliminate the need for two separate entities for voice and data in an infrastructure. Voice over WLAN provides mobility and inexpensive calls with less infrastructure. Current WLAN systems cannot accommodate large number of voice sessions; 802.11b is capable of running only three uncompressed audio streams smoothly. On the other hand, IEEE 802.11a can accommodate four times as much voice traffic as 802.11b. Still, a connected network interference could bring the performance down. IEEE 802.11e provides QoS upgrade, which could introduce some priority to voice packets, but a direct treatment could be introduced within the MAC layer by frame aggregation.

Proliferation of home and business wireless LANs is driving the demand for broadband connectivity back to the Internet, which IEEE 802.16 can fulfill by providing the connection back to the service provider. OFDM is one of the most popular multi-carrier modulation techniques. OFDM technology divides the operating bandwidth into small tones and converts the single bit stream into multiple non-overlapping parallel bit streams. The roots of OFDM date back to the 1960s. However, the popularity of OFDM increased in the 1990s with the advent in the high-speed communication. OFDM has become the hottest choice as a transport technology for the most popular standards of today, and is highly likely to maintain its ideal candidacy for the next-generation high-speed applications. Presently, in addition to WLAN, OFDM is also a strong candidate for Broadband Wireless Access (BWA) systems. OFDM modulation can be further exploited to deploy QoS for real-time data. Some of the carriers can be reserved for the QoS requesting station, in which QoS information is served from the MAC layer. In each packet, a set of

subcarriers may address only one station. This reduces the delay due to random based scheme and provides a packet delivery with a substantially reduced packet interarrival time.

Deploying multiple antennas separated in space is a new technique that leverages the antenna diversity in the physical layer. This technique is called adaptive antenna or MIMO. A high power directed beam can be formed towards the destination in order to improve the signal-to-noise ratio. The beam can be formed by adjusting the power levels in each antenna according to the reception level. Current studies in MIMO are at the PHY layer. With state-of-the-art simulators (OPNET) [4], throughput performance can be measured in the MAC layer. Adaptive antenna systems work in conjunction with the MAC layer because control signals are sent in isotropic antennas and data signals are sent in directional beams. As stated above, bandwidth is a limited resource, thereby limiting the number of non-overlapping channels. In multiple access point networks, there is considerable interference because some access points are hampered by using the same channel. Adaptive antennas are expected to increase the performance by reducing the interference and thus increasing the signal-to-noise.

1.2 General Framework and Contributions

The view taken throughout this dissertation is to first formulate the problem then support it with experimental results. The contributions of this work are as follows;

- $\surd$ CSMA/CA is a random multiple access scheme that transmits if the channel is sensed idle; otherwise, it backs off. IEEE implements CSMA/CA in distributed co-

ordination function (DCF). The theoretical characterization of these systems is an open topic and needs to be formulated. The recent work [5] in throughput analysis of DCF function is lack of proper characterization of the procedure and only considers the system in saturation. The model also fails to incorporate signal-to-noise ratio (SNR), wireless channel, interfering stations and multiple access points; and considers it only in one network. We introduce a new Markov model that incorporates data rate into the formula. The results are also fortified with the simulation. An important issue that has caused a lot of confusion in the literature concerning analytical Markov Model introduced by [5] is defining the virtual time slots. We treat this by explaining the observation of time in the process of event definition.

√ Throughput can be very difficult to observe if the stations connect at different data rates. We introduce a closed formula that provides the throughput of each individual station.

√ We then find the delay characteristic of the network. Our approach is unique in terms of finding the probable mass function of time intervals between successful transmissions of a station. These results are essential for predicting the performance of the system for some QoS implementation.

√ We introduce controllability on top of the random access protocol, CSMA/CA. It gives a base to tune the network from DCF (random control) to PCF (full control) depending on the network condition. Admission control mechanism controls the number of stations in a network in order to keep the performance at maximum.

- √ We develop a mechanism to derive a throughput formulation for an extended service set that has more than one access point. This is essential in order to implement a high quality of service to the entire network. With this scheme, a network administrator can monitor network performance offline and take therapeutic actions to improve the performance.

- √ We find that in a network where the nodes have different data rates, the throughput is confined to the data rate of the slow station. To alleviate the low performance of high data rate stations, we introduced a cross layer optimization mechanism that implements an adaptive scheme to adjust the packet size according to the data rate. With our scheme, stations reserve channels equal amount of times.

- √ We investigate the voice over IP system and implement an algorithm to improve the performance in WLAN. Our frame aggregation algorithm brings a superior performance for voice communication and increases the number of voice sessions in a WLAN.

- √ In adaptive bit loading systems, in order to provide real-time quality of service (QoS), subcarriers can be assigned dynamically to users in different modulations. We propose a scheme to allocate subcarriers within the coherence time of the channel for Broadband Wireless Access Systems. This scheme could be used with configurable radio in a station that supports both WLAN and WiMAX.

- √ In adaptive antenna systems, a directed beam can be aimed towards the destination at high power. The beam can be formed by adjusting the power levels in each

antenna according to reception level. Current studies in adaptive antenna area at the PHY layer and consider BER vs. SNR in performance. We implemented an adaptive antenna module in an OPNET network simulator. We see that in multiple access point networks there is a considerable interference due to dearth of enough non-overlapping channels. Adaptive antenna systems reduce this interference but increase the hidden and exposed terminal problem when used in an ad-hoc network. The best performance occurs when these two forces are balanced.

Chapter 2

Wireless LANs

2.1 Introduction

Wireless Local Area Networking (WLAN) has become a primary solution for mobile users in terms of high speed data connectivity and cheap voice communication. WLAN offers two kind of services. It replaces the last hop with a wireless link and connects the devices to the IP core; or makes an ad-hoc network among wireless stations away from infrastructure. WLAN offers mobility and flexibility in addition to the same features and benefits of traditional wired LANs. Much interest has been directed to its efficiency in terms of low cost and flexible installation.

WLAN offers a medium access control (MAC) layer and a physical layer (PHY). The MAC layer is responsible for providing a multiple access scheme to allow more than one user at a time and the PHY layer is responsible for transmitting and receiving packets over the air.

There are parallel converging line of activities in WLAN which are conducted by

the Institute of Electrical and Electronics Engineers (IEEE) under the standard 802.11 and European Telecommunications Standards Institute (ETSI) under project Broadband Radio Access Networks (BRAN). IEEE standardizes 802.11 as the WLAN standard and HiPERLAN is one of the BRAN standards which is designed to provide high-speed access to networks. They differ in terms of how they define their MAC and PHY layers [6, 7].

The IEEE 802.11 protocol specifies both Medium Access Control (MAC) and Physical (PHY) layers. The 1997 standard [8] was updated in 1999 with two new physical layers, IEEE 802.11b [9] and IEEE 802.11a [10]. The standard includes a contention based and a polling based medium access protocol, called Distributed Coordination Function (DCF) and Point Coordination Function (PCF), respectively. There are several sub-standards of IEEE [11, 12]. The IEEE 802.11e standard is expected to be ratified in 2006; it will provide quality of service by implementing different access to distinguished packets [13].

The ETSI HiPERLAN standard has two sub-standards. HiPERLAN/1 is designed for ad-hoc networks and operates in the 5.1-5.3 GHz bandwidth. The MAC protocol is based on a variant of carrier sense multiple access/ collision avoidance (CSMA/CA) scheme. HiPERLAN/1 does not guarantee QoS and considers only best effort service. This is what motivated ETSI to develop HiPERLAN/2. HiPERLAN/2 focuses on managed infrastructure and wireless distribution system. HiPERLAN/2 is a time division multiple access/time division duplexing (TDMA/TDD) protocol. As a result, HiPERLAN/2 schedules the access in a deterministic manner where a time slot is assigned to each station to transmit packets [14].

2.2 WLAN Topology

WLAN has two types of network elements: stations (STAs) and access points (APs). A STA is defined as the device that is equipped with a component that can communicate via a WLAN standard. A STA is called an AP if it is connected to a wired network and offers infrastructure service to mobile STAs. Networks composed of a combination of these actors are categorized as Independent Basic Service Set (IBSS) or Infrastructure Basic Service Set (Infrastructure BSS) [15].

2.2.1 Independent Basic Service Set

IBSS type of network is also referred to as an ad-hoc network. An ad-hoc network is a network where there is no infrastructure. In such a network, stations communicate directly to each other and they spontaneously establish a distributed communication mechanism. As shown in Figure 2.1, the IBSS network is not connected with the wired domain. IBSS is initiated if a station is in the vicinity of another. One of the STA selects itself as the starter and sends beacon signals indicating the service id of the IBSS (SSID).

The routing algorithm in IBSS is handled by ad-hoc routing protocols which reside in the network layer (See Figure 2.2). In ad-hoc networks, there is no fixed routers and each station is capable of functioning as a router which can discover and maintain routes to other nodes in the network.

Ad-hoc routing protocols are classified as table-driven and source-initiated on-demand. A table driven routing protocol constructs routes for each pair of stations and maintains up-to-date routing information from each node to every other node in the network regard-

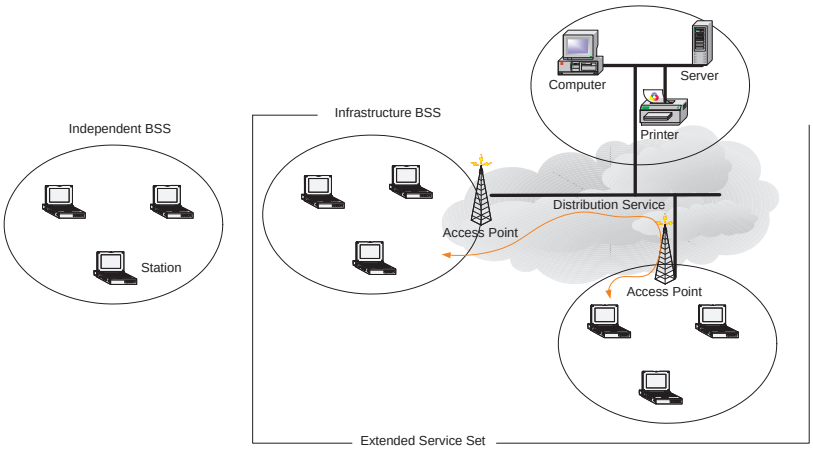


Figure 2.1: IEEE 802.11 networks

less of when and how frequently such routes are desired. The existing and famous table-driven protocols are Destination-Sequenced Distance-Vector Routing (DSDV), Cluster-head Gateway Switch Routing (CGSR), Wireless Routing Protocol (WRP). The differences arise in the number of necessary routing related tables and the methods for notification of changes [16, 17, 3].

Source-initiated on-demand routing, on the other hand, creates routes only when it is necessary. When a station has a packet to send, it initiates the route discovery protocol to establish a route to the destination. Once the route is constructed, it is maintained as long as it is no longer desired or when the destination becomes inaccessible. Ad-Hoc On-Demand Distance Vector Routing (AODV), Distance Source Routing (DSR), Temporally Ordered Routing Algorithm (TORA), Associativity Based Routing (ABR), Signal Stability Routing (SSR) are some on-demand routing protocols [16, 18, 19, 20].

Table-driven ad-hoc routing relies on the underlying routing table update mechanism since a route is always available regardless of whether or not it is needed. The limitation

occurs in route creation since it incurs significant control traffic and power consumption. In on-demand routing, on the other hand, this control traffic is minimized by initiating route discovery when needed, but now the station has to wait until a such a route can be discovered.

2.2.2 Infrastructure Basic Service Set

In contrast to ad-hoc networks, Infrastructure BSS contains a gateway to the wired domain and enables communication between a station of a BSS and a station of other BSSs and other LANs as seen in Figure 2.1. The gateway is called access point (AP) in the IEEE 802.11 standard. APs communicate with each other via either cable or radio to interconnect BSSs. The IEEE 802.11 standard specifies a distribution system (DS) to enable roaming between interconnected BSSs and to create connections to wired network resources. The interconnected BSSs and DS allow an extension to the IEEE 802.11 network which can be as large as desired. This is called extended service set (ESS) in IEEE 802.11. There handover support in the MAC layer in order to provide seamless connectivity within the ESS. Each ESS is identified by its ESS id (ESSID). The DS resides above MAC layer, which makes it homogeneous and can be implemented by any other backbone network such as ethernet, token ring, optical or wireless network. The ESS communication in the backbone network is in layer 2.

In WLAN, stations maintain their connections during their movement. If station crosses one BSS to another in the same ESS, stations make a decision and change its AP. During this change, the station associates itself with a new AP and the information is

sent to DS which notifies the old AP about the new location.

If a station crosses into another ESS, then this movement is not supported by WLAN but handled in the network layer. Mobile IP is one of the solutions for this type of movement. Mobile IP introduces two IP addresses: permanent and temporary. Basic Mobile IP is designed based on IPv4. Whenever a station changes its ESS it gets a new temporary IP address and notifies its home agent (HA). The HA is the agent located in the home network where the station's permanent IP address belongs. The HA records the temporary IP address of its stations. Whenever a message is sent by a correspondent host (CH) (any host in the Internet) to a mobile host (MH) the packet is sent to the network where the permanent address belongs. The HA is responsible for capturing the packets of the STA if away, and for sending them in encapsulation to the foreign agent (FA) where temporary IP address belongs. The FA decapsulates the packet and sends it to MH using layer 2. IPv6 brings improvement to Mobile IP with its large address space and eliminates the need to send the packets to HA. Once associated, the packets can be sent directly to MH from CH [21, 22].

2.3 Architecture

The significant element of the WLAN is its medium access control (MAC) protocol. The MAC protocol controls the transmission and lies above the physical layer. It is compliant with any physical layer introduced. The logical architecture is shown in Figure 2.2. The underlying mechanism in the MAC layer is carrier sense multiple access (CSMA) scheme like ethernet. It is hard to detect collisions right away in wireless medium due

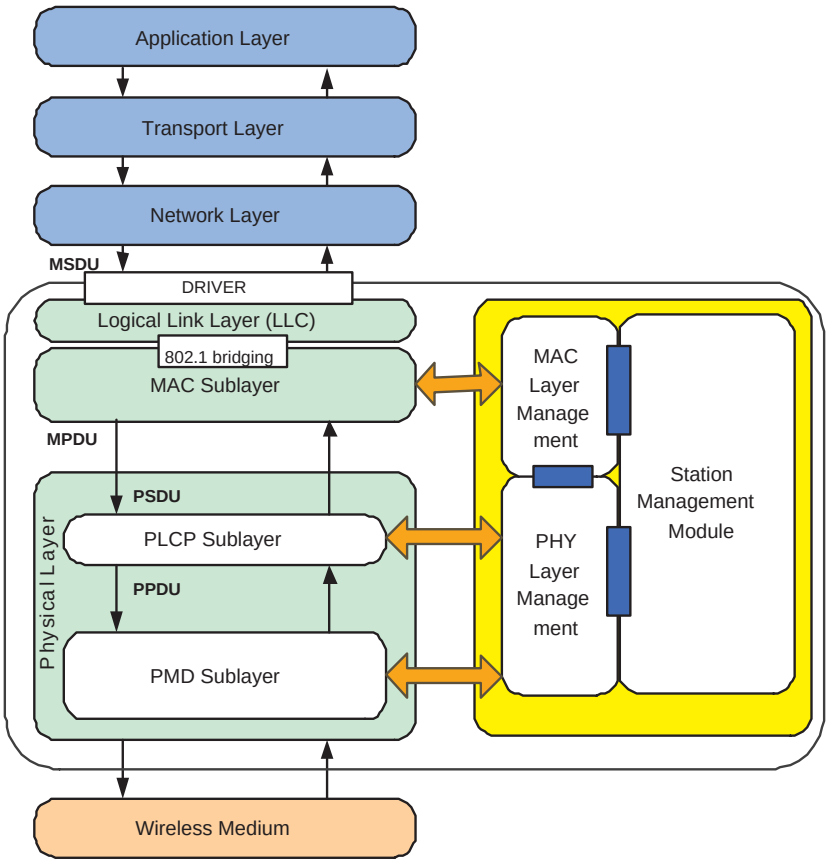


Figure 2.2: IEEE 802.11 architecture © IEEE

to wireless propagation issues. As a result, IEEE 802.11 implements collision avoidance (CSMA/CA) unlike ethernet where collision detection is implemented (CSMA/CD).

The physical layer, on the other hand, handles the transmission from one station to the other in a reliable way. It is responsible for converting data into a waveform in the transmitter, and the waveform into the data in the receiver.

A wireless medium introduces noise, multipath fading, and interference. Any transmission that is successfully terminated can be learned from the acknowledgement frame. In the basic access mechanism, all transmitted frames are acknowledged to reduce the inherent error rate of the medium. at the expense of additional bandwidth consumption. If

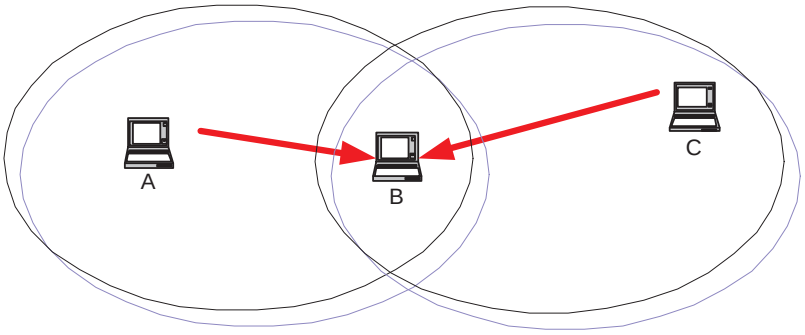


Figure 2.3: The hidden node problem

the ACK frame is not received within a ACK time-out period, retransmission takes place in the transmitter. Dealing with this issue in the MAC layer is much more efficient since higher layer timeouts are larger and often measured in seconds.

The Hidden Node Problem

The hidden node problem that is unlikely to occur in a wired LAN is another challenge for WLANs. If two stations (A,C) are unreachable and if there is a station (B) in the middle of those two that is reachable from both, transmission from A to B can be interrupted by the transmission from C to B as illustrated in Figure 2.3. IEEE 802.11 introduces two additional frames to the basic access mechanism. The source sends Request to Send (RTS) and the destination replies with Clear to Send (CTS) packets to reserve the channel in advance. Stations that hear RTS delay their transmission until the CTS frame. The stations that hear CTS suspend transmission until they hear acknowledgement. If the stations that hear RTS do not hear CTS, they continue as if they did not hear RTS. Figure 2.4 shows the RTS/CTS frame exchange sequence.

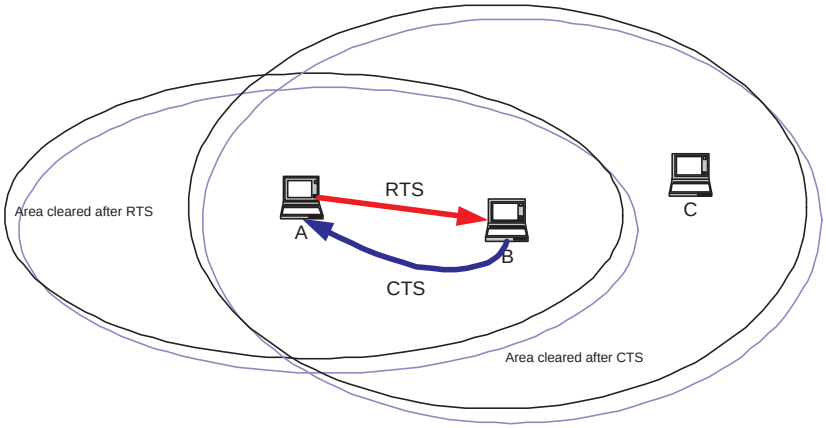


Figure 2.4: RTS and CTS solution

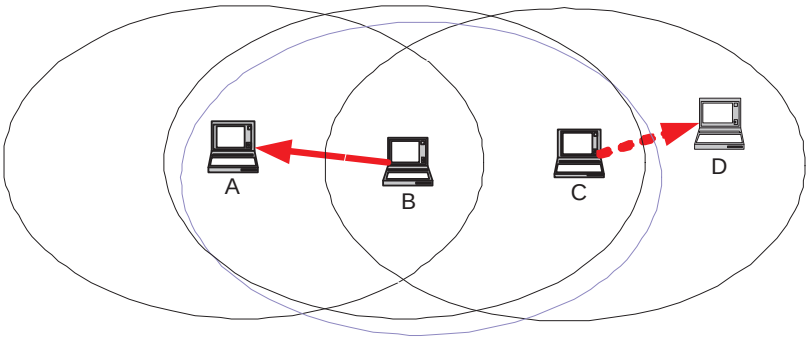


Figure 2.5: The exposed terminal problem

The RTS/CTS transmission introduces a fair amount of capacity consumption. This makes the RTS/CTS usage decidable. RTS/CTS mechanism can be disabled by an attribute (*dot11RTSThreshold*) in the management information base (MIB).

RTS/CTS can be disabled when there is low demand for bandwidth with correspondingly less contention, and all stations are able to hear the transmission of every station. When there is an AP, the RTS/CTS is unnecessary in downlink.

The Exposed Terminal Problem

Depicted in Figure 2.5, the exposed terminal problem arises when station (C) attempts to transmit to station (D) while station (B) transmits to station (A). In this case, station (C) is unnecessarily delayed. If there is a RTS/CTS exchange then the CTS frame sent by station (A) will not propagate to station (C). Thus, station (C) knows that the transmission from station (B) to station (A) might not interfere with its transmission to station (D). As a result, station (C) will initiate transmission to station (D).

2.4 Medium Access Control Layer

IEEE 802.11 introduces coordination functions. Coordination functions control the access of a station to the medium. IEEE 802.11 has two different coordination functions based on contention or polling. The distributed coordination function (DCF) based on contention is the fundamental access method and is also known as CSMA/CA. DCF can be implemented in any network and basically a STA shall ensure that the medium is idle before any transmission attempt. The point coordination function (PCF) on the other hand, provides a contention free (CFP) period. The point coordinator (PC) located in an AP controls the frame transmissions of the STAs by polling so as to eliminate contention. The PCF alternates between a contention free period (CFP) and a contention period (CP) in time. The CP is placed to give transmission rights to late coming STAs or to DCF-based STAs. PCF is only implemented in infrastructure BSS.

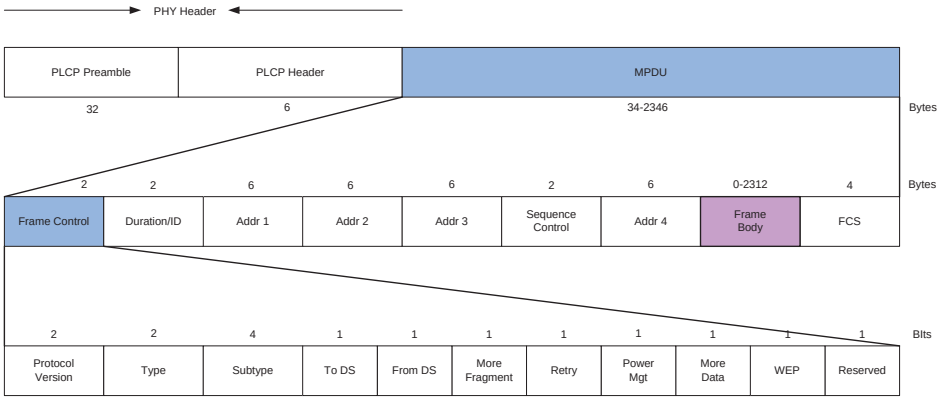


Figure 2.6: MAC frame

Frames

MAC supports three different types of frames:management, control, and data. Figure 2.6 illustrates the format of a typical MAC frame. Data frames are used to transfer data and control frames exist to make sure that the delivery is reliable. Management frames perform operations of joining or leaving a network and manage access points in handoff. This feature includes three steps: locating of an infrastructure; authenticating to the AP to make the wireless channel reliable; associating to access services of the infrastructure.

Fragmentation

The packets are partitioned into small pieces in order to reduce the interference of non-STA objects such as bluetooth devices and microwave ovens which operates also in 2.4-GHz ISM band. As a result, the interference only affects small portions of the packet and can be recovered easily in short amount of time.

Fragmentation is only applied to packets that have unicast receiver addresses when the packet length exceeds the fragmentation threshold. Broadcast and multicast frames

are not allowed to be fragmented. The fragments are sent in burst unless interrupted in wireless medium and acknowledged individually.

IFS Timing

Time intervals for control are called inter frame spaces (IFS). Starting from the shortest to the longest, the list is as follows:

SIFS Short interframe space is the smallest time required to give priority to the completion of a frame exchange sequence, since other STAs wait longer to seize the medium. SIFS[1] time is determined by the delay introduced in PHY and MAC layer.

Slot Time Time is quantized in slots. Slot[2] time is specific to PHY layers. In IEEE 802.11a, Slot time is shorter than SIFS. The backoff counter is decremented after sensing the channel idle for a slot time.

PIFS STAs operating in PCF seize the medium at least PCF interframe space (PIFS)[3] time in contention free period (CFP) to gain access to the medium.

DIFS A STA using the DCF function is allowed to transmit if its carrier sense mechanism determines that the medium is idle at least one DCF interframe space (DIFS)[4] time. This also allows STAs that are in the RTS/CTS exchange to seize the medium immediately after SIFS.

[1]SIFS=Rx RF Delay + Rx PLCP Delay + MAC Processing Delay + Rx Tx Turnaround Time
[2]Slot Time=Channel Clear Assessment (CCA) Time + Air Propagation Time + SIFS
[3]PIFS=SIFS+Slot time
[4]DIFS=SIFS+2xSlot time

EIFS The Extended interframe space (EIFS) period starts when PHY detects idle after an erroneous frame reception. EIFS gives enough time for the transmitter to acknowledge. EIFS[5] is derived from the SIFS, DIFS and the length of time it takes to transmit an ACK control frame.

Backoff Time A STA having packet to send initiates a transmission. STA invokes the carrier sense mechanism in DCF to determine the state of the medium. If the medium is busy, the STA defers until the medium is idle for at least one DIFS period when the last frame detected on the medium is received correctly; or at least one EIFS when the last frame detected on the medium is erroneous. After DIFS or EIFS, the STA backs off for a randomly chosen integer period[6] over the interval [0, CW]. CW is restricted to lie between CW_{min} and CW_{max} which are specified by the PHY layer. Backoff level i is incremented up to CW_{max} or reset if the transmission is unsuccessful or successful, respectively. The station attempts to retransmit the unsuccessful frame as long as its retry count does not reach retry count limits (*dot11ShortRetryLimit* or *dot11LongRetryLimit*).

2.5 IEEE 802.11 Distributed Coordination Function

It is essential to understand CSMA/CA in detail since this random based protocol is implemented in other standards and analytical models. The CSMA/CA scheme is composed of carrier sensing and collision avoidance.

[5]EIFS=SIFS+DIFS+ACK_time (@ lowest mandatory rate)
[6]Backoff Time=Uniform[0, $CW_{min}2^{i}$-1] x Slot Time

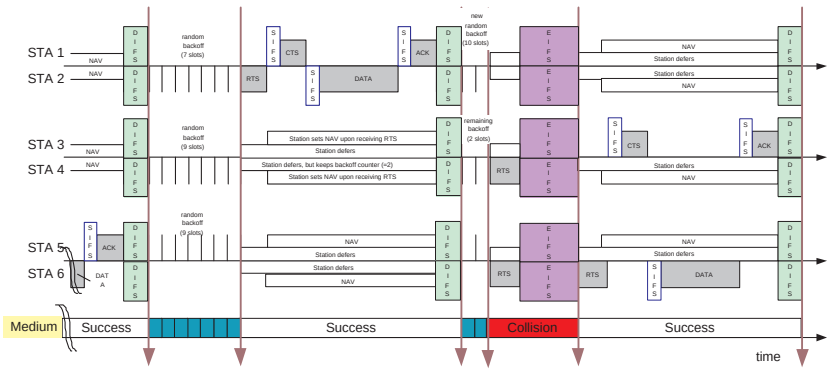


Figure 2.7: DCF Operation in IEEE 802.11 with RTS/CTS

2.5.1 Carrier Sense Mechanism

Carrier sensing mechanism senses and avoids collisions. IEEE 802.11 introduces Physical Carrier Sense (PCS) and Virtual Carrier sense mechanisms (VCS). PCS is a notification mechanism from the PHY layer to the MAC layer, whether the medium is idle or not. PCS sets a physical allocation vector (PAV). The VCS foresees that there is a transmission taking place. The VCS mechanism sets the network allocation vector (NAV) and updates it with the value in the Duration/ID field of received packets only when the new NAV value is greater then the existing NAV, and only when the STA is not the addressee.

Figure 2.8 indicates the NAV update procedure. When the packet is correctly received, *currentIFStime* attribute is set to DIFS, otherwise to EIFS. If the packet is not sent to this station, the station updates its NAV value if the current NAV value is larger than the existing NAV value, and if the packet is not RTS frame. If the packet is RTS, NAV permitted to reset if there is no start in receiving activity (*PHY-RXSTART*) for a duration of T after end time of receiving (*PHY-RXEND*). The station requires this to detect the erroneous CTS frame and reset its NAV accordingly.

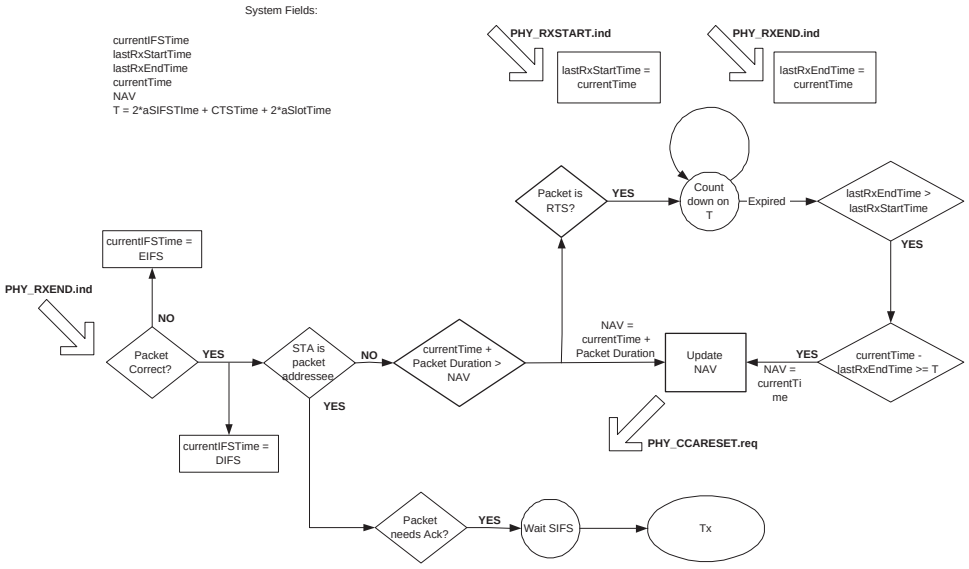


Figure 2.8: NAV submodule finite state machine

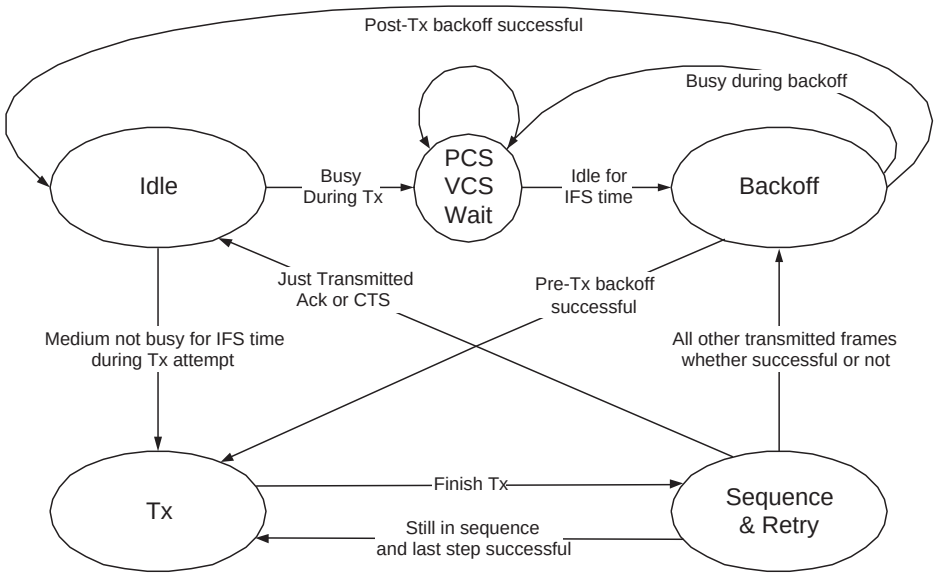


Figure 2.9: DCF finite state machine

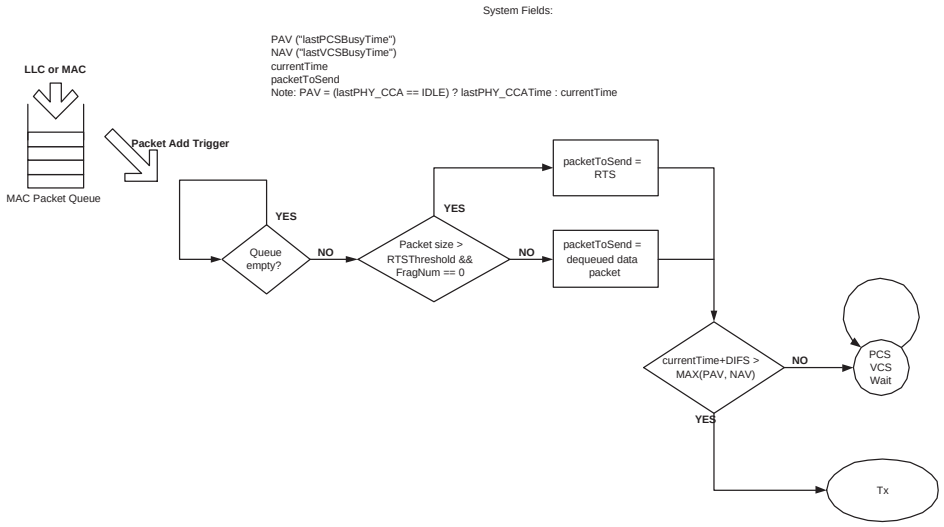


Figure 2.10: Idle submodule finite state machine

2.5.2 Collision Avoidance and Basic Access Mechanism

If the station has a packet to send, the STA attempts to transmit. For RTS, it decides if the packet size is larger than or equal to *dot11RTSThreshold*, and then checks its carrier sense mechanisms. If it detects that the medium is idle longer than DIFS, the station transmits immediately as seen in Figure 2.9 and Figure 2.10. If the medium is busy, the station needs to wait at least IFS time before backoff. If the previous frame was erroneous then IFS time is EIFS, otherwise it is DIFS.

Collision Avoidance is basically selecting a random interval and waiting to transmit until the time is up. In the beginning of the backoff procedure, the STA sets its Backoff timer to a random backoff time described in Section 2.4. If the previous frame that the STA sent was correctly transmitted then backoff level i is set to 0 indicating that it is the initial attempt of a transmission. After selecting a backoff level, the backoff counter is decremented for each idle slot. During backoff, if the medium is determined to be busy,

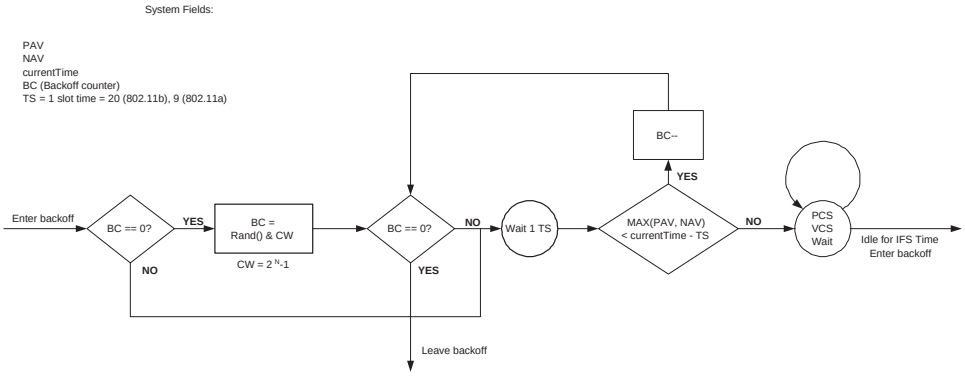


Figure 2.11: Backoff submodule finite state machine

decrementing the backoff timer is suspended and the procedure resumes after noting the medium is idle for the duration of a DIFS or EIFS period (as appropriate, see Section 2.4). The procedure is illustrated in Figure 2.11.

When the backoff timer reaches zero, the STA initiates transmission. The transmission precedes another backoff procedure, whether or not the transmission is successful. If the transmission turns out to be erroneous. Then STA waits EIFS time and sets its backoff timer from [0, $CW_{min}.2^{i}$-1] after incrementing its backoff level. Otherwise, the station resets its backoff level and sets its backoff timer from [0, CW_{min}-1]. If the transmitted frames are ACK or CTS frames then STA goes directly to idle state and does not backoff. By placing backoff between transmissions, the standard tries to implement fairness.

If the transmitted frame needs to be acknowledged then STA sets its ACK timer. If there is a timeout STA increments one of the short retry counter (SRC) (or long retry counter (LRC)) if the packet size is smaller (or bigger) than *dot11RTSThreshold*. Following Figure 2.12, if a valid ACK is received then SRC, LRC and backoff level is reset.

The station immediately continues with transmission if it is in the middle of a se-

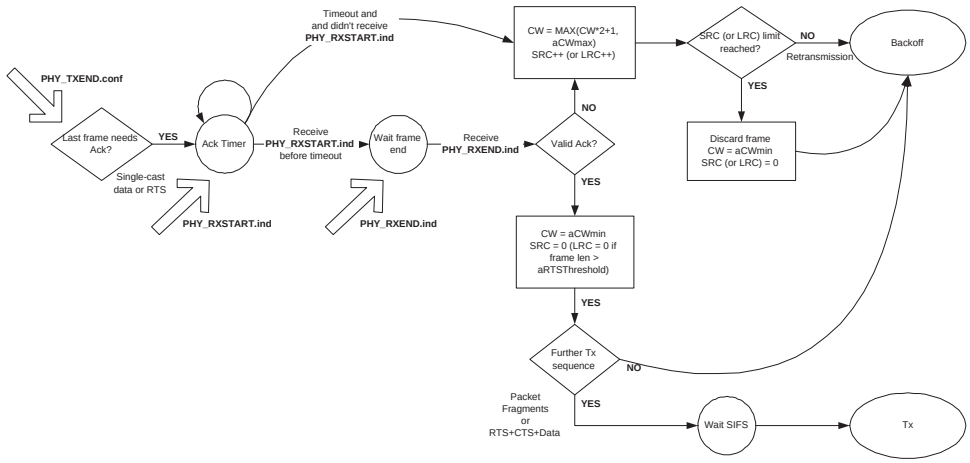


Figure 2.12: Frame Sequence and Retry submodule finite state machine

quence: RTS/CTS transmission sequence or fragmented packet stream. Otherwise, the station invokes the backoff procedure. A realization of DCF access procedure is shown in Figure 2.13.

Features

Control SIFS interval plays a major role in control of delivery. Each packet is separated with a SIFS time interval. This SIFS time corresponds to the time that is introduced by the process in PHY and MAC layer. For example, after transmission, the STA switches from transmitting to receiving state and waits for ACK; or after receiving RTS (or ACK), the STA switches from receiving to transmitting and transmits a CTS (or Fragment) frame.

Fragmentation RTS/CTS usage with fragmentation is shown with the following procedure. Each fragment is acknowledged. The RTS and CTS frames are used to set the NAV to indicate busy until the end of ACK0 which is ACK of fragment0. Both

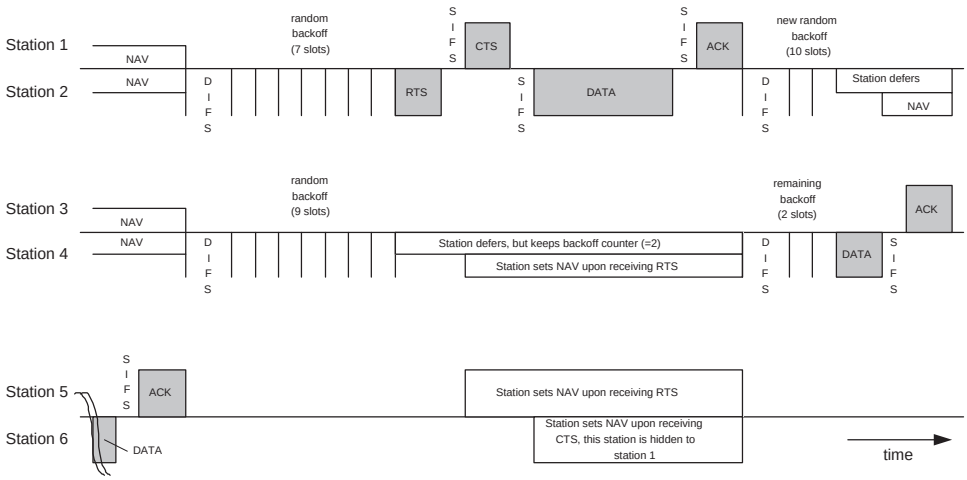


Figure 2.13: Timing of the 802.11 DCF: Note that station 6 cannot hear station 2 but station 1.

Fragment0 and ACK0 contain duration information for NAV until the end of ACK1. This continues until the last fragment. Fragment and ACK play the role of RTS and CTS implicitly.

CTS Procedure A STA who is the addressee of the RTS frame transmits a CTS frame after an idle SIFS period as shown in Figure 2.13. The duration field of the CTS frame is the subtraction of SIFS time and CTS time from the duration field of the received RTS frame. After transmitting the RTS frame, the STA waits for a *CTS Timeout* interval to recognize a valid CTS frame. The recognition of any other frame is interpreted as RTS failure and STA invokes backoff procedure.

Transfer The RTS/CTS is used for unicast frames when the packet size is not less than *dot11RTSThreshold* attribute. In the case of broadcast or multicast packet transmission, only basic access method is used (without RTS/CTS) and no ACK is transmitted by the any of the recipients of the frame. Reliability on broadcast or multicast

frames is reduced relative to the unicast frames.

ACK Procedure Upon successful reception, the recipient generates an ACK frame (See Figure 2.13). The source STA waits an *ACK Timeout* interval to conclude that transmission failed if there is not a valid ACK. Thus, STA invokes backoff procedure.

Duplicate Detection There is a possibility that a frame may be received more than once due to acknowledgements and retransmissions. Duplicate data framing is filtered out by using the sequence number and fragment number fields assigned to each frame. The receiver keeps a cache of <Address, seguence number, fragment number> tuples and if there is a duplicate frame that matches a tuple with the Retry bit set, the STA rejects the frame. On the other hand, the destination STA acknowledges all successfully received frames even if the frame is a duplicate.

2.6 IEEE 802.11 Point Coordination Function

The point coordinator function (PCF) introduces another access mechanism which can assist sessions that requires quality of service. PCF provides a contention free period that alternates with the contention period. In opposition to the DCF, PCF implements a centralized control where AP, point coordinator (PC), controls the network. The AP restricts the access to the medium. Any station, whether it agreed to operate in PCF or not, but is associated, can transmit data as long as the AP allows it to do so.

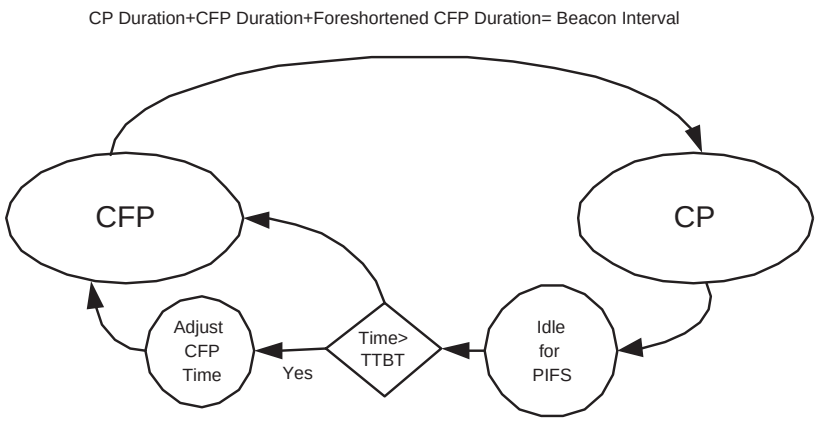


Figure 2.14: Macro states of an access point in PCF

2.6.1 CFP Timing

The PCF quantizes the time with a superframe. The superframe is repetitive and composed of a contention free period (CFP) and a contention period (CP). The superframe starts with a beacon as seen in Figure 2.15. AP and stations alternate in CP and CFP operation. CP is placed in order to give a chance to new coming stations to introduce themselves to the network. Thus, CP must be long enough to transmit at least one frame sequence. *CFPRate* is defined as a number of DTIM intervals. *CFPRate* may be longer than a beacon period. As a result, beacon frames contain a *CFPDurRemaining* field in CF Parameter Set to indicate the remaining time to the end of CFP period. In CP, *CFP-DurRemaining* is set to zero.

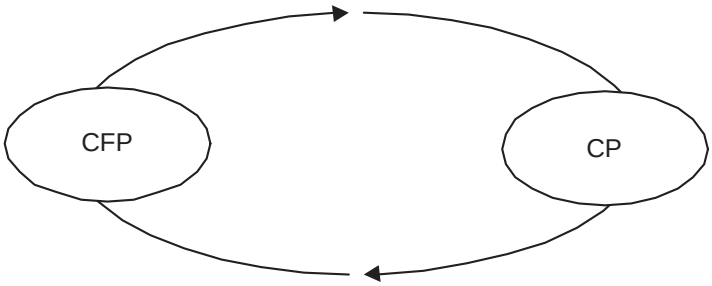


Figure 2.15: Macro states of a station in PCF

2.6.2 PCF Access

Figure 2.16 shows the timing diagram of a typical PCF operation. The beacon frame contains a CFP duration (*CFPMaxDuration*[7]) and has STAs set their NAV. Each frame transmission in CFP is interpolated by SIFS and PIFS which makes non-polled STAs lag behind since DIFS is bigger than PIFS and SIFS. When an AP seizes the medium, the AP polls STAs that agreed to be in the *polling list* during association. In response to receiving a contention-free poll (CF-Poll) frame from the AP, the STA is allowed to send only one frame. Each frame is separated by an SIFS interval. If the AP does not hear any response to its CF-Poll frame for a PIFS period, the AP polls the next station. The situation is depicted in Figure 2.16.

[7]CFPMaxDuration field is set to the minimum interval that may allow sufficient time for the AP to send one data frame to a STA and for the polled STA to respond with one data frame. CFPMaxDuration is set to the maximum interval that may allow to send at least one data frame during the CP.

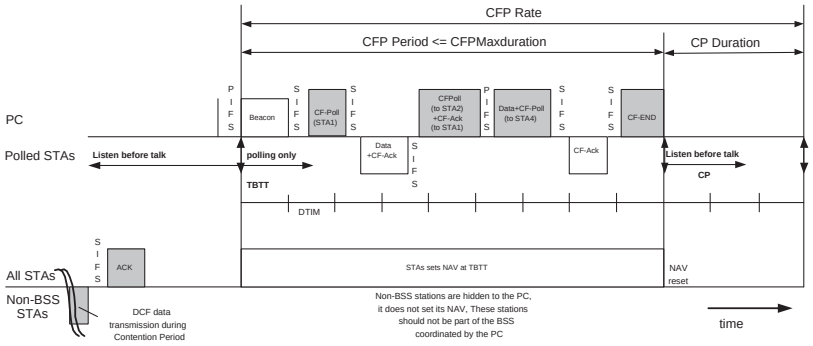


Figure 2.16: Timing diagram for PCF

2.6.3 Beacon Frame

Beacon and *Probe Response* frames contain the CF Parameter Set information element. *Probe Response* frame is sent in order to respond to the *Probe Request* frame transmitted by STAs. The STA sends a *Probe Request* frame to find and initiate association with a network before waiting to hear a *beacon*. CF Parameter Set contains attributes that delineate the outline of PCF operation.

CFP Period Length of CF Period (indicated in DTIM intervals).

CFP Count Time that is between Beacon time and start of the next CF Period (indicated in DTIM intervals).

CFP MaxDuration Maximum duration of CFP period (indicated in time units (TUs)).

CFP DurRemaining From beacon time to the end of CFP period (indicated in TUs).

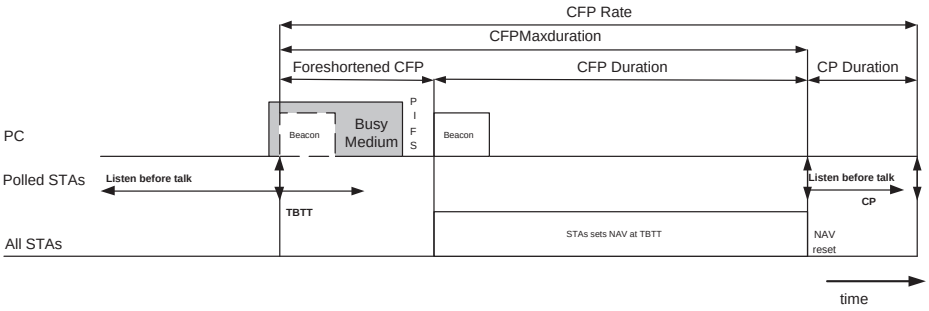


Figure 2.17: Contention Free Period determination

2.6.4 Piggybacking

In order to save bandwidth, piggybacking is introduced. Piggybacking allows the presence of multiple frames in a single frame. For example *Data+CF-Ack+CF-Poll* brings data transmission, polling feature, and acknowledgement in one frame.

2.6.5 CFP Duration

Figure 2.17 shows the situation when there is a delay in beacon transmission. The AP tries to keep the beacon transmission periodic by shortening the contention free period if there is a delay in beacon transmission due to a CP interval. The contention free period (CFP)[8] is adjusted according to Target Beacon Transmission Time (TBTT). The PC also ends the CFP period at any time by the *CF-End* frame depending on the polling list, traffic load, or any other reason.

[8]CFP Period=*CFPMaxDuration*-(Actual CFP Start-TBTT)

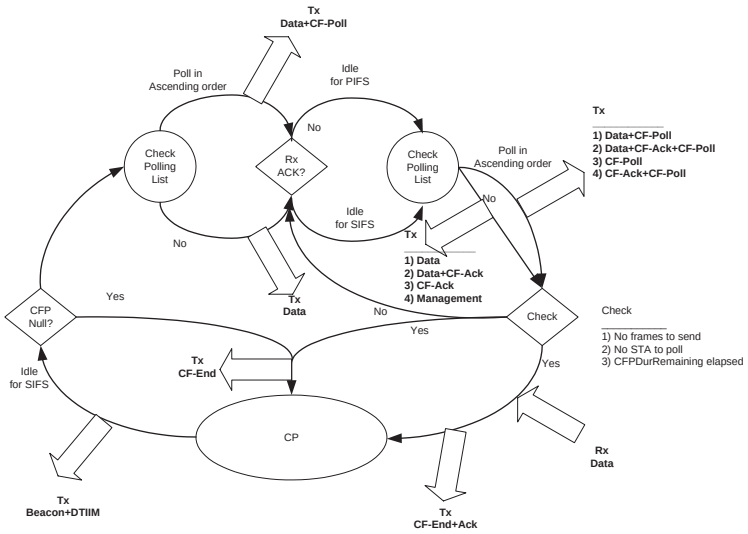


Figure 2.18: Access Point CFP Period finite state machine

2.6.6 NAV Operation

NAV operation facilitates the PCF in the case of overlapping BSSs. Each STA in BSS sets its NAV to *CFPMaxDuration* with the beacon, at each target beacon transmission time (TBTT), and updates its NAV by *CP Period* field with every other beacon frame. This operation prevents STAs from seizing the medium.

An incoming STA updates its NAV with the information in the *CFPDurRemaining* field of any received beacon and waits to hear a *CF-End* or *CF-End+ACK* frame in order to reset its NAV.

2.6.7 PCF transfer procedure

According to the direction of the transmission, whether from AP/PC or to AP/PC, *From DS* or *To DS* fields are set in the frame structure (See Figure 2.6), respectively.

A STA notifies the AP about its decision to be included in the polling list (*CF-*

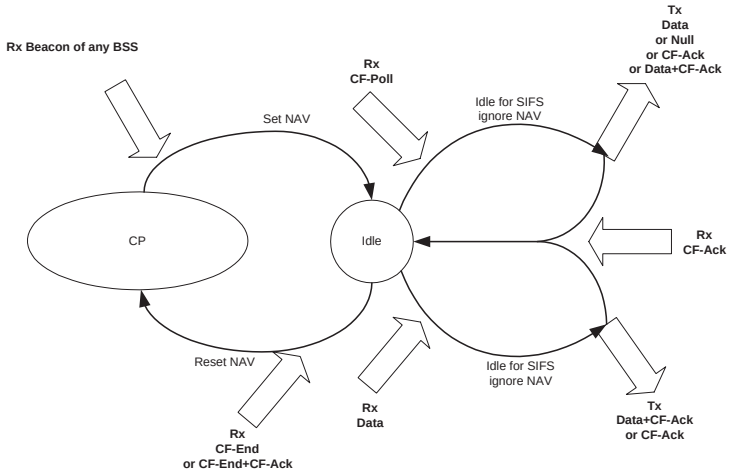


Figure 2.19: Station CFP Period finite state machine

Pollable) with association. A STA wanting to change its condition may initiate retransmission for being included (or secluded) in (out of) the polling list. When included, STA alternates between CP and CFP as AP. The CFP state machines, both for STA and AP, are shown in Figures 2.18 and 2.19. One can follow the figures easily to understand the operation and packet exchange of AP and STAs.

2.7 IEEE 802.11e MAC Protocol

IEEE 802.11e is supplementary to the MAC layer protocol. It allows support of LAN applications with Quality of Service (QoS) requirements over IEEE 802.11 wireless LANs. 802.11e is designed to be complementary to the 802.11 physical standards such as a,b and g. The classes of service are offered with a QoS management scheme, where the applications are data, voice and video.

The QoS facility is available to QoS enhanced stations (QSTAs) and QoS enhanced

access points (QAP) in QoS BSS (QBSS) or in QoS IBSS (QIBSS). The standard improves the legacy DCF and PCF and introduces Enhanced Distributed Coordination Access (EDCA) and HCF Controlled Channel Access (HCCA), where HCF stands for Hybrid Coordination function. EDCA implements a contention based access mechanism. EDCA prioritizes the traffic and varies the amount of time a station would wait before transmission depending on the traffic type. HCF introduces a hybrid coordinator (HC) which coordinates a reservation of transmission opportunities. A wireless station (WSTA) that is not attached to QAP, requests the HC for a transmission opportunity. HC schedules a transmission opportunity for the WSTA and delivers the frames based on a polling mechanism. A non-QoS STA is allowed to associate with a QBSS and interact with legacy 802.11.

Distribution service introduces a traffic differentiation service in which the packet transmission is controlled based on the traffic class to which it belongs. This QoS functionality makes IEEE 802.11e WLAN become a part of a large QoS network where end-to-end delivery may be performed, or makes it a last hop of the network where QoS is provided within its boundary.

2.7.1 MAC services

MAC service provides connectionless exchange of MSDUs. In legacy 802.11, the packet exchange is based on best effort service but in 802.11e, the QoS facility involves a per MSDU basis and identifies the traffic. Depending on the traffic priority level, the MAC delivers MSDUs with a corresponding user priority. The traffic identifier gives zero pri-

Priority	Access Category (AC)	Designation (Informative)
1	0	Best Effort
2	0	Best Effort
0	0	Best Effort
3	1	Video Probe
4	2	Video
5	2	Video
6	3	Voice
7	3	Voice

Table 2.1: Priority access category mappings for IEEE 802.11e

ority to an MSDU which belongs to a non-QoS STA.

There are eight user priorities (UP) defined in the standard as seen in Table 2.1. Either each MSDU has a UP value or a traffic specification (TSPEC). The traffic identifier (TID) field values, 0 through 7, are designated for UP. The TID field values, 8 through 15, are interpreted as traffic stream identifiers and designated to TSPEC. Outgoing MSDUs are handled with priority parameter values 8 through 15 in accordance with the user priorities 0 through 7. MSDUs are permitted to reorder in the MAC layer which allows it to implement priority. The general packet format depicted in Figure 2.20 includes a 2 octets QoS control field. Maximum MSDU size is 2304 octets which determines the maximum frame body minus any encryption overhead.

2.7.2 MAC architecture

The MAC architecture is shown in Figure 2.21. The MAC protocol introduces EDCA and HCF that can work with DCF and PCF.

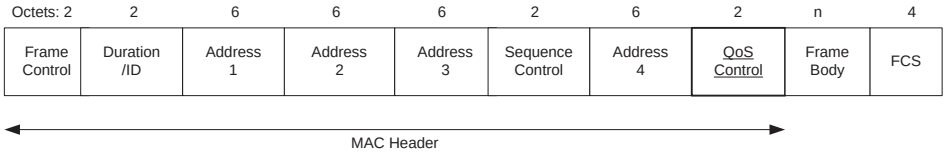


Figure 2.20: MAC frame format © IEEE

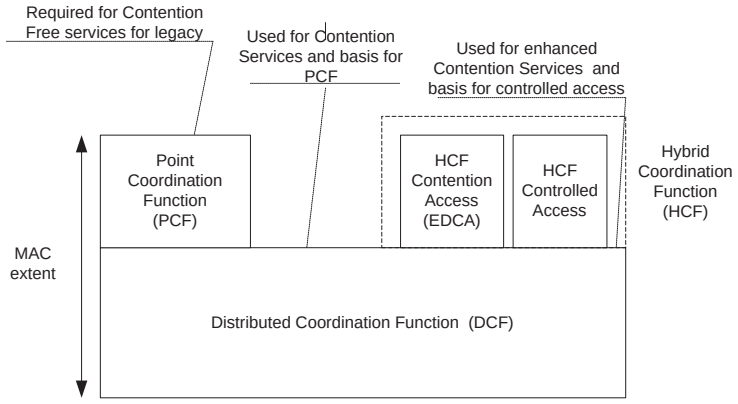


Figure 2.21: MAC architecture © IEEE

2.7.3 Hybrid coordination function (HCF)

HC is a coordination function in QBSS. HCF introduces two access mechanisms: contention based and polling based access mechanisms. Contention based channel access is called enhanced distributed channel access (EDCA) and contention-free transfer is called controlled channel access (HCCA) mechanism. Transmission opportunity (TXOP) is acquired by using one or both of the access mechanisms by QSTA. Depending on the access mechanism used, the TXOP is called either EDCA TXOP or polled TXOP.

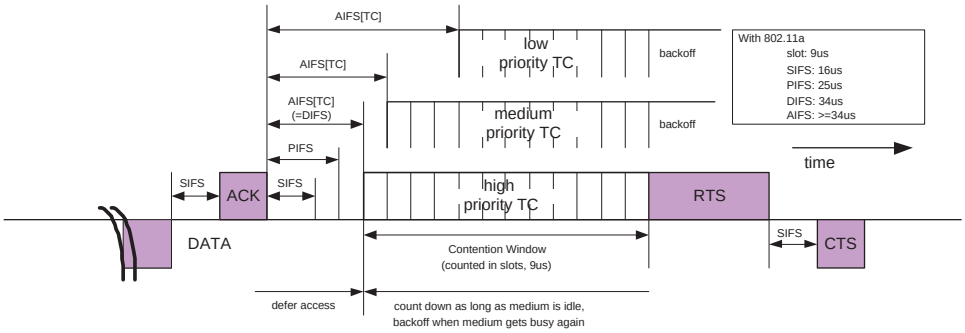


Figure 2.22: EDCA mechanism of IEEE 802.11e

HCF contention-based channel access (EDCA)

EDCA provides a contention-based differentiated access. Contention parameters vary according to user priorities. There are eight user priorities (UP) which define the access categories (AC). The idle time is now not constant (DIFS) but equals AIFS[AC]. In the same way, CW_{min} and CW_{max} are not fixed per PHY but variable and equal to CW_{min}[AC] and CW_{max}[AC] depending on the AC. In contention, the higher AC valued station receives the TXOP and transmits multiple frames as long as it does not exceed the EDCA TXOP limit [AC]. A typical situation is depicted in Figure 2.22.

Figure 2.23 is a reference model of mapping the frames into access categories. Each access category contains a queue, channel access function, and medium occupancy timer. When a channel access function gets the transmission opportunity (TXOP) it sets the medium occupancy timer to a value defined in MIB. As long as the channel access function continues to use of TXOP, the medium occupancy timer is not reloaded but continues to count down to zero. The EDCA access mechanism is similar to DCF. But now a station

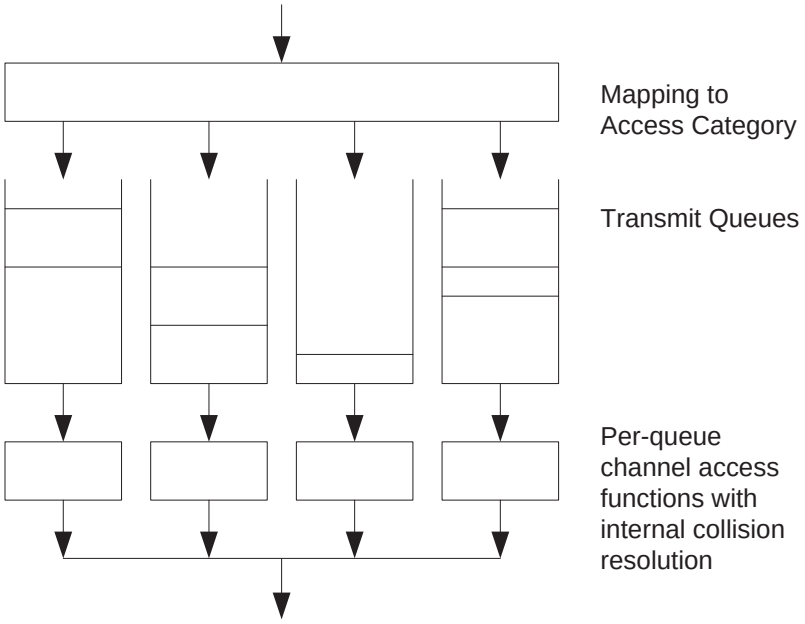


Figure 2.23: Reference implementation model © IEEE

has more than one channel access function: each one tries to contend for channel access. Before a transmission, the channel access function looks for an idle channel for a time greater than AIFSD[AC][9]+Slot time, or a time greater than EIFS-DIFS+AIFS[AC], if the previous transmission was in error. If there is transmission in the medium, the channel access function backs off. The backoff timer is decremented after detecting the medium is idle for a slot time or AIFSD[AC] from the last indicated medium busy or a period of EIFS-DIFS+AIFSD[AC] if the previously received frame was erroneous. Continuation of TXOP is given to the frame of the same access category as the transmitted frame.

The backoff procedure has finite state machine. Each channel access function maintains a state variable CW[AC], and initializes to CW_{min}[AC]. The backoff procedure is

[9]AIFSD[AC]=AIFS[AC]+Slot time+ SIFS

invoked when the frame with that AC has requested transmission and the medium is busy; or when a frame with that AC is transmitted and TXOP is terminated; or when a frame with that AC fails; or when there is higher AC, another channel access function in the same STA, which is granted access at the same time. If there is a failure or if a higher AC grants the access, the CW[AC] is updated as long as it is less than the CW_{max}[AC] and as long as the retry limits are not exceeded. The update procedure is as follows:

$$CW[AC] = (CW[AC] + 1) * 2 - 1.$$

When the backoff procedure is invoked, the backoff time is chosen randomly between (1, CW[AC]).

The backoff timer is decremented after a slot time if the medium is idle; or after AIFS[AC] if the medium is busy; or after EIFS - DIFS + AIFS[AC] + SlotTime if the last frame was erroneous or after a ACK timeout interval if there is no ACK requiring transmission initiated from the channel access functions of this QSTA. There might be internal collision between channel access functions which leads to an increase in one of the retry counts.

The retransmission procedure is different compared to the legacy retransmission procedure. In the legacy transmission procedure, if the transmission fails, the STA or AP attempts to retransmit as long as retry limits allow. In 802.11e if the STA or AP uses the non-QoS data types, the same rule applies; but if the STA or AP uses the QoS data types, it has the right to initiate a transmission attempt of a frame of an access category not belonging to the access category of the frame which failed previously.

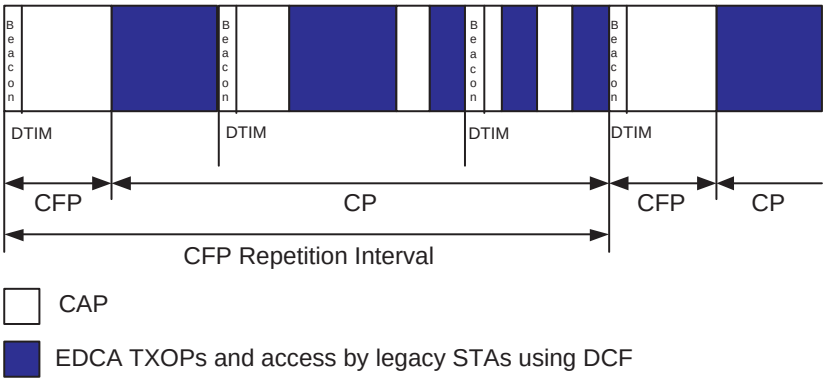


Figure 2.24: CAP/CFP/CP periods © IEEE

2.7.4 HCF controlled channel access

The HCF uses a point controller called Hybrid controller (HC) and resides in the QoS enhanced access point (QAP). HC is similar to PC but with additional functionality. HC provides TXOP to WSTAs in the called controlled access phase (CAP) to transfer QoS data. Contention free periods (CFP) and contention periods (CP) alternate in HC operation within a superframe. A non-QSTA operates in CP using the DCF access method. HC has the functionality to grant TXOP with duration specified to a non-AP QSTA. This allows a non-AP QSTA to transmit multiple frame exchanges.

HC accesses the WM by starting a CFP or a TXOP in CP. The CAP procedure in conjunction with CFP and CP is depicted in Figure 2.24.

If an expected response is received during the first slot time following SIFS, the HC may initiate by transmitting at a PIFS after the end of the last transmission, or non-AP QSTA initiates recovery by transmitting at a PIFS after the last transmission [13].

During the TXOP, NAV protects the transmission by suspending the all stations except the transmitting one. QSTA uses the QoS control fields in the packet to indicate the

traffic belonging to the packet. Non-AP QSTAs may also request TXOP by setting the appropriate fields in the packet.

2.7.5 Admission Control

The Admission Control Unit (ACU) is left open in the standard. The ACU is implemented to control the admission criteria to a given request in polling. A reference implementation calculates the Scheduled Service Interval (SI) and TXOP duration for a given SI. The calculation of the Scheduled Service Interval can be done as follows. The scheduler determines the minimum of all Maximum Service Intervals for all admitted streams and takes closest number which is multiple of beacon intervals. If ρ is mean data rate, L is nominal MSDU size and R is the physical transmission rate then

$$TXOP_i = max\left(\frac{N_i \times L_i}{R_i} + O, \frac{M}{R_i} + O\right) \tag{2.1}$$

where $N_i = \lceil \frac{SI \times \rho_i}{L_i} \rceil$, O is overhead in time units, M is maximum allowable size of MSDU, i.e., 2304 bytes.

When a new stream is added, the criteria is as follows:

$$\frac{TXOP_{k+1}}{SI} + \sum_{i=1}^{k} \frac{TXOP_i}{SI} \leq \frac{T - T_{CP}}{T} \tag{2.2}$$

where k is the number of existing streams; T and T_{CP} indicate the beacon interval and time used for EDCA traffic, respectively. If the inequality in Equation 2.2 holds, the new arriving stream is admitted.

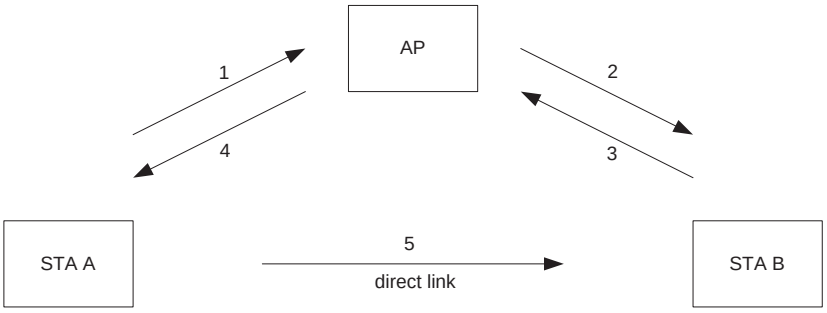


Figure 2.25: Direct Link handshake © IEEE

2.7.6 Block Acknowledgement

Block acknowledgement is introduced to save bandwidth by acknowledging multiple frames with one ACK. Block acknowledgement is granted after an exchange of request and response frames. The BlockACK control frame is sent after receiving blocks of QoS data type frames.

2.7.7 Multirate support

A dynamic rate switching capability is allowed in the standard. 802.11e defines a set of rules for possible algorithms. All control, multicast or broadcast frames except the ACK frames are transmitted at a rate that belongs to the basic rate set. Data, ACK, management with unicast receiver address are sent in any rate that is being agreed upon by both sides.

2.7.8 Direct Link Protocol

IEEE 802.11e introduces direct link protocol (DLP) which allows transmission of frames directly to a STA from another STA. DLP is established by three way handshake. STA (A) which is willing to communicate STA (B) sends request to the AP which forwards

that request to STA (B). If STA (B) accepts the direct stream, STA (B) replies back to the AP and the AP forwards the response to STA (A) (See Figure 2.25). The AP may also pair them without any request. STA (B) does not go into power save as long as there is a packet transmission, or for a duration of DLP idle time out period.

DLP is useful if the recipient is in power save mode and needs to be awakened. DLP enables more secure sessions between sender and receiver thereby rate set and other information can be exchanged.

2.8 IEEE 802.11 and 802.11b Physical Layer

The physical layer (PHY) is the layer 1 element of OSI protocol stack. IEEE 802.11 introduced three PHY standards in 1997 and two supplementary standards in 1999:

- Frequency-hopping spread-spectrum (FHSS)
- Direct-sequence spread-spectrum (DSSS)
- Infrared light (IR)
- 802.11b: High-rate Direct Sequence (HR/DSSS)
- 802.11a: Orthogonal Frequency Division Multiplexing (OFDM)

The PHY layer is composed of physical layer convergence (PLCP) and physical medium dependent (PMD) layers as seen in Figure 2.2. PLCP is an interface to MAC layer and PMD is equipped with a transmission interface to send and receive files over the air. The PLCP layer prefixes preamble and header to the MAC layer data frame as seen in Figure 2.6. The PLCP preamble provides enough time for the receiver to apply its functions

such as antenna diversity, clock, data recovery, and field delineation of the rest of the packet. There is a synchronization field in the preamble to detect the signal and there is a Start Frame Delimiter (SFD) field for the receiver to lock the start of the frame. The PLCP header is used to provide enough information for the receiver to process the packet.

Allocated spectra for wireless LAN applications are typically in 2.4 GHz and 5 GHz range where the bandwidth is scarce and in much demand. WLAN system at 2.4 GHz must meet ISM band requirements. The IEEE 802.11 standard supports both FHSS and DSSS techniques for this band [7]. Spread spectrum techniques shows robustness against interference and do not require adaptive equalization. However, for higher data rates the synchronization requirements of spread spectrum techniques are more restrictive and complex. For data rates of above 10Mbps, OFDM system shows better performance. In the next section we discuss spread spectrum to motivate the OFDM-based physical layer.

2.8.1 Spread Spectrum

Initiated from military technology, spread spectrum technology became the vital component for systems ranging from cellular to WLAN systems. Spread spectrum uses wideband, noise-like signals that are hard to detect. The spread spectrum technique spreads the narrow band signal over a wider band by pre-known or pseudonoise spreading code in both sides to increase the bandwidth of signals to be transmitted. The main parameter is processing gain which is defined as the ratio of transmission bandwidth to information bandwidth. The receiver performs the inverse operation and the spreaded signal is reconstructed in the narrow band along with narrowed noise. Spreading the signal over

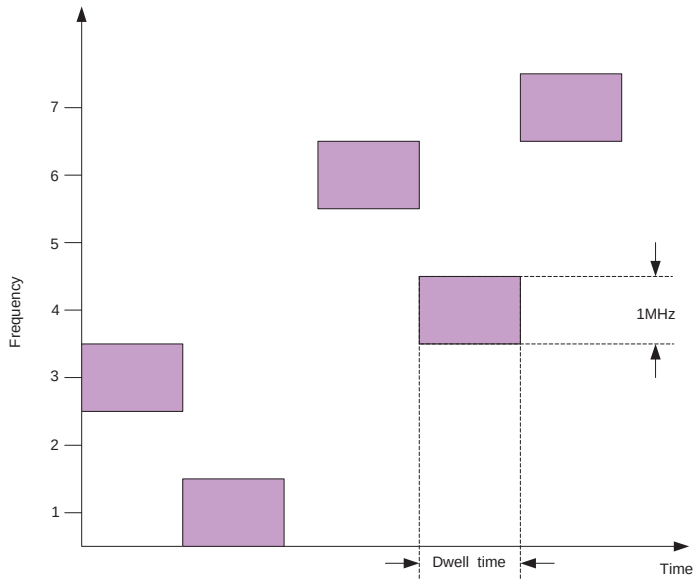


Figure 2.26: Frequency Hopping Spread Spectrum

a wider frequency makes the signal more noise like and secure. A different code in each receiver diminishes the interference considerably but not totally. It also allows access at any arbitrary time. Processing gain determines the number of users that can be allowed, the amount of reduction in multi-path effect, the difficulty to jam or detect a signal, and so on. It is advantageous to have processing gain as high as possible.

2.8.2 FHSS Physical Layer

Frequency hopping is simply changing the frequency from one to another in each timely slotted intervals. The hopping pattern is unique and available in transmitter and receiver. In this case bandwidth is increased by a factor length of the sequence. Time spent in one frequency is called *dwell time*. There is a synchronization mechanism that helps the receiver and transmitter pair operate in the same frequency in the same dwell time.

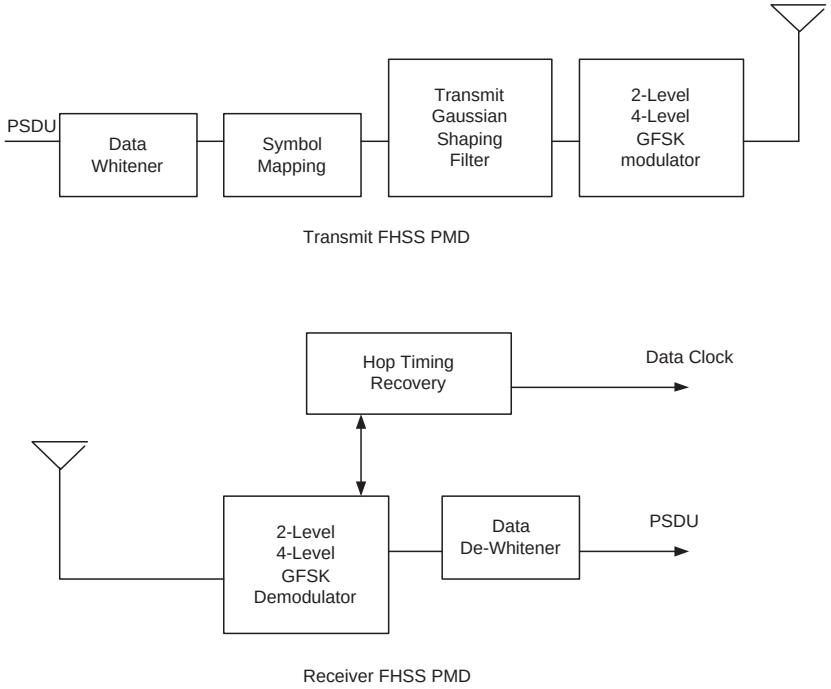


Figure 2.27: Transmit and Receive in FHSS © IEEE

Channel length	1MHz
Channel number	79 from 2.402 GHz
Dwell Time	0.4secs
Hoping Sequence Size	26
Data rate	1-2 Mbps

Table 2.2: FCC rules for IEEE 802.11 FHSS

Beacon frames include information about the pattern and time. Thus a receiver knows the hopping sequence number, hopping index and timestamp. Hopping occurs when the timestamp equals multiples of dwell time (See Table 2.2).

Gaussian Frequency Shift Keying

FHSS uses Gaussian Frequency Shift Keying (GFSK). GFSK uses frequency modulation to get rid of unwanted changes in the amplitude. The narrow band signal is filtered with a

low pass Gaussian filter (500KHz bandwidth in 3dB) and then the signal is FM modulated in GFSK. It basically deviates the frequency on either side of the carrier hop frequency depending on whether the value of binary symbol being transmitted equals 1 or 0. 1Mbps and 2Mbps data rate is achieved by 2-level GFSK and 4-level GFSK, respectively. 4GFSK uses four symbols instead of two as in 2GFSK and also brings a complex and high-cost transceiver pair.

FHSS PLCP and PMD Sublayer

The PLCP state machine contains transmit, receive and carrier sense / clear channel assessment (CS/CCA) states. CS/CCA supports CSMA/CA mechanism of MAC layer. CS/CCA informs the MAC layer about the medium whether it is idle or busy.

The PLCP Header contains PSDU Length Word (PLW), PLCP Signalling Field (PSF) and Header Error Check (HEC) fields. They represent the length of the MAC frame, data rate, frame check respectively. The PLCP Header is always sent with 1Mbps. Data whitening is applied to the PSDU before transmission to minimize DC bias. The PMD layer is responsible for controlling the hopping sequence and transmitting the whitened PSDU by hopping from channel to channel in a pseudo-random fashion. The transmitter and receiver are illustrated in Figure 2.27.

Pros and Cons

The wireless environment introduces frequency selective fading and fades are correlated in adjacent frequencies. The hopping sequence determination should try to avoid selection of the adjacent channels. FHSS shows resistance to jamming unless the jammer jams all

frequencies. Collision in one or two frequencies can be recoverable. Spending the time to change the frequency introduces delay in transmission time. The FHSS systems are very cheap. Once they were very important in the history of communication but the higher rate systems have overtaken FHSS technology due to their ability to achieve higher processing gain.

2.8.3 DSSS Physical Layer

Direct Sequence is the best known Spread Spectrum Technique. A data signal at the point of transmission is multiplied with a higher data-rate bit sequence (also known as a chipping code) that divides the data according to a spreading ratio and spreads the spectrum out while dropping the power spectral density. This chipping code introduces redundancy which enables correct symbol recovery even if some of the symbol is damaged. Increasing the rate of the chipping signal increases the throughput and consequently requires higher bandwidth and a complex transmitter receiver pair. A good chipping sequence is called a pseudo noise (PN) sequence. A given PN sequence is orthogonal with all the other PN sequences of the same length, and a given PN sequence is almost uncorrelated to a shifted version of the given sequence. This is very important in synchronization.

A receiver begins by despreading the signal. This is done by the help of the same PN sequence. It collapses the desired signal to its original bandwidth due to the correlation of PN sequence and the transmitted signal. The signal power is also increased by the amount of the processing gain. DSSS allows multiple access to the medium since each transmission has its PN sequence and each DS receiver collapses only correlated signals

Channels	1 to 11 (2.412-2.462 GHz)
Channel bandwidth	5MHz
Channel Spread	25MHz
Chipping Sequence	Barker Sequence
Data rate	1-2 Mbps
Min. Processing Gain	10dB

Table 2.3: FCC rules for IEEE 802.11 DSSS

to data. The block diagram is illustrated in Figure 2.28

DSSS PLCP and PMD Physical Layer

The PLCP header has signal, service, length and CRC fields. Signal fields notify the receiver of the applied modulation. The Service field is reserved for future use and the length field indicates the length of MDPU. The CRC checks whether or not the frame is received correctly. The PMD scrambles all the bits transmitted in order to randomize the data in the SYNC field and data patterns where there may be long strings of repeated 1s or 0s. PLPC Preamble and PLCP Header are modulated using differential binary phase shift keying (DBPSK). The MPDU is transmitted with either 1 Mbps DBPSK or 2Mbps DQPSK (Differential quadrature phase shift keying).

In the transmitter, the 11-bit Barker word is applied to each transmitted bits via modulo-2 adder. The information bits modulated with a rate of 1Mbps are extended to data rate 10x higher than the information rate when 11-Barker word is at 11Mbps. This spreading ratio is also called processing gain.

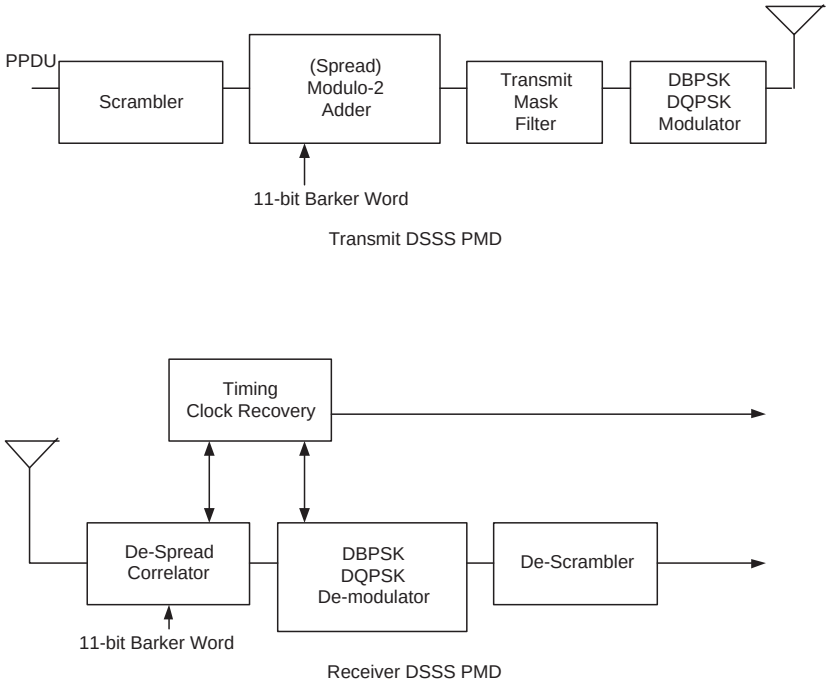


Figure 2.28: Transmit and Receive in DSSS © IEEE

Pros and Cons

The main problem with DSSS is the Near-far effect. This problem is present when an interfering transmitter is much closer to the receiver then the intended transmitter. The correlation of received signal from interfering transmitter and the code can be higher than the correlation of received signal from intended transmitter and the code. The DSSS system also consumes more power than the FSSS system since it is more complicated. On the other hand, the DSSS is more resistant to interference than FHSS. Unless the interference level passes, the noise floor, the signal is recoverable. Transmission time in DSSS is shorter than FHSS since it does not need to wait to change the frequency.

2.8.4 IR Physical Layer

The infrared (IR) based physical layer is another specification introduced by IEEE 802.11. Infrared transmission requires line of sight and is very susceptible to reflection. As a result it has very low range and can be confined to a room since infrared signals can not pass through the walls. IR uses pulse position modulation (PPM) where it keeps the amplitude and pulse width constant, and varies the position of the pulse in time. The position in time represents a symbol. The IR transmits binary data with 1Mbps and 2Mbps with 16-PPM and 4-PPM respectively. IR PHY have never received interest and no product has been developed.

2.8.5 HR/DSSS Physical Layer

IEEE 802.11b, introduced in 1999, implements high rate (HR) extension of the physical layer of Direct Sequence Spread Spectrum (DSSS). HR/DSSS provides two functions. First, the HR/DSSS builds 5.5 and 11Mbps data rate capabilities in addition to the 1 and 2 Mbps rates. Secondly, the HR/DSSS provides a rate shift mechanism to introduce the functionality of falling back to 1 and 2 Mbps so that it can interoperate with legacy IEEE 802.11. IEEE 802.11b operates in 2.4GHz ISM band along with cordless phones, microwave ovens, other adjacent WLAN networks, and personal area networks (PANs). Being susceptible to interference is the primary concern for IEEE 802.11 products. 802.11b uses the same configuration as in DSSS except for the modulation and chipping sequence.

Complementary Code Keying (CCK) Modulation

IEEE 802.11b uses eight-chip complementary code keying (CCK) as the modulation scheme to achieve the higher data rates. The DSSS system uses the Barker sequence which is static; but HR/DSSS uses a code word that is derived partially from data. Thus, it carries information as well as spreads the signal. CCK has mathematical properties that allows them to be correctly identified from one another in the presence of noise and interference.

The spreading code length is 8 and is based on complementary codes. The chipping rates is 11Mchip/s. 1-2 Mbps is achieved with DQPSK same as in DSSS. Transmission at 5.5Mbps with CCK is transmitting 4 bits per symbol. Two of them are carried by DQPSK and the other two are injected to the code words. In the same way for 11 Mbps transmission, two of them are carried in DQPSK and the other 6 are injected to the code words [7, 9].

2.8.6 RF Interference

The presence of unavoidable interfering RF signals disturbs IEEE 802.11 operation. When the power of interferer is significant, it may cause the STA to become inactive for indefinite periods until the interferer disappears.

An interfering signal basically causes an energy jump in radio and may start abruptly in the middle of a transmission causing a collision. This is because the CS/CCA mechanism sends a busy signal to the MAC layer through the carrier sense mechanism. If the interference continues, the STA either automatically switches to a lower rate or holds off

until the medium is clear.

The unlicensed 2.4 GHz bandwidth contains several technologies that can interfere with IEEE 802.11 or 802.11b network: wireless phones, microwave ovens, Bluetooth. For a WLAN, another WLAN that operates in the same channel is also an interferer. This situation is likely occur in IEEE 802.11 and 802.11b since there are only three non-overlapping channels. IEEE 802.11a avoids interference for the near future since most potential RF interference today is in the 2.4GHz, and IEEE 802.11a has twelve non-overlapping channels and operates in 5 GHz.

2.9 IEEE 802.11a Physical Layer

IEEE 802.11a has been developed to provide high data rate service at 5GHz U-NII bands. IEEE 802.11a selects multi-carrier modulation. Multi carrier modulation is a strong candidate for packet switched wireless applications and offers several advantages over single carrier approaches. The OFDM system is a viable solution to accommodating 6-54 Mbps data rates for the following reasons.

Robustness against delay spread: Data transmission in a wireless environment experiences delay spreads of up to 800 ns which covers several symbols at baud rates of 10 Mbps and higher. In a single carrier system, an equalizer handles detrimental effects of delay spread. When delay spread is beyond 4 symbols, use of maximum likelihood sequence estimator structure is not practical due to its exponentially increasing complexity. Linear equalizers are not suitable for this application either since in a frequency selective channel it amounts to significant noise enhancement.

Hence, other equalizer structures such as decision feedback equalizers are used. Number of taps of the equalizer should be enough to cancel the effect of inter-symbol interference, and must act as a matched filter too. In addition, equalizer coefficients should be trained for every packet, as the channel characteristics are different for each packet. A large header is usually needed to guarantee the convergence of adaptive training techniques. A multi-carrier system is robust against delay spread and does not need a training sequence. Channel estimation is required however.

Fall-back mode: Depending on the delay spread for different applications a different number of carriers is required to null the effect of delay spread. The structure of FFT lends itself to simple fallback mode by using the butterfly structure of FFT as shown in [6].

Computational efficiency: Use of FFT structure at the receiver reduces the complexity. As the number of carriers grows higher efficiency can be achieved.

Fast synchronization: OFDM receivers are less sensitive to timing jitter as compared to spread spectrum techniques.

Near Far affect Unlike spread spectrum techniques where the signal is detected by correlation, the near-far problem does not apply to OFDM receivers since the detection is not based on correlation.

There are 12 channels where each set of four are in U-NII lower band (5.15-5.25 GHz), mid-band (5.25-5.35 GHz), and upper band (5.725-5.825 GHz) with a channel

Parameters	Slot Time	SIFS	PIFS	DIFS	CWmin	CWmax
802.11a	9μs	16μs	25μs	34μs	15	1023

Table 2.4: IEEE 802.11a OFDM PHY characteristics

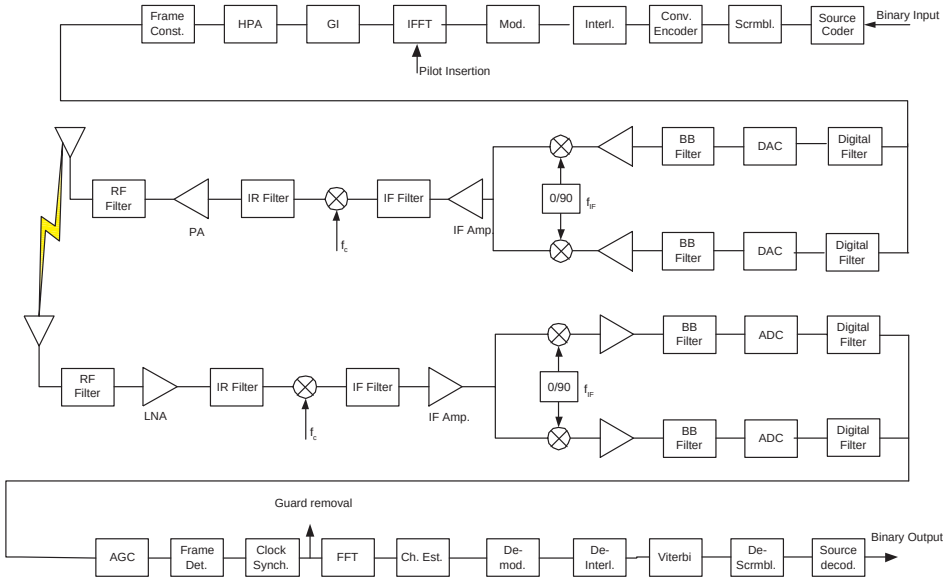


Figure 2.29: IEEE 802.11a PHY architecture

spacing of 20 MHz. IEEE 802.11a parameters are seen in Table 2.4.

2.9.1 OFDM Architecture

A typical OFDM wireless LAN transceiver is shown in Figure 2.29. We explain the process of IEEE 802.11a transceiver starting from the binary input to the binary output. First, an OFDM symbol is constructed then an OFDM frame in PLCP layer is put into together with the OFDM symbols. The PLCP/PMD layer sends the frame into air. On the receiver side, first the frame is detected and then the receiver converts the radio signal into OFDM symbols. OFDM symbols are demodulated and binary data is extracted. Logical

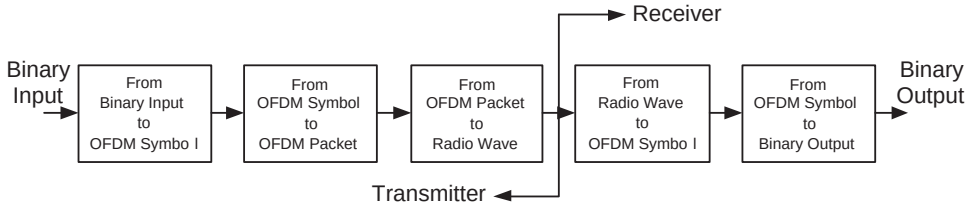


Figure 2.30: Logical Block diagram of an OFDM architecture

block diagram is represented in Figure 2.30. Each block is set of modules that states the changes of binary input up to binary output.

2.9.2 Transmitter

(1) From Binary Input to OFDM Symbol

Source Coding

A number can be represented in binary with 1s and 0s. Binary data is converted into an efficient representation where the redundancy of the binary data is reduced by leveraging the correlation between the numbers. Information theory defines entropy as the smallest achievable average length after removing redundancy [23].

Scrambler

Data may contain long sequence of 1s or 0s. This degrades the performance of receiver when resolving the data. A scrambler helps change the bit pattern with a 127 pseudo random sequence. The same scrambler is used to descramble the data. When transmitting, the initial state of the scrambler will be set to a pseudo random non-zero state. The 7 least

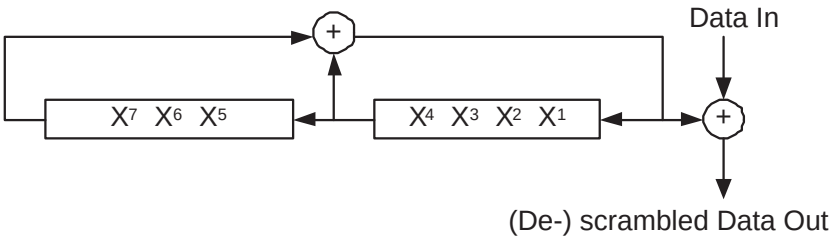


Figure 2.31: Scrambler/Descrambler

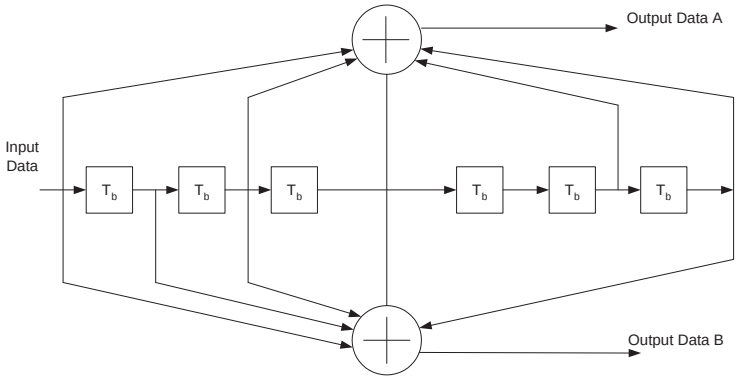


Figure 2.32: Convolutional encoder (k=7) © IEEE

significant bits of the SERVICE field are set to all zeros prior to scrambling to enable estimation of the initial state of the scrambler in the receiver (See Figure 2.31).

Convolutional Encoder & Puncturing

Channel coding is the crucial part of any communication system. Coding brings reliability at the expense of adding redundancy. Same bit error rates can be achieved with a lower signal-to-noise ratio.

There are two types of coding: block and convolutional. IEEE 802.11a and HIPERLAN/2 WLAN use convolutional error coding. The convolutional encoder is a technique which is suitable for continuous streams. Variable data rates resulting from different mod-

ulation schemes require variable coding gain for each data rate. Different coding rates for convolutional coding are proposed depending on required protection. Obviously, more crowded constellations need higher coding protection. A combination of error detection, such as CRC, and error correction, such as convolutional coding, in conjunction with interleaving are used to provide better coding gain in the presence of frequency selective fading.

The convolutional encoder is a kind of finite state machine and has registers to represent the state of the system. The coding rate ($\frac{k}{n}$) is defined as the number of registers (k) over the number of output (n). The length of the shift registers makes the system more reliable but on the other hand increases the decoding complexity of the Viterbi algorithm [24].

For a 1/2 rate coder, each input bit is processed as two coded output bits. Thus, the convolutional coding increases the length of the original message. Figure 2.32 shows the coder used in IEEE 802.11a [10].

Puncturing is a method to achieve high data rates by just omitting some of the encoded output bits in the transmitter. Higher coding ratios of 2/3 and 3/4 are obtained by puncturing the 1/2 rate code. When 2 out of 6 bits are omitted the resulting rate becomes 3/4 and when 1 out of 4 bits is omitted the result gives a code with a rate 2/3. The Viterbi algorithm is the best possible choice for decoding since the algorithm tries the best decoding by maximum likelihood code word estimation. The data is coded with a convolutional encoder of coding rate R, corresponding to the desired data rate as seen in Table 2.5.

Mode	Modulation	Code Rate	Data Rate	BpS
1	BPSK	1/2	6 Mbps	3
2	BPSK	3/4	9 Mbps	4.5
3	QPSK	1/2	12 Mbps	6
4	QPSK	3/4	18 Mbps	9
5	16-QAM	1/2	24 Mbps	12
6	16-QAM	3/4	36 Mbps	18
7	64-QAM	2/3	48 Mbps	24
8	64-QAM	3/4	54 Mbps	27

Table 2.5: Eight PHY modes of the IEEE 802.11a PHY

Interleaving

Performance of codes designed for AWGN channel rapidly degrades in frequency selective channels with correlated channels. Interleaving is a method to separate bits in time and frequency. Block interleaving is used for a length of one OFDM symbol since the channel is basically be quasi-static and assumed to stay the same for duration of a symbol. OFDM systems leverage frequency diversity since they are wideband systems. Block interleaving is done by a matrix where the input is written in columns and the output is read in rows. For 6×8 Matrix, the interleaving depth is 48 [10].

Modulation

Instead of sending one bit in one symbol, multiple bits can be sent by converting more than one bit into a complex number bigger than 1 or 0. Modulation schemes have constellation diagrams that map multiple bits into complex numbers. The type of constellation diagram affects the bit error rate (BER), peak to average power ratio (PAPR) and RF spectrum shape. IEEE 802.11a uses coherent modulation where there is phase lock be-

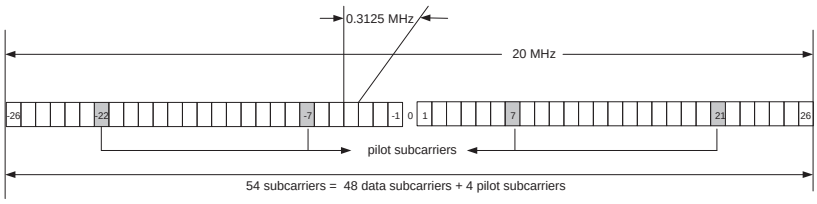


Figure 2.33: OFDM subcarrier allocation

tween transmitter and receiver. This helps the receiver convert complex numbers into bits easily with a complex receiver architecture. The number of data bits per OFDM symbol is calculated from desired data rate and coding rate. For instance for a rate of 54Mbps 64-QAM is used with a rate of 3/4. This corresponds to 6 bits in each subcarrier and 6x48x3/4 (27bytes) for an OFDM symbol when number of data subcarriers is 48. Refer to Table 2.5 for other data rates.

OFDM Symbol

In an OFDM symbol, IEEE 802.11a uses subcarriers of an IFFT block of length 64. Forty-eight subcarriers are occupied with modulated symbols and 4 subcarriers are used for pilot symbols to make the coherent detection overcome frequency offsets and phase noise problems. The pilots are BPSK modulated and inserted in -22, -7, 7, 22 subcarriers and the center subcarrier at 0 is not used as seen in Figure 2.33. The remaining 12 subcarriers are not used due to bandwidth limitations.

In the data stream, each 48 modulated symbol is used to make an OFDM symbol. The output of IFFT is prepended with a circular extension to create a guard interval (GI). The guard interval is used to avoid ISI from the previous frame. The total OFDM symbol is a guard interval plus IFFT block. The guard interval is one fourth of FFT length in IEEE

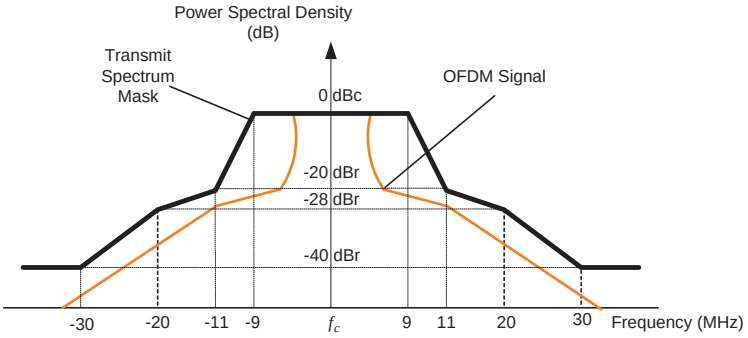


Figure 2.34: Transmitter spectrum mask

802.11a. Table 2.6 shows the key parameters of IEEE 802.11a. An adequate number of pad bits are appended to create data in multiples of OFDM symbols.

Pulse shaping has an important role in robustness of the system against phase noise and frequency offset as explained in [6]. In addition, pulse shaping provides a smoother transmit signal spectrum which is critical in an OFDM system due to higher peak-to-average ratio. Transmitter output has to meet strict out-of-band leakage criteria. A spectrum mask, such as the one shown in Figure 2.34, is usually defined to control the effect of bandwidth regrowth caused by amplifier clipping and limits the signal band to 30 MHz. A typical amplifier non-linearity model used in simulation and analysis is:

$$\begin{aligned} v_o &= \frac{v_i}{(1+(\frac{|v_i|}{v_s})^{2p})^{1/2p}} \\ back-off &= -10log\frac{Avg(|v_i|^2)}{|v_s|^2} \end{aligned} \tag{2.3}$$

where p is smoothness factor and v_s is saturation voltage. A higher power amplifier (HPA) is used in wireless systems [25]. The HPA is modeled by a memoryless nonlinear subsystem; it is shown in Figure 2.35.

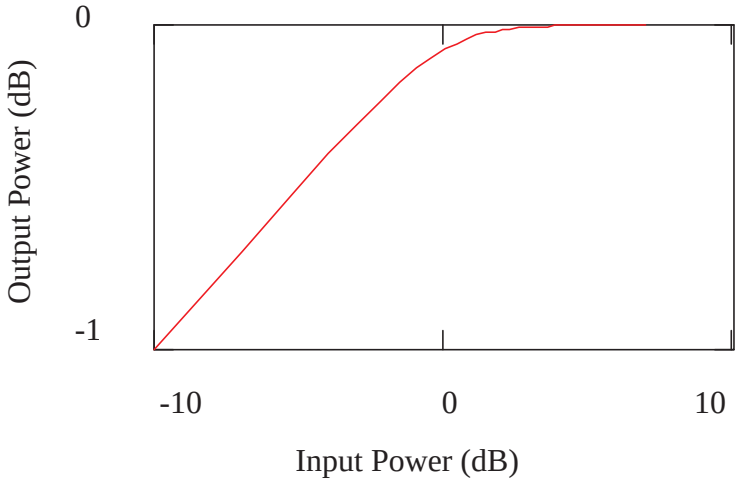


Figure 2.35: Nonlinear HPA model

Parameter	Value
N_{SD}:Number of data subcarriers	48
ND_{SP}: Number of pilot subcarriers	4
N_{ST}: Number of subcarriers, total	52
Δ_F: Subcarrier frequency spacing	0.3125 MHz (=20MHz/64)
T_{FFT}: IFFT/FFT period	3.2μs (1/Δ_F)
$T_{PREAMBLE}$: PLCP preamble duration	16 μs (T_{SHORT}+T_{LONG})
T_{SIGNAL}: Duration of the SIGNAL	4.0 μs (T_{GI}+ T_{FFT})
T_{GI}: GI duration	0.8 μs (T_{FFT}/4)
T_{GI2}: Training symbol GI duration	1.6 μs (T_{FFT}/2)
T_{SYM}: Symbol interval	4 μs (T_{GI}+T_{FFT})
T_{SHORT}: Short training sequence duration	8 μs (10x$T_{FFT}/4$)
T_{LONG}: Long training sequence duration	8 μs (T_{GI2}+2xT_{FFT})
W:Signal Bandwidth	16.66 MHz

Table 2.6: Key Parameters of the IEEE 802.11a

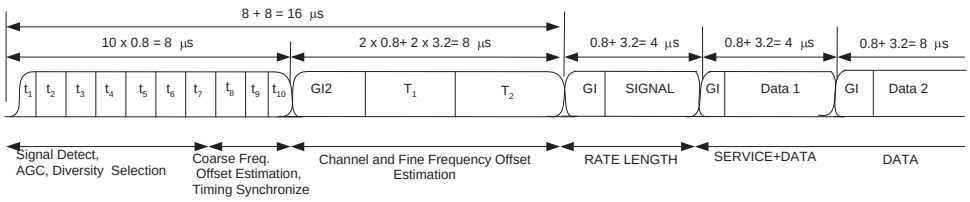


Figure 2.36: Format of an OFDM frame © IEEE

(2) From OFDM Symbol to OFDM Packet

The resulting OFDM symbols are appended one after another and data field of a frame is constructed. A typical frame structure is shown in Figure 2.37. Figure 2.36 shows the frame in OFDM symbols. Framing occurs in the PLCP layer and the frame structure includes a PLCP preamble, PLCP header, and PLCP data (PSDU and tail and pad bits).

PLCP Preamble

The PLCP preamble contains 10 repetitions of a "short training sequence (STS)" and two repetitions of a " long training sequence (LTS)". STS is used for automatic gain control (AGC) convergence, diversity selection, timing acquisition, and coarse frequency acquisition. LTS is used for channel estimation and fine frequency acquisition.

STS consists of 12 subcarriers as follows:

$$\begin{aligned} S_{-26,26} = \sqrt{52/(2 \cdot 12)} \times \{ \quad & 0,0,1+j,0,0,0,-1-j,0,0,0,1+j, \\ & 0,0,0,-1-j,0,0,0,-1-j,0,0,0, \\ & 1+j,0,0,0,0,0,0,0,-1-j,0,0, \\ & 0,-1-j,0,0,0,1+j,0,0,0,1+j, \\ & 0,0,0,1+j,0,0,0,1+j,0,0\} \end{aligned} \tag{2.4}$$

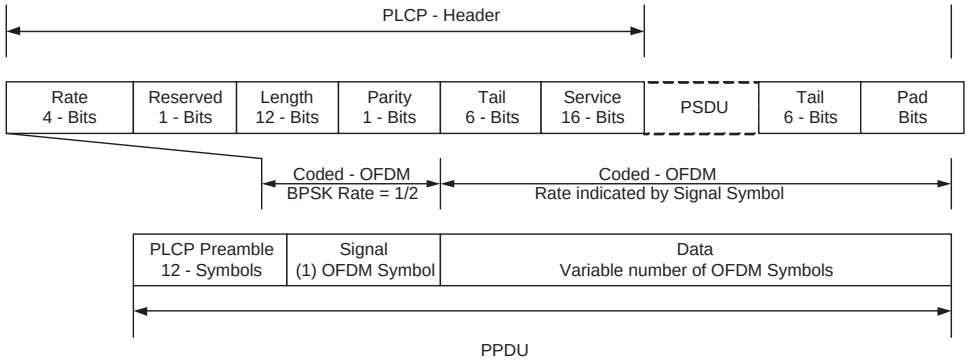


Figure 2.37: Logical representation of an OFDM frame © IEEE

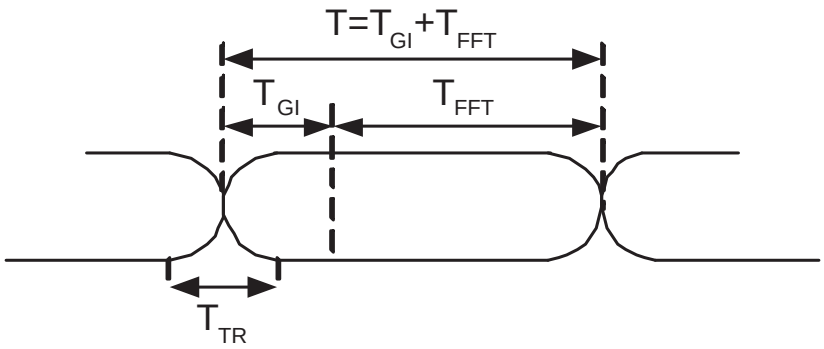


Figure 2.38: Format of an OFDM frame © IEEE

where $\sqrt{52/(2 \cdot 12))}$ is used for energy normalization. Periodicity of $T_{FFT}/4 = 0.8\mu s$ in the time domain is achieved after taking a 64-point IFFT of the sequence $S_{-26,26}$.

LTS consists of 53 subcarriers and is generated directly by applying the IFFT to the following training sequence:

$$\begin{aligned} L_{-26,26} = \{ \quad & 1,1,-1,-1,1,1,-1,1,-1,1,1,1,1,1, \\ & 1,1,-1,-1,1,1,-1,1,-1,1,1,1,1,1,0, \\ & 1,-1,-1,1,1,-1,1,-1,1,-1,1,-1, \\ & -1,-1,1,1,-1,1,-1,1,-1,1,1,1,1\} \end{aligned} \tag{2.5}$$

Time-windowing function, $w_T(t)$ is applied to provide smoothed transition. $w_T(t)$ is given by:

$$w_T(t) = \begin{cases} sin^2(\frac{\pi}{2}(0.5 + \frac{t}{T_{TR}})) & -\frac{T_{TR}}{2} < t < \frac{T_{TR}}{2} \\ 1 & \frac{T_{TR}}{2} < t < T - \frac{T_{TR}}{2} \\ sin^2(\frac{\pi}{2}(0.5 - \frac{(t-T)}{T_{TR}})) & T - \frac{T_{TR}}{2} < t < T + \frac{T_{TR}}{2} \end{cases} \quad (2.6)$$

where T_{TR} is illustrated in Figure 2.38 and continuous time short OFDM training symbol is defined as:

$$r_{short}(t) = w_T(t) \sum_{k=-N_{ST}/2}^{N_{ST}/2} S_k \exp(j2\pi k\Delta_F t). \quad (2.7)$$

PLCP Header

The PLCP header field contains rate, length, parity, tail, and service fields. Except for the service field, the PLCP header is sent in one OFDM symbol called SIGNAL and coded with BPSK with a coding rate of R=1/2. Rate and length fields help CCA to determine the busy time of the station even if the station is not the addressee.

PLCP Data

Service field and PSDU (with tail and pad bits) are sent in multiple OFDM symbols at the data rate described in the rate field.

The finite state machine of the PLCP/PMD structure is shown in Figure 2.39. From the figure, the MAC layer notifies PLCP/PMD about a packet transmission. The PLCP/PMD layer confirms the transmission and the MAC layer sends the data, and the PLCP/PMD layer notifies the MAC layer after finishing transmission. This ending shifts the receiver

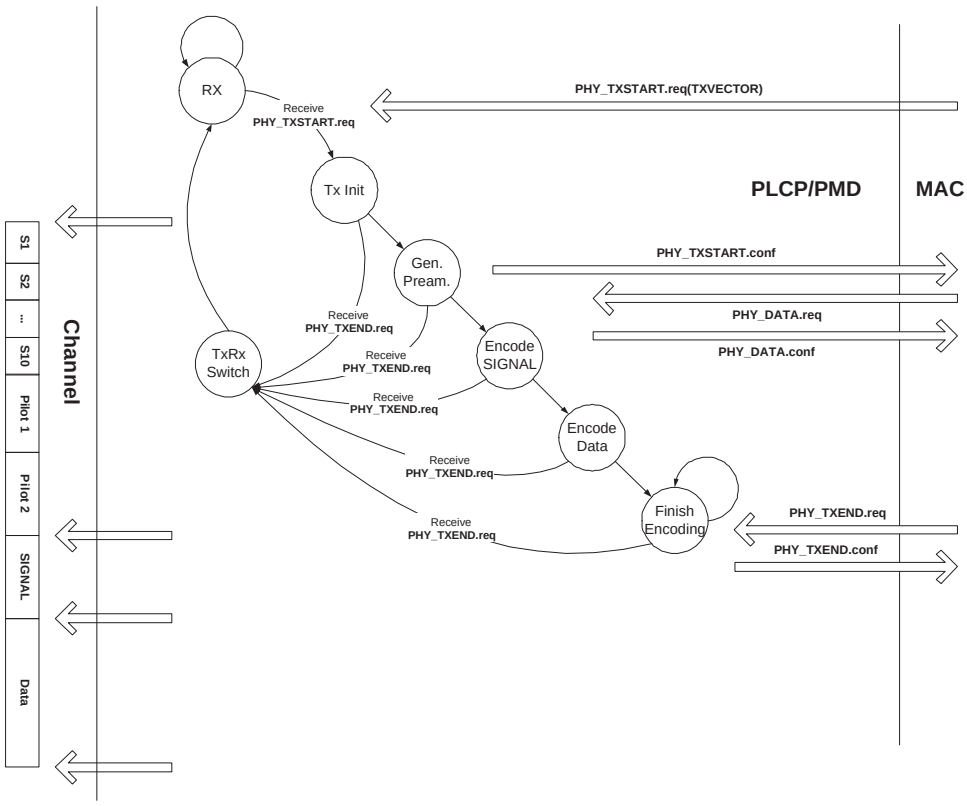


Figure 2.39: IEEE 802.11a PLCP/PMD transmitter finite state machine

from transmit state to the receive state.

(3)From OFDM Packet to Radio Waveform

The baseband signal is converted into a bandpass signal in the RF system. The digital data stream is split into two and each stream is converted to analog with a digital-to-analog converter (DAC) and a baseband (BB) filter is applied. Analog streams are multiplied by sine and cosine waveforms to create the modulated intermediate frequency (IF) signal in in-phase and quadrature (IQ) modulator. The IF signal is brought up to the RF signal by a local oscillator of f_c frequency and a multiplier. After passing the signal through RF filter to make the signal reside only in the approved bandwidth, the signal is relieved into air.

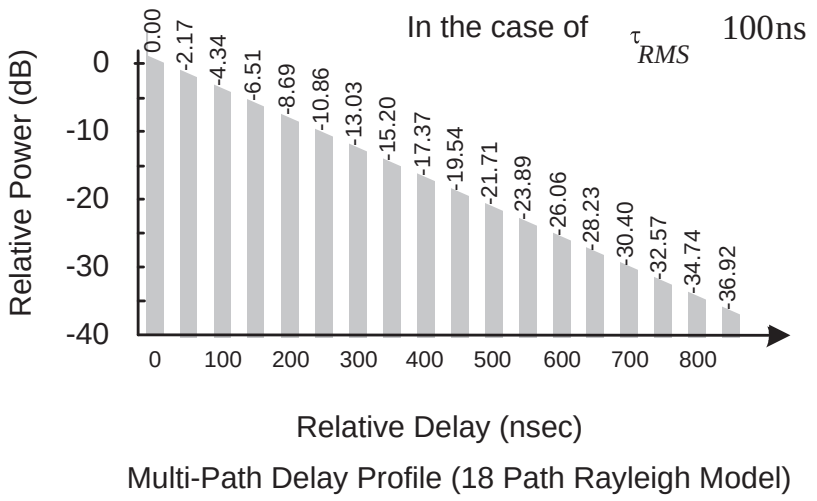


Figure 2.40: Channel impulse response for typical wireless LAN medium

The wireless channel model considered for WLAN system is illustrated in Figure 2.40.

2.9.3 Receiver

(4) From Radio Waveform to OFDM Symbol

Frame Detection

Initially, the receiver must detect the frame, adjust AGC to the proper level, utilize any diversity capability, and adjust for coarse frequency offset. Fast synchronization is critical in this application. High phase noise and frequency offset of low cost receiver oscillators (with about $\pm$10 ppm frequency stability) interfere with the synchronization process and require robust algorithm design.

A typical AGC block consists of a correlator followed by a confirmation block. Since the received signal level could vary significantly due to shadowing, a low threshold should

be used followed by a confirmation block, which utilizes the repetitive pattern of preamble sequence to reduce false alarm probability.

Probability of false alarm and detection is a function of threshold and signal-to-noise ratio. Probability of detection versus false alarm for different thresholds is referred to as Receiver Operation Characteristics (ROC). If more than one training sequence is used for synchronization, the probability of false alarm decreases while probability of detection improves. The overall P_F^t and P_D^t will be:

$$\begin{aligned} P_F^t &= \sum_{i=0}^{t} \begin{pmatrix} N \\ i \end{pmatrix} p_f^i (1-p_f)^{N-i} \\ P_D^t &= \sum_{i=t+1}^{N} \begin{pmatrix} N \\ i \end{pmatrix} p_d^i p_m^{N-i}, \end{aligned} \tag{2.8}$$

where N is the number of correlations and t is the threshold determined by required probabilities of false alarm and detection.

The preamble field of OFDM packet is dedicated to facilitate packet detection. The double sliding window algorithm is used to leverage the periodicity of short training symbols. The receiver keeps two windows and observes energy. The first window is used to keep the cross-correlation between the signal and its delayed version, and the second window is used to keep the power level of the received signal. When there is a packet, the ratio of first to second shows a jump otherwise it is approximately equal to zero since the cross-correlation of noise is 0.

Figure 2.41 shows the finite state machine of PLCP/PMD receiver. Energy detection

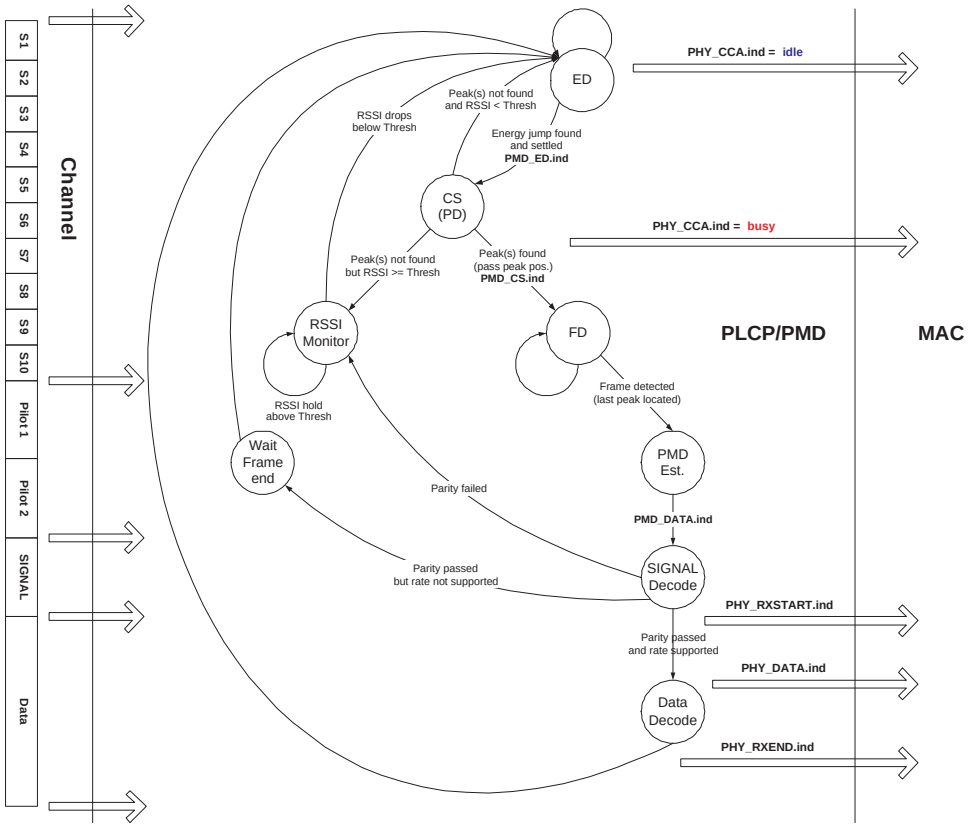


Figure 2.41: IEEE 802.11a PLCP/PMD receiver finite state machine

(ED) looks for a peak and goes to the carrier sense (CS) field. If the observed energy jump is higher than the threshold, the PLCP/PMD layer sends notification to MAC layer to inform it that the channel is busy. If the CS field continues to observe peaks than it concludes that this is a frame and the finite state machine passes to frame detection (FD) state. Otherwise, it goes to RSSI state wherein the signal strength of the received signal is observed until the peak goes below threshold. The PLCP/PMD layer starts decoding the signal and sends an ending notification to MAC layer when the process is done.

Symbol Detection

Symbol timing synchronization has the task of finding the precise moment of the OFDM symbol after the packet detector has provided an estimate of the start edge of the packet. Correlation is further exploited to achieve precise OFDM symbol detection. Since the receiver knows the training symbols, any high correlation can be expected as the start of the first OFDM symbol.

Clock Synchronization

Mismatch in sampling instances of the digital-to-analog converter (DAC) and analog-to-digital converter (ADC) causes slightly shifted versions of the symbols to be detected which degrades the performance. Pilot carriers inserted symmetrically into the FFT block are used to detect the sampling frequency error and the necessary adjustment is done in the receiver structure.

There are two approaches to do frequency offset estimation: first one uses the short and long training symbols and the second one uses cyclic prefix. Frequency offset esti-

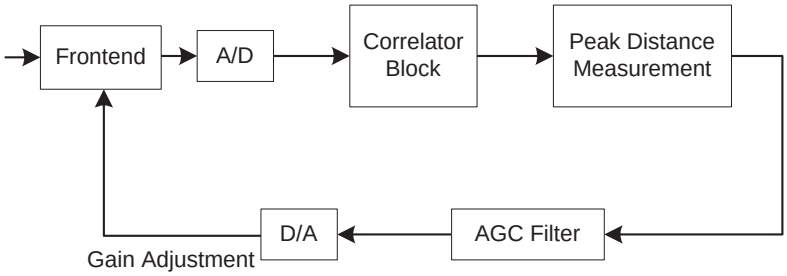


Figure 2.42: Frame synchronization and AGC

mation cannot perfectly eliminate the offset; there remains some residual frequency error resulting in constellation rotation. The main purpose of inserting four pilots is to help the receiver track the carrier phase.

The frequency offset estimator has an operating range which determines how large a frequency offset can be estimated. The limit can be defined as follows:

$$|f_\Delta| \leq \frac{N}{2D}\Delta_F \tag{2.9}$$

where $\Delta_F = 0.3125$ MHz is the subcarrier frequency spacing and D is the delay between the identical samples of two repeated symbols. For the short training symbols (D=16), the limit is 625 kHz (coarse) and for the long training symbols (D=64), the limit is 156.25 kHz (fine). In IEEE 802.11a, the standard specifies a maximum oscillator error of 20 parts per million (ppm) and if the channel is 5.3 GHz then corresponding frequency offset is:

$$|f_\Delta| = 40 \times 10^{-6} \times 5.3 \times 10^9 = 212kHz \tag{2.10}$$

which is 156.25 kHz (fine) $<$ 212 kHz $<$ 625 kHz (coarse). If the initial frequency offset

is very high, it may reduce the magnitude of peak and increase probability of false alarm. A frequency offset of f_Δ amounts to peak reduction of:

$$\left| \sum_{n=1}^{N} e^{j2\pi nTf_\Delta} \right| = \frac{sin(\pi NTf_\Delta)}{sin(\pi Tf_\Delta)} \tag{2.11}$$

In this case, a peak detector can be implemented as a bank of correlators tuned to different frequency offset values. Therefore, frame synchronization, coarse frequency offset estimation and AGC can be implemented using a common structure of Figure 2.42.

(5) From OFDM Symbol to Binary Output

Once initial acquisition is accomplished, the detection process begins. The detection process requires channel estimation, prefix removal, frequency domain equalization and soft decoding.

Channel Estimation

In WLAN, the channel is assumed to be stationary for the duration of the packet. Long training symbols (LTS) in the preamble are used to estimate the channel. This type of estimation is called block type channel estimation where all subcarriers are pilot symbols. The channel for each subcarrier can be estimated easily and since there are two LTS, the average of the two gives the possible applicable channel to the frame. Typically, the channel estimation block has the structure of Figure 2.43.

After channel estimation and fine synchronization, guard intervals should be removed. The position of the OFDM block should be adjusted carefully to make sure that the effect

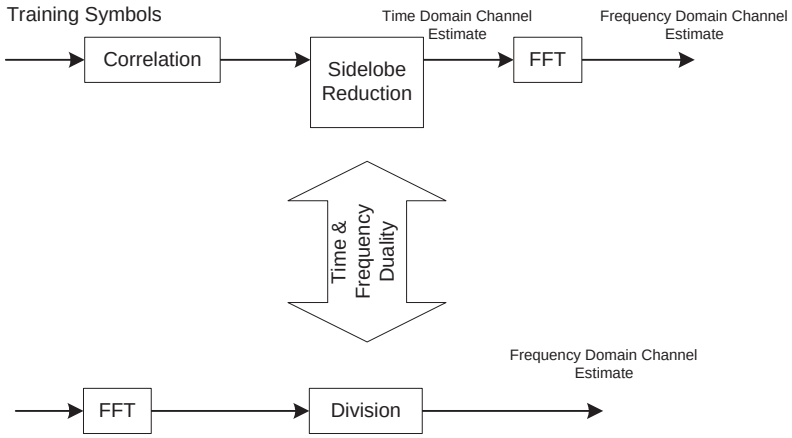


Figure 2.43: Channel estimation block

of pulse shaping is preserved and interference from adjacent blocks is minimum.

A reliability index concept is used to perform soft decoding of the demodulated signal. Due to the presence of an interleaver block a classic trellis decoding is not possible. A reliability parameter is assigned to each bit and used in decoding. The reliability index for each bit is derived from the difference in the metric value caused by bit flipping.

After deinterleaving and viterbi decoding, descrambling and source decoding is applied. This sequence of processes gives the binary output.

2.9.4 Degradation Factors

Some typical degradation factors at the transmitter and receiver are briefly discussed here and performance results are shown in the next section. Quadrature phase error and phase imbalance of transmitter modulation result in a C/N degradation and higher packet error rate. Another source of degradation is the fixed-point error caused by truncating the output of the IFFT butterfly. After each stage of IFFT (or FFT) butterfly, the required number of

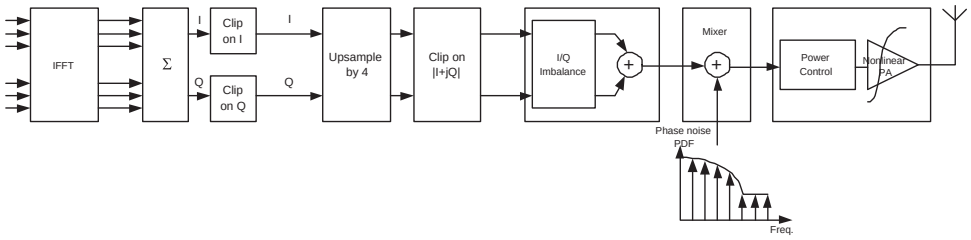


Figure 2.44: Simplified schematic of the IEEE 802.11a simulation model for the transmitter [26]

bits grows. Truncation of bits can also increase the packet error rate. Maximum tolerable transmitter degradation is restricted by vector error requirements set forth in wireless LAN standards.

Typical degradation sources in the receiver are phase noise of the local oscillator and D/A quantization error. Overall, transmitter power amplifier non-linearity and receiver phase noise are dominant factors for total vector error.

2.10 Performance of 802.11a Transceivers

IEEE 802.11a is highly sensitive to the impairments that are introduced by front-end components and channel. The components that cause distortion are phase noise, channel, frequency offset, IQ imbalance, quantization and clipping, power amplifier, narrowband systems, co-channels, ultrawide band systems. These impairments can lead to an irreducible error floor. They can be investigated with a MATLAB simulator provided by IMEC [27]. The simulator provides a generic front-end digital baseband and a set of multipath channels corresponding to an indoor environment. Figure 2.44 shows a simplified schematic of the transmitter simulation model and a description of the simulator obtained

from [26] is described as follows:

"Signals coming out of the IFFT is clipped to reduce large crest factors inherent to DMT modulation techniques. The spectrum of the OFDM signal occupies 16.8MHz in a 20MHz channel and the digital modem produces I and Q signals sampled at 20MHz [26]. A typical power spectrum obtained from the simulator is shown in Figure 2.45 and a FIR filter that is designed to provide 30dB power at the receiver is shown in Figure 2.45. To avoid high sample rates when simulating this bandpass signal at RF frequency of 5.2GHz, simulations are conducted at baseband (BB) using its complex low-pass equivalent representation. Oversampling by 4 is added in the interface between the modem and the front-end so that nonlinear distortion can be seen. In the final steps, the analog front-end model consists of three blocks: an I/Q modulator adding I/Q imbalance, a mixer with a local oscillator signal defined by its phase noise power density function, and power amplifier (PA) exhibiting a cubic nonlinearity. Similar BER degradation is produced in the receiver side also."

All the simulation results, unless otherwise noted, are for 8-bit quantization level, 4σ (rms value of the signal power) clipping level, 115dB amplifier input, AWGN channel, and 100 dB backoff.

2.10.1 Data Rate

The data rate performance of the IEEE 802.11a system is shown in Figure 2.47. Multipath channel is used to evaluate the modulation schemes in different coding gains where the channel changes in every symbol. Figure 2.48 shows the comparison of the performance

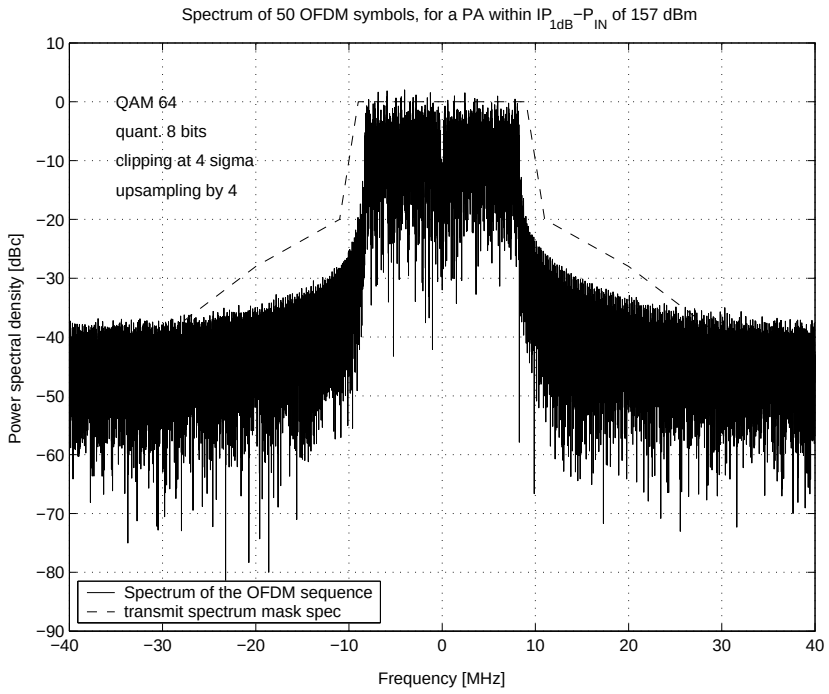


Figure 2.45: An example of power spectrum

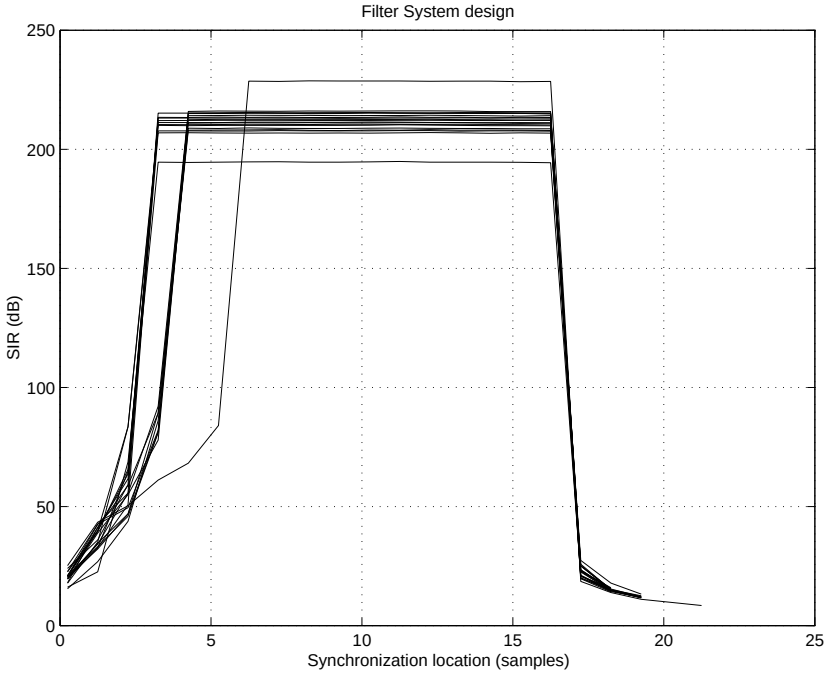


Figure 2.46: An example of FIR filter where the required SIR at the receiver is 30dB

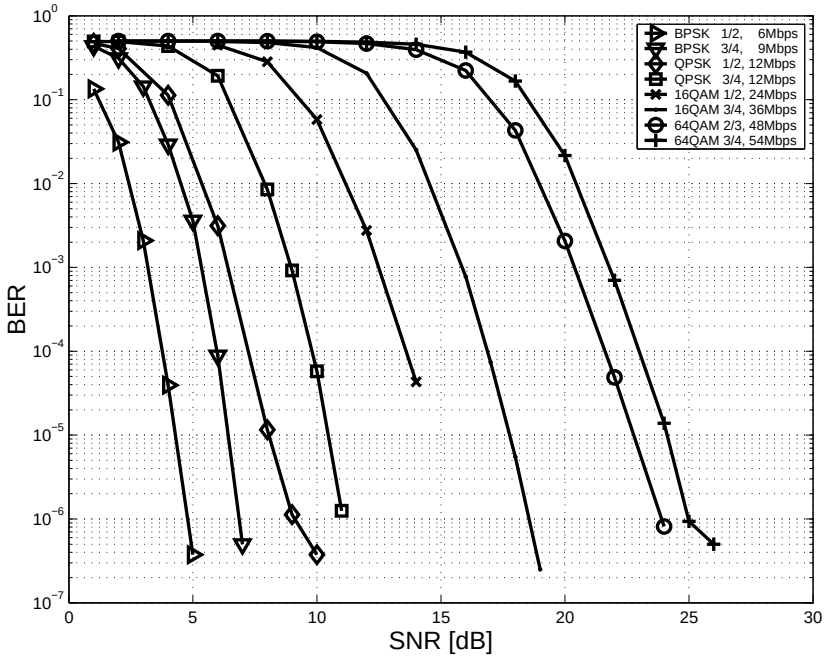


Figure 2.47: Bit error rate of IEEE 802.11a in Rayleigh channel

of an uncoded situation with theoretical values.

2.10.2 Phase Noise

Phase noise has two effects in OFDM. First, a random phase disturbance occurs in each symbol which is known as common phase error (CPE). All carriers suffer the same CPE. Second, inter-carrier interference (ICI) is introduced which takes the form of an additive noise-like signal. ICI is peculiar but CPE error needs to be compensated. CPE generates a rotation of the constellation and the error is compensated in the opposite direction if the level of rotation is known.

Figure 2.49 shows the BER performance with phase noise with the following frequency shaping: -65, -75 dB till 10 kHz offset, slope -20 dB/dec, and -135 dBc noise

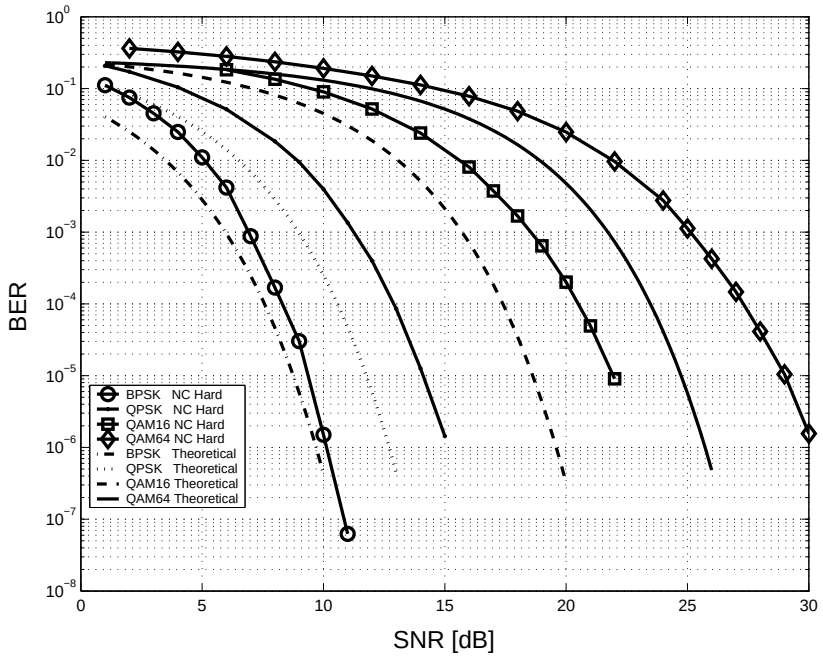


Figure 2.48: Bit error rate comparison of modulation schemes in AWGN

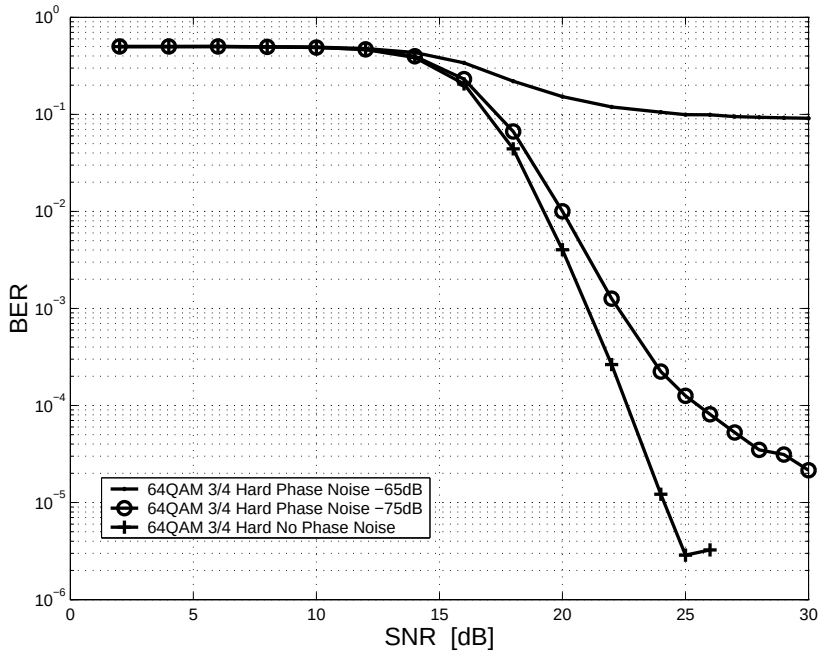


Figure 2.49: Phase noise

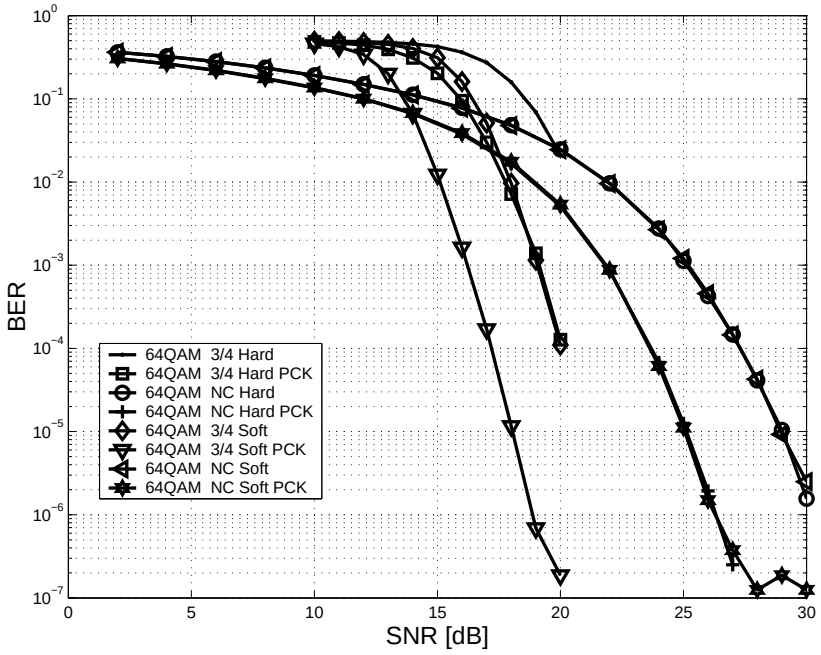


Figure 2.50: Channel estimation compared to Perfect Channel Knowledge (PCK)

floor. One can see the effect of phase noise from the figure.

2.10.3 Channel Estimation

Channel estimation in the IEEE802.11a is block type. Block type estimation estimates the channel in the first symbol and use that channel information for the rest of the packet.

Figure 2.50 shows the BER performance with or without perfect channel knowledge (PCK). Hard or Soft decision types are investigated under 64QAM (3/4) coded or no coded (NC) modulation. One can observe the effect of impairment that the channel introduces from the figure.

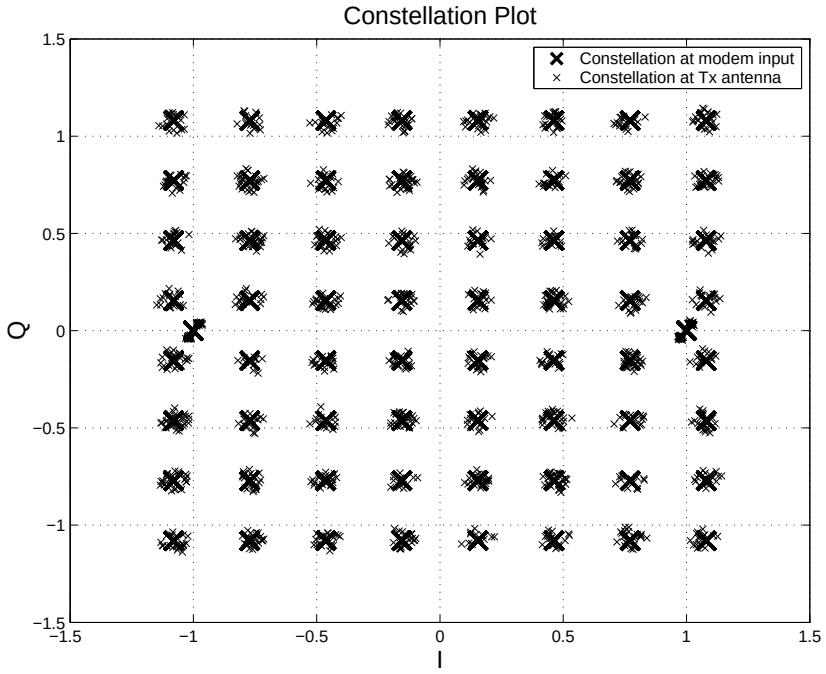


Figure 2.51: An example of constellation diagram with IQ Imbalance

2.10.4 Frequency Offset

One of the main drawback of OFDM is its sensibility to frequency offset which is caused by oscillator inaccuracies and Doppler shift induced by the channel. The effect of frequency offset is two fold. It reduces the signal power of the desired subchannel outputs and also introduces ICI. The ICI power can be significant even for small frequency offsets in OFDM due to the high sidelobe power of the subchannels.

2.10.5 IQ Imbalance

OFDM is very sensitive to receiver IQ mismatch since. A constellation diagram with IQ mismatch is shown in Figure 2.51 [28].

Figure 2.52 shows the BER performance when there is no coding and no perfect chan-

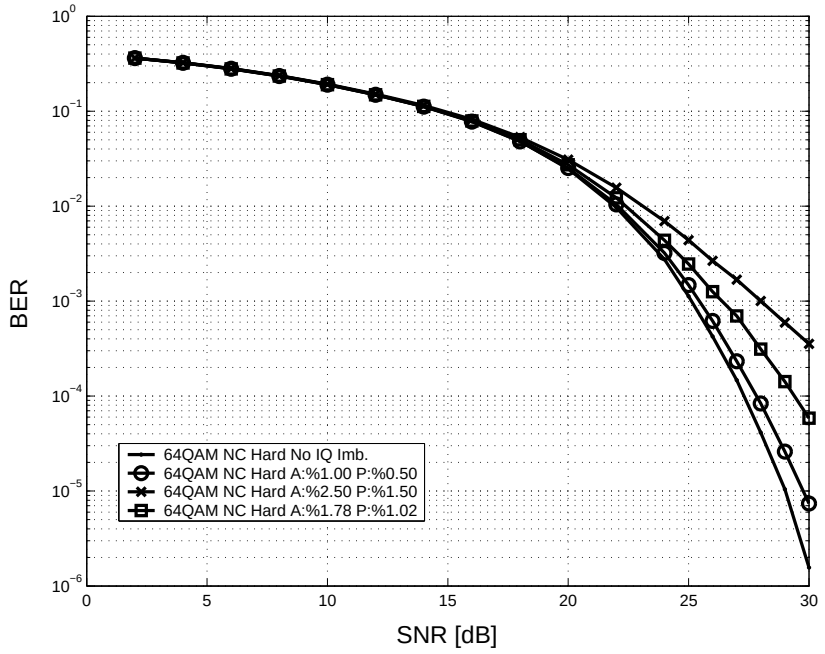


Figure 2.52: IQ imbalance when there is no coding and no perfect channel knowledge

nel knowledge. Amplitude and phase mismatch is represented in percentages. Figure 2.53 shows the BER performance when there is coding and perfect channel knowledge. As it can be seen, IQ mismatch effects are more severe when there is no coding.

2.10.6 Quantization and Clipping Error

The Quantization level (QL) has a major impact both on implementation cost and performance limitation. As QL decreases, the power consumption and the complexity of the D/A and A/D converters decreases at the expense of the quantization noise [26]. QL can be limited by the clipping level. Optimal QL and clipping levels can be found after performing a joint optimization process [26].

Figure 2.54 shows the BER performance for different clipping levels with 8-bit QL.

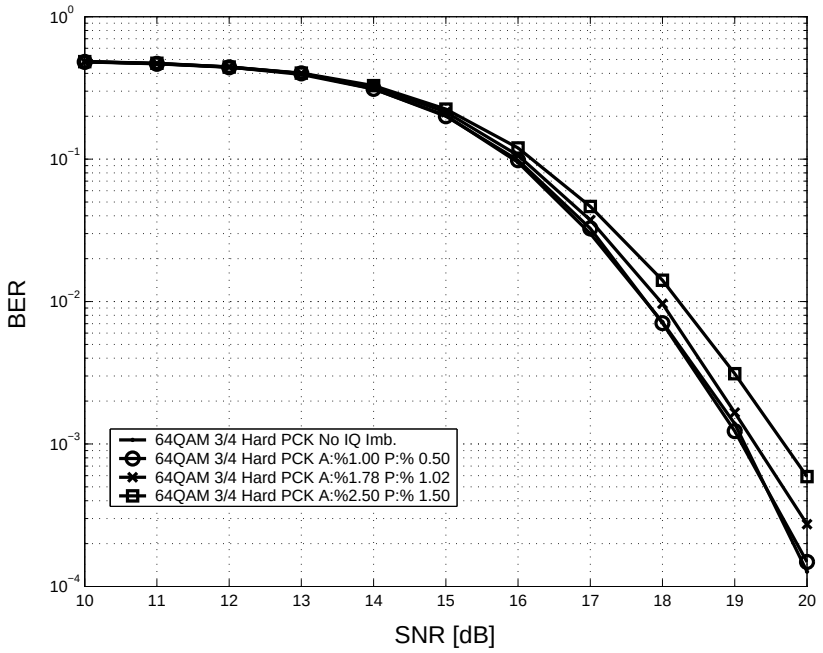


Figure 2.53: IQ imbalance when there is coding and perfect channel knowledge

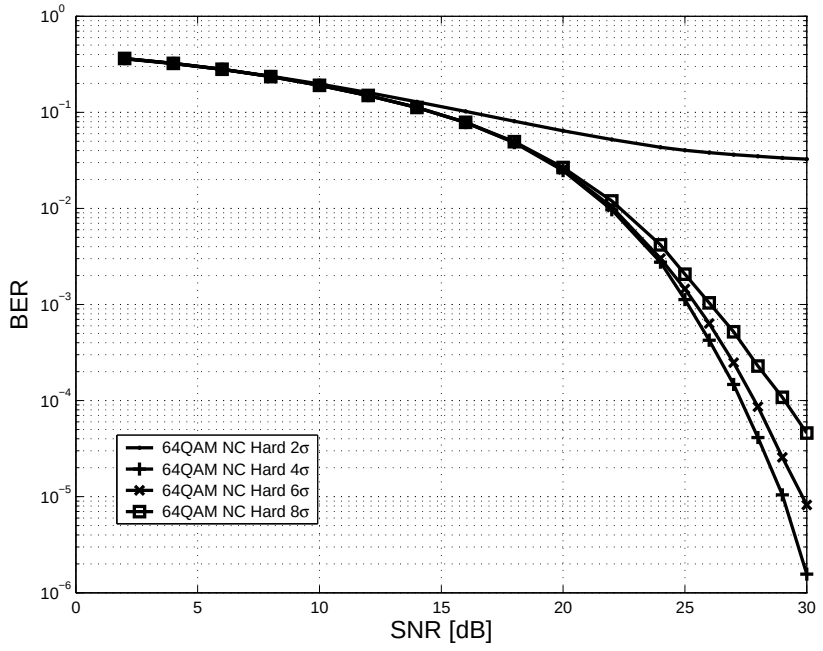


Figure 2.54: Impact of clipping threshold on performance

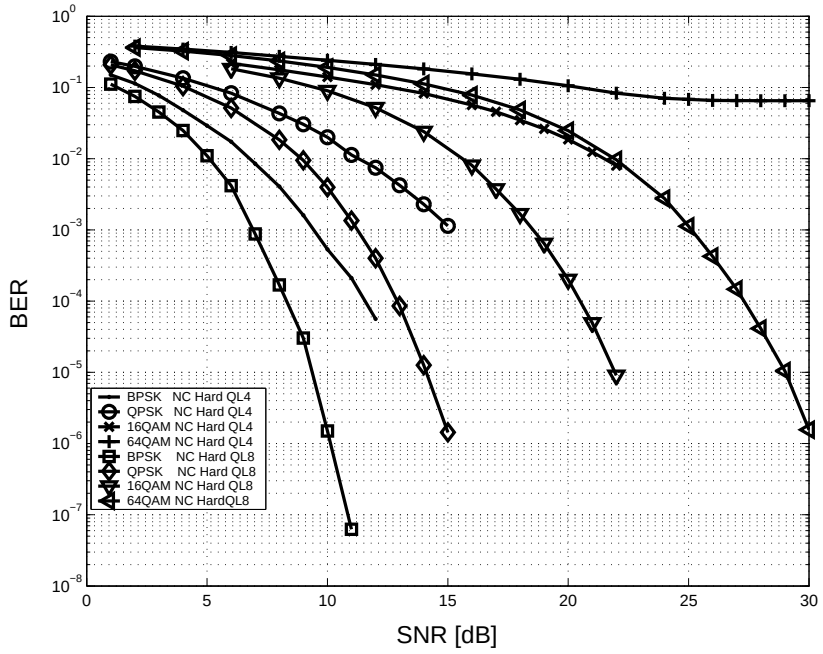


Figure 2.55: Quantization performance in IEEE 802.11a

The Clipping level is defined in terms of the rms amplitude σ of the time domain signal. From the figure, one can see that 4σ shows superior performance than 2σ, 6σ and 8σ which fortifies the need to do joint optimization.

Figure 2.55 shows the BER performance under different QL levels and modulation schemes with 4σ clipping level. Quantization error can be observed in the lower QL level. The 8-bit QL is found to be optimum with 4σ clipping level [26]. Figures 2.56 and 2.57 shows the BER performance of quantization and clipping with no quantization and clipping (NoQC) for QPSK and 16QAM modulations.

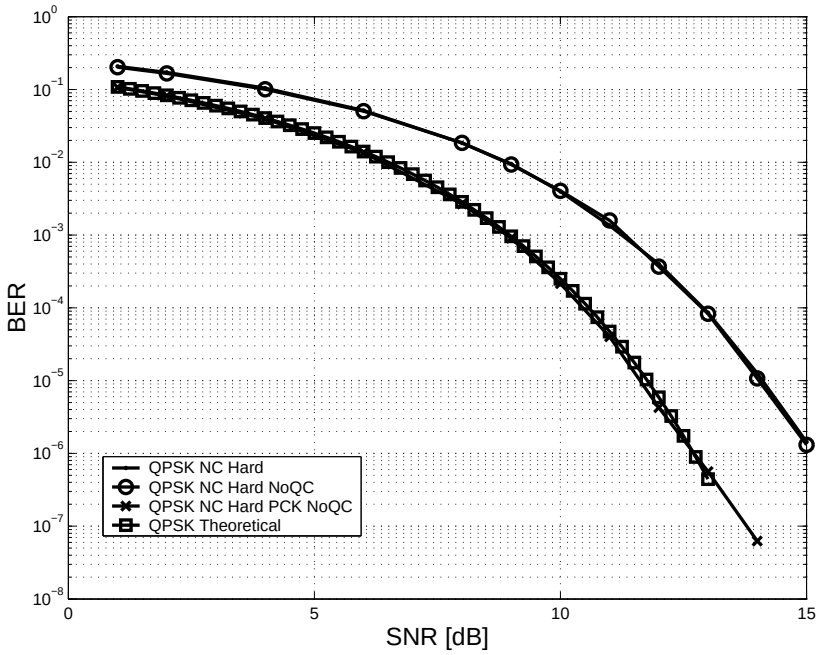


Figure 2.56: Quantization effect in QPSK

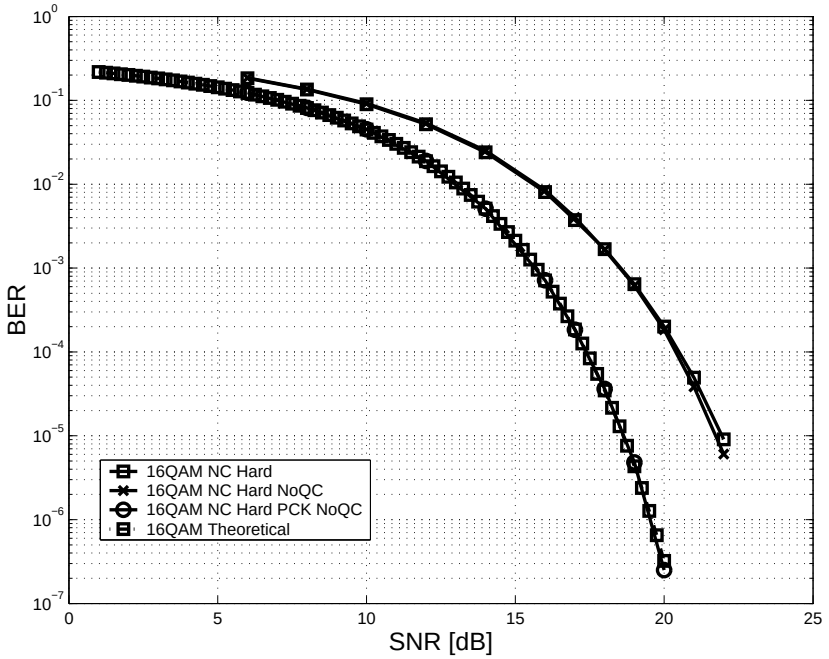


Figure 2.57: Quantization effect in 16QAM

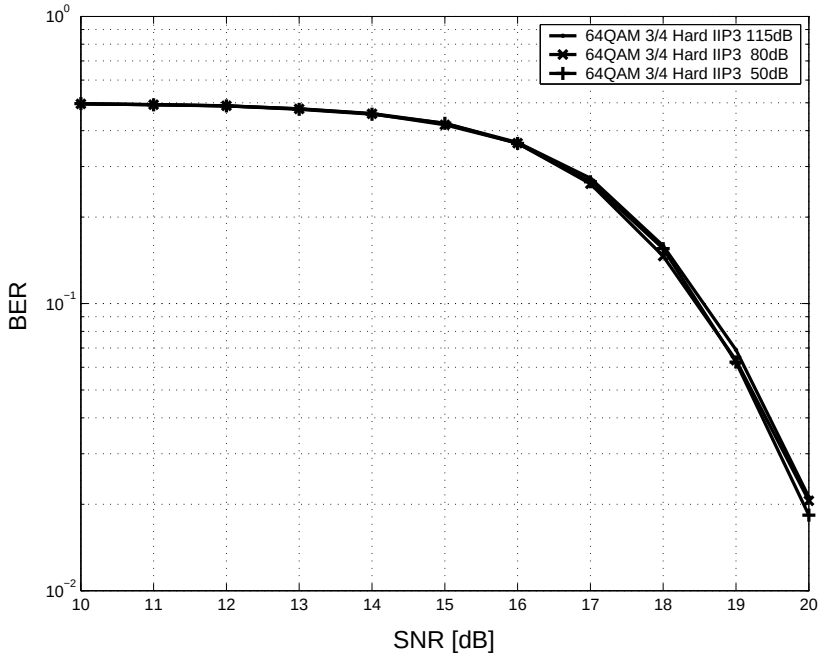


Figure 2.58: Nonlinear distortion in IEEE 802.11a

2.10.7 Power Amplifier Nonlinearity

OFDM has a sensitivity to nonlinear distortion, which causes crosstalk between subcarriers. The linearity of power amplifier (PA) is a significant metric for power consumption and BER. The nonlinear amplitude transfer function of a PA with a linear gain G, assuming that the PA is kept out of saturation, was modelled by a cubic nonlinearity:

$$y = G.(x - \alpha.x^3)\ for x < x_{sat} \tag{2.12}$$

where x_{sat} is the limit of the saturation region and α equals $\frac{4}{(3.IIP3^2)}$ [26]. Figure 2.58 shows the BER performance for different IIP3 values. If IIP3 is bigger than 100dB then a linear amplifier is applied.

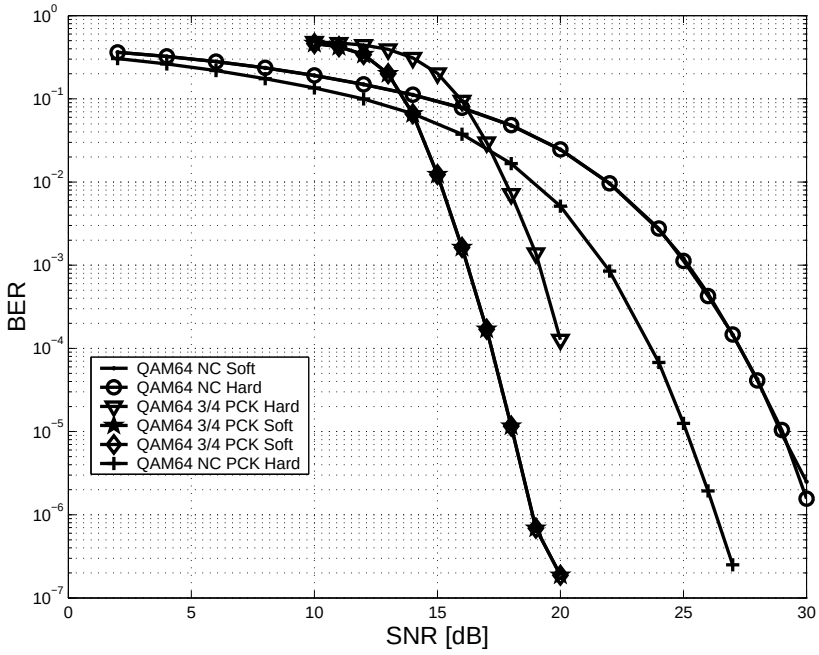


Figure 2.59: Hard or Soft decision coding

2.10.8 Hard or Soft Decision Decoding

Hard or soft decision comparison is shown in Figure 2.59. The performance of coded and uncoded schemes with or without perfect channel knowledge (PCK) is compared in this figure. As shown in the figure, when there is no coding, hard and soft decision decoding has almost no effect. When there is coding and perfect channel knowledge, soft decision decoding has superior performance.

2.10.9 Co-channel Interference

Co-channel interference degrades the performance of an OFDM system. Figure 2.60 shows the BER performance of the IEEE 802.11a system when there are 10 adjacent channels. Relative power of adjacent channels to the signal (AdjCh) is used to compare

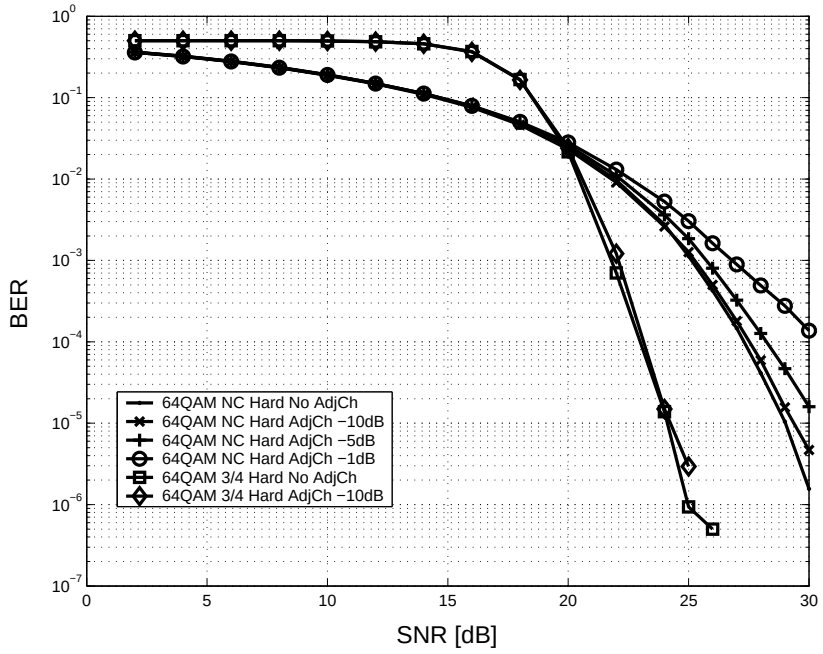


Figure 2.60: Co-channel interference

the effect of co-channels.

2.10.10 Narrowband Interference

The benefit of OFDM is its resistance to narrowband interference. Since the OFDM waveform is composed of many narrowband tones, a narrowband interferer will degrade the performance of a limited portion of the spectrum. The performance degradation can be corrected by forward error correction coding.

2.10.11 UWB Interference

Ultra Wide Band (UWB) operates in an unregulated band where some portion of it also overlaps the spectrum of IEEE 802.11a. Several interference mitigation methods are pro-

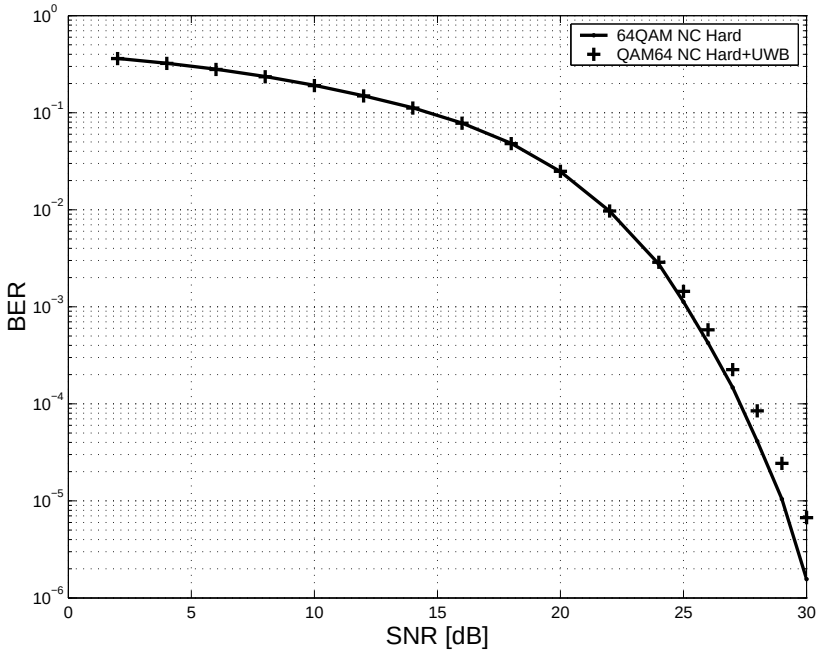


Figure 2.61: UWB interference

posed.

Using a notch filter filters out the spectrum of UWB pulses to suppress the content in certain bands. Impulsive noise is mitigated more effectively with OFDM. Since the OFDM symbols are long in duration, energy from low-level impulse noise can be integrated to a level benign to system performance [29].

UWB interference has an almost negligible effect on the OFDM receiver when it is transmitted in a time-hopping pattern. Time hopping patterns distribute the power among the spectrum and the power spectral density of the signal becomes almost equivalent to white Gaussian noise when the number of UWB users is large, due to the central limit theorem (See Figure 2.61). Concurrently, the degradation is higher when there is a small number of UWB signals [30]. The time-hopping sequence can only be applied to impulse

radio systems.

A Multiband approach is another solution. A Multiband approach divides the spectrum into subbands (See [6] for detail). As a result, sub-bands can be turned off to avoid interference; or it can be combined with the above-mentioned methods in each sub-band [31].

2.10.12 Performance of 64QAM

Figures 2.62, 2.63 and 2.64 show BER performance of various components of an IEEE 802.11a system: (3/4) coded or non coded (NC) schemes compared with or without perfect channel knowledge (PCK). (Soft) or (Hard) decision decoding are used to investigate the effect of quantization and clipping with no quantization and clipping (NoQL). IQ imbalance (IQimb.) is characterized by amplitude mismatch of 1.78% and phase mismatch of 1.02 degrees. In the figure, (Lorenz.) indicates the Lorenzian phase noise which is used in the local oscillator by a total integrated noise of -32dBc; (BO54) refers to back-off between the average power and the input power 1dB-compression point of 5.4dB; (pnc) is for phase noise compensation.

2.11 Rate Adaptation

Rate adaptation is a mechanism to select one out of available transmission rates at a given time. It is required to improve the performance of WLAN. IEEE 802.11b operating in 2.4GHz supports four PHY rates: 1, 2, 5.5 and 11Mbps; and IEEE 802.11a operating in 5GHz supports eight PHY rates: 6, 9, 12, 18, 24, 36, 48, and 54 Mbps. There is a tradeoff

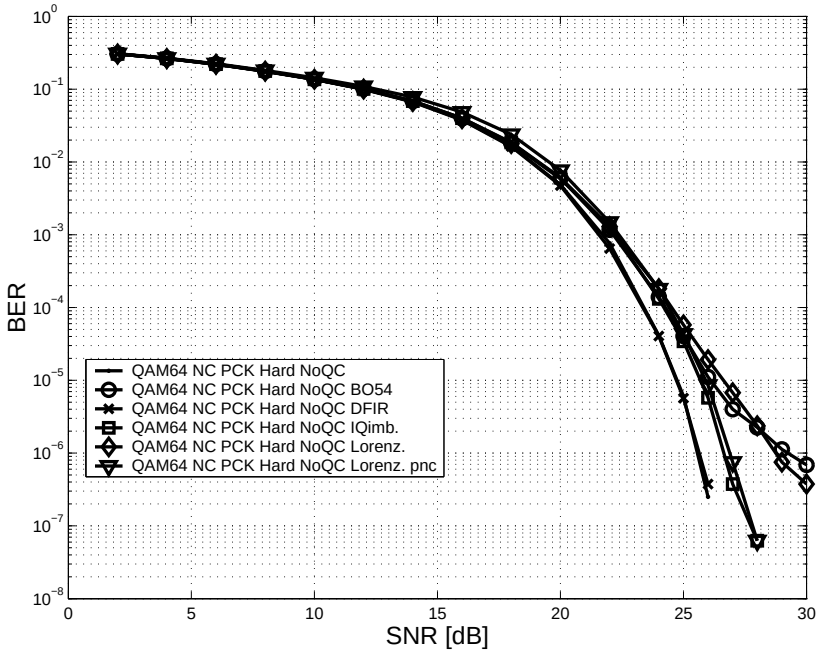


Figure 2.62: Impairments in 64QAM when there is no coding, and perfect channel knowledge

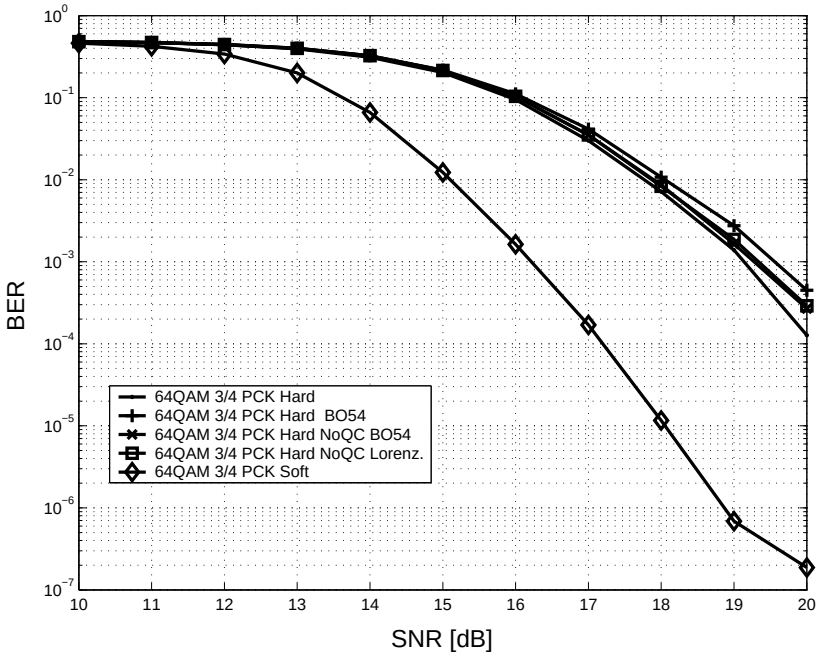


Figure 2.63: Impairments in 64QAM when there is coding and perfect channel knowledge

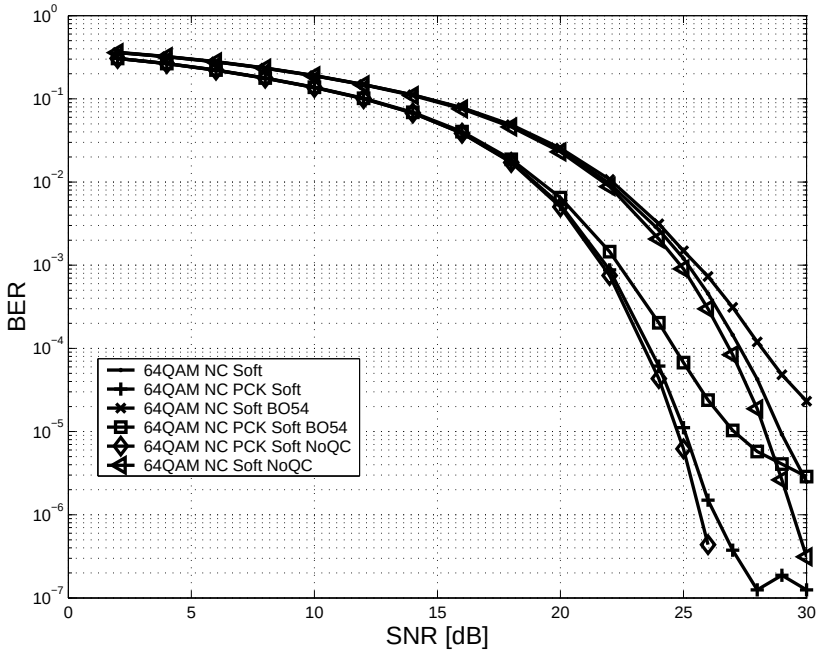


Figure 2.64: Impairments in 64QAM when there is no coding and soft decision decoding

between robust communication at lower rates and high speed communication at higher rates. The optimum performance is achieved by making the rate selection adaptive with regard to channel conditions.

The standard leaves the rate adaptation out of scope. It is useful to show some examples. Auto Rate Fallback (ARF) protocol alternates between 1 to 11 Mbps based on the timing function and missed ACK frames. After not receiving two consecutive ACKs, the data rate is reduced from the default 11 Mbps to 5.5 Mbps and the timer is reset. If there is a failure, the data rate falls back from 5.5 Mbps. When the timer expires or the received number of ACKs reach 10, the transmission rate is raised to the higher data rate and the timer is canceled. However, if an ACK is not received for the very next packet, the transmission rate is lowered again and the timer is reset.

Receiver-Based Auto-Rate (RBAR) is a proprietary mechanism which is based on RTS. The receiver selects a rate based on the signal strength of received RTS and notifies the transmitter by a CTS frame. Then, the transmitter responds with a data packet at the chosen rate [32].

The control, broadcast, and multicast frames are sent with rates defined in basic rate set and it should be compliant with all transceivers.

Remark

This chapter is published in Chapter 9 of "Multi Carrier Digital Communications: Theory and Applications of OFDM." Authors are Ahmad R.S. Bahai, Burton R. Saltzberg and Mustafa Ergen and the book is published by Springer-Verlag, New York, September 2004.

Chapter 3

Understanding the Analytical Markov Model

3.1 Introduction

Distributed Coordination Function (DCF) is the crucial component of the IEEE 802.11 standard. DCF employs Collision Avoidance/ Carrier Sense Multiple Access Scheme (CSMA/CA). This type of scheme is present in future standards such as IEEE 802.11e. Performance analysis of a Distributed Coordination Function has been investigated through analytical Markov models. The most famous model was introduced by Giuseppe Bianchi [5]. Most later models are built on that model [33, 34, 35, 36, 37, 38, 39]. In this chapter we present a clear understanding of the analytical model and the critical "independent assumption." We also present a new Markov model that closely follows the IEEE 802.11 standard.

The model in [5] considers only saturation throughput, meaning that the stations al-

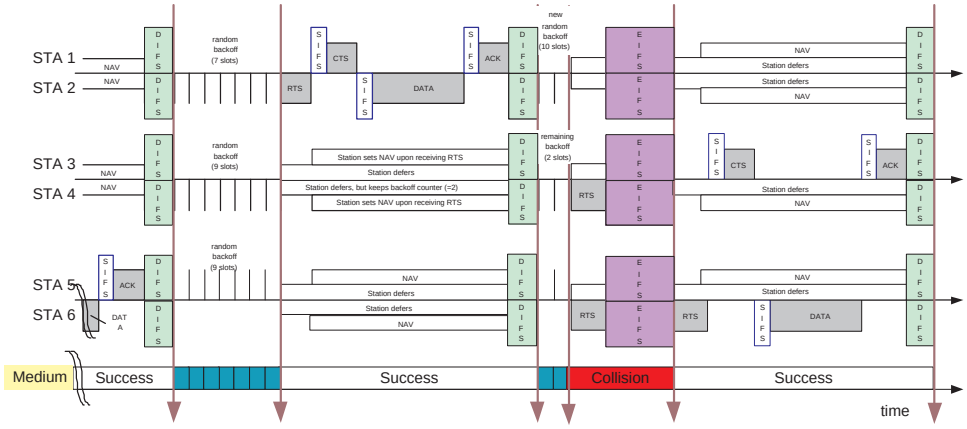


Figure 3.1: An illustration of DCF Mechanism: STA 6 selects a backoff counter 0 and transmits immediately.

ways have packets to transmit; non-saturated conditions are not considered. Furthermore, the models of [5] and [34] do not take into account freezing of the backoff counter due to consecutive transmissions: According to the standard, STA stops decrementing the backoff counter if there is an ongoing transmission, and resumes after waiting for at least DIFS time.

The model in [33, 40] approximates freezing of the backoff counter and also introduces an additional state for the carrier sense period. However, the carrier sense model in [33, 40] is inaccurate: since event definition is required and amount of time passed during a state change is not a fixed number but a random variable.

In this chapter we propose a Markov chain model that is closer to the standard, and compare it with previous models. Unlike previous models, which only consider the saturated case, we augment the chain with an additional module to consider non-saturated conditions. Our model also reflects the behavior when stations have different signal-to-noise ratio (SNR) levels, hence transmit at different data rates.

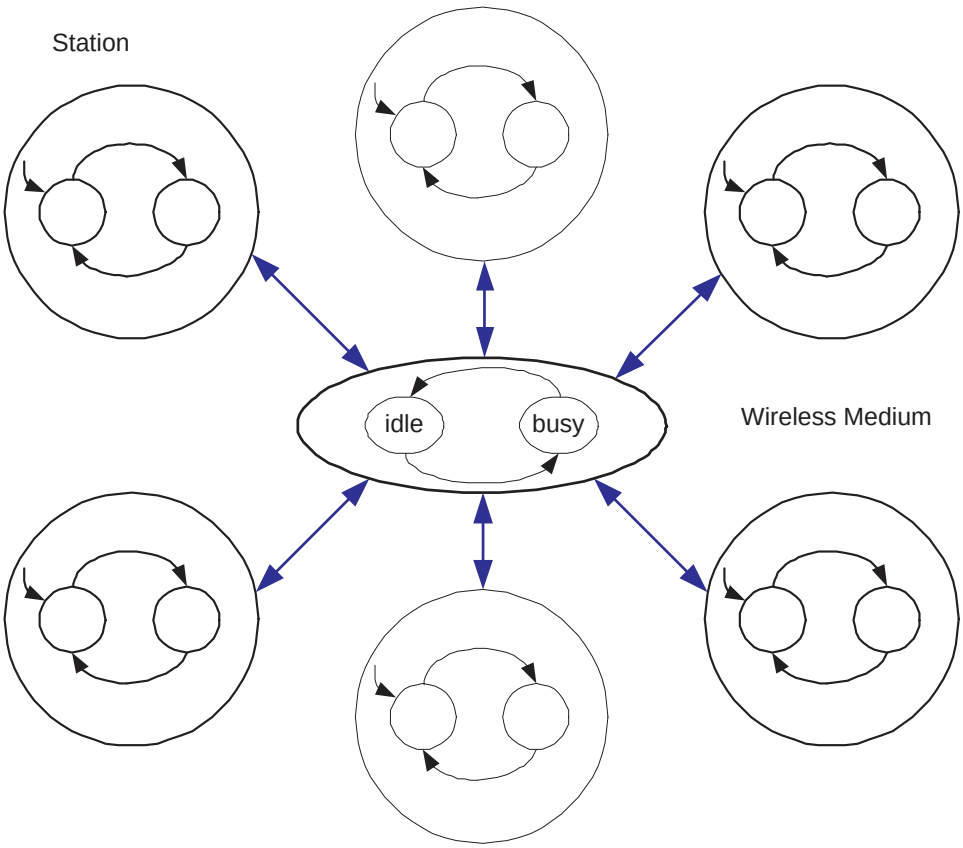


Figure 3.2: An illustration of a joint model of a DCF network

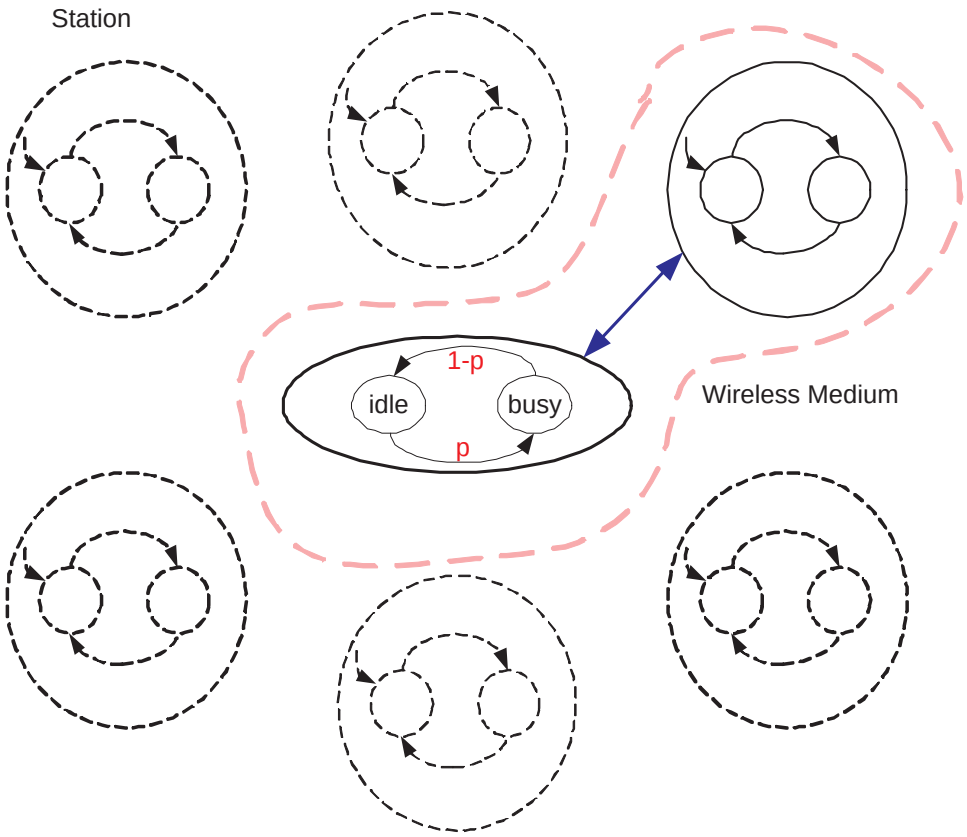


Figure 3.3: An illustration of an independent model of a DCF network

3.2 Markov Model

A good understanding of CSMA/CA is crucial to developing the Markov model. In CSMA/CA, stations change their state according to the condition of the wireless channel. The wireless channel is affected mostly by transmissions of stations. The wireless channel is modeled as "idle" or "busy", and is considered "idle" when stations do not transmit and "busy" if there is at least one transmission. If the channel is busy then the station clearly waits until the channel is idle. Hence, one station is dependent on others and the total behavior of all stations determines the wireless medium as seen in Figure 3.2. This implies that the finite state machine of a station is dependent on the states of other stations. Therefore, a joint model is required to understand network behavior. Unfortunately, this becomes very complex since a possible Markov model will have N^n states if each station has N states and there are n nodes in the medium.

Rather than constructing a complex Joint Markov model (JM), an abstraction of a wireless channel can be used to find a closed expression for the throughput and delay. If we model the channel as a two state Markov process, an individual station can use the channel information to determine its behavior independently from others.

Let's say the channel goes to the "idle" state from the "busy" with probability p and from "busy" to "idle" state with probability $q = (1 - p)$. Of course with p probability it stays in the "busy" if it is in the "busy" and with probability q it stays in the "idle" if it is in the "idle" state. Stationary probabilities for "idle" and "busy" states are $(1 - p)$ and p, respectively. The probability p represents the *probability of detecting the channel busy*. It also represents the *conditional collision probability* and we assume that it is independent

and constant, regardless of the number of retransmissions experienced.

This model represented in Figure 3.3 shows medium utilization can be observed by a station's behavior since it reflects the medium. The response it gets from the medium is given by this two state Markov model of the channel in which the condition is determined by the contribution of all stations.

If a station attempts to transmit, its transmission will be successful with probability $(1 - p)$ and unsuccessful with probability p. Depending on the response of the channel the station will alter its state. For example, if station is in backoff state it will decrement its counter if the wireless medium is idle with probability q, and stay in the same backoff counter with probability p.

Let's revisit the DCF mechanism with the assumption that each station has a packet to transmit all the time, this means each station operates in the saturation condition[1] [5]. The station backoffs if the channel is not idle for a DIFS time, and then it decrements its counter every time it experiences an idle slot time. When the backoff timer is zero the station transmits; if the transmission results in a collision then the station doubles its contention window and selects another backoff counter.

Now it is time to construct the Markov Model. Figure 3.1 shows the medium utilization from which we need to identify an event to trigger the state changes. Different observations may identify different events. The previous model introduced in [5] defines

[1] In the simulation, the saturated traffic condition is taken as the condition when a single node network is saturated. The stations operate in the saturated traffic condition which is obtained in OPNET by decrementing the packet arrival time of each station. They operate with 1Mbps data rate and saturation throughput is obtained when load is around 1.6 Mbps which is shown in Figure 3.4. As it can be seen, when the network saturates the stations obey the same characteristics and decreasing packet inter-arrival time brings the saturation point closer to the less number of stations.

an event as a virtual slot which could be an empty slot or a transmission-plus-an-empty slot which means that after each event the backoff counter is decremented. Of course this clustering assumes that there is no consecutive transmission and each transmission, whether it is successful or not is intervened by at least one empty slot.

The probability of having two consecutive transmission is equivalent to the ability of selecting zero backoff counter right after a transmission. Figure 3.5 shows the probability of having two consecutive transmissions. The results are obtained with OPNET simulation tool and as one can infer from the figure, the probability decays as the number of stations increases; the increase is considerable if the number of stations is low. The figure shows two sets of data for the total network and a station.

We take another clustering scheme in which a virtual event is defined as an empty slot or a transmission. In our case, we take the consecutive transmission probability into account. Figure 3.6 depicts the virtual events for Case I defined in [5] and for Case II defined above. From now on we call 802.11^b the model based on Case I and 802.11^+ the model based on Case II.

As we explained above the complexity of joint Markov model grows exponentially with the number of stations in the network. It is essential to investigate the difference between joint and independent models. We look at a simplified backoff scheme: it has one-level backoff and the contention window size is only four.

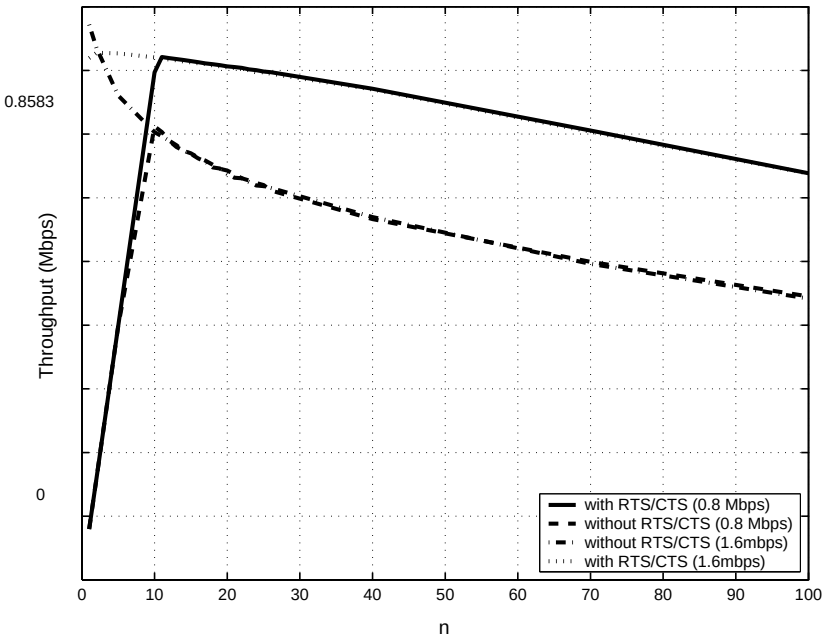


Figure 3.4: Saturation throughput

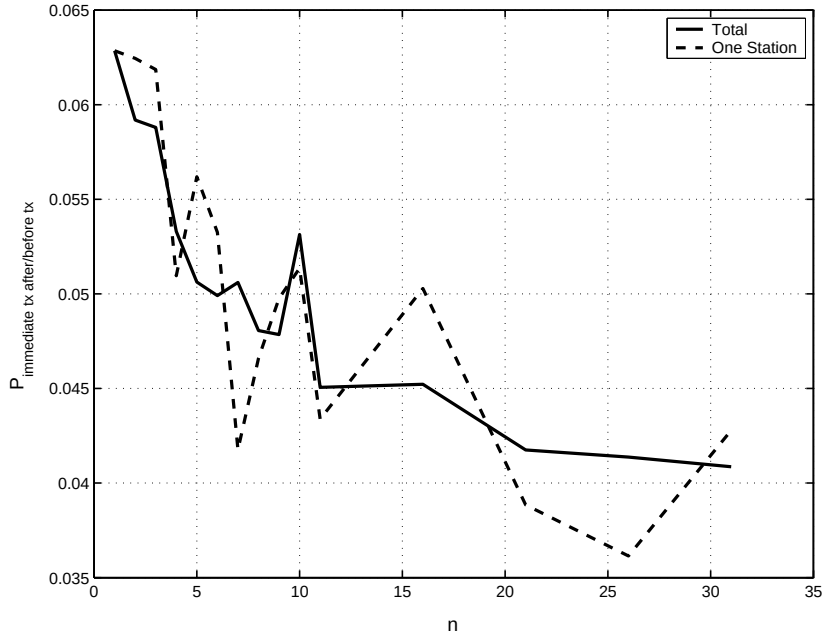


Figure 3.5: Zero backoff probability

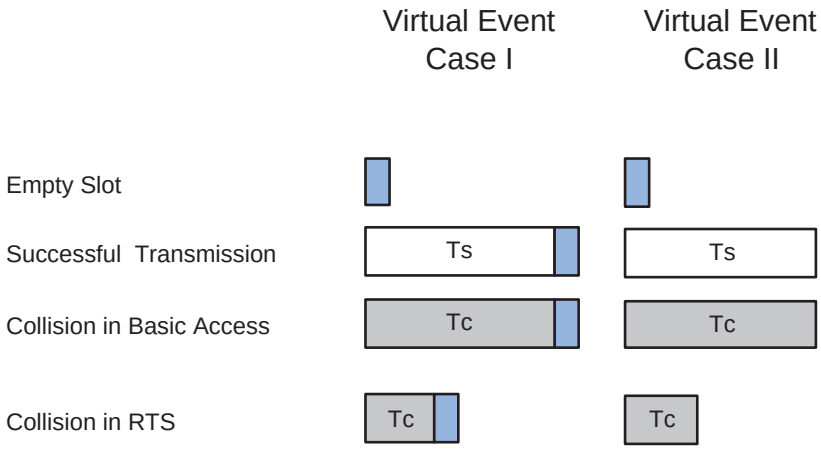


Figure 3.6: Virtual events

3.3 One-Level Backoff

We can see from Figure 3.7 that 802.11^b model matches the model introduced in [5] and the 802.11^+ model is our model. As we stated above a virtual event in 802.11^b requires a state change after each event resulting in the probability of state change in the backoff procedure to be one. On the other hand, 802.11^+ has a loop in the backoff states to ensure that the counter freezes when there is a transmission and resumes after sensing that the channel is idle at least one DIFS time. If there is another transmission after the DIFS time, then the backoff counter stays the same. This is represented in Figure 3.7. The stations wait in the current backoff state with probability p and advance to next state with probability $(1 - p)$. When they come to the backoff state 0, they transmit no matter what the result of the transmission is and they select another backoff level.

Let's call τ *the probability of transmission of a station*, P_{tr} is *the probability of transmission in the medium*, and P_s is *the probability of successful transmission in the medium*. If is defined as [5] "throughput is the fraction of time the channel is used to

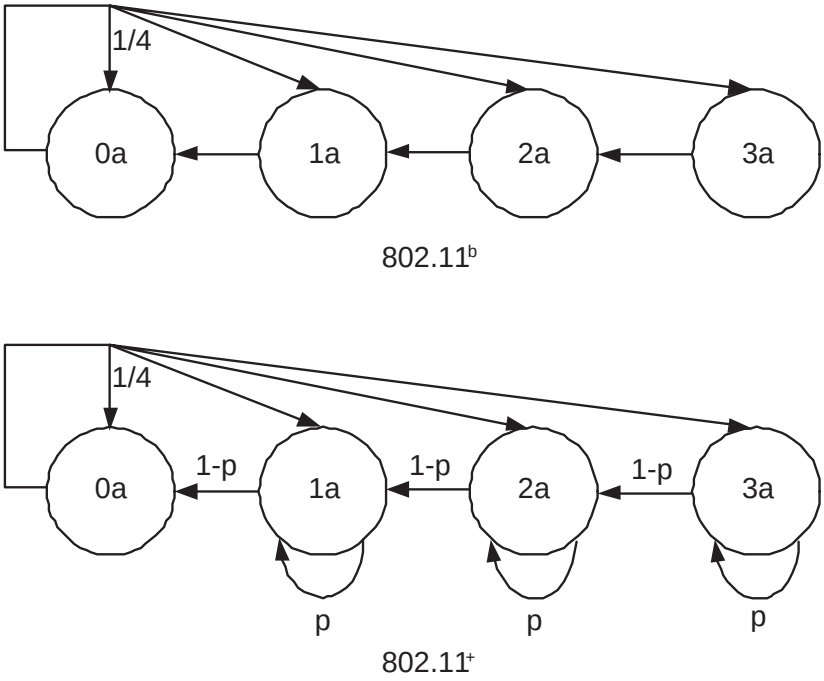


Figure 3.7: Throughput analysis with one-level backoff (CW=4)

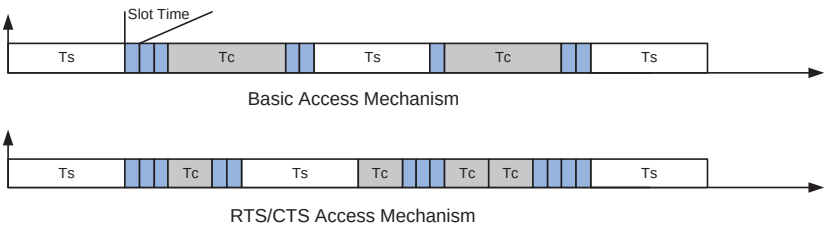


Figure 3.8: Time scale

successfully transmit payload bits",

$$Throughput = \frac{E[payload]}{E[length\ of\ virtual\ event]}, \tag{3.1}$$

and numerically the throughput S is the ratio:

$$S = \frac{P_s E[P]}{(1 - P_{tr})\sigma + P_s T_s + (P_{tr} - P_s)T_c}, \tag{3.2}$$

in which $E[P]$ is the average packet payload size. The denominator is the average duration of a virtual event, which may be an idle slot (of duration σ), a successful transmission (of duration T_s), or a collision (of duration T_c). These durations for the basic and RTS/CTS mechanisms are given below:

$$\begin{aligned}
T_s^{basic} &= T_{DATA} + SIFS + \delta + T_{ACK} + \delta \\
&+ DIFS \\
T_c^{basic} &= T^*_{DATA} + \delta + EIFS \\
T_s^{rts} &= T_{RTS} + SIFS + \delta + T_{CTS} \\
&+ SIFS + \delta + T_{DATA} + SIFS + \delta \\
&+ T_{ACK} + \delta + DIFS \\
T_c^{rts} &= T_{RTS} + \delta + EIFS.
\end{aligned} \tag{3.3}$$

In (3.3), T_{DATA} is the duration of a packet of size $E[P]$, and T_{RTS}, T_{CTS}, T_{ACK} are the durations of the corresponding frames. T^*_{DATA} is the average time to send $E[P^*]$ bytes, which is the average length of the longest packet payload involved in a collision. When

all packets have the same size: $E[P] = P = E[P^*]$. The propagation delay is denoted as δ. Unlike in basic access, T_c^{rts} only contains T_{RTS} since a collision can only occur in the RTS frame transmission. This difference is represented in Figure 3.8.

3.3.1 Independent Markov Model

The crucial difference between the independent and joint Markov model is how P_{tr} and P_s relate to τ and p. We will first start with the Independent Markov (IM) Model and then give the analysis for the Joint Markov Model. Figure 3.3 shows that each station is independent since they only see a two-state Markov model of the channel. As a result, the probability of having a busy medium is

$$P_{tr} = 1 - (1 - \tau)^n, \tag{3.4}$$

and the probability of having a successful transmission is equivalent to the probability of having only one transmission which is:

$$P_s = n\tau(1 - \tau)^{n-1}, \tag{3.5}$$

where n is the number of stations in the medium. The probability τ can be found by solving the Markov model: The value τ is the probability of being in state 0.

Solving the Markov model is easy for the 802.11^b model of Figure 3.7 and probability

		0a0b	0a1b	0a2b	0a3b	1a0b	1a1b	1a2b	1a3b	2a0b	2a1b	2a2b	2a3b	3a0b	3a1b	3a2b	3a3b
p1	0a0b	1/16	1/16	1/16	1/16	1/16	1/16	1/16	1/16	1/16	1/16	1/16	1/16	1/16	1/16	1/16	1/16
p2	0a1b		1/4				1/4				1/4				1/4		
p3	0a2b			1/4				1/4				1/4				1/4	
p4	0a3b				1/4				1/4				1/4				1/4
p5	1a0b					1/4	1/4	1/4	1/4								
p6	1a1b	1															
p7	1a2b		1														
p8	1a3b			1													
p9	2a0b									1/4	1/4	1/4	1/4				
p10	2a1b					1											
p11	2a2b						1										
p12	2a3b							1									
p13	3a0b													1/4	1/4	1/4	1/4
p14	3a1b									1							
p15	3a2b										1						
p16	3a3b											1					

Figure 3.9: Joint model for 802.11$^+$ (n=2)

of being in state 0 is $\tau^b = 0.4$. For the 802.11^+ model τ^+ is:

$$\tau^+(p) = \frac{2-2p}{5-2p}. \tag{3.6}$$

Thus, τ is dependent on p and τ and p are related

$$p = 1 - (1-\tau)^{n-1}. \tag{3.7}$$

From equation(3.7) we can find a τ [5]. Then we can find the throughput S after calculating P_{tr} and P_s.

3.3.2 Joint Markov Model

For the Joint Markov (JM) model, we need to construct a Markov chain that has 4^n states. We adopt a notation that each state is represented by binary numbers. For instance, for

		0a0b	0a1b	0a2b	0a3b	1a0b	1a1b	1a2b	1a3b	2a0b	2a1b	2a2b	2a3b	3a0b	3a1b	3a2b	3a3b
p1	0a0b	1/16	1/16	1/16	1/16	1/16	1/16	1/16	1/16	1/16	1/16	1/16	1/16	1/16	1/16	1/16	1/16
p2	0a1b	1/4				1/4				1/4				1/4			
p3	0a2b		1/4				1/4				1/4				1/4		
p4	0a3b			1/4				1/4				1/4				1/4	
p5	1a0b	1/4	1/4	1/4	1/4												
p6	1a1b	1															
p7	1a2b		1														
p8	1a3b			1													
p9	2a0b					1/4	1/4	1/4	1/4								
p10	2a1b					1											
p11	2a2b						1										
p12	2a3b							1									
p13	3a0b									1/4	1/4	1/4	1/4				
p14	3a1b									1							
p15	3a2b										1						
p16	3a3b											1					

Figure 3.10: Joint model for 802.11^b (n=2)

two nodes we can represent 16 states with two bits.

The difference arises in the backoff scheme between 802.11^b and 802.11^+ models. 802.11^b requires a state change in every event but 802.11^+ changes the backoff state if nobody is transmitting. Figure 3.9 represents the 802.11^+ model for two nodes. If two stations are at the zero state they just select a backoff interval with $1/16$ probability. For instance, if Station a is in State 1 and Station b is in State 0 (1a0b)-which means that Station b is transmitting-then the next state for Station b could be any state between 0 and 3 with $\frac{1}{4}$ probability; but Station a stays in State 1 in 802.11^+. On the other hand, in 802.11^b model, Station a advances to State 0 as seen in Figure 3.10.

Following the definition above, P_{tr} is defined as the sum of the probabilities of indices which have zero bit; and P_s is defined as the sum of the probabilities of indices which have only one zero bit which means that there is only one station is transmitting.

According to Figures 3.9 and 3.10 for two nodes, P_{tr} and P_s are calculated as follows;

$$P_{tr} = p_1 + p_2 + p_3 + p_4 + p_5 + p_9 + p_{13}$$

$$P_s = p_2 + p_3 + p_4 + p_5 + p_9 + p_{13}.$$

3.3.3 Performance of One-Level Backoff

In order to compare the performance we setup the scenario in OPNET. We counted the number of collisions and number of successful transmission; and measured the throughput. Stations operate at saturated throughput and with only one-level backoff of four states. They are arranged in a way that they all hear each other. The FHSS physical layer is chosen and data rate is 1 Mbps. In OPNET, the parameters are achieved as follows:

$$\begin{aligned} P_{tr} &= \frac{\#Total\ ACK\ received + \#Collision}{\#Total\ ACK + \#Collision + \#Back-off\ slot} \\ P_s &= \frac{\#Total\ ACK\ received}{\#Total\ ACK + \#Collision + \#Back-off\ slot} \end{aligned} \tag{3.8}$$

We verified the simulation by equation 3.9 since the number of events multiplied by their associated time values gives the simulation time. This means that the events we consider are complete and there is no missing periods that we have not considered.

$$\begin{aligned} Simulation\,Time \ &\approx \sigma * \#Back - off\ slots \\ &+T_s * \#Total\ ACK\ received + T_c * \#Collision \end{aligned} \tag{3.9}$$

Figure 3.11 shows the performance of 802.11^+ and 802.11^b model for P_{tr} and P_s. The solid line represents the OPNET simulation and circled line represents the 802.11^+ model for joint Markov model. We found the same values for joint and independent Markov

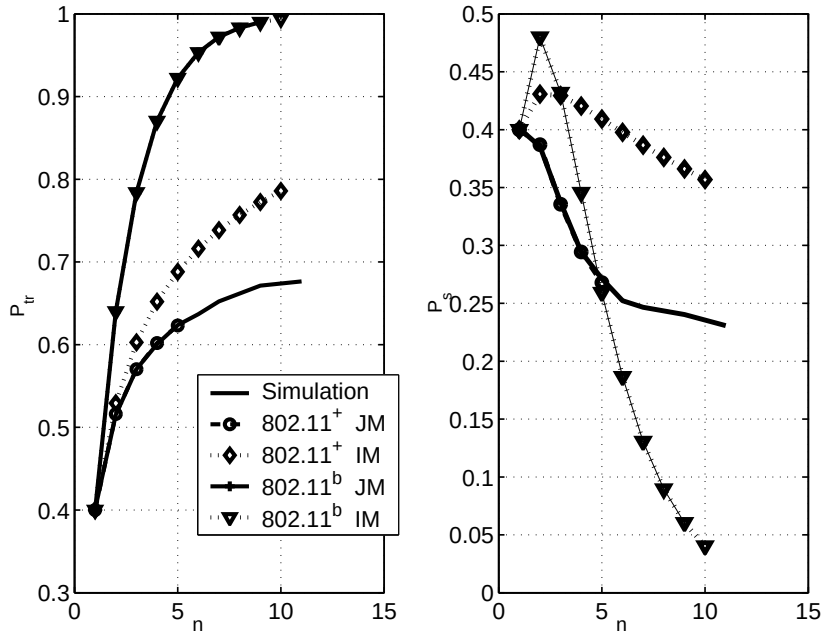


Figure 3.11: P_{tr} and P_s for one-level backoff

model of 802.11^b and the performance of JM$^+$ of 802.11^+ gives exactly the performance as the simulation. Of course, the data set for the joint model is only up to five nodes due to the increasing complexity. IM$^+$, on the other hand, approximates the simulation.

Now let's look at the throughput performance. Figure 3.12 shows the performance for RTS/CTS and basic access mechanisms. We find the duration values for the basic access mechanism as follows:

$$\begin{aligned} T_s^+ &= T_{DATA} + SIFS + \delta + T_{ACK} + DIFS + \delta \\ T_c^+ &= T_{DATA}^* + \delta + EIFS \\ T_s^b &= T_{DATA} + SIFS + \delta + T_{ACK} + DIFS + \delta + \sigma \\ T_c^b &= T_{DATA}^* + \delta + DIFS + \sigma \end{aligned} \tag{3.10}$$

and for the RTS/CTS access mechanism as follows:

$$\begin{aligned} T_s^+ &= T_{RTS} + SIFS + \delta + T_{CTS} + SIFS + \delta \\ &\quad + T_{DATA} + SIFS + \delta + T_{ACK} + DIFS + \delta \\ T_c^+ &= T_{RTS} + \delta + EIFS \\ T_s^b &= T_{RTS} + SIFS + \delta + T_{CTS} + SIFS + \delta \\ &\quad + T_{DATA} + SIFS + \delta + T_{ACK} + DIFS + \delta + \sigma \\ T_c^b &= T_{RTS} + \delta + DIFS + \sigma \end{aligned} \tag{3.11}$$

T^b values are updated values of [5] according to the standard. Figure 3.12 shows that lines which are for basic access and RTS/CTS access mechanisms are most closely followed by the JM$^+$ model of 802.11^+.

As one can infer easily throughput characteristic of a network can be found exactly with JM$^+$ model but the limitation is in the computing power which leads us to use the independent model for closer approximations. We proved here that 802.11^+ model performs better than 802.11^b model. Of course, the reason why 802.11^b showed inferior performance is that with four backoff states consecutive transmission probability is $\frac{1}{4}$ which is very high compared to the actual backoff structure (See Figure 3.5). We see that this probability is low with the multi-level backoff structure and decreasing with the increase in the number of stations.

This analysis is worth investigating further to prove the methodology of the Markov model. Since we have the P_{tr}, P_s and S values from the OPNET simulation, we can

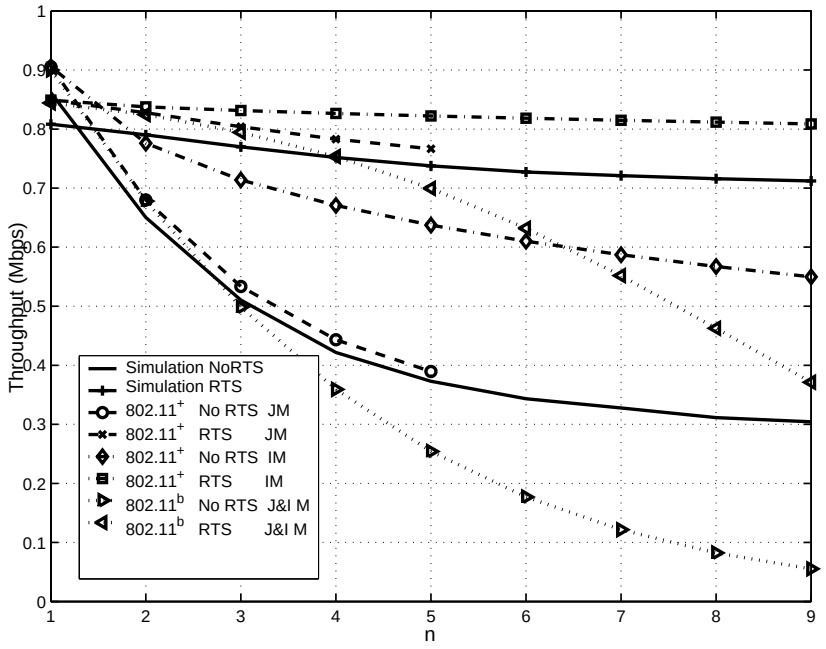


Figure 3.12: Throughput for one-level backoff

revisit the throughput formulation in reverse and obtain the T_s and T_c as follows;

$$P_s T_s + (P_{tr} - P_s) T_c = \frac{P_s E[P]}{S} - (1 - P_{tr})\sigma, \tag{3.12}$$

where we can use LLSE to find unique duration values. We found $T_s = 0.0088sec$ and $T_c = 0.0088sec$ for basic access mechanism and $T_s = 0.0087sec$ and $T_c = 0.0007sec$ for RTS/CTS access mechanism where $E[P]$ is $1000bytes$. Duration values are almost same as what we found from the formulas depicted in Figure 3.13 for IEEE 802.11b. As a result, we can conclude that the throughput formulation is correct and accurate.

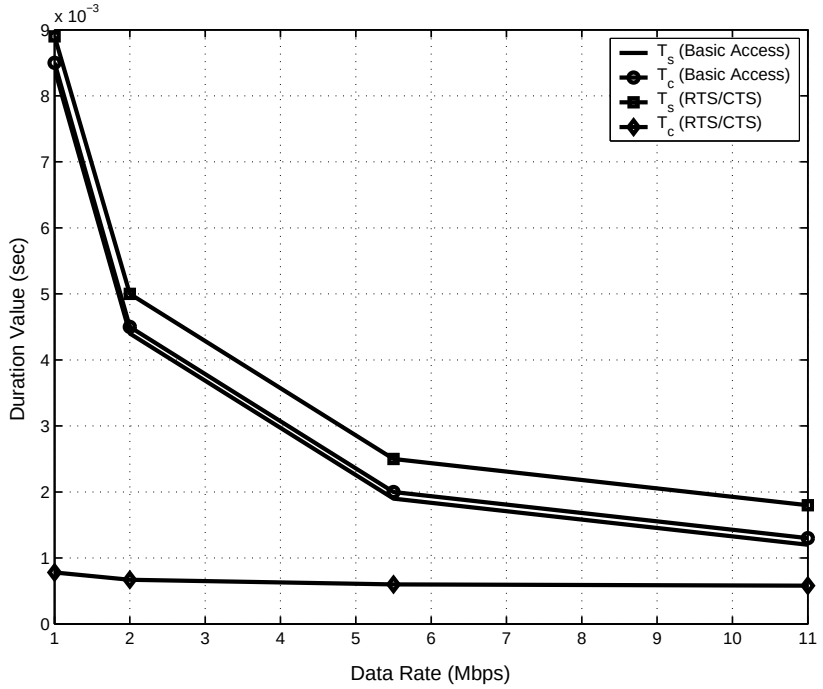


Figure 3.13: Duration values for IEEE 802.11b

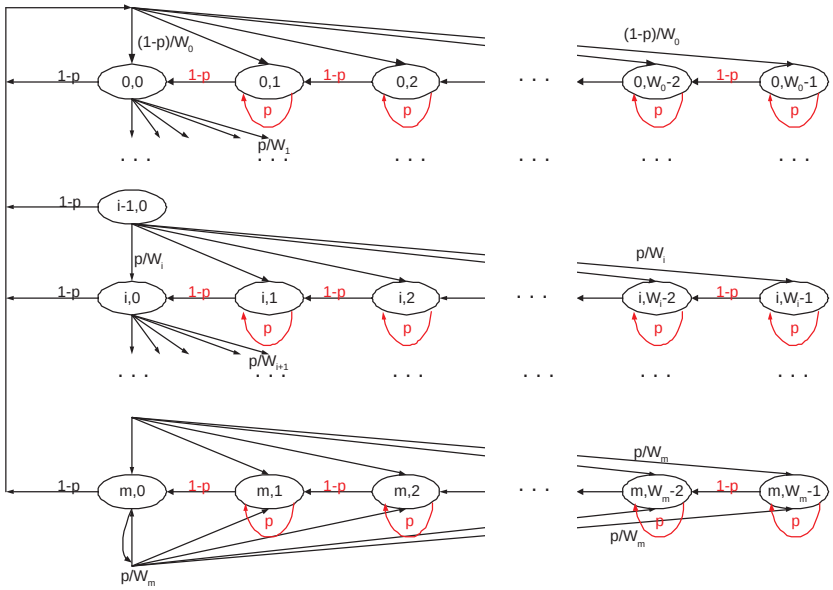


Figure 3.14: Independent Markov model of 802.11$^+$ for multi-level backoff

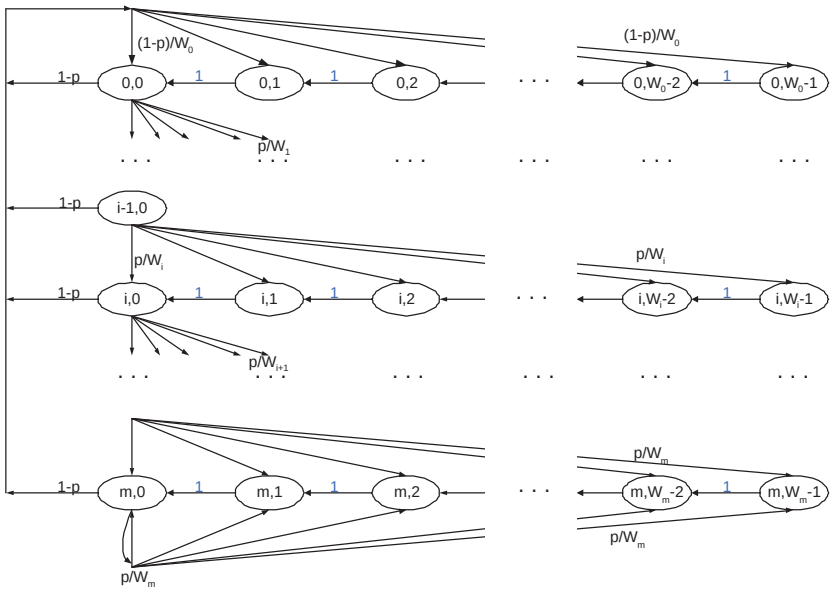


Figure 3.15: Independent Markov model of 802.11^b for multi-level backoff

3.4 Multi-Level Backoff

The Multi-level backoff scheme complies with the 802.11 standard. There are 7 backoff levels, the minimum contention window (CW_{min}) is 15 and the maximum contention window (CW_{max}) is 1024. The difference between 802.11^+ and 802.11^b model is illustrated in Figures 3.14 and 3.15. For 802.11^+ there is either an empty slot (of duration σ) during which the backoff counter is decremented, or a transmission (of duration equal to a data or RTS/CTS packet) during which backoff is frozen. So the events that cause a state transition in the Markov chain are an empty slot or a transmission. This is different from [5], in which a slot is either an empty slot or a transmission-plus-empty-slot. (Thus every slot in [5] includes an empty slot.) For this reason, our Markov model (Figure 3.14) has the self loop in the backoff stages to model freezing of the backoff counter when the medium is busy.

The Markov chain has two dimensions, $s(t)$ and $b(t)$, representing the state number and backoff stage, respectively. The derivation of the throughput is carried out in two steps. We first obtain τ^+ for 802.11^+ and then use τ^+ to calculate throughput S.

$$\begin{cases} P\{i,k|i,k+1\} = (1-p) & k \in [0, W_i - 2] \quad i \in [0, m] \\ P\{i,k|i,k\} = p & k \in [1, W_i - 1] \quad i \in [0, m] \\ P\{i,k|i-1,0\} = p/W_i & k \in [0, W_i - 1] \quad i \in [1, m] \\ P\{m,k|m,0\} = p/W_m & k \in [0, W_m - 1] \end{cases} \tag{3.13}$$

The two-dimensional chain $(s(t), b(t))$ is governed by the one-step transition probabilities (3.13) of 802.11^+. The first and second equations respectively indicate that at the beginning of each slot, the backoff counter is decremented if the channel is sensed idle (which happens with probability with $(1-p)$) and frozen if the channel is sensed busy (which happens with probability p).

The third and fourth equations respectively indicate that following an unsuccessful transmission, the station in backoff stage $(i-1)$ selects a backoff interval uniformly in the range $(0, W_i - 1)$ and when the backoff stage reaches m, W_m stays constant.

We can solve the balance equations to obtain the stationary distribution denoted by $b_{i,k}, i \in [0, m], k \in [0, W_i - 1]$.

$$\begin{aligned} b_{i-1,0} \cdot p = b_{i,0} \quad &\longrightarrow \quad b_{i,0} = p^i b_{0,0},\ 0 < i < m \\ b_{m-1,0} \cdot p = (1-p) b_{m,0} \quad &\longrightarrow \quad b_{m,0} = \tfrac{p^m}{1-p} b_{0,0} \end{aligned} \tag{3.14}$$

From (10.2), the stationary distribution is:

$$b_{i,k} = \frac{W_i - k}{W_i(1-p)} \cdot \begin{cases} (1-p) \cdot \sum_{j=0}^{m} b_{j,0} & i = 0 \\ p \cdot b_{i-1,0} & 0 < i < m \\ p \cdot (b_{m-1,0} + b_{m,0}) & i = m \end{cases} \tag{3.15}$$

or

$$b_{i,k} = \frac{W_i - k}{W_i(1-p)} b_{i,0} \;\; i \in [0, m], \; k \in [0, W_i - 1]. \tag{3.16}$$

$$1 = \sum_{i=0}^{m} \sum_{k=0}^{W_i - 1} b_{i,k} = \textit{backoff} \tag{3.17}$$

All $b_{i,k}$ can be expressed in terms of $b_{0,0}$, which can then be obtained because all probabilities add to one (10.6). This finally yields $b_{0,0}$ in terms of p, W, m in equation (10.8).

$$\begin{aligned} \textit{backoff} &= \sum_{i=0}^{m} \sum_{k=0}^{W_i - 1} b_{i,k} = \\ &\frac{b_{0,0}}{(1-p)} \left(\sum_{i=0}^{m} p^i \left(\frac{W_i + 1}{2}\right) + \frac{p^{m+1}}{(1-p)} \left(\frac{W_m + 1}{2}\right)\right), \end{aligned} \tag{3.18}$$

$$b_{0,0} = \frac{1}{\frac{1}{(1-p)} \sum_{i=0}^{m} p^i \left(\frac{W_i+1}{2}\right) + \frac{p^{m+1}}{(1-p)^2} \left(\frac{W_m+1}{2}\right)}. \tag{3.19}$$

A packet is transmitted in states $b_{i,0}, i \in [0, m]$, so τ, the probability of transmission in a plot, is given by (3.20),

$$\begin{aligned} \tau^+ &= \sum_{i=0}^{m} b_{i,0} = \frac{b_{0,0}}{(1-p)} \\ &= \frac{1}{\sum_{i=0}^{m} p^i \left(\frac{W_i+1}{2}\right) + \frac{p^{m+1}}{(1-p)} \left(\frac{W_m+1}{2}\right)}. \end{aligned} \tag{3.20}$$

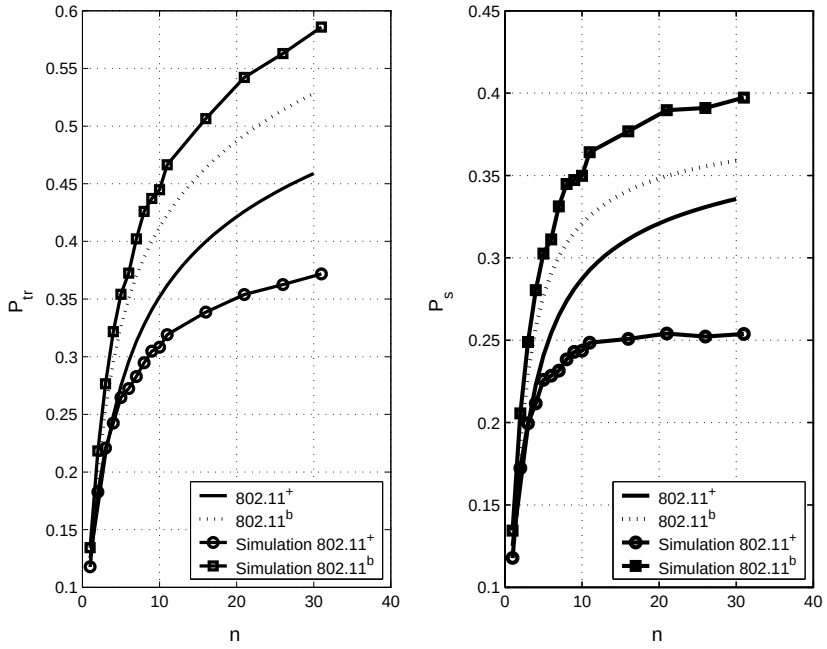


Figure 3.16: P_{tr} and P_s for multi-level backoff

Taking the contention window $CW_{max} = 2^m CW_{min}$, so $W_i = 2^i W, i \in [0, m]$, and $W = CW_{min}$, gives a simpler expression for τ^+,

$$\tau^+ = \frac{1}{\frac{(1-2p)(W+1)+pW(1-(2p)^m)}{2(1-2p)(1-p)}}. \tag{3.21}$$

For purposes of comparison, the transmission probability τ^b of [5] is

$$\tau^b = \frac{\tau^+}{1-p}. \tag{3.22}$$

The comparison with respect to the simulation shows that in Figure 3.16 the independent multi-level Markov model approximately follows the simulation results. We set the maximum retry count to 255 in the simulation and we use the FHSS physical layer with

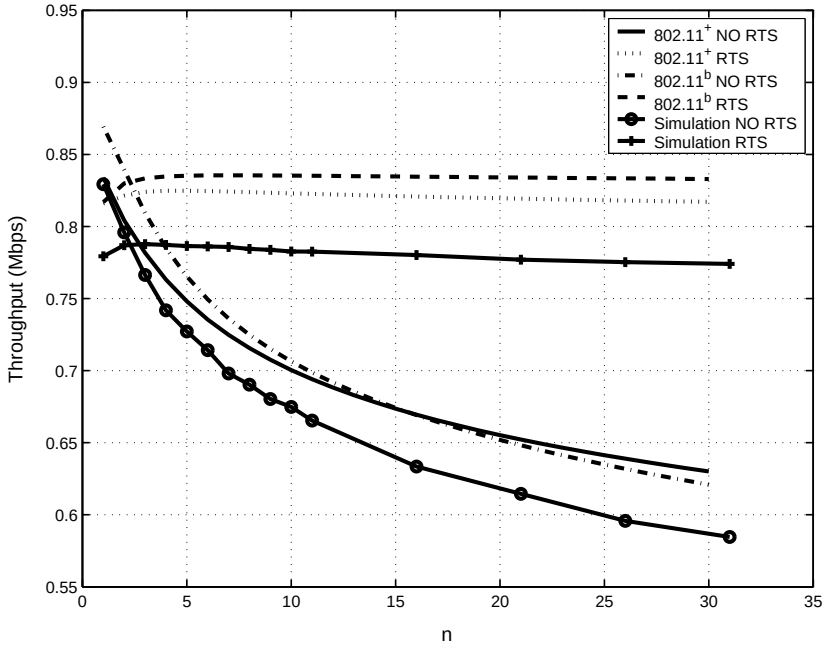


Figure 3.17: Throughput for multi-level backoff

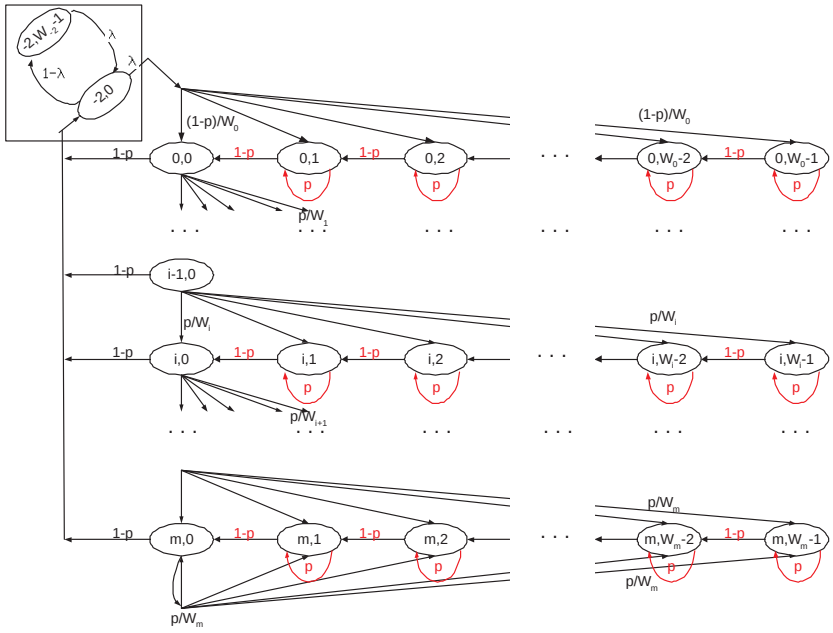


Figure 3.18: Markov model for the IEEE 802.11 DCF in normal traffic condition

1 Mbps data rate. We verified the duration values again and obtain the following;

$$\begin{array}{ccc} & Basic & RTS/CTS \\ T_s & 0.0088sec & 0.0090sec \\ T_c & 0.0088sec & 0.0007sec \end{array}$$

which are very close to the analytical findings. Another point we consider in simulation is in finding P_{tr} and P_s for 802.11^b model since, due to the high number of stations, consecutive transmission probability is low. As a result, the equation (3.8) is modified for 802.11^b as follows:

$$\begin{aligned} P_{tr}^b &= \frac{\#Total\ ACK\ received + \#Collision}{\#Back{-}off\ slot} \\ P_s^b &= \frac{\#Total\ ACK\ received}{\#Back{-}off\ slot} \end{aligned} \tag{3.23}$$

Since each back-off slot for 802.11^b is either a transmission+backoff slot or an empty back-off slot, it is enough to consider only the number of back-off slots for the total number of events. The results show that 802.11^b underestimates the probabilities; on the other hand, 802.11^+ overestimates the probabilities. Figure 3.17 shows the throughput performance and it can be seen that 802.11^+ gives better performance with lower number of nodes.

3.4.1 Formulation for Unsaturated Traffic

To model normal, non-saturated conditions, we introduce additional states, giving the chain of Figure 3.18. The assumption we adopt is omitting the post-backoff state which is taken into account right after a transmission as seen in Figure 2.9. We leave the discussion

to the last section in this chapter.

We introduce only two states by taking $W_{-2} = 2$. The one-step transition probabilities are slightly changed:

$$\begin{cases} P\{-2,1|-2,1\} = (1-\lambda) \\ P\{0,j|-2,0\} = \lambda/W_0 \qquad j \in [0, W_0 - 1] \\ P\{-2,1|-2,0\} = (1-\lambda) \\ P\{-2,0|-2,1\} = \lambda \\ P\{-2,0|i,0\} = (1-p) \qquad i \in [0,m] \end{cases} \tag{3.24}$$

Under non-saturated conditions, a station may now wait in the idle state for a packet from upper layers. This corresponds to a delay in the idle state, represented by the box in Figure 3.18. The delay in the idle state is geometric with parameter λ. The transition probabilities in (3.24) are straightforward modifications of those previously obtained for the saturated case.

The stationary probabilities add up to 1,

$$1 = \sum_{i=0}^{m} \sum_{k=0}^{W_i - 1} b_{i,k} + \sum_{k=0}^{W_{-2}-1} b_{-2,k} = \textit{backoff} + \textit{idle}. \tag{3.25}$$

The probabilities $b_{-2,0}$ and $b_{-2,1}$ can be expressed in terms of $b_{0,0}$ using (10.2), (3.24), and by representing the probability *idle* in terms of $b_{0,0}$ by:

$$idle = \sum_{k=0}^{W_{-2}-1} b_{-2,k} - 1 = \frac{b_{0,0}}{\lambda^2} - 1.$$

The new τ^+ is given by (3.26), which reduces to (3.21) for the saturated case ($\lambda = 1$),

$$\tau^+ = \frac{1}{\frac{(1-2p)(W+1)+pW(1-(2p)^m)}{2(1-2p)(1-p)} + (1-p)(\frac{1}{\lambda^2} - 1)}. \tag{3.26}$$

From (3.26) we see that $\tau^+ = \tau^+(p, m, W, \lambda)$ depends on the unknown p. Also, as in [5],

$$p = 1 - (1-\tau^+)^{n-1} \text{ or } \tau^+(p) = 1 - (1-p)^{\frac{1}{(n-1)}}. \tag{3.27}$$

Equations (3.26) and (3.27) together determine τ and p. Figure 3.19 and 3.20 plot the collision probability p and transmission probability τ as the number of stations varies, for five cases: the 802.11^+, and the proposed model 802.11^+ for four different $\lambda = 1, 0.2, 0.1, 0.05$. The 802.11^b gives higher values of p and τ than our model for $\lambda = 1$. In general, as expected, p increases and τ decreases with n.

Also, as expected, τ increases with load λ, which is readily appreciated by taking $n \to 1$, $p \to 0$, for which

$$\lim_{p \to 0} \tau = \frac{1}{\frac{(W+1)}{2} + (\frac{1}{\lambda^2} - 1)}. \tag{3.28}$$

For the saturated case, $\lambda = 1$, and $m = 0$ (no exponential backoff), we can compare τ^+ with τ^b in (3.22),

$$\tau^+(p, 0, W, 1) = \frac{2(1-p)}{W+1} < \tau^b = \frac{2}{W+1}. \tag{3.29}$$

Unlike in 802.11^b, τ^+ depends on the collision probability p (and hence on n). Intuitively of course, τ should depend on n: if there are more stations, the medium will be busy more often, and a station will transmit less frequently.

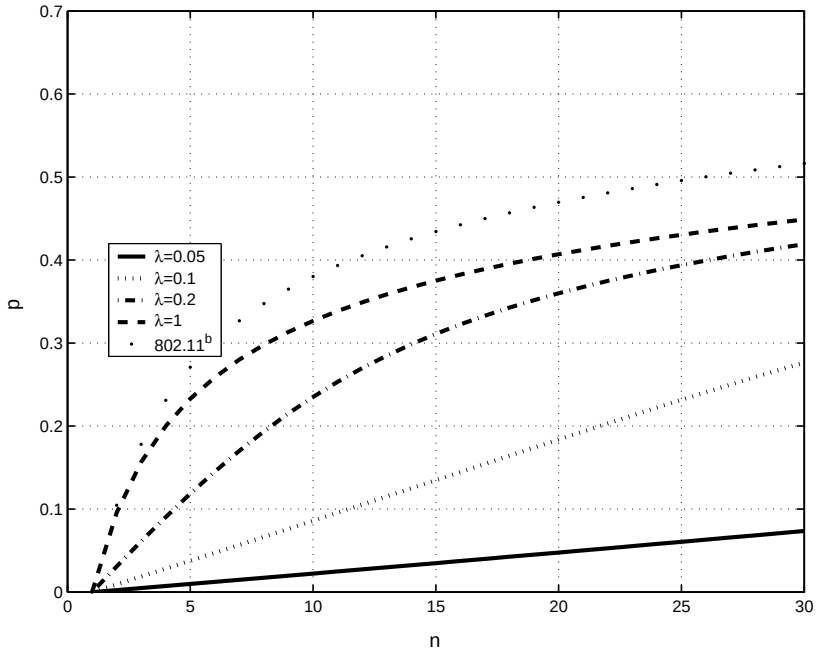


Figure 3.19: Probability of collision (p)

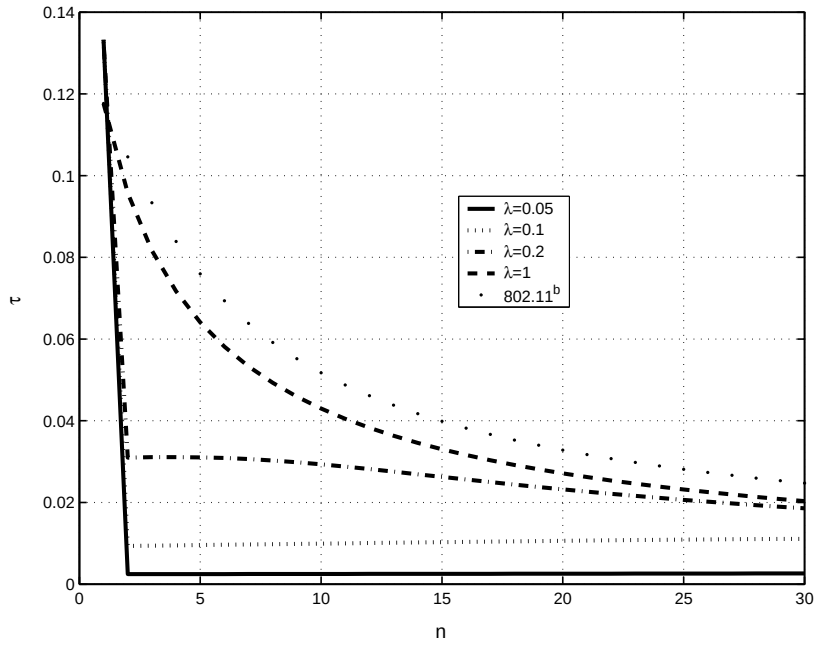


Figure 3.20: Probability of transmission (τ)

Mode	Modulation	Code Rate	Data Rate	BpS	SNR
1	BPSK	1/2	6 Mbps	3	25
2	BPSK	3/4	9 Mbps	4.5	27
3	QPSK	1/2	12 Mbps	6	30
4	QPSK	3/4	18 Mbps	9	32
5	16-QAM	1/2	24 Mbps	12	35
6	16-QAM	3/4	36 Mbps	18	40
7	64-QAM	2/3	48 Mbps	24	42
8	64-QAM	3/4	54 Mbps	27	45

Table 3.1: Eight PHY Modes of the IEEE 802.11a PHY

DATA

Frame Control	Duration/ ID	Address 1	Address 2	Address 3	Sequence Control	Address 4	Frame Body	FCS
Octets: 2	2	6	6	6	2	6	0-2312	4

RTS

Frame Control	Duration	RA	RA	FCS
Octets: 2	2	6	6	4

CTS & ACK

Frame Control	Duration	RA	FCS
Octets: 2	2	6	4

Figure 3.21: Frame formats

3.4.2 IEEE 802.11a OFDM Physical Layer

The IEEE 802.11a PHY uses OFDM modulation, and provides eight modes with different modulation schemes and coding rates. Table 3.1 shows the supported rates depending on SNR (SNR values are vendor-proprietary). As shown in Figure 3.21, each MAC data frame or MAC Protocol Data Unit (MPDU), consists of the MAC Header, Frame Body, and Frame Check Sequence (FCS). The MAC header and FCS together are 28 octets, the RTS frame is 18 octets, and the CTS and ACK are 12 octets long.

When a MPDU is passed to the PLCP layer it is called PSDU. In order to create a PLCP Protocol Data Unit (PPDU), PLCP headers are added. Figure 3.22 shows the

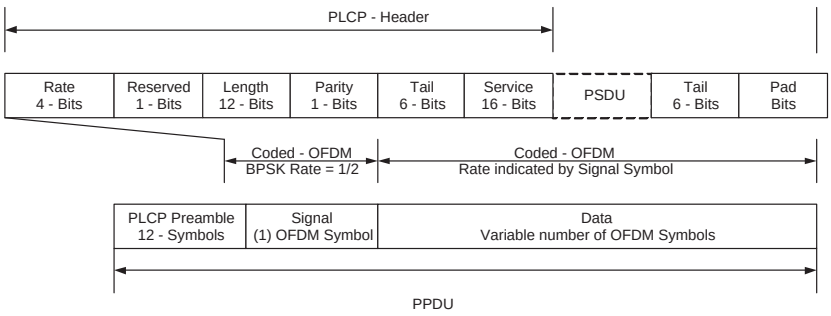


Figure 3.22: IEEE 802.11a OFDM packet

PPDU format. During transmission, a PLCP preamble and a PLCP header are added to a PSDU to create a PLCP Protocol Data Unit (PPDU). The PLCP preamble field, with the duration of $tPLCPPreamble$, is composed of 10 repetitions of a short training sequence (0.8μs) and two repetitions of a long training sequence (4μs). The PLCP header except the SERVICE field, with the duration of $tPLCP\_SIG$ constitutes a single OFDM symbol. Each OFDM symbol interval is denoted by $tSymbol$ and its duration is 4μs. The 16-bit SERVICE field of the PLCP header and the MPDU (along with six tail bits and pad bits), represented by DATA, are transmitted at the data rate specified in the RATE field. The BSS basic rate set is $\{$6 Mbps, 12 Mbps, 24 Mbps$\}$ and each station should support these rates and control information should be sent at these rates [32]. We assume that all rates specified in Table 3.1 are supported and control signalling can be in any rate.

We can thus obtain the duration of each packet. The time to transmit a frame with

$E[P]$ octets of data payload with the IEEE 802.11a PHY 3.1 is given below:

$$T_{DATA}(m) = tPLCPPreamble + tPLCPHeader$$
$$+MACHeader + E[P] + FCS + Tailbits + PadBits$$
$$= 20\mu s + [\frac{28+(16+6)/8+E[P]}{BpS(m)}] \cdot 4\mu s$$

$$T_{RTS}(m) = tPLCPPreamble + tPLCPHeader$$
$$+MACHeader + FCS$$
$$= 20\mu s + [\frac{20+(16+6)/8}{BPS(m)}] \cdot 4\mu s$$

$$T_{CTS}(m) = T_{ACK}(m) = tPLCPPreamble$$
$$+tPLCPHeader + MACHeader + FCS$$
$$= 20\mu s + [\frac{14+(16+6)/8}{BPS(m)}] \cdot 4\mu s.$$

The Bytes-per-Symbol for PHY mode m is $BpS(m)$ in Table 3.1. The BpS value in our example depends on SNR. Duration values are illustrated in Figure 3.23.

3.5 Throughput Characteristics

The performance results below are based on the physical layer parameters specified in Tables 3.1 and 3.2. The payload is constant $E[P] = 1024bytes$. Figure 3.24 shows the throughput for the basic access scheme for the model of 802.11^b and our model for four

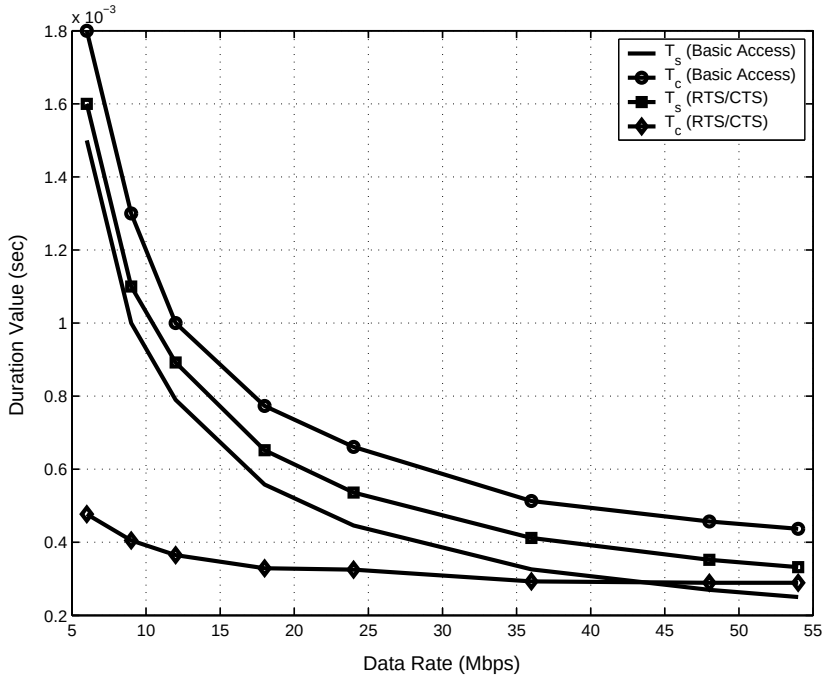


Figure 3.23: Duration values for IEEE 802.11a

Characteristics	Value	Definition
Slot Time	9μs	Slot time
SIFS	16μs	SIFS Time
DIFS	34μs	DIFS=SIFS+2xSlot
CWmin	15	min CW
CWmax	1023	max CW
tPLCPPreamble	16μs	PLCP preamble duration
tPLCP_SIG	4μs	PLCP SIGNAL field duration
tSymbol	4μs	OFDM symbol interval

Table 3.2: IEEE 802.11a OFDM PHY Characteristics

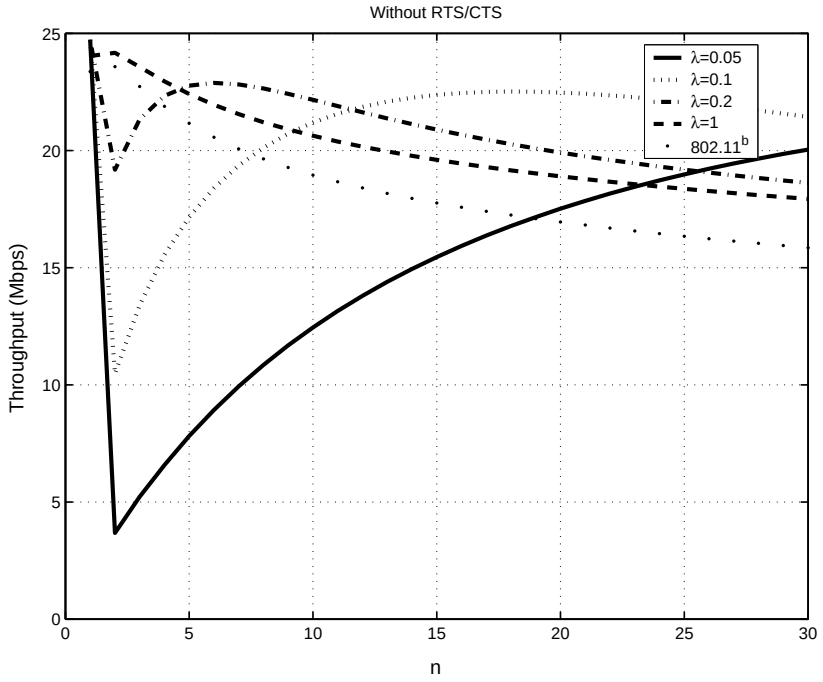


Figure 3.24: Throughput for IEEE 802.11a: Data rate=54Mbps (w/o RTS/CTS)

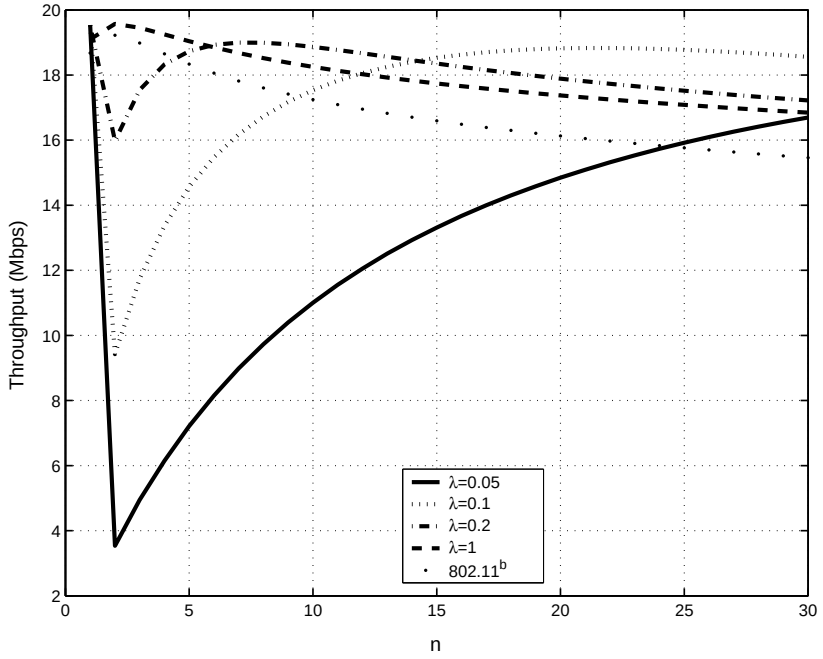


Figure 3.25: Throughput for IEEE 802.11a: Data rate=54Mbps (with RTS/CTS)

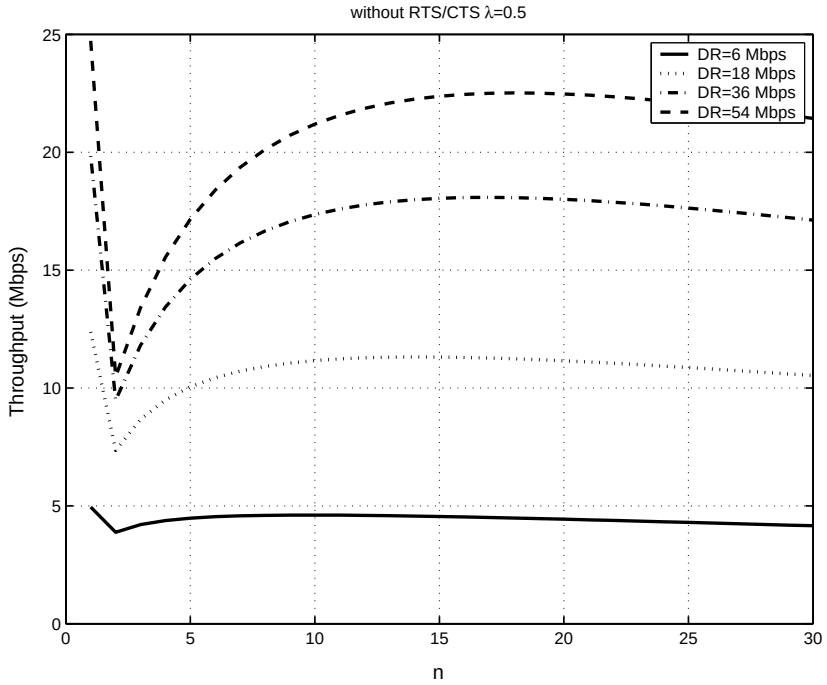


Figure 3.26: Throughput for IEEE 802.11a: Different data rate (w/o RTS/CTS)

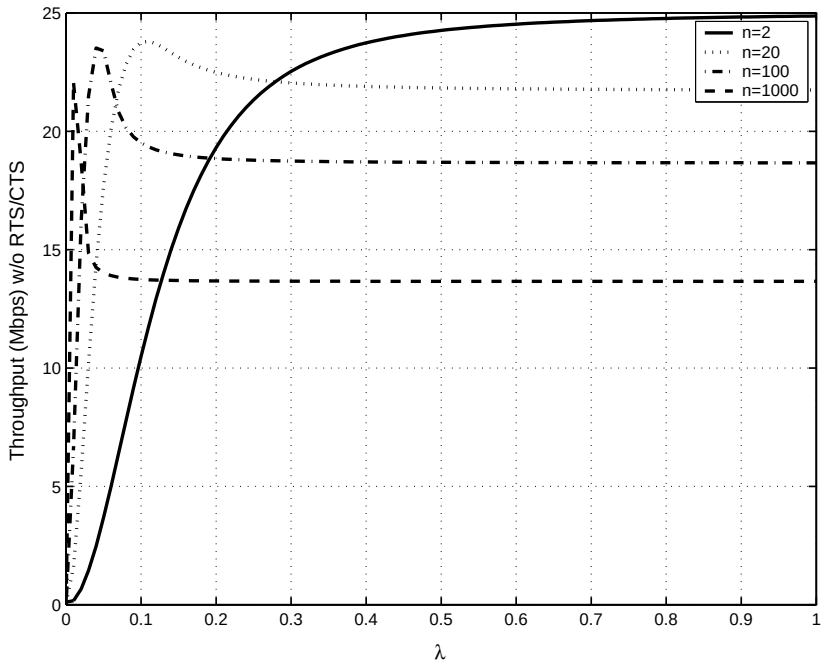


Figure 3.27: Throughput for IEEE 802.11a: Data Rate=54Mbps (w/o RTS/CTS) when n fixed

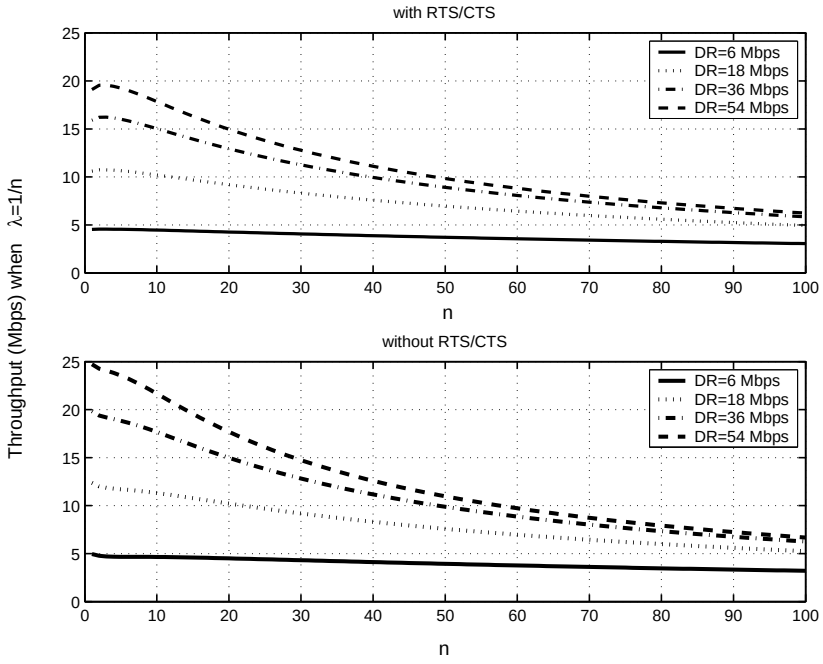


Figure 3.28: Throughput in IEEE 802.11a with constant total load $\lambda = \frac{1}{n}$

different values of λ. The throughput first increases with the number of stations until congestion sets in, after which throughput decreases. As the traffic intensity decreases, the maximum throughput is reached with a larger number of active stations. Figure 3.25 reports the throughput for RTS/CTS access mechanism. The effect of congestion is now less severe. However, as with basic access, the throughput decreases and goes to 0, as $n \rightarrow \infty$. Figure 3.26 shows the effect of different SNR levels which determines the data rate (all stations have the same SNR level). Figure 3.27 shows the effect of traffic intensity on throughput for a fixed number of stations. We again observe onset of congestion. Figure 3.28 shows the performance for a fixed load. Traffic intensity λ decreases with the number of stations and throughput settles down as λ goes to 0 and n goes to ∞.

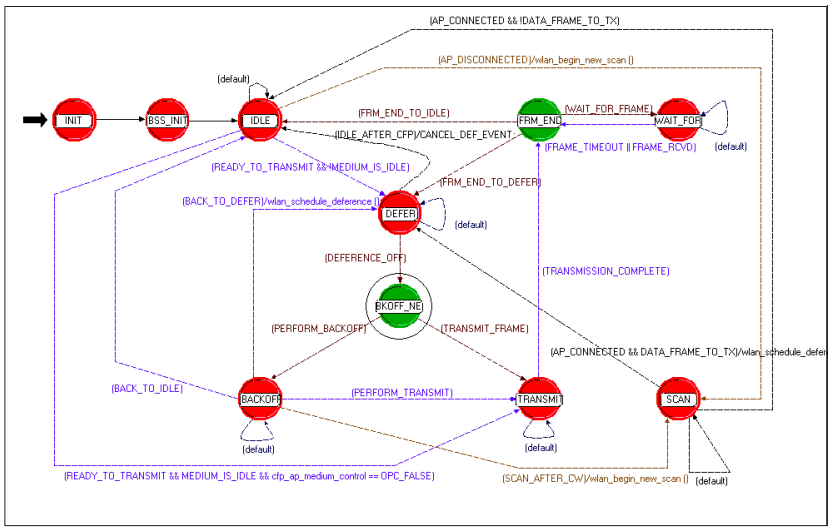


Figure 3.29: 802.11 MAC module in OPNET

3.6 Discussion

The IEEE 802.11 standard specifies a post-backoff state which introduces additional random delay after a transmission. This post-backoff state has been considered in some works in literature but it is often neglected in the implementations. It is easy to configure the Markov Model with an additional backoff level right after transmission but we take the approach without using it. The OPNET simulation platform also considers a non-post backoff state whose MAC finite state machine is shown in Figure 3.29. After transmission, stations go to the IDLE state if they do not have any packet to transmit. When λ is 1, that means the station goes to the backoff state to transmit right after the transmission and when λ is 0, the station waits in the IDLE state forever. Of course, the traffic model is not realistic but virtually it gives us a tool to estimate unsaturated traffic conditions. The standard specifies a mandatory backoff state between two transmissions which is also taken into account in our model. Two transmissions are interleaved with a backoff stage.

Remark

Part of this chapter will appear in ACM-Kluwer MONET Special Issue on WLAN Optimization at the MAC and Network Levels titled as “Throughput Analysis and Admission Control in IEEE 802.11a” and authors are Mustafa Ergen and Pravin Varaiya.

Chapter 4

Mixed Data Rate Formulation

4.1 Introduction

Performance of IEEE 802.11 network under the DCF mode degrades when some stations use a lower data rate than the others. This is common in a network environment since multi-rate cards have an adaptation mechanism to select a rate. If the link to the destination is under severe fading and interference the adaptation mechanism reduces the rate by changing its modulation scheme. Rate adaptation mechanisms are proprietary and may consider successfully received ACKs or signal-to-noise ratio (SNR). The reason why throughput degrades is hidden under the basic CSMA/CA channel access method cited in [41, 42] and easy to see from the Markov model. If a station acquires the transmission opportunity, it uses it as long as it is needed. As a result, if a station operates with a lower data rate it takes longer to transmit with the same payload, which is an under utilization of the channel.

We made a simulation in OPNET and configured a network as illustrated in Figure 4.1.

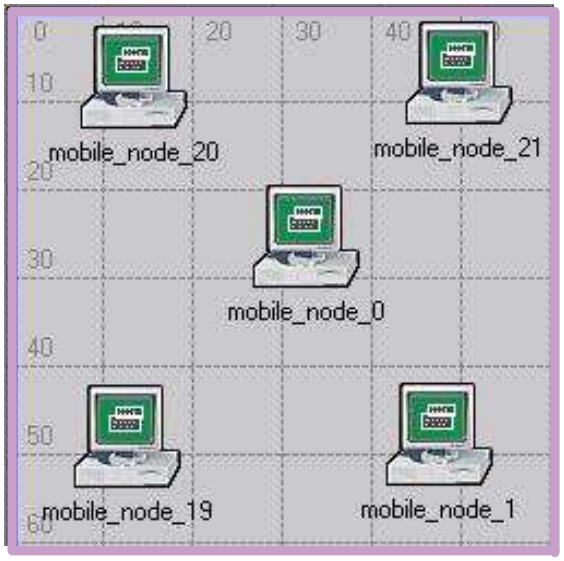


Figure 4.1: Network

The network has five nodes, all with 11 Mbps data rate but one which changes its rate with time. The upper plot of Figure 4.2 shows the total throughput of the network. The lower plot of the figure shows the activity of the channels for the station that changes its data rate. The station starts with 11Mbps, and at time 100sec it drops down the rate to 1Mbps until 200sec. Between 200sec and 300sec, the station operates with 2 Mbps and for the next 100sec it operates with 5.5 Mbps. Between 400sec and 600sec, the station again operates with 1Mbps and after 600sec, the station is with 11Mbps. The lower data rate severely affects the network; this is shown in the first plot of Figure 4.1. A single station with 1 Mbps decreases the total throughput by half. This kind of data rate arrangement is very common in a typical network since stations can change their data rate: or it can be limited by the standards; for instance, IEEE 802.11b and 802.11g stations can exist together. It is important to analyze and formulate this behavior.

Figure 4.3 on the other hand shows another key characteristic which contains only

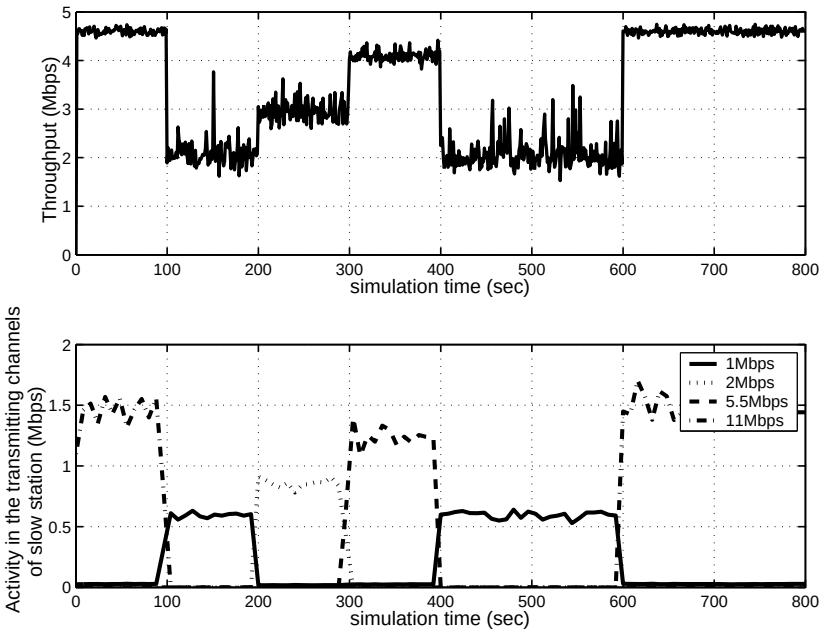


Figure 4.2: Throughput in a mixed data rate environment

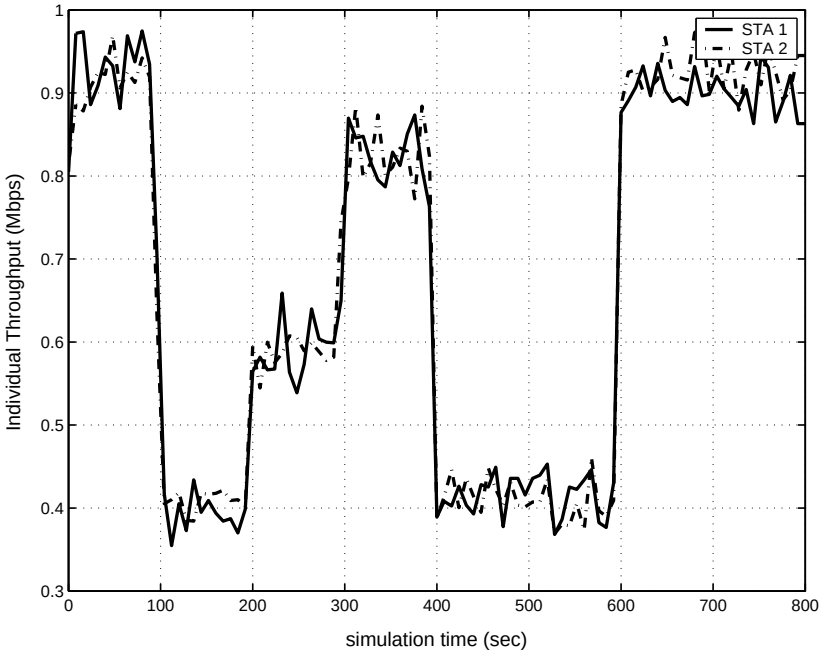


Figure 4.3: Individual throughput in a mixed data rate environment

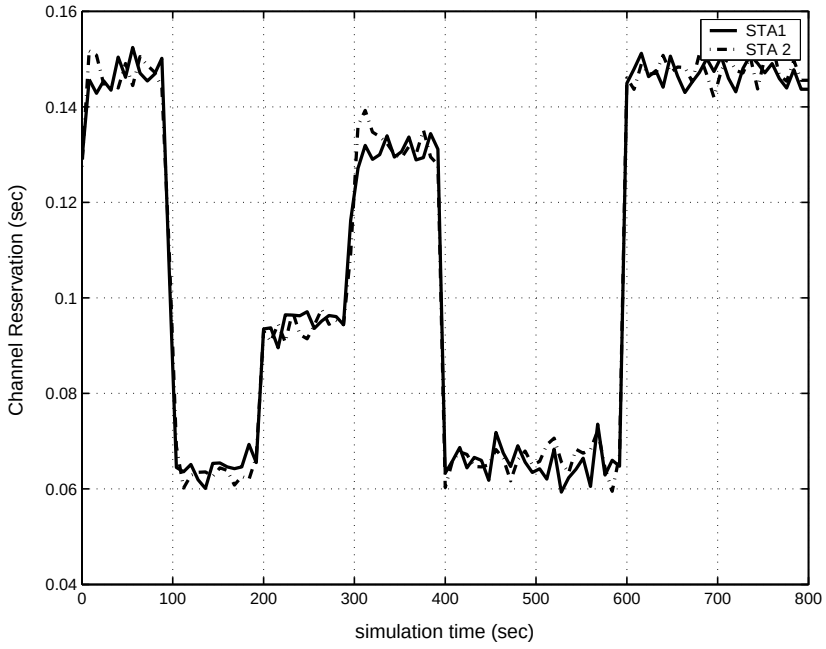


Figure 4.4: Channel reservation in a mixed data rate environment

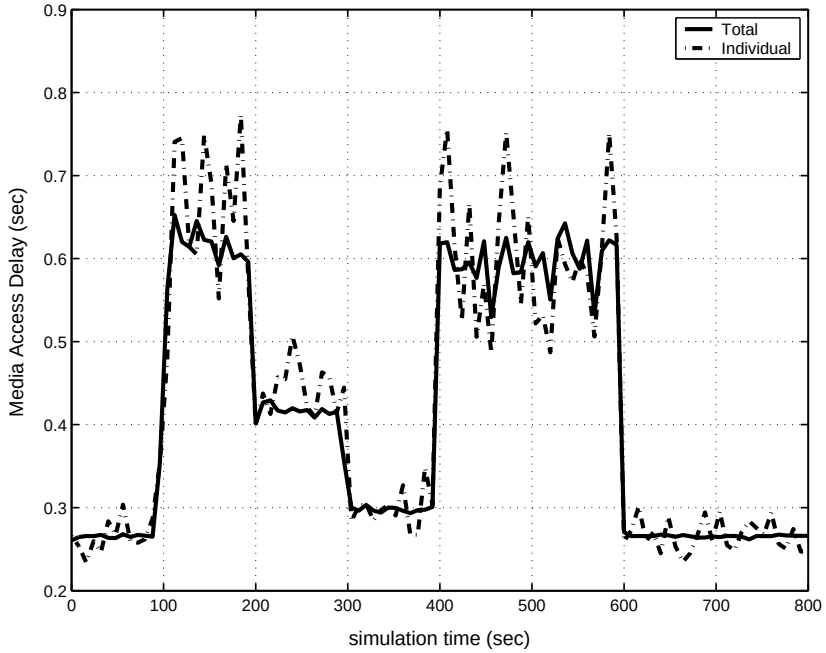


Figure 4.5: Media access delay in a mixed data rate environment

two traces for two stations. It can be inferred that stations observe the same individual throughput. As a result, stations with higher data rate observe the same throughput with the slow station.

Figure 4.4 proves our previous claim about the equal channel access. The figure shows channel reservation time of two stations which is also equal for other stations. Figure 4.5 shows the media access delay for the network and for a station. Media access delay is the waiting time for a packet in the MAC. It is easy to infer that when the throughput is low media access delay is high.

4.2 Individual Throughput Formulation

If we revisit the Markov Model we can analyze the individual throughput in detail. Wireless medium has two state Markov chain in which the states are either busy or idle. The wireless medium traverses from idle to busy with probability p. Therefore, the channel is not busy with probability $(1 - p)$. States in the Markov model can be defined as $b_{i,j}$ for $0 \leq i < m$ and $0 \leq j < W_i$. A station transmits when it is in $b_{0,j}$. The total probability is represented as follows,

$$1 = \sum_{i=0}^{m} \sum_{j=0}^{W_i - 1} b_{i,j} \tag{4.1}$$

and probability of transmission τ can be found after solving the balance equations as:

$$\tau = \frac{1}{\frac{(1-2p)(W+1)+pW(1-(2p)^m)}{2(1-2p)(1-p)}} \tag{4.2}$$

$$
\begin{array}{ll}
 & Basic Access Mechanism \\
T_s(R^i,P) & = T_{DATA}(R^i,P) + SIFS + \delta + T_{ACK}(R^i) + \delta + DIFS \\
T_c(R^i,P) & = T^*_{DATA}(R^i,P) + \delta + EIFS \\
 & RTS/CTS Access Mechanism \\
T_s(R^i,P) & = T_{RTS}(R^i) + SIFS + \delta + T_{CTS}(R^i) + SIFS + \delta \\
 & + T_{DATA}(R^i,P) + SIFS + \delta + T_{ACK}(R^i) + \delta + DIFS \\
T_c(R^i,P) & = T_{RTS}(R^i) + \delta + EIFS.
\end{array}
\tag{4.4}
$$

where τ depends on p. If there are n stations, $p = 1 - (1-\tau)^{n-1}$ if we make the independent assumption (See Chapter 3).

Following [5], "the throughput is the fraction of time the channel is used to successfully transmit payload bits." With probability P_{tr}, there is at least one transmission in the considered slot time and with probability P_s, a transmission occurring on the channel is successful:

$$
\begin{aligned}
P_{tr} &= 1 - (1-\tau)^n \\
P_s &= n\tau(1-\tau)^{n-1}.
\end{aligned}
\tag{4.3}
$$

Let's express the total throughput S as the ratio:

$$
S = \frac{P_s E[P]}{(1-P_{tr})\sigma + P_s T_s(R^i) + (P_{tr} - P_s)T_c(R^i)}. \tag{4.5}
$$

where $E[P] = P$ is the average packet payload size and constant for each station. P_s is the probability for successful transmission. The average length of an empty slot time is with probability $(1 - P_{tr})$ and with probability $(P_{tr} - P_s)$, there is a collision. If the station have data rate R^i, $T_s(R^i, P)$ is the average time the channel is sensed busy due to a successful transmission; and $T_c(R^i, P)$ is the average time the channel is sensed busy

by each station during a collision, represented in (4.4).

$T_{DATA}(R^i, P)$ is the time taken to send a packet with size P; and $T_{RTS}(R^i)$, $T_{CTS}(R^i)$, $T_{ACK}(R^i)$ are the times taken to send the corresponding frames. $T^*_{DATA}(R^i, E[P^*])$ stands for average time taken to send $E[P^*]$, which is the average length of the longest packet payload involved in a collision. When all packets have the same size, $E[P] = E[P^*] = P$ [5]. The propagation delay is given by δ.

We can also find the throughput in a different way. Let's say stations operate at different data rates. Let's consider the first K event for a network with n nodes which could take T time. If L is number of empty slots, M is number of successful transmissions and Y is the wasted time in collisions, T is:

$$T = L\sigma + \sum_{i=1}^{n} X(i)T_s(R(i)) + Y \tag{4.6}$$

where $\sum_{i=1}^{n} X(i) = M$. $X(i)$ is the number of successful transmission and $R(i)$ is the data rate for station i; If $P(i)$ is the packet size for station i the throughput S_i is:

$$S_i = \frac{X(i)P(i)}{T}. \tag{4.7}$$

Now it is important to find $\frac{X(i)}{M}$ since we know that total throughput S is:

$$S = \frac{ME[P]}{T} \tag{4.8}$$

If there are M successful transmissions and assuming that the stations have the equal chance of transmission, we can conclude that $X(i) = \frac{M}{n}$. This is also followed by the

independent assumption since it is clear that epochs are independent and we assume a two state Markov channel. In each virtual event, the probability of having that virtual event empty is $(1-\tau)^n$. Thus, $L = K(1-\tau)^n$ and in the same way, finding that epoch occupied for a successful transmission of station i is $\tau(1-\tau)^{n-1}$. Consequently $X(i) = K\tau(1-\tau)^{n-1}$. Since there are n stations, then $M = Kn\tau(1-\tau)^{n-1}$. We can conclude that stations access the channel equally since channel access is obtained when the opportunity arises and it is used as long as it is needed. Hence, the duration is determined by the station.

As a result, the individual throughput is $S_i = \frac{1}{n}S$ since the throughput is fairly distributed regardless the station's data rate. This is proven with simulation.

4.3 Formulation for Mixed Data Rate

We evaluate the throughput when different stations have different data rates but same packet size. The protocol gives each station the same chance to transmit, and different data rates only affect average slot duration. Suppose there are n stations, D different data rates, $R^1 < \cdots < R^D$, and n^i stations have rate R^i with corresponding slot durations $T_s(R^i)$ and $T_c(R^i)$. For example if there are eight stations $N = 8$ and four data rates $D = 4$,

S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8
R^1	R^2	R^1	R^3	R^4	R^1	R^4	R^1
$T_s(R^1)$	$T_s(R^2)$	$T_s(R^1)$	$T_s(R^3)$	$T_s(R^4)$	$T_s(R^1)$	$T_s(R^4)$	$T_s(R^1)$
$T_c(R^1)$	$T_c(R^2)$	$T_c(R^1)$	$T_c(R^3)$	$T_c(R^4)$	$T_c(R^1)$	$T_c(R^4)$	$T_c(R^1)$

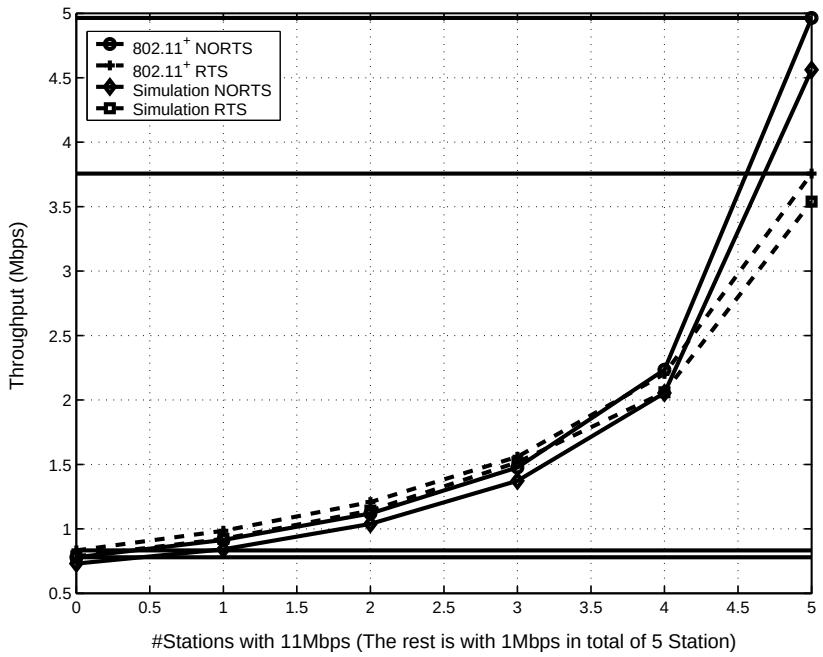


Figure 4.6: Throughput verification for the mixed data rate formulation

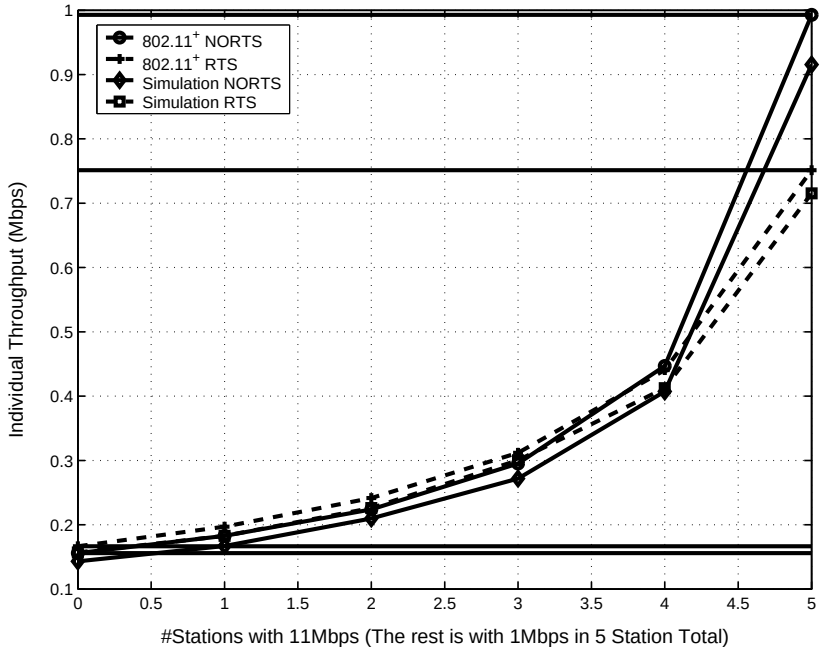


Figure 4.7: Individual throughput verification for the mixed data rate formulation

where for instance $R(5) = R^4$ and

$n^1 = 4$	T_s^1	$T_s(R^1)$	T_c^1	$T_c(R^1)$
$n^2 = 1$	T_s^2	$T_s(R^2)$	T_c^2	$T_c(R^2)$
$n^3 = 1$	T_s^3	$T_s(R^3)$	T_c^3	$T_c(R^3)$
$n^4 = 2$	T_s^4	$T_s(R^4)$	T_c^4	$T_c(R^4)$

which is clustering from the highest duration to the lowest duration. Since we consider same packet size P, duration values only dependent on rate R.

The successful duration value can be evaluated by averaging the successful duration values of each station as only one station is involved in a successful duration. Since we know that each station has P_s/n as the probability of having a successful transmission, then the new successful duration value $\bar{T}_s$ is given by (4.10). $\bar{T}_s$ is basically represented as

$$\bar{T}_s = E[T_s^i]. \tag{4.9}$$

When calculating the collision duration, we have to consider the stations that are involved in the collision and how many times they are involved. During a collision, duration value for that collision is determined by the station that has the lowest data rate. The average collision duration $\bar{T}_c$ is given by (4.11). This formulation considers the lowest data rate stations first. As they are grouped with n^i, the second step is to determine how many times the stations are involved in the collision with the corresponding probability, and the number of stations in the collision. This process is repeated from the lowest data rate to the highest. The collusion occurrences with the lower data rate is subtracted from the

count.

Note that the formula (4.11) is present for all sets of data rate choices. We know that 802.11b has 4 and 802.11a has 8 data rate options. Another key characteristic of the formula is it is independent of the Markov model applied. Any model that can give the transmission probability τ of a station could use the formula to obtain the duration values. The formula works if the stations have different packet sizes. In that case clustering will be based on duration. A formulation is presented in Chapter 9.

The throughput of a station is now given by (4.12). Note that the individual throughput S_i is the *same* for all stations and total throughput S is nS_i since they have the same packet size and they equally access the channel.

$$\bar{T}_s = \frac{P_s}{n}\sum_{i=1}^{D} n^i T_s^i \tag{4.10}$$

$$\bar{T}_c = \sum_{i=1}^{n-1}\sum_{j=1}^{D}\sum_{k=1}^{n^j} \begin{pmatrix} n-k-\sum_{l=1}^{j-1} n^l \\ i \end{pmatrix} \times T_c^j \tau^{i+1}(1-\tau)^{n-1-i} \tag{4.11}$$

$$S_i = \frac{1}{n}\frac{P_s E[P]}{(1-P_{tr})\sigma + \bar{T}_s + \bar{T}_c} \tag{4.12}$$

4.4 Verification

In the performance analysis, we have 5 nodes and $D = [\ 1 \quad 2 \quad 5.5 \quad 11\ ]$Mbps. $E[P]$ is 1000 bytes and stations operate in saturation throughput. Note that RTS and CTS packets are sent with 1Mbps all the time.

We started each station with 1Mbps and shifted one of them in each step to 11Mbps.

At the fifth iteration we have 5 stations each with 11Mbps. We compare our results with a simulation in the OPNET platform.

Figures 4.6 and 4.7 show that the analytical formula (4.12) closely approximates the actual performance. Solid lines represent the case when all stations have the same data rate. The individual throughput of the stations is also found to be equal in the simulation, and the throughput distribution among stations is verified to be equal. One of the individual realization is plotted in Figure 4.7.

As can be inferred from the graph, when there are 4 stations with 11Mbps and one station with 1Mbps, the throughput performance is almost half that of when all are with 11Mbps. A lower data rate causes a considerable degradation for all the stations.

4.5 Discussion

This performance anomaly has been studied in [42] with a simple model. In contrast, we used an analytical Markov model which gave a closed formula. We came across a preprint paper [35] which builds its model on top of [5] with an additional post-backoff state which we discussed in Section 3.6. The model in [35] considers only two data rate options low and high and it is not as complete as ours.

Remark

This chapter is published in the Proceedings of IEEE GLOBECOM 2004 Conference titled “Throughput Formulation and WLAN Optimization in Mixed Data Rates for IEEE 802.11 DCF Mode” and authors are Mustafa Ergen and Pravin Varaiya.

Chapter 5

Delay Analysis

5.1 Introduction

Characterizing the delay of IEEE 802.11 is crucial since the quality of service of any implementation depends on a tight delay budget [43, 44, 45]. Since 802.11 DCF is a random access mechanism, the inter-arrival time between successful packet transmission of a station is a random variable. The important parameters we seek are expected value and variance of the jitter with the increase in the number of stations and load. We provide an important characterization of the network which could be used to tune the network parameters accordingly. We consider a network with n stations without any hidden or exposed terminals.

5.2 Delay Model

We can easily infer from the analytical Markov Model of the DCF that the random variable $(t_i - t_{i-1})$ and $(t_{i+1} - t_i)$ are independent if we define t_i as the i^{th} time of a station when it sends a packet successfully. Observing only one interval can give the delay characteristic of the network. As stated earlier, the interval starts and ends with a successful packet transmission of a station. After a successful transmission, a station resets its backoff counter to CW_{min} and selects a backoff counter value in random from $[0, CW_{min}]$. Then it waits in the backoff process and decrements the counter whenever there is a idle slot time (σ). It suspends decrementing the counter when there is activity in the medium. Let's call random variable $\top_i$ the time taken to come to backoff counter 0 from the selected backoff counter value in backoff level i. When the backoff counter is 0, the station transmits the packet. The transmission may be successful with probability q, or it can experience a collision with probability p, where $p + q = 1$. We know that the transmission time is fixed and given by T_s and T_c, for successful and unsuccessful transmission, respectively. If the transmission is unsuccessful, then the station doubles its contention window and selects another backoff counter. Of course after the 6^{th} try the contention window reaches maximum CW_{max} and stays there. Therefore, the random variable for waiting time in backoff level is $\top_6$ for $i > 6$.

Figure 5.1 illustrates the delay structure. From the successful transmission to another, all possibilities are presented. Note that in the 802.11^+ model, the virtual slot is defined differently from 802.11^b. Transmission times must included times since we know that when backoff counter is 0, a transmission time is definitely consumed. On the other hand,

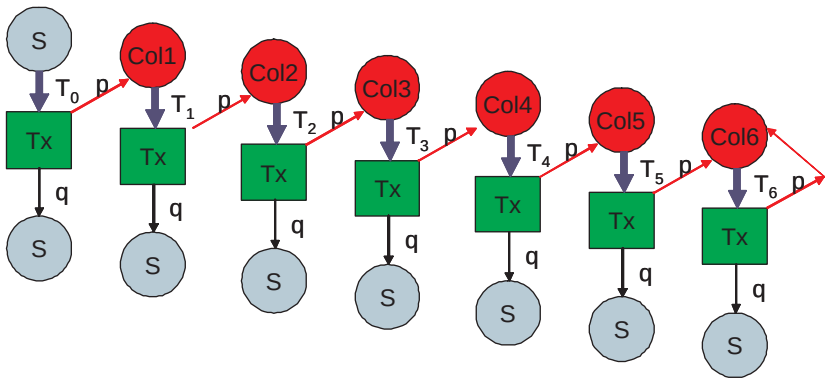


Figure 5.1: Delay Analysis

in 802.11^b the transmission time is included in the virtual slot.

If station transmits successfully after its X^{th} unsuccessful attempt the total delay Γ is

$$\Gamma = \sum_{i=0}^{\infty} 1(X = i)\Gamma_i \tag{5.1}$$

and Γ_i is

$$\Gamma_i = T_s + iT_c + \sum_{k=0}^{i} \top_k,\ \top_k = \top_6\ for\ k > 6. \tag{5.2}$$

Probability mass function of delay $P(\Gamma)$ is

$$P(\Gamma) = \sum_{i=0}^{\infty} P(X = i)P(\Gamma|X = i) \tag{5.3}$$

where $P(X = i)$ is qp^i and $P(\Gamma|X = i)$ requires convolution of $P(\top)$s because they are independent. Expected value $E(\Gamma)$ can be represented by

$$E[\Gamma] = \sum_{i=0}^{\infty} P(X = i)E[\Gamma_i] \tag{5.4}$$

where $E[\Gamma_i]$ is

$$E[\Gamma_i] = T_s + iT_c + \sum_{k=0}^{i} E[\top_k],\ \top_k = \top_6\ for\ k > 6 \tag{5.5}$$

and $Var[\Gamma]$ is

$$Var[\Gamma] = \sum_{i=0}^{\infty} P(X = i)E[\Gamma_i^2] - (E[\Gamma])^2. \tag{5.6}$$

The main element of these equations is finding the $\top$ for each backoff level. We look at $\top$ for 802.11^+ and 802.11^b models. As it can be inferred easily $\top$ has uniform distribution for the 802.11^b model since average length of a virtual event is used in order to identify the time that is passed for a station to decrement its backoff counter by one. On the other hand for 802.11^+ the time to decrement the backoff counter by one can range from a slot time to forever. This is because of the self loop in the backoff states.

We start with our model and compare first order statistics. Let's first consider Figure 5.2 where a simple model with four backoff levels is represented.

From the 802.11$^+$ model, immediate transmission occurs with probability $1/4$, and transmission after an empty slot is with probability $1/4q$. The probability of an empty slot is q and transmission is p. Before looking at different combinations, let's define a counter "$Tick$" and increment it with either an empty slot or transmission where we assume $T_s = T_c = T_{tx}$. Then the probabilities for each $Tick$ is as follows:

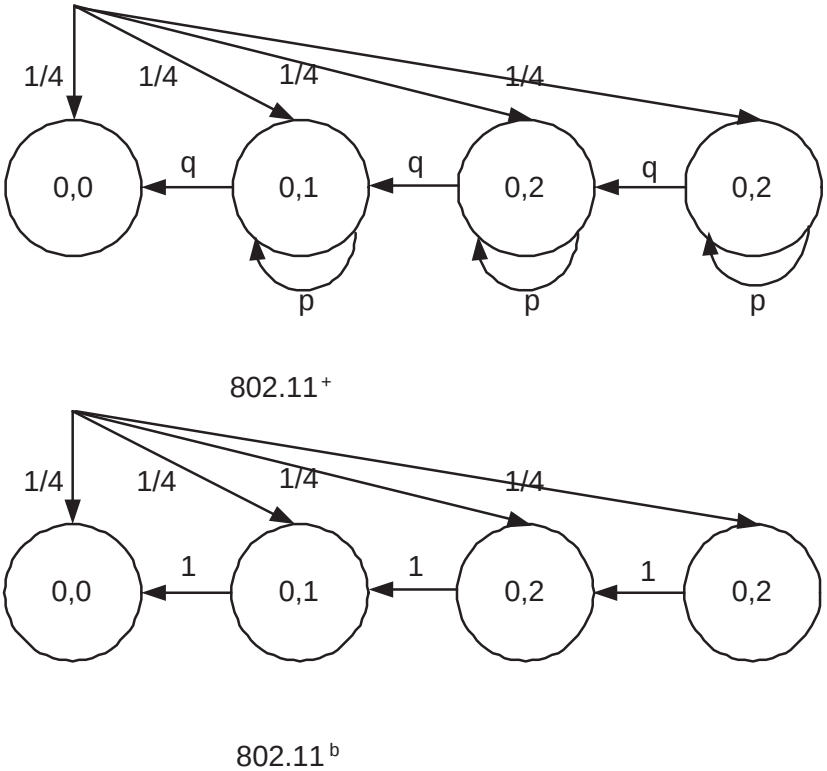


Figure 5.2: Delay analysis with one-level backoff (CW=4)

Tick	0	σ+kT_{tx}	2σ+kT_{tx}	3σ+kT_{tx}
0	$\frac{1}{4}$			
1		$\frac{1}{4}$q		
2		$\frac{1}{4}$qp	$\frac{1}{4}$q^2	
3		$\frac{1}{4}$qp^2	$2\frac{1}{4}$q^{2}p^1	$1\frac{1}{4}$q^3
4		$\frac{1}{4}$qp^3	$3\frac{1}{4}$q^{2}p^2	$3\frac{1}{4}$q^{3}p
5		$\frac{1}{4}$qp^4	$4\frac{1}{4}$q^{2}p^3	$6\frac{1}{4}$q^{3}p^2
6		$\frac{1}{4}$qp^5	$5\frac{1}{4}$q^{2}p^4	$10\frac{1}{4}$q^{3}p^3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k+1		$\frac{1}{4}$qp^k	k$\frac{1}{4}$q^{2}p^{k-1}	$\begin{pmatrix} k \\ 2 \end{pmatrix}\frac{1}{4}$q^{3}p^{k-2}
$\vdots$		$\vdots$	$\vdots$	$\vdots$
∞		0	0	0

Examining this structure shows that the pattern obeys *Pascal's triangle* which is as follows;

$$a_{nr} \equiv \frac{n!}{r!(n-r)!} \equiv \begin{pmatrix} n \\ r \end{pmatrix} \tag{5.7}$$

starting with $n = 0$. Of course, when $n > W$, r stays at W since W is the number of backoff levels which is 3 right now. We can construct the probability mass function with this information. The probability of completing a successful transmission in $(k\sigma + iT_{tx})$ time frame is $\frac{1}{W}\begin{pmatrix} i+k-1 \\ i-1 \end{pmatrix} q^k p^i$. This formula helps us to find probability mass function $P(\top^+)$.

Since average length for a virtual slot is fixed, it is easier to look at $\top^b$, call this T^b_{vs}:

Tick	0	T^b_{vs}	$2T^b_{vs}$	$3T^b_{vs}$
0	$\frac{1}{4}$			
1		$\frac{1}{4}$		
2			$\frac{1}{4}$	
3				$\frac{1}{4}$

Then the next step is finding the convolution of $\top$'s since as we know that each $\top_i$ is independent of $\top_k$ where $i \neq k$.

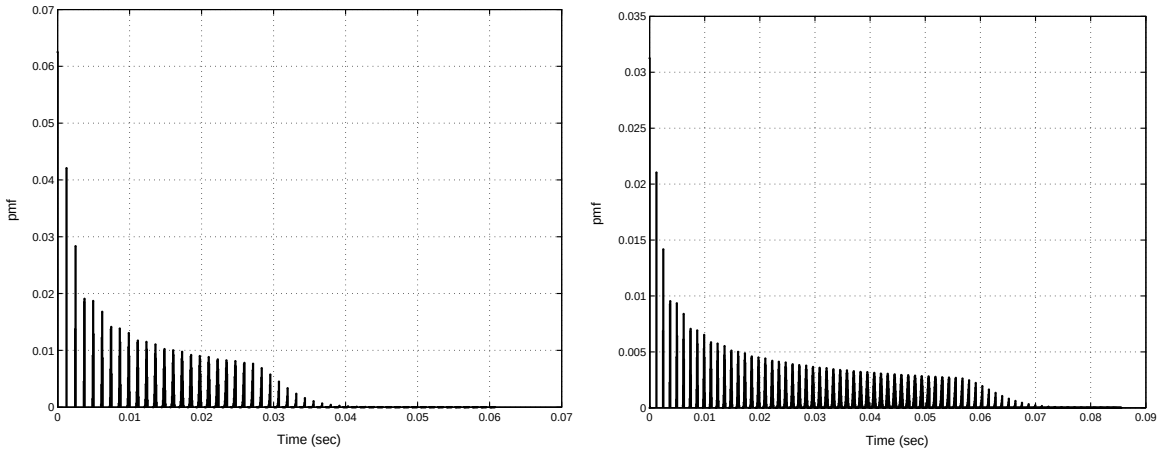
If we consider the backoff level probabilities of 802.11^+ in Figures 5.3(a)-5.6, we can approximate that they resemble a uniform distribution except at the head and tail. We can find from the first order statistics that they both give approximately similar results.

5.3 Performance Analysis

The following results are obtained with IEEE 802.11b PHY parameters where stations are with 1Mbps data rate and slot time, σ, is $50\mu sec$ and transmission time is $T_{tx} = 0.0012sec$.

Figure 5.7 shows the average delay for a successful transmission. As one can infer from the figure, average delay is increasing with the number of stations. This is expected since probability of successful transmission decreases with the number of stations which affects the collision probability. As a result stations wait longer and experience more collisions than before.

Figure 5.8 illustrates the variance of the delay. As expected variance increases with the



(a) $P(\top_0)$ (b) $P(\top_1)$

Figure 5.3: pmfs for Backoff Levels 0 and 1 ($P(\top_0)$ and $P(\top_1)$)

number of stations. The main reason why 802.11^+ model shows a higher delay is because of the self-loop in backoff since it takes consecutive transmission into consideration which introduces more randomness than before. It is not guaranteed that the station decrements the backoff counter in every event.

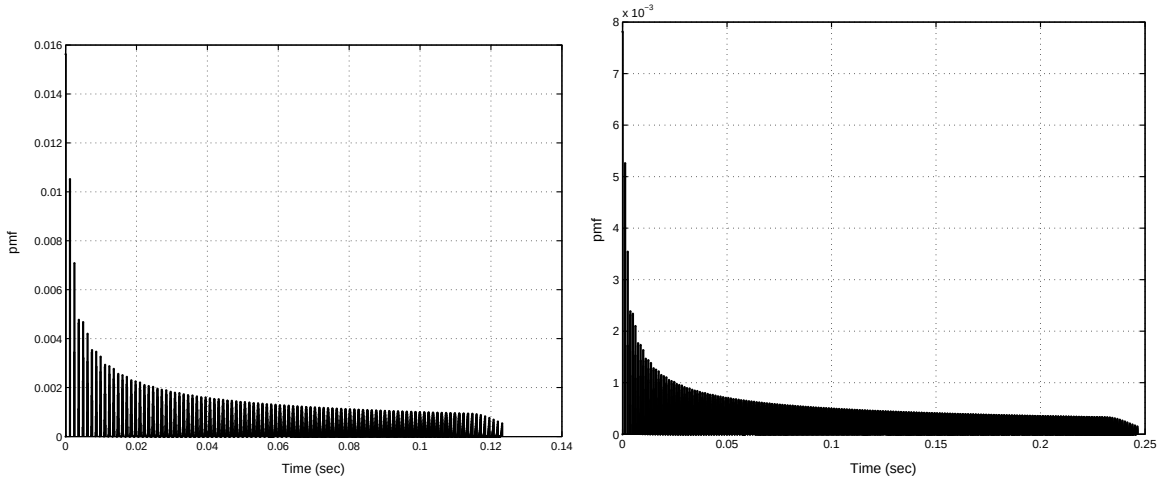
Figure 5.9 shows the probability mass function for a network with 30 stations. The result is semi-averaged meaning that average time for a virtual event is used. The actual length of a virtual event is a random variable and has a pmf as follows:

$$pmf(virtual - event) = (1 - P_{tr})\delta(t - \sigma) + P_s\delta(t - T_s) + (P_{tr} - P_s)\delta(t - T_c). \quad (5.8)$$

The average delay is found to be $0.03sec$ as shown in Figure 5.7. Let's modify the formula for unsaturated traffic. Then the first order statistics are as follows:

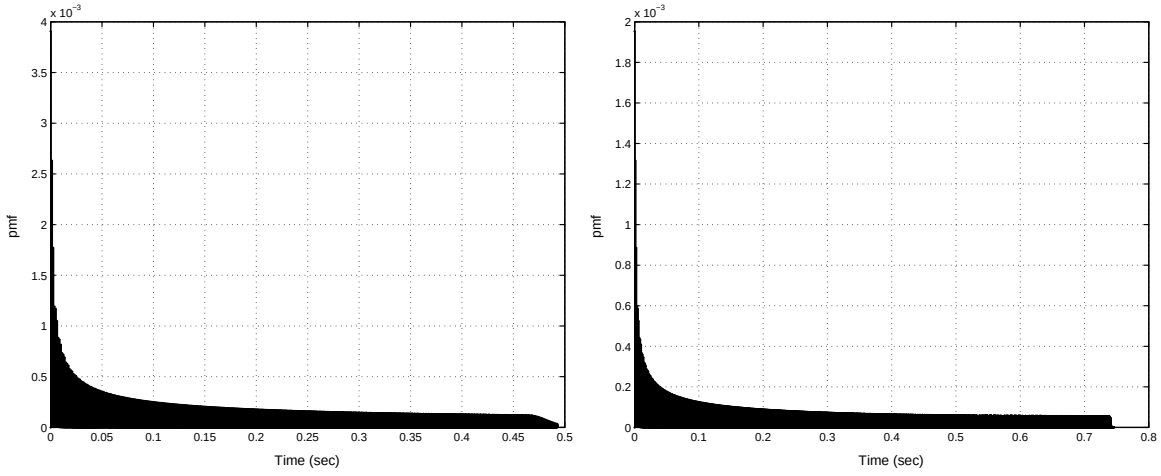
Expected value $E[\Gamma_{US}]$ can be represented by

$$E[\Gamma_{US}] = \frac{1 - \lambda}{\lambda}\sigma + E[\Gamma] \quad (5.9)$$



(a) $P(\top_2)$ (b) $P(\top_3)$

Figure 5.4: pmfs for Backoff Levels 2 and 3 ($P(\top_2)$ and $P(\top_3)$)



(a) $P(\top_4)$ (b) $P(\top_5)$

Figure 5.5: pmfs for Backoff Levels 4 and 5 ($P(\top_4)$ and $P(\top_5)$)

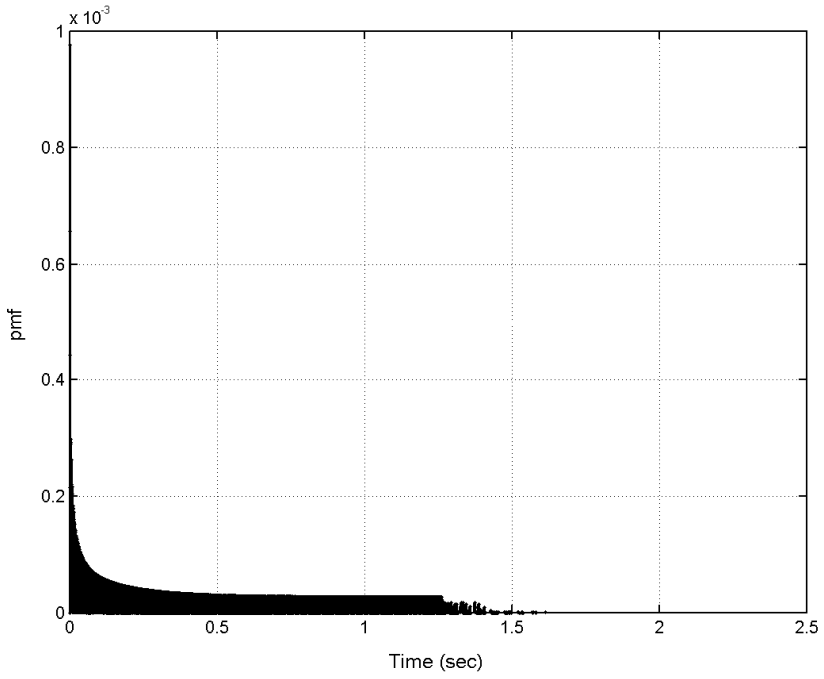


Figure 5.6: pmf for Backoff Level 6 for $n = 30$ ($P(\top_6)$)

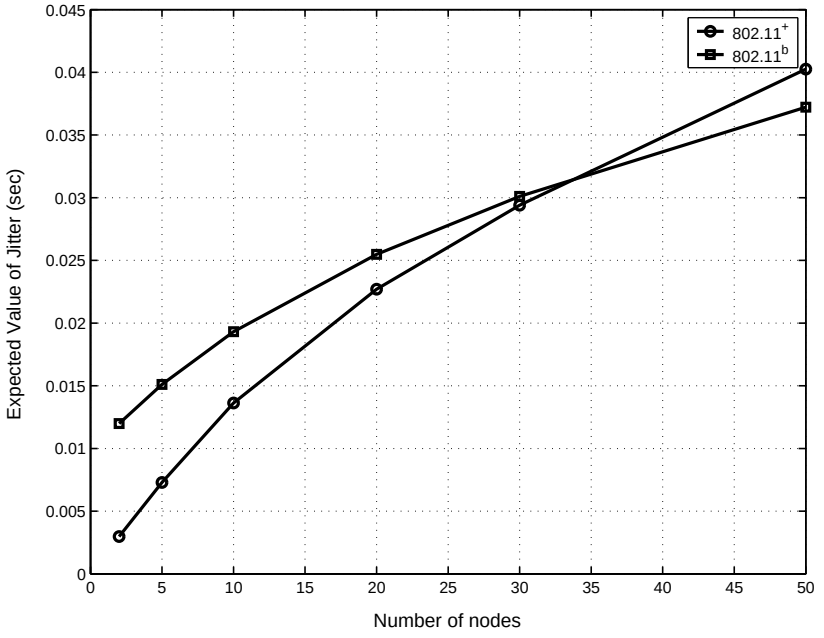


Figure 5.7: Average delay with respect to number of stations ($E[\Gamma]$)

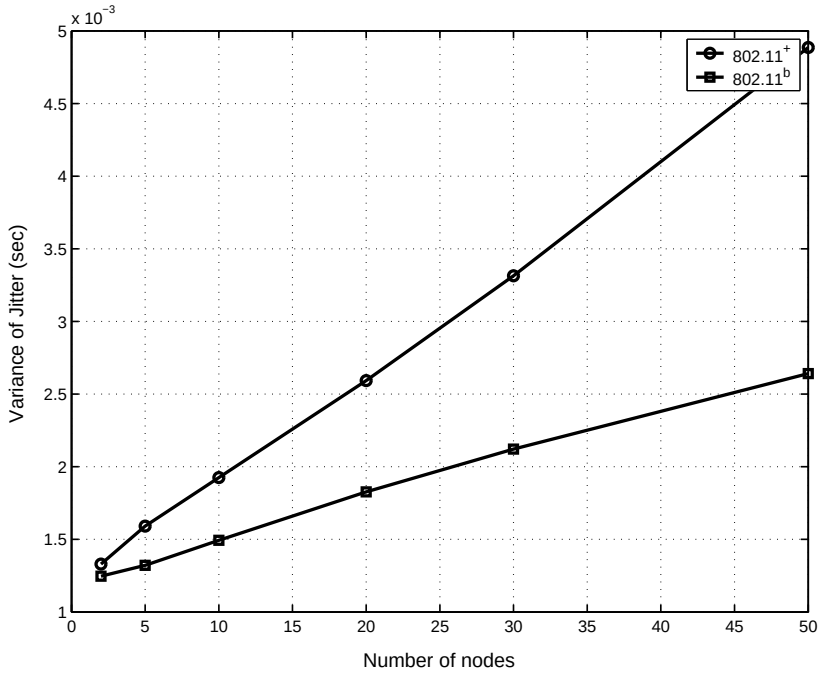


Figure 5.8: Variance of the delay with respect to number of stations ($Var[\Gamma]$)

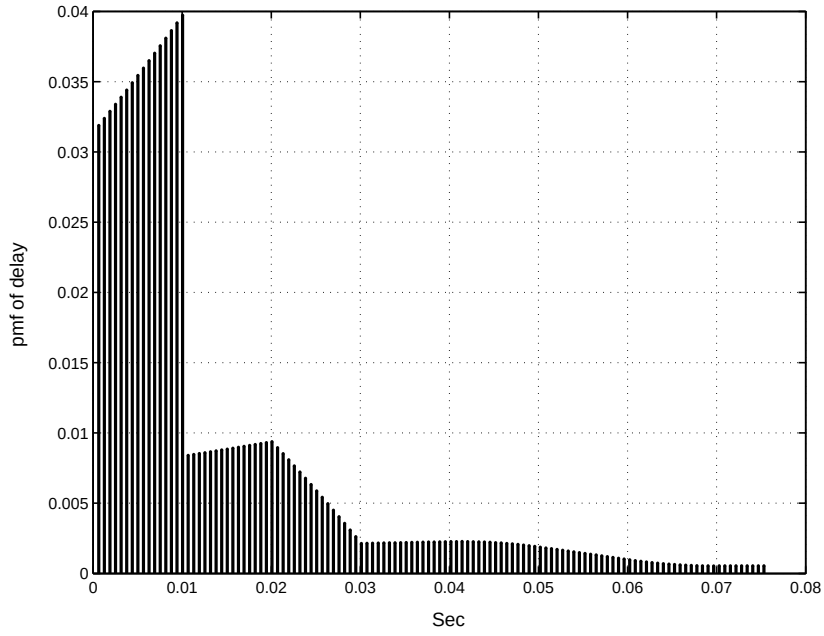


Figure 5.9: Probability Mass Function of the Delay ($P(\Gamma)$) for $n = 30$

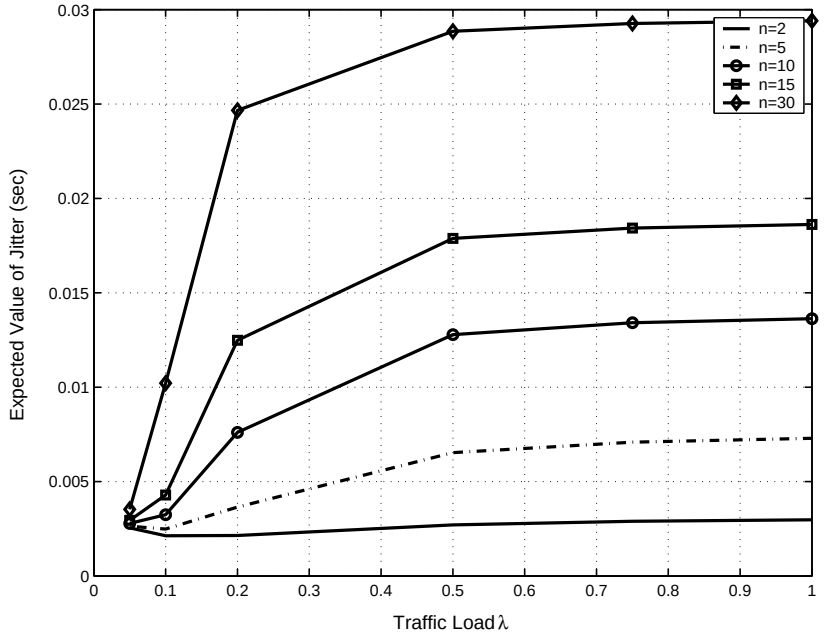


Figure 5.10: Average delay with respect to load ($E[\Gamma_{US}]$)

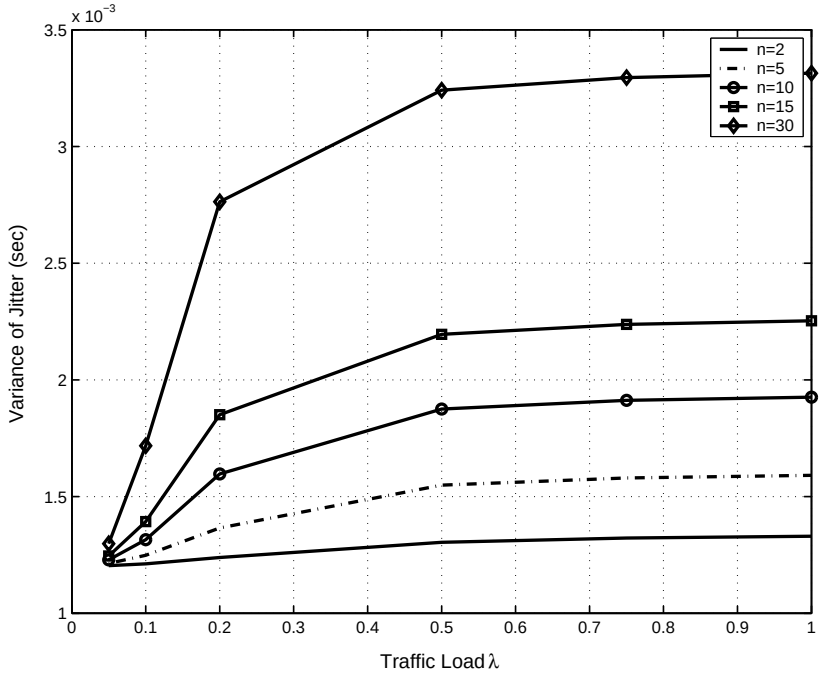


Figure 5.11: Variance delay with respect to load ($Var[\Gamma_{US}]$)

and $Var[\Gamma_{US}]$ is

$$Var[\Gamma_{US}] = \frac{1-\lambda}{\lambda^2}\sigma^2 + Var[\Gamma] \tag{5.10}$$

Figures 5.10 and 5.11 show the delay characteristics with respect to increasing load for different network sizes. One can see that delay increases with the load and with the number of stations. In order to implement quality of service a network can be configured based on the delay budget.

Chapter 6

Application: Admission Control

6.1 Introduction

The CSMA/CA scheme which is the underlying mechanism in DCF of IEEE 802.11 offers random based access. This random access scheme as represented in Chapter 3 shows a behavior where the overall throughput at first increases and later decreases with the number of stations. Chapter 4 shows that the total throughput with different individual data rates exhibits the same behavior. Thus there is a need for an admission control (AC) mechanism that restricts access in order to maintain high system throughput. We introduce one AC mechanism.

We consider AC within the BSS using DCF, so we formulate our problem for a network with one access point (AP) connected to several stations. The goal is to maintain maximum throughput as the number of active stations increases. In inter-BSS admission control, there is more than one AP and in addition to restricting access, the mechanism may assign stations to different APs.

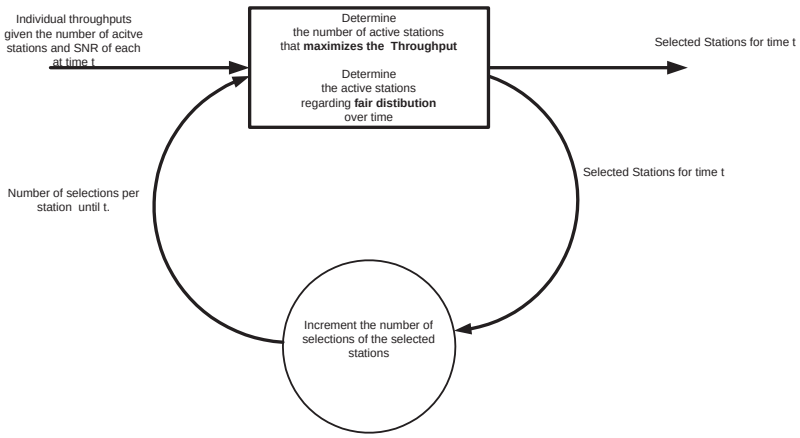


Figure 6.1: Control algorithm for admission control

6.2 Intra-BSS Admission Control

The goal is to maximize system throughput, while maintaining fairness in the sense of equalizing the chance each station has to transmit. As suggested in Figure 6.1, the algorithm makes two decisions. It first determines the number of active stations in each period to maximize throughput; it then selects the stations to achieve fairness.

We use the following notation to describe the algorithm. The time interval is divided into periods indexed $t \in [1, \textit{Endtime}]$. For each station $i \in [1, N]$, $x_i(t) = 1$ or 0, accordingly as station i is or is not selected in period t, and

$$x_i^{Total}(t) = \sum_{s \leq t} x_i(s).$$

For $I \subset \{1, \cdots, N\}$, let $S(I)$ be the system throughput if subset I of stations is selected.

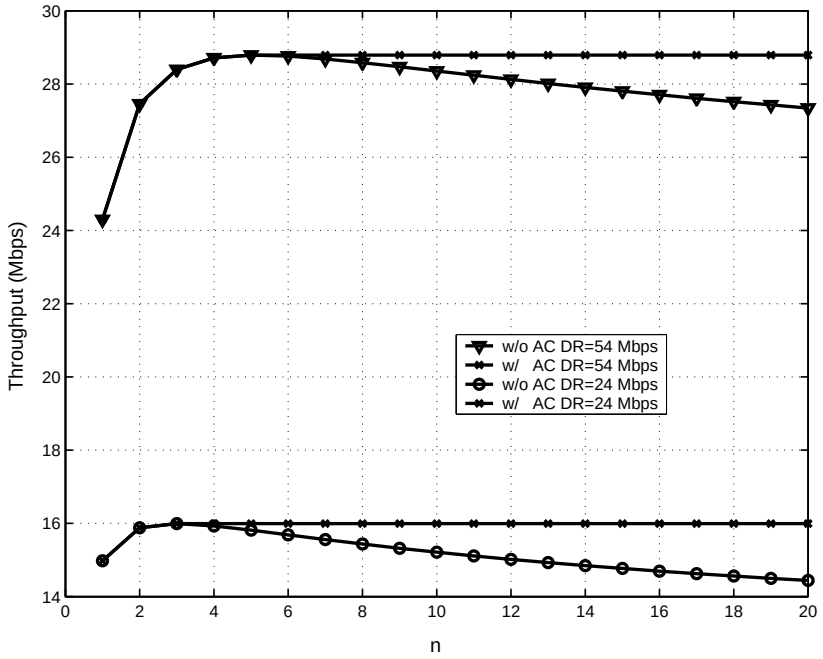


Figure 6.2: Admission Control when data rates are same: Throughput

Given the data rates of every station, $S(I)$ is given by as follows.

$$S(I) = \frac{P_s(I)E[P(I)]}{(1 - P_{tr}(I))\sigma + \bar{T}_s(I) + \bar{T}_c(I)} \tag{6.1}$$

At each t, the mechanism activates the subset $I(t)$ of stations that solves the following optimization problem:

$$\begin{aligned} \max_{I(t)} \quad & S(I(t)) - KC(t) \\ \text{s. t.} \quad & C(t) = \max_i \ x_i^{\textit{Total}}(t) - \min_i \ x_i^{\textit{Total}}(t) \end{aligned} \tag{6.2}$$

Here $K > 0$ is a constant. $C(t)$ is the inequality among stations at time t, so the objective function strikes a balance between throughput and fairness, depending on K.

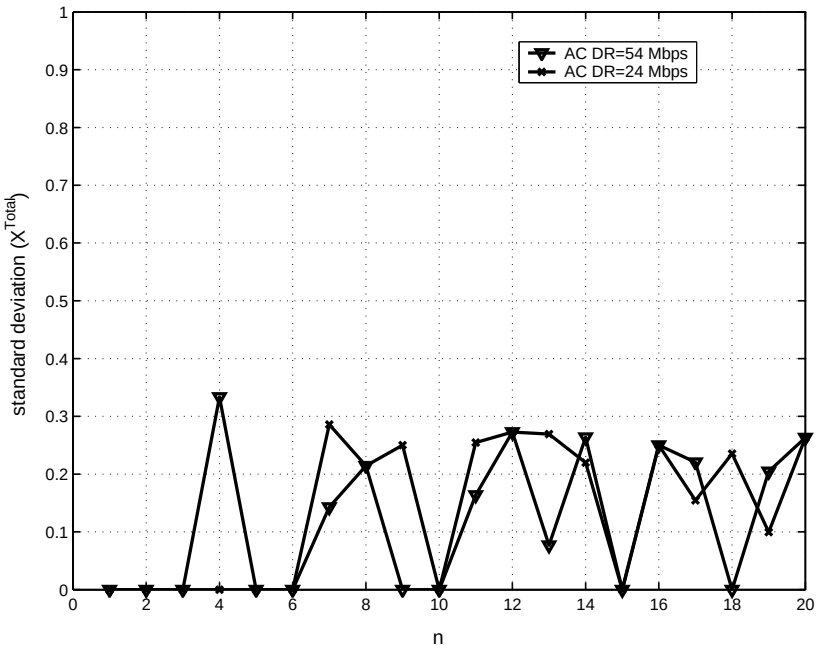


Figure 6.3: Admission Control when data rates are same: Fairness metric

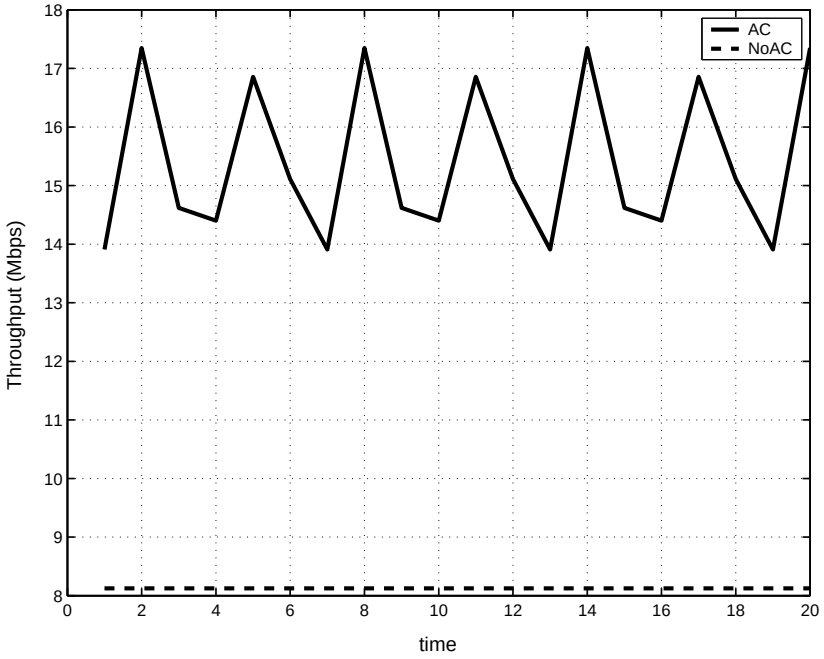


Figure 6.4: Admission Control when there is no mobility: Throughput

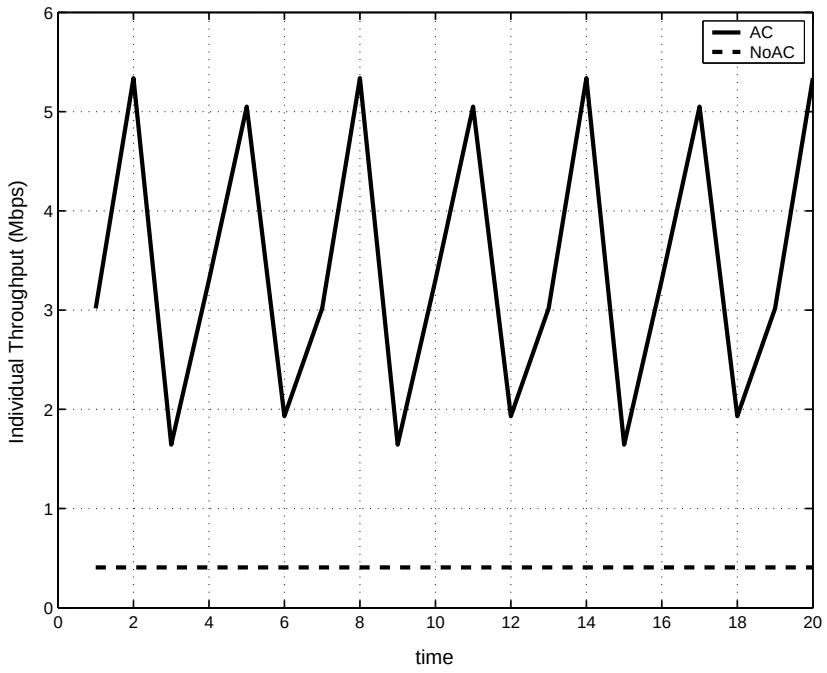


Figure 6.5: Admission Control when there is no mobility: Individual throughput

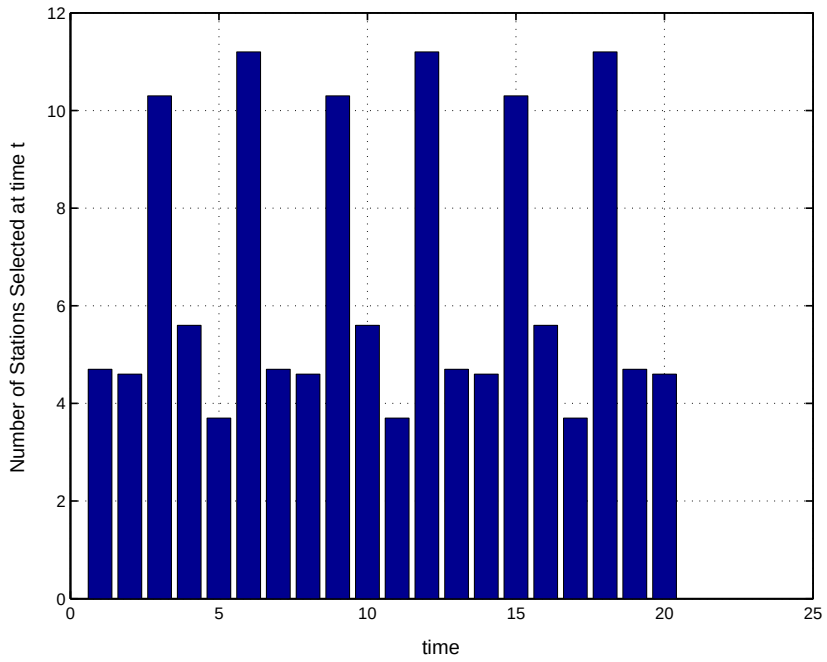


Figure 6.6: Admission Control when there is no mobility: Number of active stations over time

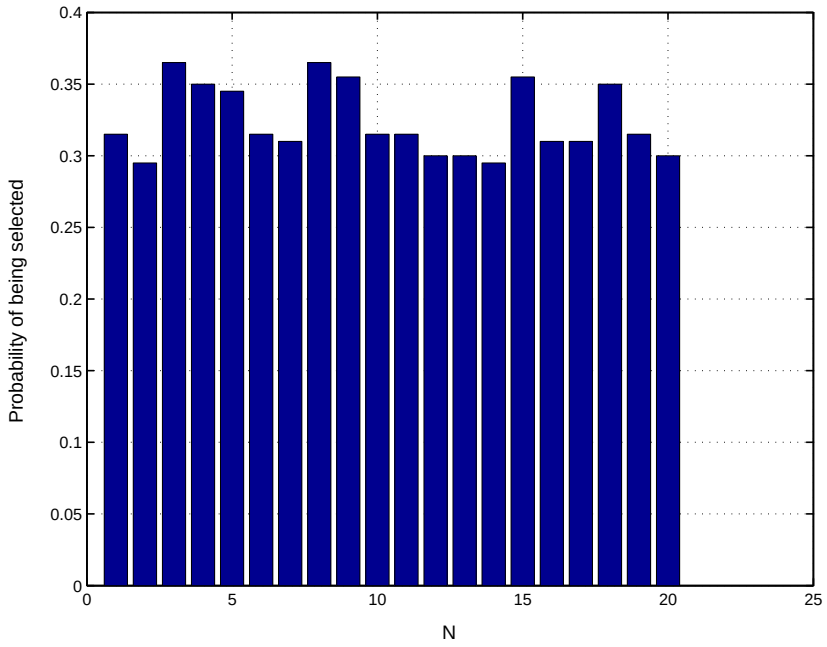


Figure 6.7: Admission Control when there is no mobility: Probability of being selected at time t

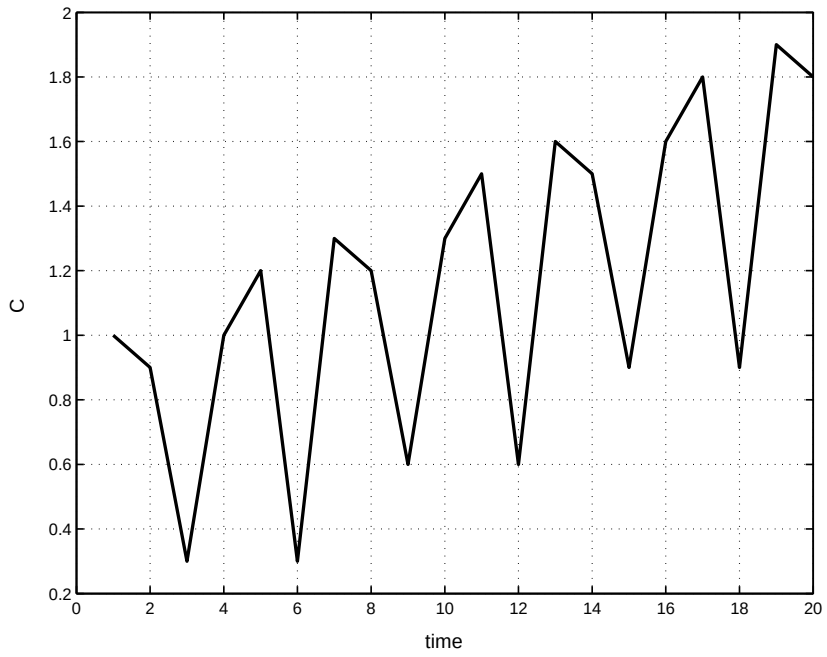


Figure 6.8: Admission Control when there is no mobility: Fairness Constraint

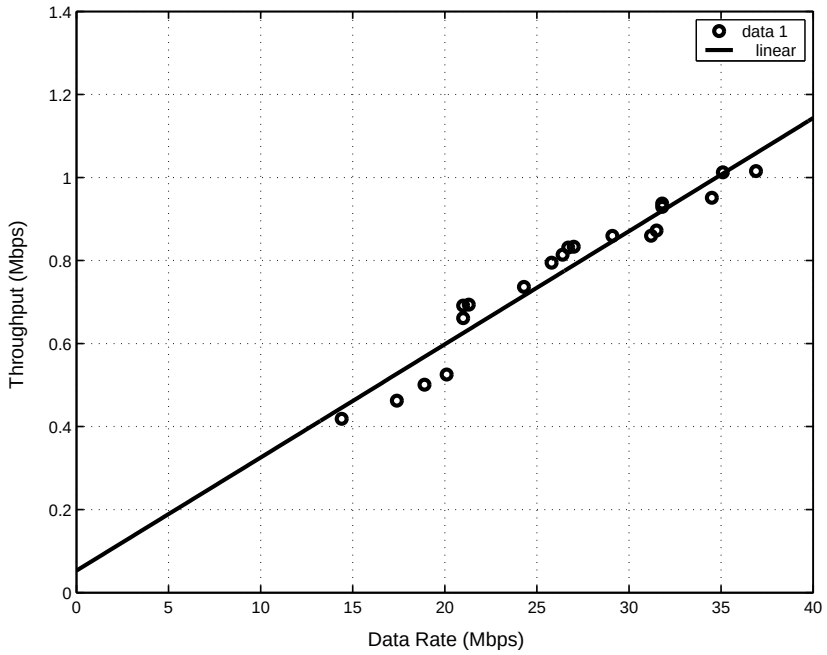


Figure 6.9: Admission Control when there is no mobility: Throughput vs Data rate

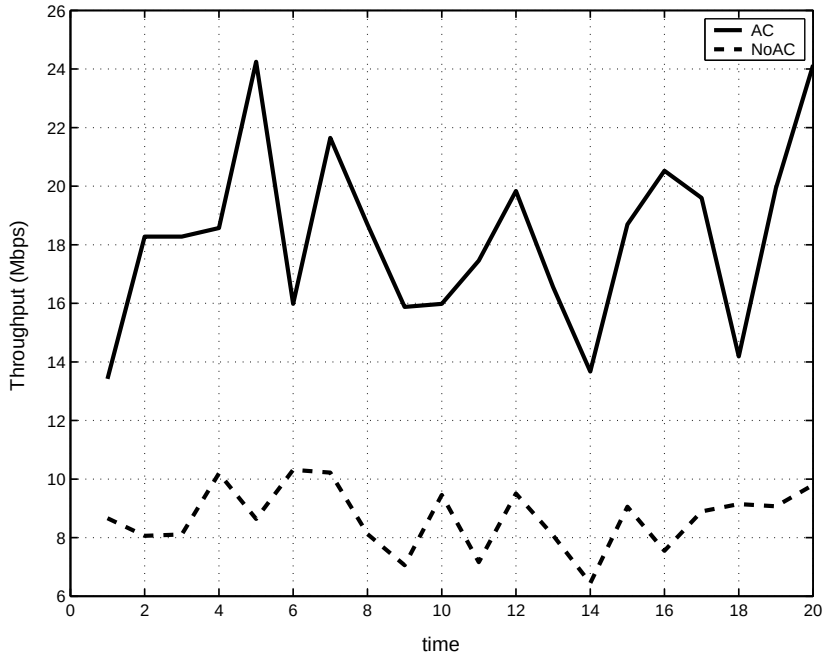


Figure 6.10: Admission Control when there is mobility: Throughput

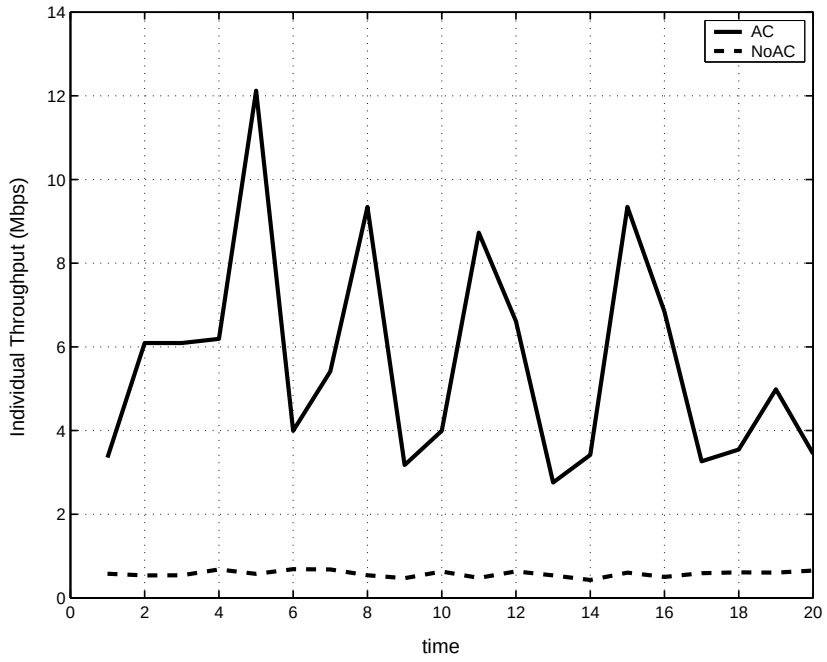


Figure 6.11: Admission Control when there is mobility: Individual throughput

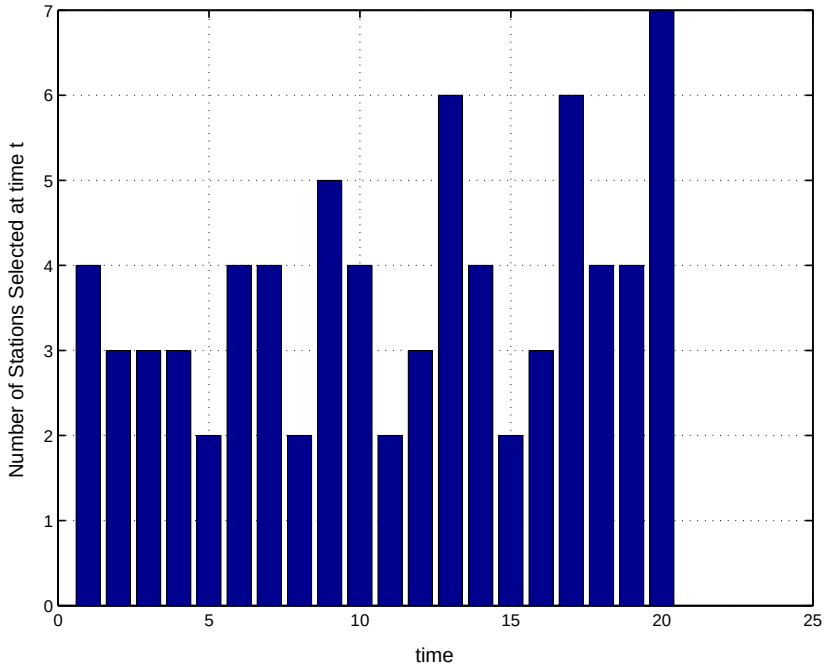


Figure 6.12: Admission Control when there is mobility: Number of active stations over time

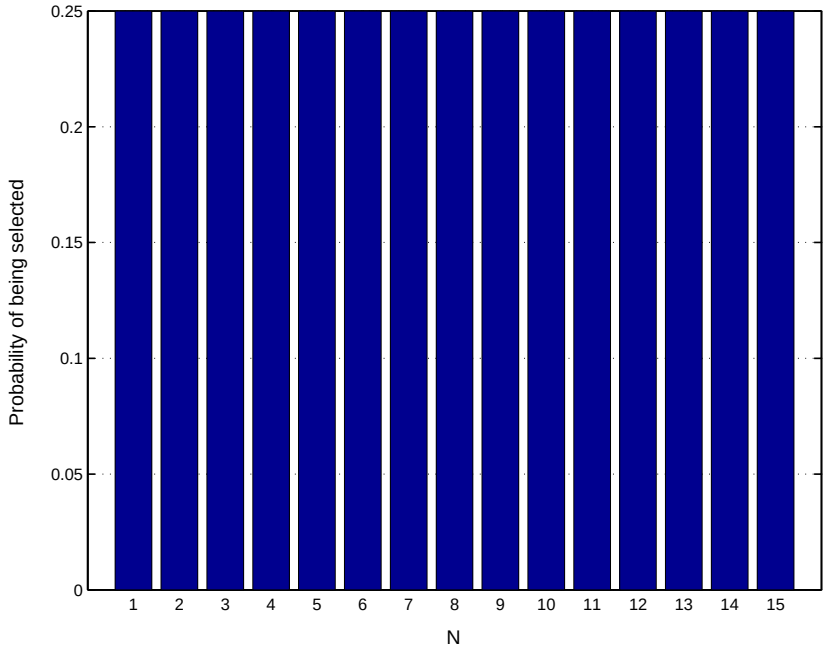


Figure 6.13: Admission Control when there is mobility: Probability of being selected at time t

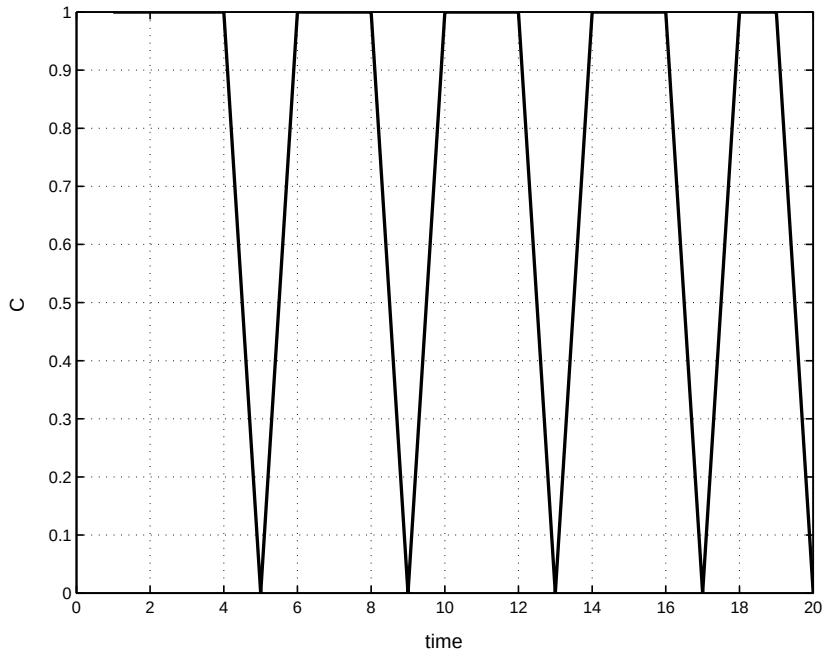


Figure 6.14: Admission control when there is mobility: Fairness constraint

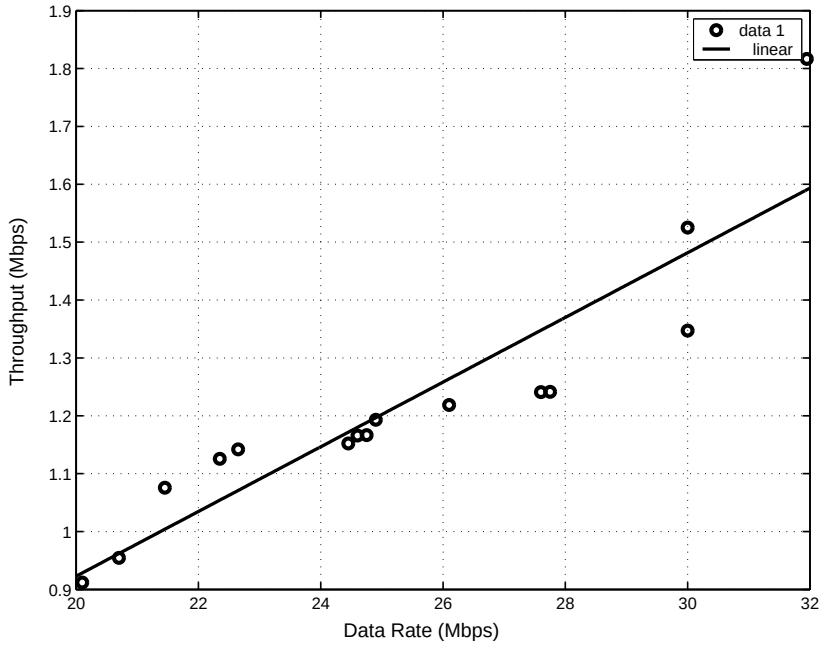


Figure 6.15: Admission control when there is mobility: Throughput vs Data rate

6.3 Performance Results

The performance of the algorithm is evaluated in three scenarios, for basic access and with $\lambda = 0.2$: (1) All stations have the same data rate; (2) Stations have different (but fixed) data rates; and (3) Station data rates change with time, suggesting that they are moving.

6.3.1 Same Data Rate

Figure 6.2 indicates that the algorithm correctly determines the number of active stations that maximizes throughput. As N increases, the number of active stations remains constant. This constant depends on the data rate.

Figure 6.3 shows that the stations are selected to ensure fairness. It plots the stan-

dard deviation of the samples $\{x_1^{Total}, \cdots, x_N^{Total}\}$. Since *Endtime* = 20, the standard deviation could be as large as 10, but the algorithm keeps it well below 1.

6.3.2 Stationary Scenario

When stations have different SNR values, the throughput depends not only on the number of active stations, but on their data rates. From (4.12) we know that if we wish to maximize the system throughput, only stations with large SNR would be selected. Thus in this case, as K increases, maintaining fairness occurs at the cost of reduced system throughput.

Figures 6.4 and 6.5 show the algorithm triples individual throughput and doubles total throughput compared with the situation with no admission control.

As stations have different data rates, the number of selected stations varies significantly as seen in Figure 6.7. The mechanism follows a pattern and movement is first to higher throughput and then to fairness. Figure 6.6 shows that fairness is achieved among stations since they are all selected with equal probability.

In Figure 6.8, the fairness metric C is around 1 meaning that the stations are selected almost equally. Figure 6.9 represents the throughput distribution for different data rates. As can be seen, as the data rate increases, the throughput also increases.

6.3.3 Mobile Scenario

When the stations move, the SNR of each station changes with time. The movement model selects a random data rate set each time. Hence, previous data rates do not impose

any constraint in future selections.

The difference between Figures 6.4 and 6.5 and Figures 6.10 and 6.11 is due to the fact that when there is no mobility the initial SNR values shape the selection and impose strong constraints that can not be disregarded by the AC mechanism. For this reason, when there is no mobility the shape of the curve is almost periodic.

Figure 6.12 and 6.13 show consistent behavior. The average number of stations selected at time t is not smooth as before because it depends on the data rate vector of the time. It is important to note that probability of being selected for each station is almost equal.

Figure 6.14 shows that fairness constraint is around one and individual throughput allocation increases as the increase in the data rate. The points in Figure 6.15 are interpolated by a linear function and one can compare the plot with the no mobility case and infer that stations with higher data rate are favored as the time passes, since the data rate changes all the time in the mobility scenario, throughput allocation is not linear with data rate.

6.4 Implementation Issues

The optimization problem (6.2) requires selection of one of 2^N subsets I, so a large data base is needed to store all the values of $S(I)$. A more scalable method might use clustering or on-demand scheduling.

The proposed AC mechanism is a centralized approach, but DCF is designed for decentralized networks. In infrastructure BSS, the access point or an independent sniffer are

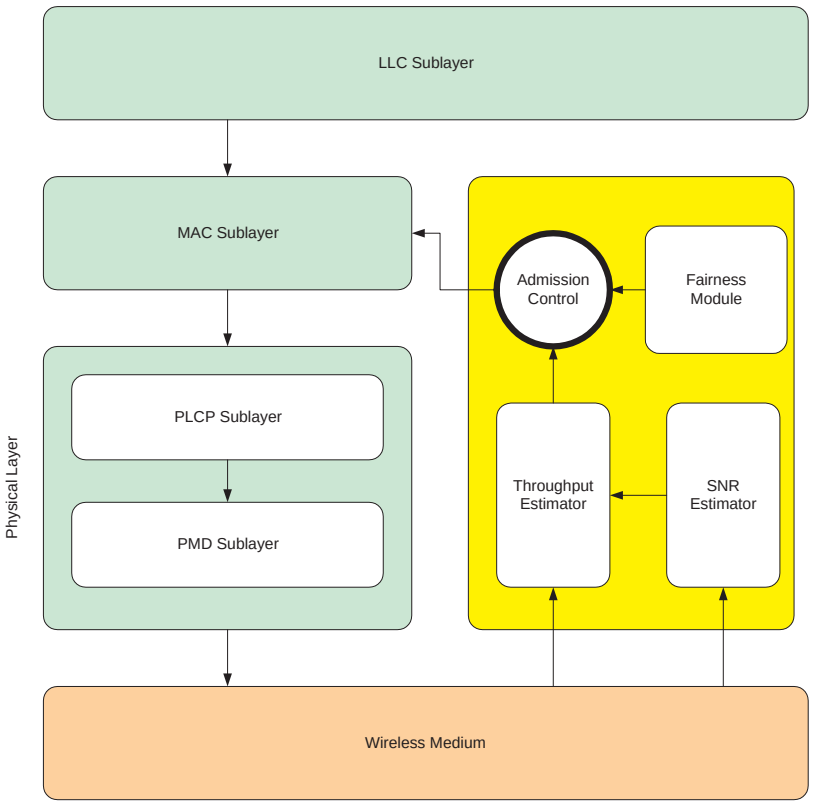


Figure 6.16: System architecture for Admission Control

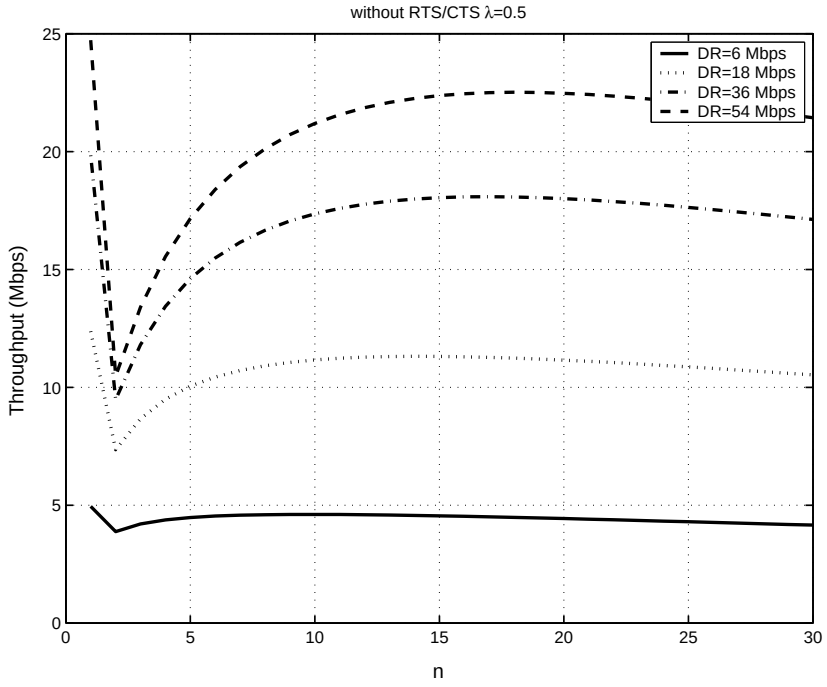


Figure 6.17: Throughput for different data rates (w/o RTS/CTS)

where the AC mechanism can be implemented. The AC mechanism monitors the channel and notifies the stations through the same channel or a dedicated channel. A dedicated channel is being proposed for inter-AP communication.

If the AP does not belong to the same ESS, the communication can be carried in a dedicated channel without any interference. This method could be extended to ad-hoc networks, in which a station is selected to monitor and control the network in terms of only selecting the stations.

The system architecture in Figure 6.16 introduces three modules. The SNR estimator estimates the SNR of each station. The throughput estimator estimates the total throughput and individual throughput. The selection mechanism uses the throughput estimator and the fairness module, which keeps each station's history.

6.5 Discussion

If we examine the Figure 6.17 we can see that if there is a single node in the network it achieves the highest throughput. For this reason AC is configured to select at least two stations ($\sum_i x_i(t) \geq 2$). When we configure AC selection to one station ($\sum_i x_i(t) \geq 1$) the DCF mechanism with AC converges to PCF mechanism. In this way a random access mechanism can be controlled and tuned for quality of service.

Remark

Part of this chapter will appear in ACM-Kluwer MONET Special Issue on WLAN Optimization at the MAC and Network Levels titled as "Throughput Analysis and Admission Control in IEEE 802.11a" and authors are Mustafa Ergen and Pravin Varaiya.

Chapter 7

Application: Network Management

7.1 Introduction

Wireless Local Area Networking has been proliferated after the standardization; with the rapid decrease in the price of the 802.11 radios, 802.11 chips have been integrated to most of the gadgets of modern life. Wireless network performance is becoming more and more important for enterprises deploying business-critical wireless networks. As the networks grow in size and become an integral part of our daily life, throughput intensive applications like voice over WLAN, content delivery are also gaining popularity. In such applications, where network throughput performance is critical, detailed network planning, monitoring and management become essential. Network planning includes detailed site surveys, determining the number of access points, locations, access point configurations. Currently, choosing how many access points to deploy, where to place them are ad-hoc processes and can be very time consuming. In this chapter, we introduce a throughput estimation of a network which estimates the individual throughput of the mobiles and

APs using RF prediction data and analytical Markov model of IEEE 802.11 [46, 5]. The configuration can be used in the detailed planning of a wireless LAN or it can be used as part of a real-time network management system which monitors the network in real-time and takes therapeutic actions depending on the state of the wireless network.

Access points must be placed to proper locations to provide good coverage so that offered services, user needs and hence expected traffic load requirements are satisfactorily met. The estimation technique aims to guarantee service with good quality and offers capacity with a sufficient low congestion. A static network configuration is not adequate for WLANs because it does not take into account how unexpected behavior of the user. Users might cluster in a location for sometime and then move to another location. In this case, access points must be properly configured in real-time to meet traffic demand. If a WLAN is not effectively managed, benefits quickly diminish and it becomes more of a cost burden than savings.

In this chapter, we introduce a real-time network management operation to optimize the throughput by estimating the throughput of an IEEE 802.11 network in advance. Our model predicts the number of stations that fall into coverage of access points, and integrates throughput formulation of a DCF network with an RF propagation model to determine the coverage and interference of access points.

Coverage, which consists of small islands, is determined by the wireless channel characteristics. Stations attached to an access point are determined by the number of stations that fall into the island of the access point. Data rates are also determined by the relative SNR to access point and the coverage is often tailored according to the users own need and can also be temporary.

7.2 Realtime Network Management

Network management constantly monitors every access point in a WLAN, giving instant feedback so a decision maker can instantly reconfigure the wireless network for optimum performance. Network monitoring reports data rate and access point for every station. Management system uses this information to compare the performance of the existing system with the possible reconfiguration. The management module uses an analytical throughput model together with an indoor propagation model to estimate the performance of a network and then fine-tunes the AP parameters such as channel, Tx power level and load balancing between APs as seen in Figure 7.1.

7.2.1 Indoor RF Propagation

Efficient propagation models are key to the successful deployment of wireless LAN systems. For propagation inside buildings, it is impossible to account for every interaction as a radio signal propagates through the buildings or to model the signal variations on a wavelength. Therefore, we focus on a semidefinite single ray based method which is not as computationally expensive as 3-D ray tracing methods. This method produces accurate and deterministic results for common indoor environments. In this model we will only focus on the attenuation from the walls and floors.

Due to the site specific characteristics of indoor environments radio planning tools have to take into account the location, orientation, electromagnetic properties of individual walls and objects; radiation patterns and locations of the antennas. Phenomena related to RF propagation, like multi-path propagation, reflection, diffraction and shadowing have

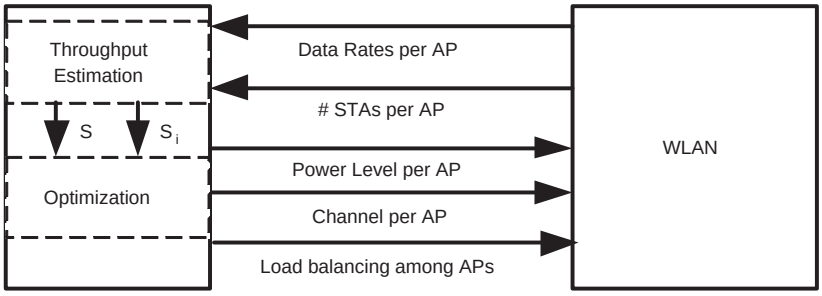


Figure 7.1: Network management module

a significant influence on the received power. So, the propagation models should consider these phenomena to obtain accurate results [6].

The prediction model used in the simulator is a variant of the COST 231 Multi Wall Model with some additions for optimizing the ray-tracing implementation. This model calculates the path loss according to both the distance between the transmitter and receiver and the penetration losses through walls and ceilings. The equation is as follows:

$$L = L_{FS} + L_e + \sum_{i=1}^{I} k_{wi} L_{wi} + k_f^{\left[\frac{k_f+2}{k_f+1} - b\right]} L_c \tag{7.1}$$

where L_{FS} is free space loss of the path in dB, L_e is constant excess attenuation of the path in dB, L_{wi} is loss of wall type i in dB, L_c is loss between adjacent floors in dB, k_{wi} is number of walls of wall type i penetrated by the path, k_f is the number of ceilings penetrated by the path, I is number of wall types, b is multi-floor parameter.

In addition to advanced modeling techniques it is essential that the performance of the propagation models is verified by comparison with measurements. Several extensive measurement campaigns have therefore been carried out using the Site Survey tool which was developed to validate the accuracy of the RF propagation model used in the simulator

and to calibrate the model if needed.

7.2.2 Channel Selection

Adjusting the non-overlapping channels in access points is a coordinated effort among access points since any overlapping channel assignment to adjacent access points causes interference and degrades the performance. If there are more access points than the number of non-overlapping channels then any adjustment takes into consideration the assignment of non overlapping channels to the congested area.

7.2.3 Power Control

The WLAN standard limits the maximum transmission power to be 100 mW and minimum power shall to be no less than 1 mW [47]. The transmission power can be adjusted by the access point to extend or shrink the coverage. Of course, a station's data rate also changes according to the signal power of corresponding access point. The network management system employs this tuning parameter to adjust the coverage according to user's capacity requirement. The user's capacity depends on the application preferred; for instance, for a wireless voice over IP, the minimum adequate throughput should be satisfied among all stations. This may require a rearrangement of power levels and estimation the throughput of a station in advance.

7.2.4 Load Balancing

APs balance their load when there is nonuniform distribution among APs. Each AP monitors its own traffic load in real-time; when the overall traffic load is over a certain threshold, a negotiation procedure will be triggered. The overloaded AP creates some procedures to reduce its own traffic load: it could adjust the power level or refuse some stations. At the same time an adjacent AP increases the power level to take over the stations refused by the AP. This causes some stations to handover traffic to lighter loaded APs.

7.3 Throughput Analysis

We introduced the formula to find individual throughput if the data rate is different for each user in Chapter 4. The 802.11 standard gives the same amount of chance to each user even if the data rate is different. The data rate only changes the duration values which changes average duration values. Successful duration is the average of the duration of each user and collision duration is determined by the longest duration of the station involved in the collision. If T_s^i and T_c^i are duration values for data rate i and if n^i is the number of nodes with data rate i, the expected duration values are calculated in equation 7.2; where D is the total number of data rates supported and $n = \sum_{i=1}^{D} n^i$.

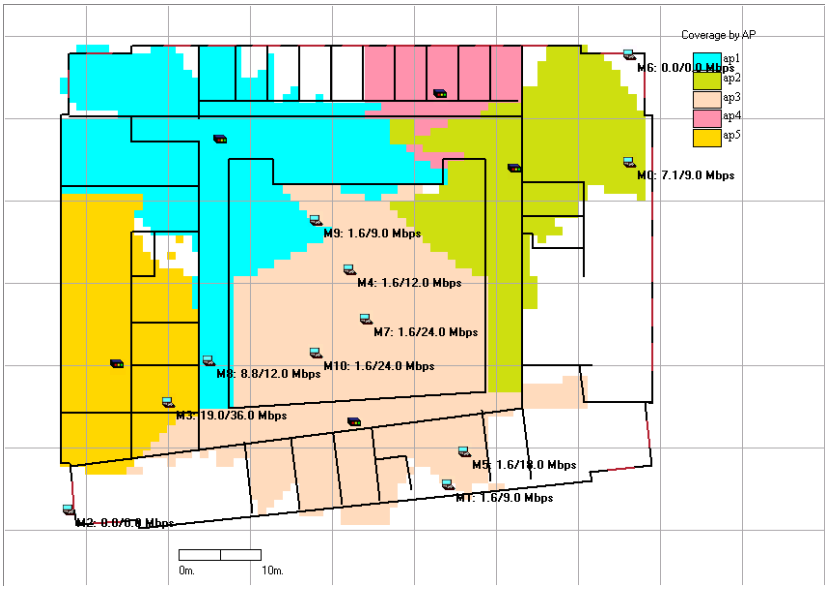


Figure 7.2: Access Point coverage map (Individual throughput/ Data rate)

Throughput of STA S_i is as follows:

$$\bar{T}_s = \frac{P_s}{n} \sum_{i=1}^{D} n^i T_s^i \tag{7.2}$$

$$\bar{T}_c = \sum_{i=1}^{n-1} \sum_{j=1}^{D} \sum_{k=1}^{n^j} \begin{pmatrix} n-k-\sum_{l=1}^{j-1} n^l \\ i \end{pmatrix} \times T_c^j \tau^{i+1} (1-\tau)^{n-1-i} \tag{7.3}$$

$$S_i = \frac{1}{n} \frac{P_s E[P]}{(1-P_{tr})\sigma + \bar{T}_s + \bar{T}_c} \tag{7.4}$$

where each station have the same packet size $E[P] = P$.

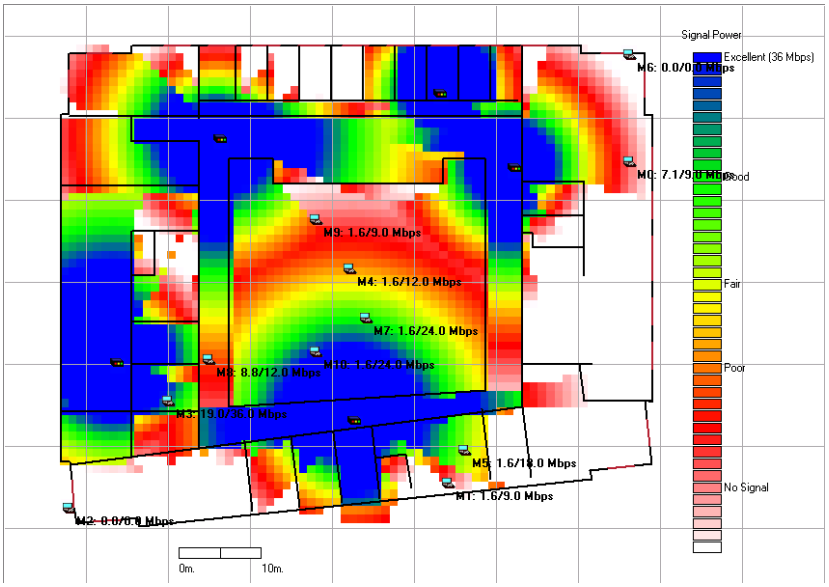


Figure 7.3: Signal power map (Individual throughput/ Data rate)

7.4 Performance Analysis

We used a fixed floor plan to determine the throughput. Access point power levels are pre-selected between 20mW and 100mw. E[P] is constant and 1024 bytes. Stations are distributed randomly and stations that fall out of any AP coverage are considered *inactive*. Figure 7.2 shows the access point coverage and Figure 7.3 shows the Signal Power Map which is used to determine the data rate. The numbers next to the station stands for throughput and data rate of that particular station. All access points are set to the same channel to see the negative effect of interference. Variables are number of access points, number of stations and the traffic intensity.

Figure 7.4 depicts the system when there are 5 fixed APs. The number of stations scattered are varied and analysis is repeated for two different traffic intensity which are 0.5 and 1. The figure depicts that the total network throughput is also almost concave as

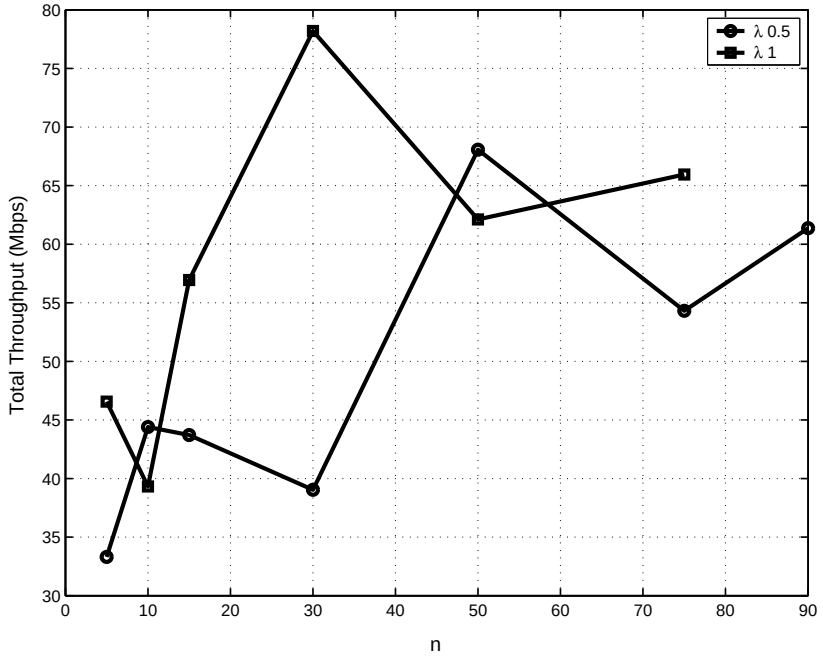


Figure 7.4: Total throughput for the whole network

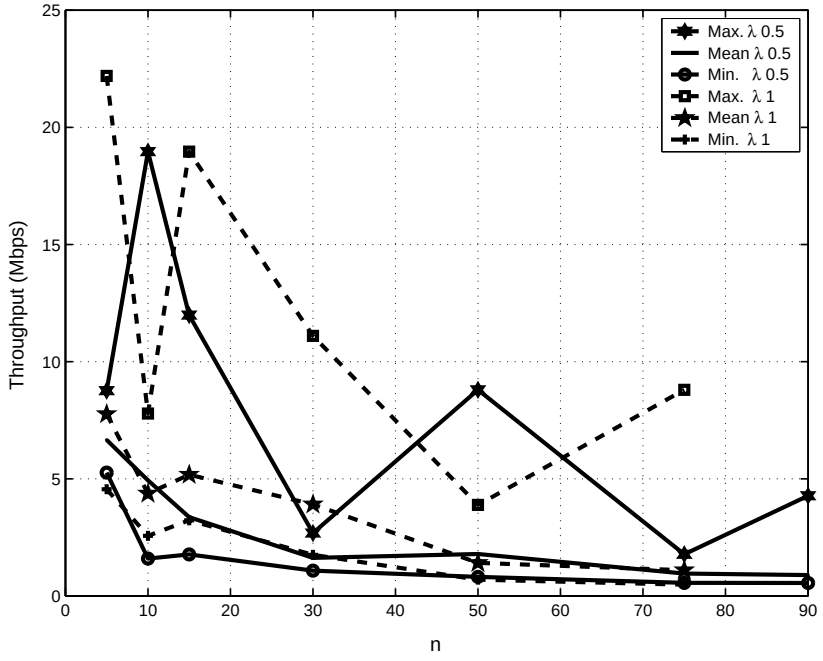


Figure 7.5: Individual throughput for the whole network

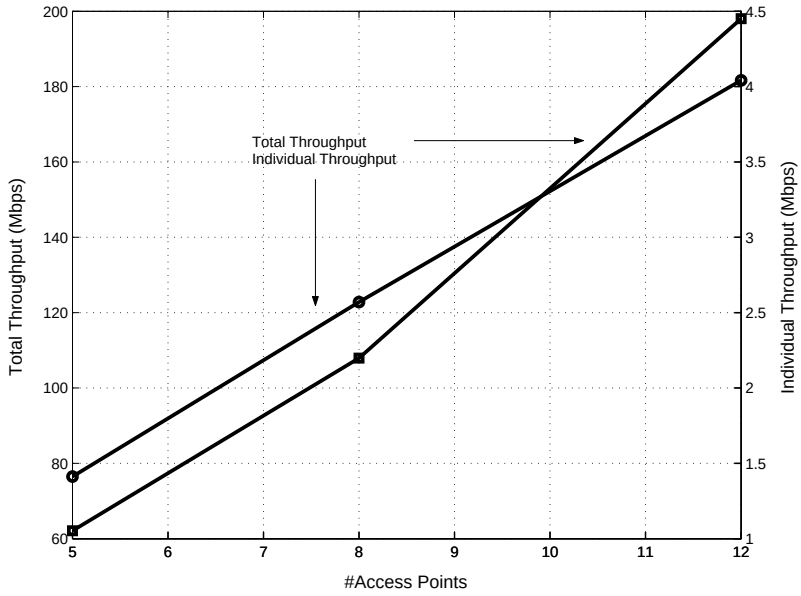


Figure 7.6: Total throughput when there are 50 STAs

in Figure 3.24 where a single access point performance is illustrated.

Figure 7.5 illustrates the individual throughput of the stations when the *inactive* stations are deducted. From this figure we can observe the minimum individual throughput of the network.

Figure 7.6 shows the throughput performance of the system when the number of access points are 5, 8 and 12. As the access points increase in number, the total throughput of the system and individual throughput also increase. This is because of the increase in the signal strength in each point and the decrease in the number of stations associated with an access points. Of course we assumed the condition that the data rate is determined by the signal power map, but in the throughput analysis, stations that fall into coverage of an access point experience the same interference. This means that stations at the border are only affected by the stations with the same access point giving the same probability

of collision to each station.

The network management function takes into account the number of stations attached per access point and performs power level or channel adjustment. It is necessary to constantly monitor access point settings to ensure that they remain in compliance with minimum capacity requirements. When there is a need for configuration, the algorithm automatically implements the changes over large groups of access points.

We choose an extreme scenario on mobility where stations move together move from one quadrant to another in a fixed floor plan. Access point power levels are selected on runtime between 1mW and 100mW. Traffic intensity is 0.5 for all stations. The following algorithm is applied;

STEP 1: Begin with getting the current situation.

STEP 2: Estimate throughput $S_{current}$ with current situation.

STEP 3: Find optimized configuration.

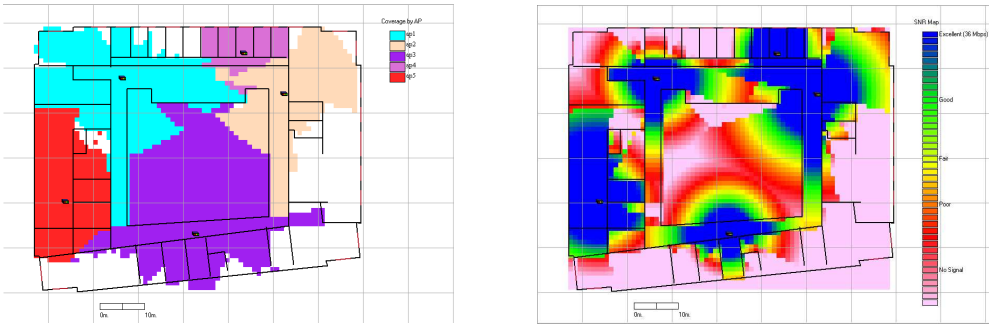
STEP 4: Estimate throughput $S_{optimized}$ with optimized situation.

STEP 5: If $S_{optimized} < S_{current}$ then GOTO STEP 1.

STEP 6: Do one step adjustment.

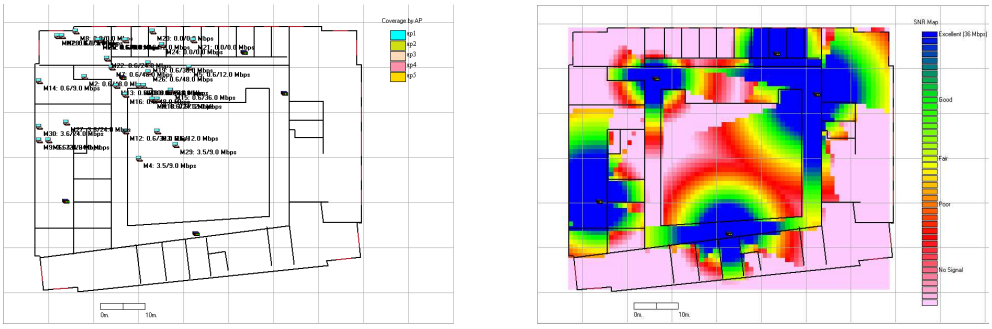
STEP 7: GOTO STEP 1.

Figure 7.7 depicts the system when there are 5 fixed APs. The APs' coverage is illustrated in Figure 7.7(a) and RSSI map for that setting is illustrated in Figure 7.7(b). As illustrated in Figures 7.8(a), 7.9(a), 7.10(a) the stations move from first to second



(a) Access points coverage (b) RSSI map

Figure 7.7: Coverage before optimization



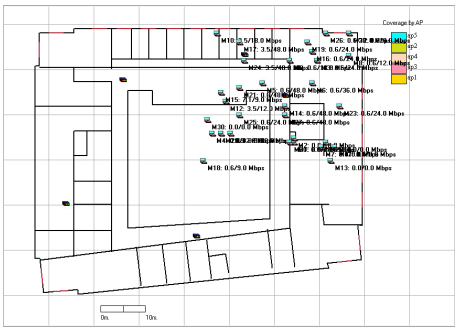
(a) Station distribution (b) Optimized RSSI map

Figure 7.8: Network management for first quadrant

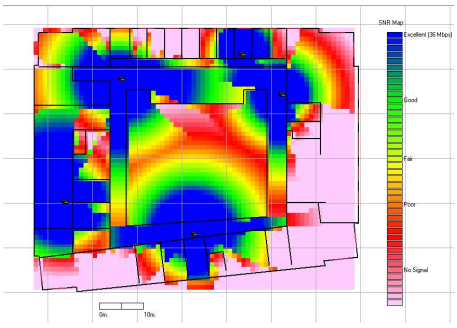
and second to third quadrants. In all cases the configuration adjusts itself to distribute the stations evenly among access points. The changes in RSSI map is shown in Figures 7.8(b), 7.9(b), and 7.10(b).

The total throughput is around { 33.37, 29.18, 37.07 } Mbps, respectively, for first, second, and third movement for 30 nodes. The network management protocol introduces around 30% increase in throughput.

This system can be implemented with or without a localization tool. If there is a position estimation of the stations, the possible access points of stations can be estimated after the power adjustment; and configuration of the access point can be found without iterations.

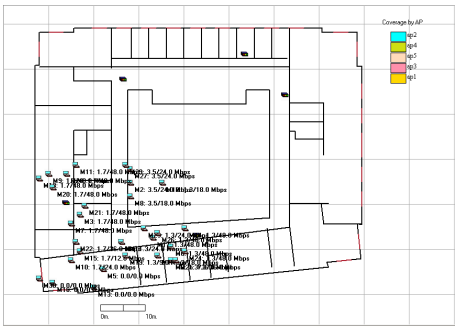


(a) Station distribution

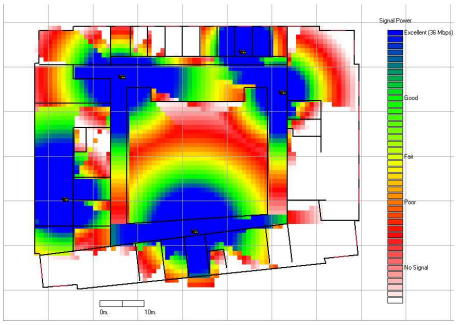


(b) Optimized RSSI map

Figure 7.9: Network management for second quadrant



(a) Station distribution



(b) Optimized RSSI map

Figure 7.10: Network management for third quadrant

Remark

Part of this chapter is published in the Proceedings of IEEE GLOBECOM 2004 Conference titled “Throughput Analysis of an Extended Service Set in 802.11” and authors are Mustafa Ergen, Baris Dundar, and Pravin Varaiya.

Chapter 8

Application: Optimization for Mixed Data Rate

8.1 Introduction

Wireless clients connect at different rates due to the mix of 802.11b and 802.11g WLAN standards. In addition, the data rate is determined by the signal-to-noise ratio. Low rate communication channels reserve relatively more time to transmit which tends to dominate the channel. We analyzed this behavior in Chapter 4. As it can be inferred, a single low data rate station brings down the individual throughput to be lower than its data rate which is very unfair for the high data rate stations. This motivates the development of an algorithm that utilizes the channel fairly.

8.2 Algorithm

Let us revisit the individual station formula represented in 4.10. Optimal packet size for a given station can be found by taking the derivative of S_i with respect to $P(i)$ where duration values are $T_s(R(i), P(i))$, and $T_c(R(i), P(i))$ and $E[P]$ is $P(i)$ for station i. We find that highest throughput is achieved when the stations with lower data rate are turned off. Although it achieves the maximum throughput, it is not fair since fair allocation is defined as giving equal amount of usage of the channel. This could be achieved by reducing varying packet size with respect to data rate. As the data rate goes low, the stations can send lower size packets prevent suffering of high data rate stations.

We round down the duration values to the duration value of highest data rate because the throughput increases monotonically with the increase in packet size (See Figure 8.1). Intuitively, the optimal packet size for the lower data rate station is the packet size that gives the same duration values of the highest data rate station. If highest data rate is R^h with packet size P_{R^h} then packet size of station with R^i data rate is given as follows for IEEE 802.11b:

$$P_{R^i} = floor(\frac{R^i * P_{R^h}}{R^h} - 30 * \frac{(R^i - R^h)}{R^h}) \tag{8.1}$$

As a result the data rate and packet size relation is as follows for $P_{R^h} = 1000bytes$:

$R(i)$ (Mbps)	$P(i)$ (bytes)
11	1000
5.5	515
2	206
1	118

This channel access mechanism not only distributes the number of chances to access the channel, it also gives equal amount of channel usage. Stations can hear each other and if they hear a station with higher data rate then stations decrease their packet size to equalize duration values with high data rate stations. As explained earlier, increasing the data rate is not a solution since it is determined by the wireless channel and confinement of the physical layer standard. If P_{R^h} and $T_s(R^h, P_{R^h})$, $T_c(R^h, P_{R^h})$ represent the packet size and duration values of the station with the highest data rate then;

$$\begin{aligned} T_s(R(i), P(i)) &= T_s(R^h, P_{R^h}) \; for \; \forall i \\ T_c(R(i), P(i)) &= T_c(R^h, P_{R^h}) \; for \; \forall i \end{aligned} \tag{8.2}$$

As a result the individual throughput formula represented in equation 4.10 is modified as follows:

$$\bar{T}_s = P_s T_s(R^h, P_{R^h}) \tag{8.3}$$

$$\bar{T}_c = (P_{tr} - P_s) T_c(R^h, P_{R^h}) \tag{8.4}$$

$$S_i = \frac{1}{n} \frac{P_s P(i)}{(1 - P_{tr})\sigma + \bar{T}_s + \bar{T}_c} \tag{8.5}$$

Figure 8.2 shows the improvement in throughput. Stations with 1 Mbps data rate decrease their packet size from 1000 bytes to 118 bytes if there is a 11 Mbps station.

The advantage is in total throughput. On the other hand, in individual throughput, high data rate stations are better off but the low data rate stations are worse off as seen in Figure 8.3

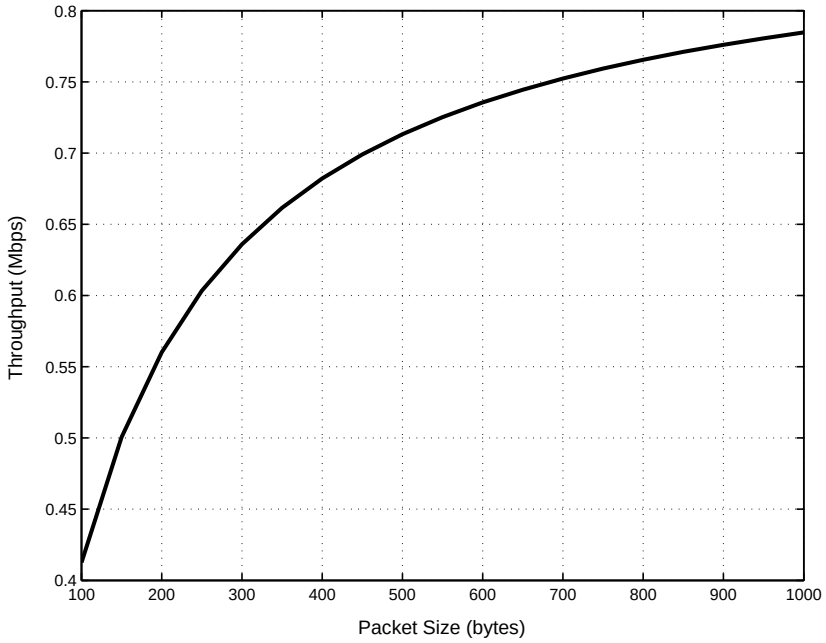


Figure 8.1: Throughput with packet size

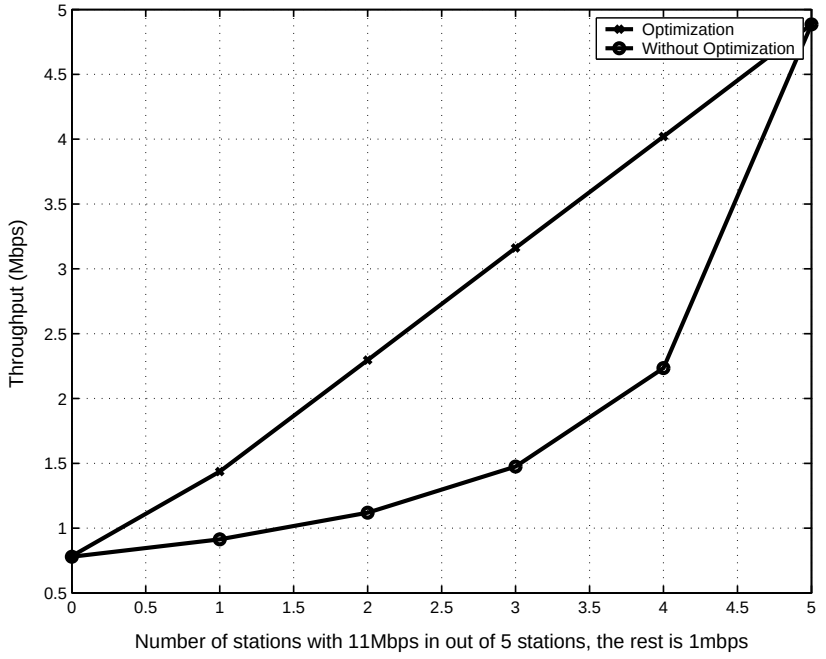


Figure 8.2: Throughput after optimization (w/o RTS/CTS)

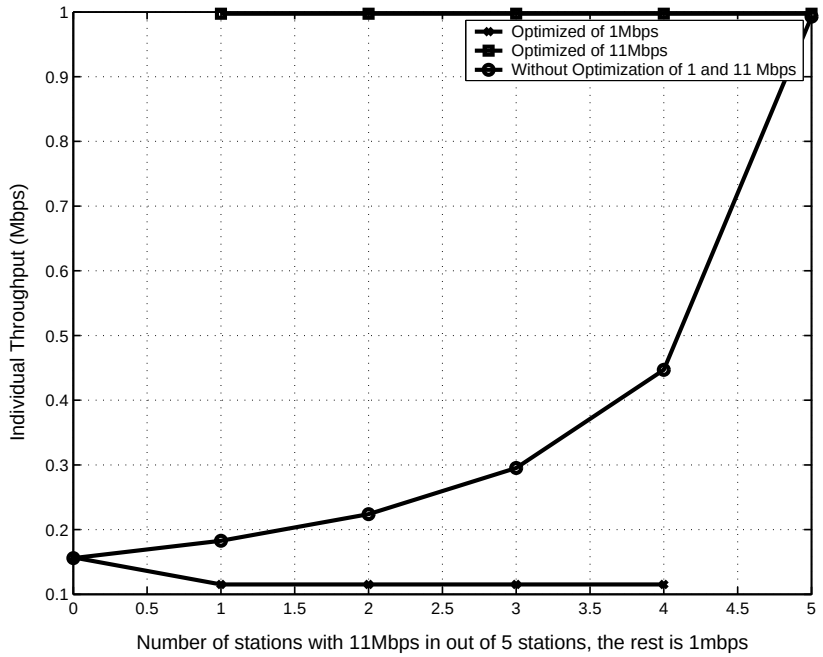


Figure 8.3: Individual throughput after optimization (w/o RTS/CTS)

8.3 Performance Analysis

Here, we perform a simulation with settings as in Chapter 4. There are five stations and all operate with 11 Mbps data rate. Only one station changes its data rate throughout the simulation and we apply the optimization algorithm. Figure 8.4 shows the transmitted data. The upper figure shows the transmitted data for fast stations and the slow station; and the lower figure shows the channel activity of the slow station. One can see that transmitted data for the slow station is the envelope of the lower figure.

Figure 8.5 shows the total throughput compared with the situation before optimization. Performance improvement decreases with the increase in the data rate of the slow station.

Note that the station's throughput in OPNET is the number of packets it received; let's call this "received throughput" (S^r). According to this definition and observing Fig-

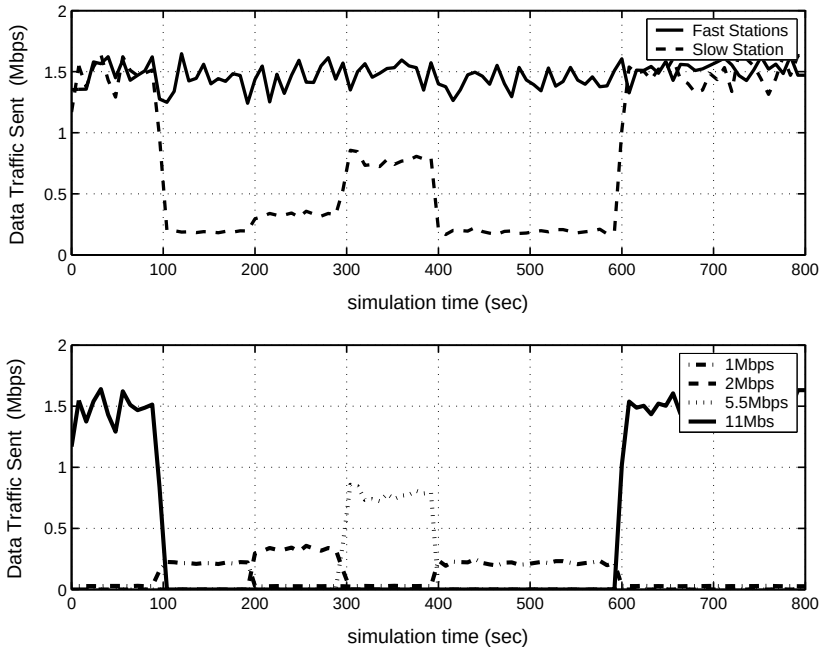


Figure 8.4: Data traffic sent

ure 8.6, we see that the slow station receives more packets than the fast stations. This is obvious because the optimization algorithm suppresses the transmission of the slow station and fast stations receive less number of packets than the slow station. This is because the slow station is a sender of packets to the fast stations. However, the throughput is still greater than prior to optimization; with optimization fast stations use the channel more than they do when there is no optimization.

If we consider throughput as the number of successfully transmitted packets and call it "transmitted throughput" (S^t) as in our Markov model analysis, we can extract this information from the received throughput. We have 4 fast stations and 1 slow station. If we look at the slow station's received throughput we know that it comes from 4 fast stations, each contributing one fourth of traffic. If we look at the fast stations throughput, it is con-

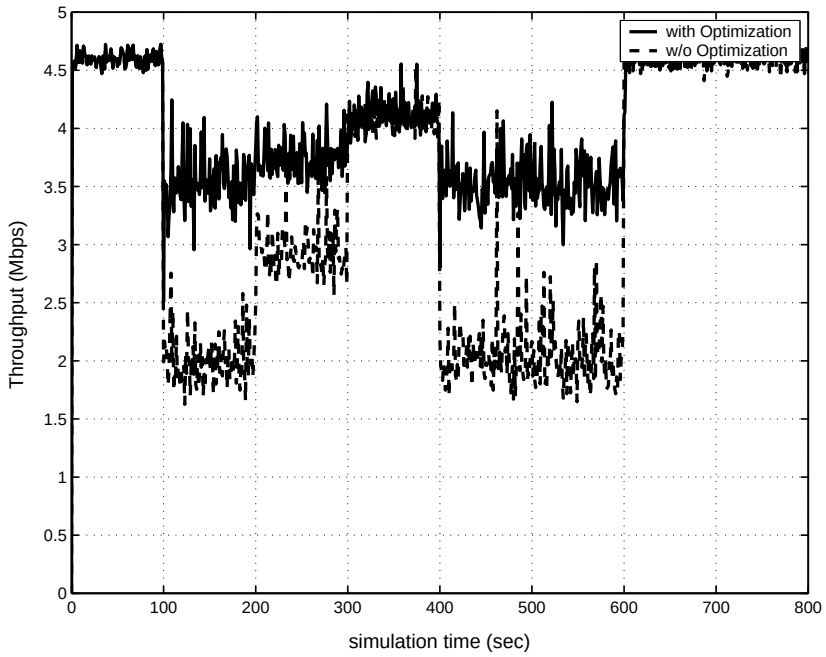


Figure 8.5: Total throughput

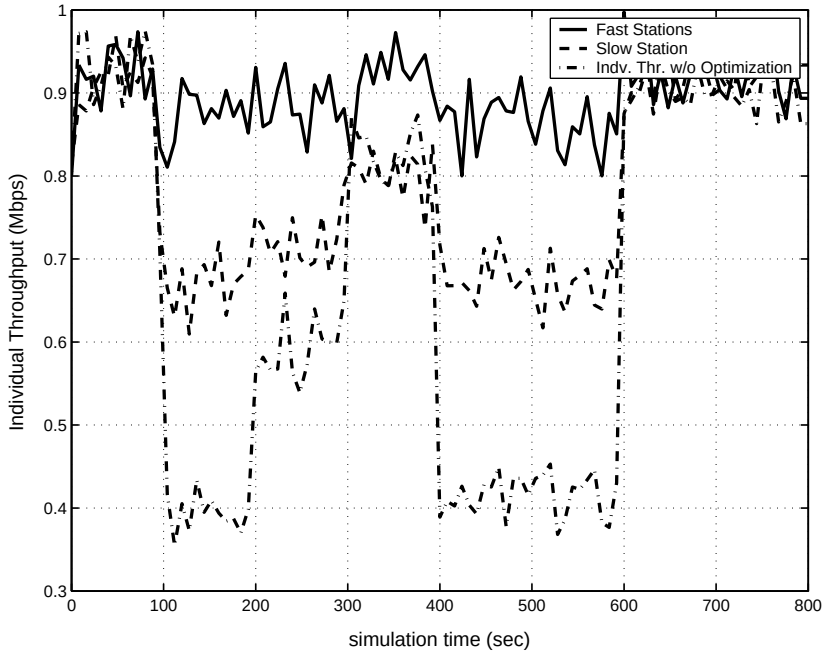


Figure 8.6: Individual throughput (Received throughput (S_i^r))

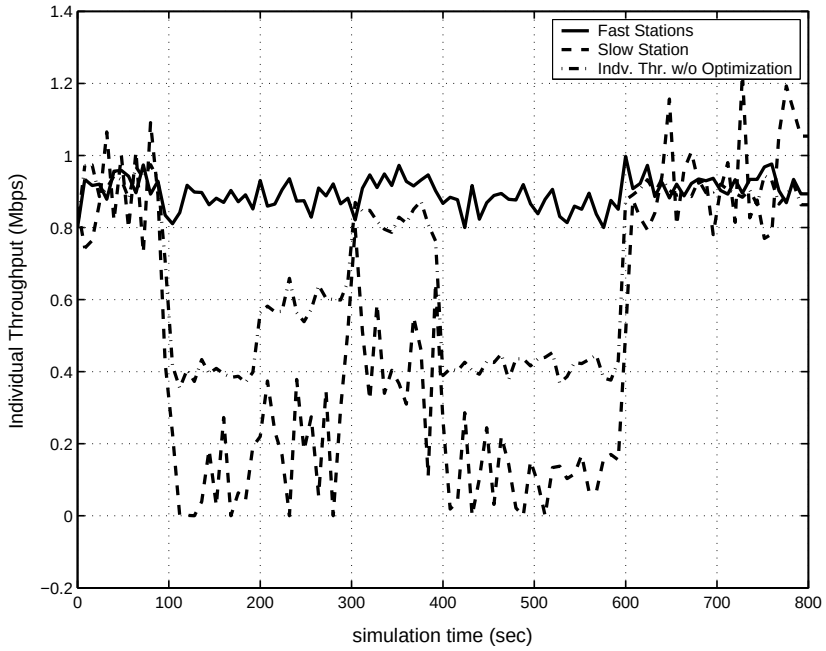


Figure 8.7: Individual throughput (Transmitted throughput (S_i^t))

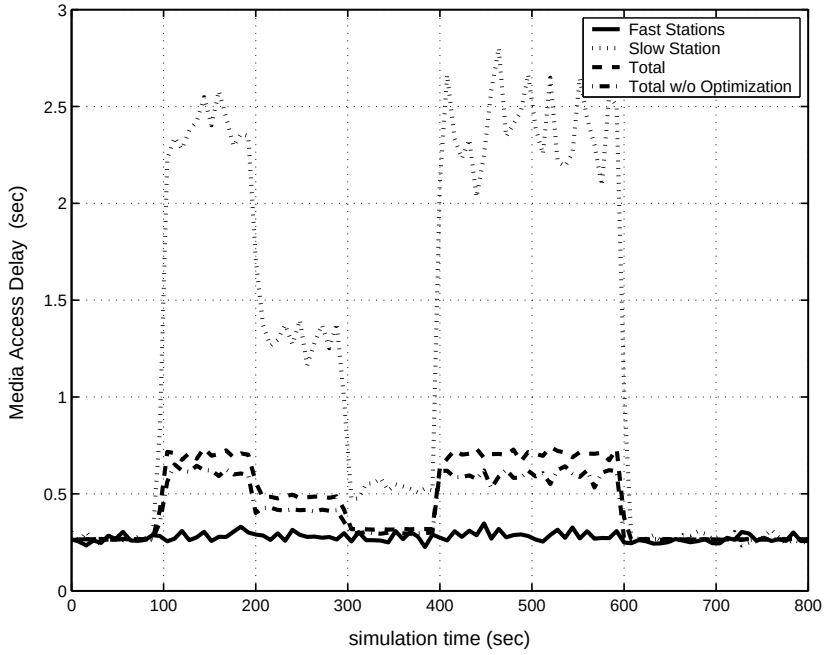


Figure 8.8: Media Access Delay

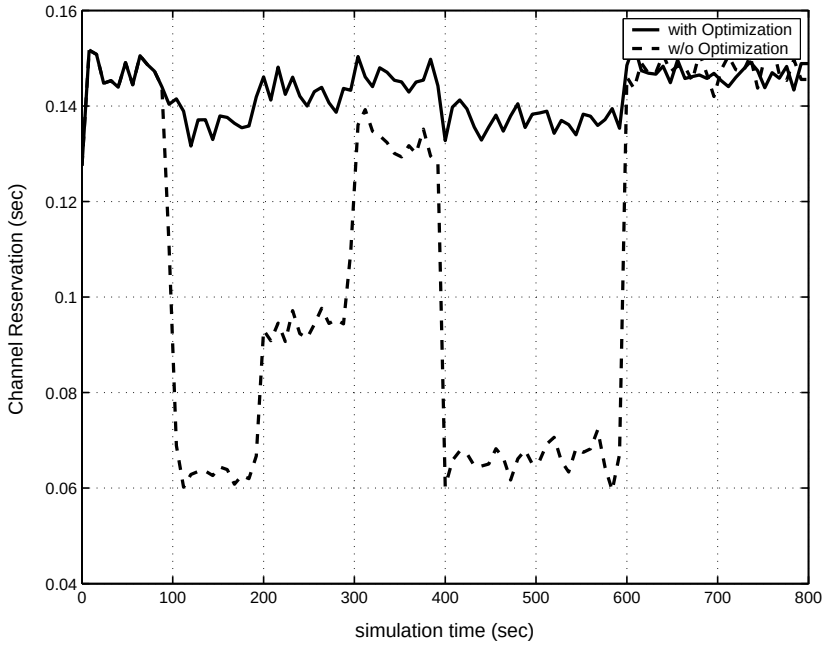


Figure 8.9: Channel reservation

tributed by 3 fast stations and 1 slow station. It is easy to conclude that transmitted and received throughput is the same for fast stations and for the slow station. The following formula gives the transmitted throughput S^t:

$$S^t_{Fast} = S^r_{Fast}$$

$$S^t_{Slow} = |4S^r_{Slow} - 3S^r_{Fast}|$$

and the result is depicted in Figure 8.7.

If we consider throughput as the number of packets that has been sent correctly we infer that slow station should have the lower throughput as compared to case prior to optimization. This is because the slow station sends less bits which is shown in Figure 8.3. We make the same conclusion by looking at the media access delay in the OPNET sim-

ulation. Figure 8.8 shows the media access delay for the stations. As expected, the slow station has the highest media access delay and fast stations have the lowest media access delay. Figure 8.9 shows the channel reservation. As can be seen, channel reservation has been increased considerably with optimization.

Chapter 9

Application: Wireless Voice over IP with Frame Aggregation

9.1 Introduction

In recent years, there has been a lot of interest in using WLAN for voice communications with the expanded coverage of hot spots. The aim of the protocol described in this chapter is to improve the performance of voice over IP (VoIP) operating on IEEE 802.11.

The goal of VoIP is to transport speech signals over an IP network. What makes VoIP challenging is that the IP network is not designed for real-time applications. Real-time aspects of conversation must be respected; the overall delay between both ends of the conversation should be low to avoid irritably long gaps of silence. This must be addressed differently from broadcast programs which are not interactive and have less stringent requirements [48, 49, 50, 51, 52, 53].

VoIP is an alternative to the telephony network and could be seen as a replacement

technology. A computer network may replace the telephony network by installing a VoIP system and the only connection to the telephony network might be a gateway that does address translation from IP to phone number or vice versa. The real benefit arises in the capacity. The capacity of a computer network could be utilized better and bandwidth is cheaper as compared to the telephony network. Long distance calls could be possible along with certain applications like whiteboarding, application sharing, file transfer, and video imaging [48].

Conversation within the LAN is possible without effort. Unfortunately delay issues arise when going across the LAN and WAN is involved. End to end delay is important and when it gets too large, the conversation experiences distortion. This could happen due to the heavily loaded roads and packets maybe lost during routing inside the WAN. Also, a high percentage of packet loss will decrease the quality of the conservation. Moreover, there is a difference in the delay of packet arrivals to receivers due to variations in the congestion level of the network over time.

We now describe problems arising from the current standards in terms of supporting the requirements of VoIP applications. Then we propose and analyze a frame aggregation algorithm that tries to increase performance for downlink communication.

9.2 Delay Components

The total delay falls into one of two categories: fixed delay or variable delay. Fixed delay is the total delay introduced by sampling, compression/decompression, transmission and jitter buffering. The variable is caused by packet queueing at the routers.

Compression delay can be divided into two categories. The first part is introduced by the calculations. This amount of delay depends largely much on the capabilities of the system performing the compression. The second part is the delay resulting from waiting for the entire information to arrive which will be included in the speech data. For instance, if the vocoder operates on a 20ms segment of speech, additional 20 ms is included. Then, 160 bits (20bytes) is required as a payload, and additionally, RTP header (16 bytes), UDP header (8 bytes), IP header (20 bytes) are needed. The total number of bytes sent to MAC layer as MSDU is then 64 [48].

Network delay is the delay to transport this data over the network. It is not possible to make a general claim about that delay in IP networks although one-way transmission delays rarely tend to exceed 100ms. However, it is possible that this delay exceeds 200ms, which is the minimum tolerable value [22]. Jitter buffering is typically set to 10-20ms.

9.3 Voice Communication

The components of VoIP include grabbing/reconstruction, compression/decompression, transmission/ reception over IP in sender/receiver parts, respectively. The analog speech signal is first encoded into a digital representation to be able to be transferred over IP network. At regular small intervals, blocks of digitized speech information is sent over the network from the transmitter to the receiver. On the receiver side, this digitized block is transformed back to an audio signal, which is then output to speakers.

Digitization of voice data includes sampling and quantization. The sampling rate and the number of bits used in quantization determines the rate of data transmission be-

fore data compression. The speech signals of humans can contain frequencies of beyond 12kHz. However, high quality communication is attained in the telephone system by transmitting only frequencies below 4kHz. Nyquist theorem then suggests that a sampling rate of 8kHz is enough for the digitization of speech.

The range of amplitudes of the voice signal is covered by at least 12 bits when a uniform quantization is used. A uniform quantization with 8-bit quantization also gives telephone quality. The required bandwidth for telephone quality conversation is 64kbps, which is 8kHz times 8 bits per sample.

To avoid the delay jitter, which is the difference in packet arrivals, a buffer is used at the receiver. Instead of playing the voice data immediately after the reception of the packet, the packets are buffered. Although the buffer slightly increases the delay, it increases the probability of playing the packets consecutively without interruption.

The number of bits that each packet contains is very important in terms of delay and packet loss. To reduce the amount of lost information, a packet should contain only a small amount of the voice signal. If a packet is lost, this will only be a small fraction of the conversation. However, as the length of data in each packet decreases, the overhead in the network increases due to the low ratio of the data length to the length of the header fields.

9.3.1 Compression Techniques

The bandwidth used by the digitized information can be reduced by compression schemes. Compression fall into one of two categories: general compression techniques and com-

pression techniques that exploit the fact that we are dealing with voice information. The first is known as wave form coding and the second is known as vocoding. Waveform coding encodes the waveform itself whereas vocoding encodes the information about how the speech signal is produced by the human vocal system.

The simplest form of waveform coding is PCM. Differential PCM (DPCM) tries to exploit the fact that the value of the samples of the audio signal can be predicted from previous values. DPCM calculates the prediction of the sampled value and uses a fixed number of bits to store only the difference between the predicted and actual values of a PCM signal. Adaptive DPCM (ADPCM) uses some of the bits to store the difference to adjust the resolution of the difference.

Vocoding is a combination of “voice” and “coding”. Instead of trying to encode the waveform itself, vocoding techniques try to determine the parameters about how the speech was created and use these parameters to encode the signal. To reconstruct the signal, these parameters are fed into a model of the vocal system that outputs a speech signal. Since the vocal tract and excitation signal change relatively slowly, the signal that has to be analyzed is split into several short pieces. A piece of the signal is then examined. If the signal is voiced, the pitch period is determined and the excitation signal is modeled accordingly as a series of periodic pulses. If the signal is unvoiced, the excitation is modeled as noise. The effect of vocal tract is recreated through the use of a linear filter. This filter contains certain parameters that have to be determined by the vocoder. Several types of vocoders exist. The main difference between these methods is the vocal tract model used.

Linear predictive coding (LPC) uses a vocal tract model as an approximation of a se-

ries of concatenated acoustic tubes. It examines the input and estimates the parameters to use in the vocal tract filter. It then applies the inverse of this filter to the signal. The result of this is called the residue signal and it basically describes which excitation signal should be used to model the speech signal as closely as possible. The parameters of the filter is found by a difference equation which describes each sample as a linear combination of the previous ones. Such an equation is called a linear predictor.

Waveform coders do not perform well at data rates below 16kbps. Vocoders produce intelligible communication at very low data rates, usually below 4.8kbps. However, the reproducted speech often sounds quite synthetic and the speaker is often unrecognizable. Hybrid decoders try to exploit the advantages of both techniques. Residual Excited Linear Prediction (RELP) uses residual signal directly as the excitation for speech synthesis instead of checking whether the signal is voiced or unvoiced and tries to model the excitation signal. Codebook Excited Linear Prediction (CELP) allows a wide range of excitation signals, which are all captured in the CELP codebook. To determine which excitation signal to use, the coder performs an exhaustive search. The excitation signal is encoded by the index of the corresponding entry [48].

The standards for voice communication have been established to make inter-operability between the applications possible. The most widely known standards in the VoIP domain are the G. standards of the ITU-T. Table 9.1 gives the list of the standards [51].

Codec	GSM 6.10	G.711	G.723.1	G.726-32	G.729
Bit rate (Kbps)	13.2	64	5.3/6.3	32	8
Framing Interval (ms)	20	20	30	20	10
Payload (Bytes)	33	160	20/24	80	10
Packets/sec	50	50	33	50	50

Table 9.1: Codec specifications for the standards

9.3.2 Transmission Protocols

Since the IP network only offers best effort service, no guarantee on the delay can be provided. Similarly, the amount of lost packets can be very high during congestion. Somehow, the sender should know whether the receiver can handle the incoming stream or not.

Over an IP network, UDP together with Real-Time Transport Protocol(RTP) and RTP Control Protocol (RTCP) are used. TCP cannot be used since waiting for retransmitted packets adds extra delay to the communication. Also, TCP has no support for multicasting, which can be used to distribute speech data to several users at the same time as explained in the following sections.

UDP is not enough by itself to support the transmission of real-time data since it provides no mechanism for synchronization and there is no support for flow or congestion control. RTP and RTCP add extra information to the speech data and use UDP to distribute this control and speech information. RTP includes field **sequence number** to deliver received packets to the application in the correct order and **timestamp** to include the synchronization information for a stream of packets. RTCP on the other hand periodically

send RTCP packets to observe the number of participants to each session and to provide feedback on the quality of data distribution.

These protocols, however, by themselves do not provide any mechanism to ensure timely delivery that can give QoS guarantees. Both IPV4 and IPV6 have a way to specify the priority of a datagram. In the IPV4 header, some level of QoS can be specified in the Type of Service (TOS) field whereas the IPV6 header uses traffic class field for the same purpose. If all routers take such priorities into account, this could help real-time data to be delivered with low delay. The only thing that needs to be done is to adjust the routers so that they take the priorities of packets into account. These mechanisms however can only help to give a better service, they cannot give any guarantees. If all the packets in the network are of high priority, the quality will still be poor.

The protocols such as Resource Reservation Protocol (RSVP) can be used to reserve resources in the network so as to give some guarantees on the delay. If a host is going to transmit data which should arrive with a certain QoS, it periodically sends a path message to the destination of the data. The path message contains information about the characteristics of the traffic generated by the sender. Then the exact reverse path is followed in response to the path message to make the reservations.

9.3.3 VoIP Problems in IEEE 802.11

The main problems of VoIP over WLAN are low VoIP capacity and increasing delay over WLAN. The number of VoIP sessions that can be supported in WLAN is much lower as compared to when the protocol overhead is excluded. We have seen in Section 9.3.1

that each VoIP typically requires a bandwidth of around 10Kbps. In principle, WLAN operated at 11 Mbps should be able to support more than 500 VoIP sessions. However, in reality, this is expected to be much lower due to the overhead of MAC protocol, which includes packet overhead and backoff before the transmission for successful delivery of the packets. Waiting for random backoff time also increases the delay in the network. As the number of users in the network increases, the backoff time increases severely.

9.3.4 Design Requirements

The main design requirement for the algorithm is compatibility with current implementations of IEEE 802.11 and 802.11e. Our goal is to provides system that leverages the nodes supporting the algorithms over those that do not support the algorithm while working concurrently in the network. Since PCF is not supported in most 802.11 products while DCF is mostly used as an IEEE 802.11 protocol, we focus on DCF and EDCF in IEEE 802.11 and 802.11e respectively.

9.4 Definition of the HEARW¡P

HEARW¡P (En**H**anced M**E**dium **A**ccess Cont**R**ol Protocol for **Wi**reless Voice over **IP**) is a protocol that has two main functionalities: sending packets back-to-back and dropping-packets-as-needed. The first functionality aims to increase throughput and decrease delay by decreasing contention attempts [51, 53, 54, 55]. The second aims at decreasing the load in the network with the goal of decreasing the delay of the remaining packets.

9.4.1 Sending Packets Together

The problem in DCF and EDCF modes is that the AP has the same access to the channel as any other node in the network although it is expected to have as many packets as the total number of packets of all the nodes in the network based on the assumption that the communication between the nodes and the network is symmetric. *Sending Packets Together functionality* aims at decreasing the overhead of backoff at the beginning of each packet transmission by allowing AP to acquire the channel and send the packets back-to-back to the other nodes. It can also decrease the overhead of headers in small VoIP packets by joining them in multicast packets.

We showed in Section 5 that we can plot the average wait time during backoff for a successful transmission. It increases considerably as the number of nodes increases. The packet overhead is also considerable as shown in Section 9.2. There is a 44 byte header overhead for the data of 20 bytes, which corresponds to a 20ms segment at 8kbps.

Sending packets back-to-back can be done with multicast transmission. Station whoever subscribes to be in multicast could get an identifier which decreases both the packet and backoff overhead. However, it requires extra functionality in the network. An alternative to this is to send the packets back-to-back. It can be realized very easily with slight changes in IEEE 802.11e MAC although it eliminates only the effect of backoff overhead not packet overhead.

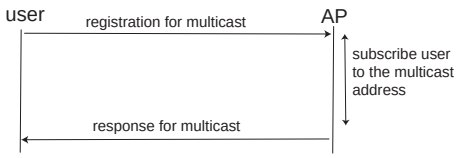


Figure 9.1: Handshaking in VMSL for registering to the multicast address

Implementation of Multicast on Top of IEEE 802.11 MAC

Multicasting VoIP packets requires an additional layer, called “VoIP Multicast Support Layer (VMSL)”, on top of MAC. The nodes that have this functionality subscribe to a multicast address where the AP is the transmitter. Upon reception of packets of these nodes, the AP stores the payload and length of the packets in order to send them in one packet. It decides to send the packets upon reception of a certain number of packets or after a certain time, whichever occurs first.

If a mobile has VMSL, it informs the AP of this functionality by using the reserved bits in the association request frame. Upon reception of this request, if the AP also has VMSL, it informs the user that it also has this functionality by using the reserved bits inside the association response frame. Then it sends another packet to this user, that includes the multicast address associated with this node. At the end of this registration process, the user is subscribed to the multicast address such that it will transfer the packets with destination address containing this multicast address to the VMSL layer. The AP also includes the address of the user inside the list of the nodes that subscribed to the multicast address. Figure 9.1 shows this process.

When the AP receives a packet for a node which is included inside the list of the nodes in the multicast group, it first checks whether it is a VoIP packet based on the

4	IHL	ToS	16-bit total length
16-bit identification	flags	13-bit fragment offset	TTL
protocol	16-bit checksum	AP address	multi-cast address
voice packet 1	voice packet 2	...	voice packet n

Table 9.2: Multicast packet sent from AP to the users subscribed to VMSL

length of the packet. VoIP packets usually contain 20-40ms duration of voice data, which corresponds to 44 byte IP+RTP+UDP header plus 20-40 byte data payload, 64-84 byte. The data packets are expected to be much larger than these packets. If the node is in the multicast group and the packet is voice packet, then the complete IP packet is stored in an array. If this is the first packet to be included in the multicast packet, the AP also starts a timer. Then after a certain time or when the maximum number of node data that can be included inside a packet have been received, the AP includes the packet with source address as its own address and destination as multicast address and including all the IP packets in the data payload as shown in Table 9.2. Multicast packets also do not require to be acknowledged in IEEE 802.11.

Implementation in IEEE 802.11e

The functionality of “Sending-Packets-Together” can be implemented in a simpler way in IEEE 802.11e. In IEEE 802.11e, TXOP is an interval of time when a station has the right to initiate transmissions, defined by a starting time and a maximum duration. Therefore, once the node acquires the channel, it can send the packets to different nodes during the TXOP duration.

The implementation of the functionality in IEEE 802.11e is similar to the implemen-

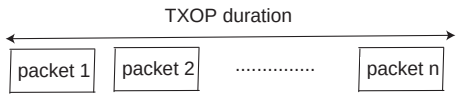


Figure 9.2: Sending packets back-to-back in one TXOP duration

tation of VMSL. In IEEE 802.11, the association request and response frames do not need to use any of the reserved bits. When the AP receives a packet destined to one of the users associated with itself, it first checks whether it is a voice packet based on the length of the packet, as described in VMSL. Then it stores the packets in the same way. If this is the first packet to be included in one frame, AP also starts a timer. Then after a certain time or the maximum number of node data that can be included inside the TXOP limit have been received, the AP acquires the channel and send the packets back-to-back during the TXOP duration as shown in Figure 9.2.

9.4.2 Dropping-Packets-as-Needed

Normally, in both IEEE 802.11a and 802.11e, if the channel is busy, the node does not decrement the backoff value if the channel is not idle for the last IFS time. Also, if the transmission is not successful, the nodes keep increasing the size of the contention window until the CWmax value. Therefore, these protocols are designed to send the packets successfully without caring about the delay.

In **HEARW¡P**, we propose to change the protocols such that the nodes make decisions about whether to send the packet or not after some time. In VoIP applications, there is a maximum tolerable RTT value of 200ms. If the RTT delay is greater than 200ms, then the packet is dropped at the receiver. Therefore, if the packet delay increases too much while

contending for the channel in WLAN, it is better to drop it beforehand without increasing the load of the network any further.

The idea is therefore to drop packets with some probability if the node cannot get the opportunity to transmit its packets. The delay in getting the transmission opportunity is expected to increase as the number of users contending for the channel in WLAN increases. It is shown that in this case the throughput of the network starts decreasing and the delay in the network increases. In order to decrease the delay of a specific percentage of the packets, we claim that we should drop another percentage of the packets generated in the network as they wait to acquire the channel. The adjustment of this percentages is based on the length of the wait time of the packets in the queue.

The algorithm is as follows. The node decides not to send the packet with some probability p_1 at the end of each time period as shown in Figure 9.3. Also, if the packet transmission is not successful, the node will decide whether to start the retransmission with some other probability p_2. The implementation of this functionality in IEEE 802.11b and 802.11e is described next.

Implementation in IEEE 802.11b

The data packets in IEEE 802.11b cannot be dropped. However, the voice packets are not useful after a certain time. In IEEE 802.11b, the channel access is not different for different types of traffic. If the voice packets drop their packets to avoid the WLAN throughput decreasing any further, then the data packets can get the channel and increase the delay of VoIP traffic further while also increasing the packet loss. Therefore, this

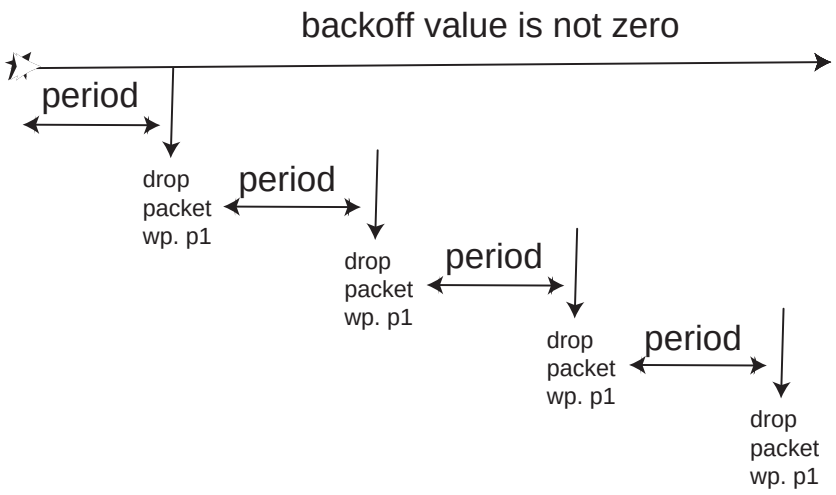


Figure 9.3: Dropping packets periodically

functionality may not be successful in IEEE 802.11b.

Implementation in IEEE 802.11e

In IEEE 802.11e, each traffic category (TC) has its own CWmin, CWmax, TXOPLimit and AIFS. Since TCs achieve higher priority by choosing smaller window sizes, we may not need to worry about the low priority traffic such as TCP connections. Then we can implement *Dropping-Packets-as-Needed* functionality.

Before starting a transmission, a station has to keep sensing the channel for an additional random time after detecting the channel as being idle for a minimum duration called AIFS. However, if the channel is busy most of the time, it means that there are a lot of active users in the network. Therefore, the node may choose to drop the packets to give the right of transmission to the other nodes. The node may decide whether to wait longer or drop the packet periodically with some probability p_1. This way, if the channel is busy most of the time, the expected number of packets that are transmitted will decrease. This

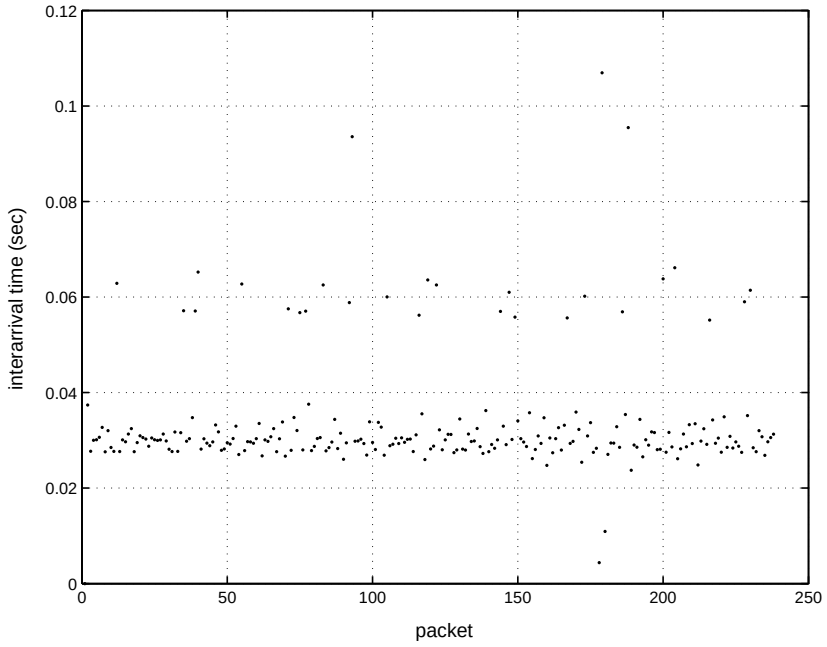


Figure 9.4: Inter-arrival time of a node in reception

may be preferred since the delay will increase so much that they will be dropped anyway.

If there is a collision, the corresponding retry counter increments and the backoff interval increases. In every transmission is intervened with a backoff procedure for fairness. However, this will also increase the delay a lot. It may not be worth retransmitting the packet. Therefore, for each retransmission, the node decides whether to continue trying to retransmit or drop the packet with some probability p_2.

9.5 Experimentation

We experimented with VoIP in the current system. The VoIP delay analysis consists of delay in a wired domain and in a wireless domain. We focus on the wireless domain since the tuning parameters are limited in the wired domain. In order to get a good estimate for

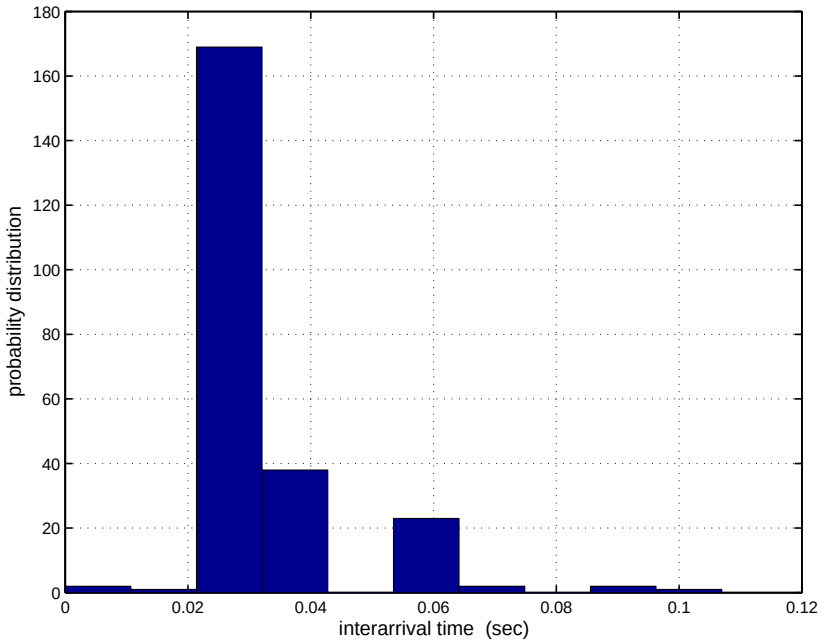


Figure 9.5: Distribution of interarrival time for a node

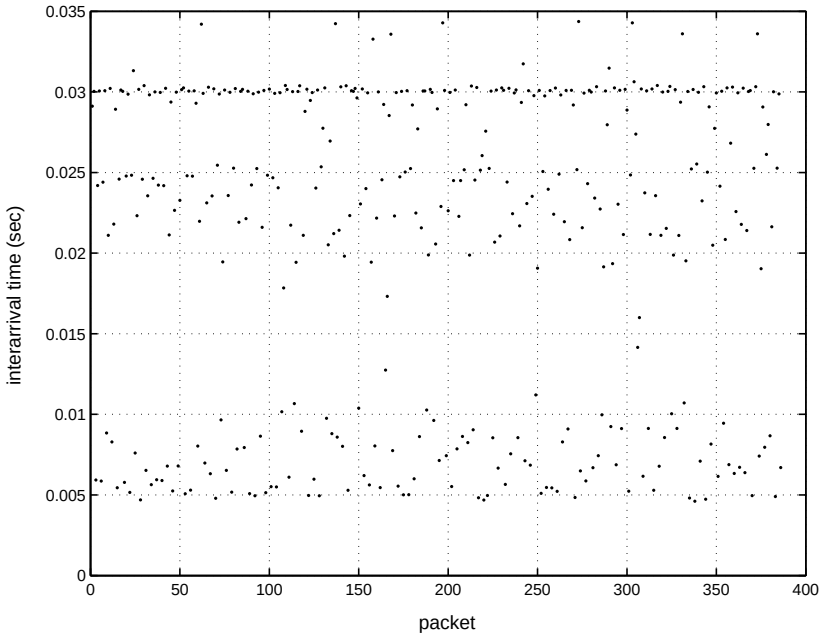


Figure 9.6: Inter-arrival time of two nodes

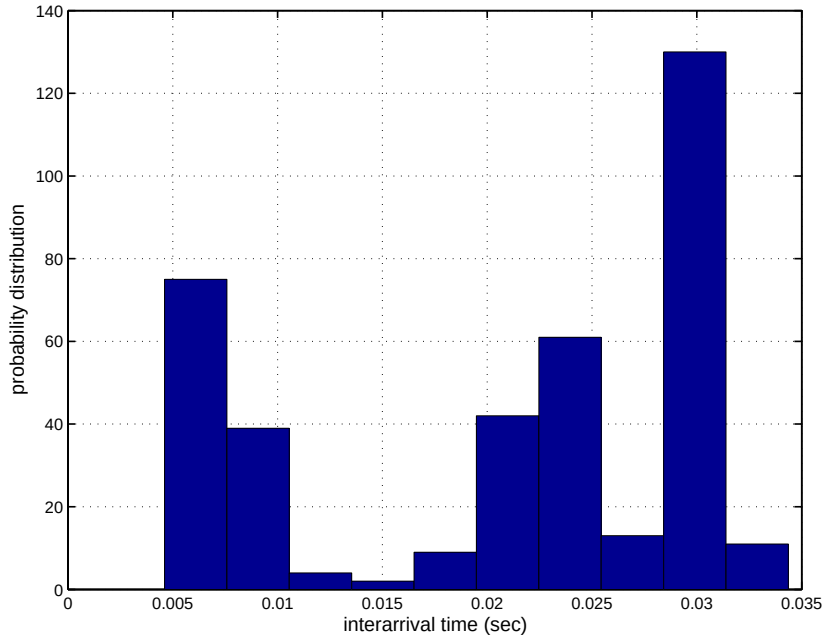


Figure 9.7: Distribution of interarrival time for a node

the wireless domain we located the source and destination to the same subnet. They are all attached to the same access points. As a result, in this setting, the inter-arrival delay is the delay budget for the wireless domain.

We look at a station and take the difference of receiving times of the ACK. Figure 9.4 and 9.5 shows the plot for the instantaneous values; as it can be seen the jitter is around 30ms. This experiment was performed using the department network which includes another regular traffic.

Another test was carried out using Skype VoIP software. This software use Peer-to-Peer methodology. Therefore, in our setting calls are confined within a subnet which means that the wired domain part of the call is very limited and negligible.

Figures 9.6 and 9.7 show the case when a pair of laptops are connected with Skype software for a voice communication. Now we see that we get jitter less than 30ms but the

30ms is still relevant in the majority of the interarrival times. The calls traverse through the wireless domain to the access point and then from access point to wireless domain. As a result, for each session there are two accesses. For (n) stations, they access the wireless domain $(2n-2)$ times in this setting since $(n-1)$ stations perform one session and the access point accesses the channel $n-1$ times in order to deliver the packets for $(n-1)$ stations. **HEARW¡P** algorithm removes one wireless domain access per session and reduces the access to $(n-1)+1$ by aggregation in order to decrease the jitter significantly.

9.6 Performance Analysis

We can formulate and analyze the VoIP system as follows. Assume that there are $(n-1)$ stations and an access point in a network. Each station performs a voice session with a node outside the network. Let's mark STA 1 as the access point. We can infer that in the legacy system packet sizes are equal and are given as follows:

$$P(1) = P(i)\ for\ i \in\ [2, n]$$

$$E[P] = P(1)$$

The packet sizes with **HEARW¡P** are:

$$P(1) = \min((n-1)(P(2)+I), 2024)$$

$$P(2) = P(i)\ for\ i \in\ [2, n]$$

$$E[P] = \tfrac{1}{n}\textstyle\sum_i^n P(i)$$

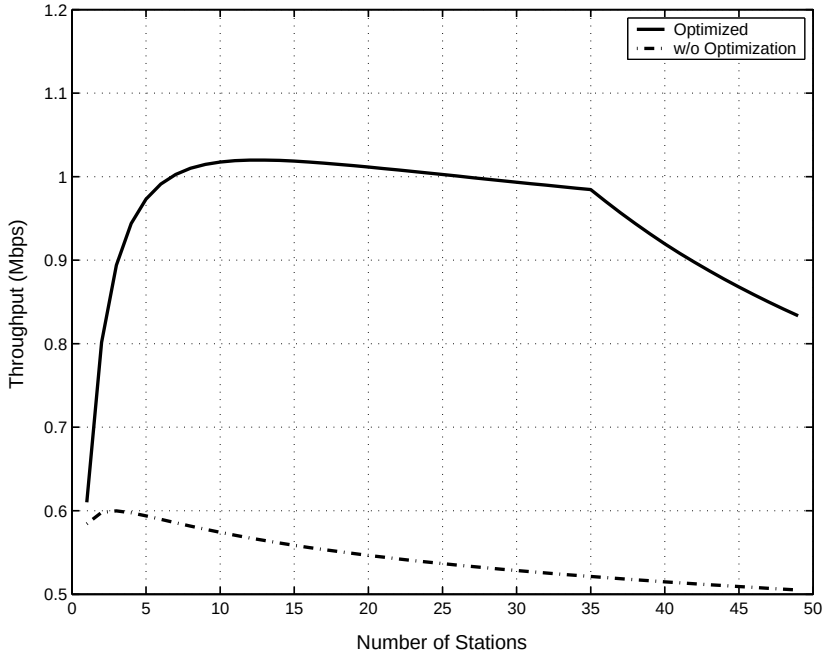


Figure 9.8: Total throughput (w/o RTS/CTS)

where I is the number of identifier bytes and 2024 bytes is the maximum allowed payload. We considered 20ms segment at 8kbps which corresponds to 64 bytes MSDU ($P(2)$) for a single voice session and each station has a packet to transmit all the time with 11 Mbps data rate. In addition, the access point does not expect a ACK packet since it transmits in multicast.

Figure 9.8 shows the total throughput S_T of the system found with our formula explained in Section 4.3:

$$S_T = \frac{P_s E[P]}{(1 - P_{tr})\sigma + \bar{T}_s + \bar{T}_c} \tag{9.1}$$

where $E[P]$ is the average packet size and $\bar{T}_s$ and $\bar{T}_c$ are the average duration values for successful transmission and collision respectively. One can see the substantial throughput

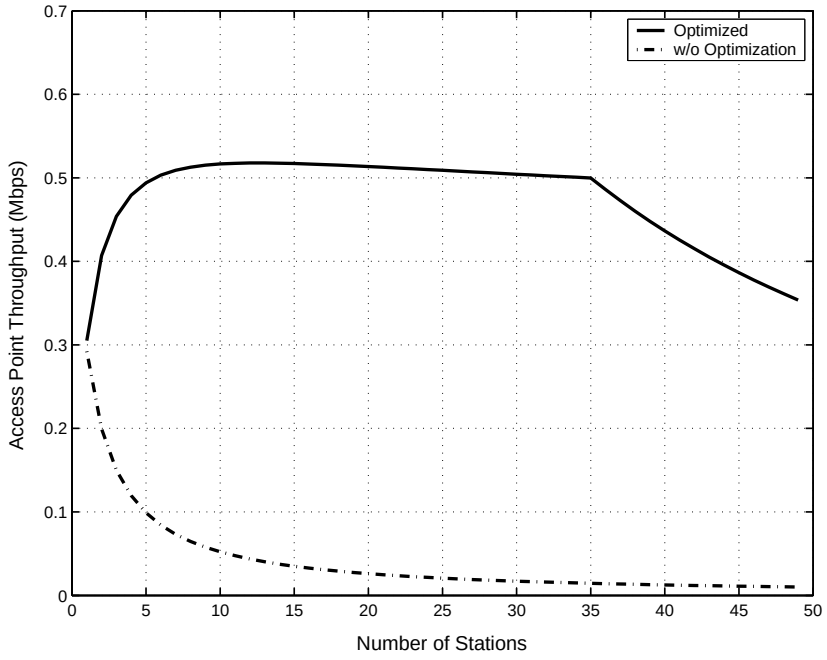


Figure 9.9: Access point throughput (w/o RTS/CTS)

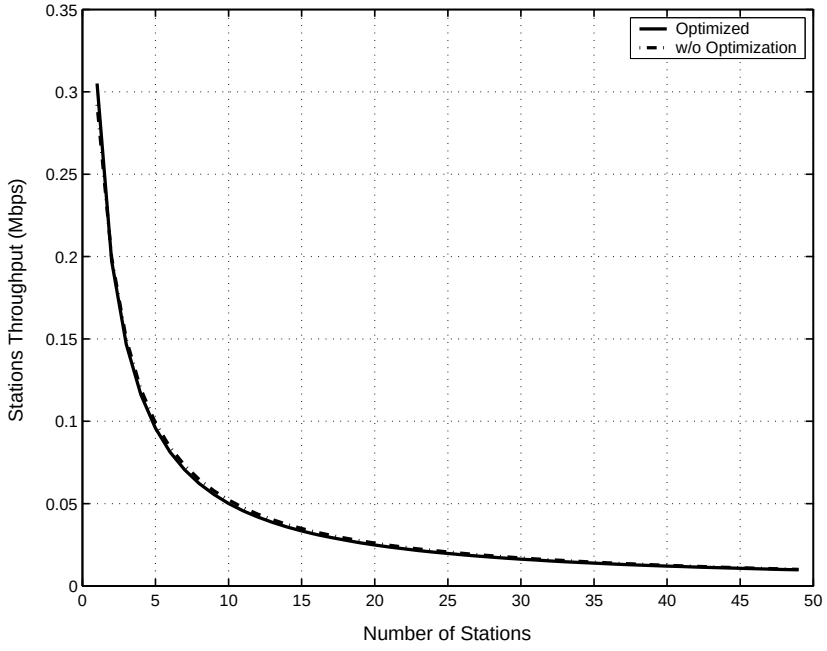


Figure 9.10: Station throughput (w/o RTS/CTS)

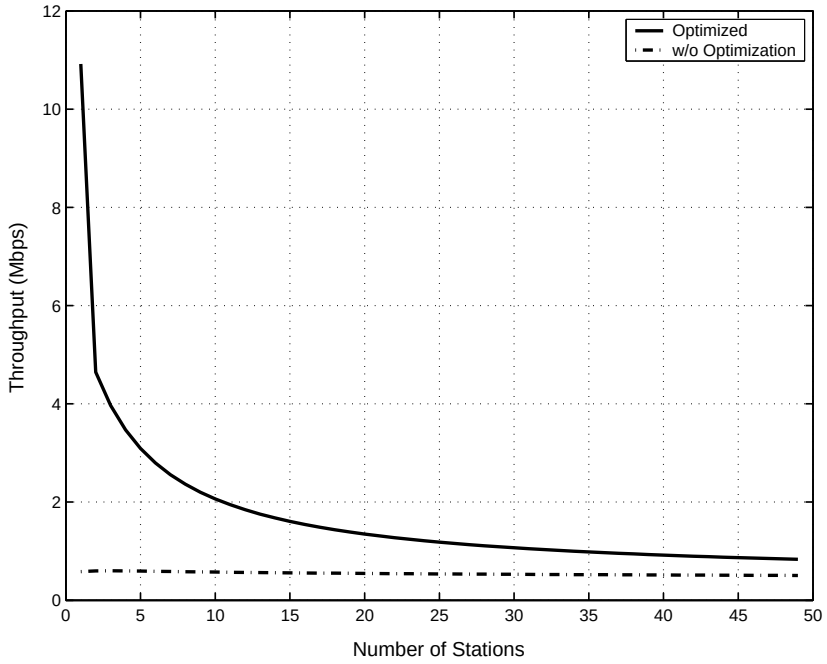


Figure 9.11: Total throughput for frame aggregation (w/o RTS/CTS)

increase from the figure. As can be inferred, the packet size of the access point increases with the number of stations, which increases the throughput. On the other hand, increase in the number of stations decreases the throughput because of the increased probability of collision. This comes to a balance for a while and when the maximum allowed packet size is reached, the throughput starts to decrease.

Figure 9.9 shows the access point throughput S_A which is found with the following formula:

$$S_A = \frac{1}{n} \frac{P_s P(1)}{(1 - P_{tr})\sigma + \bar{T}_s + \bar{T}_c} \tag{9.2}$$

Sending the packets in burst increases the throughput of the access point. This throughput calculation considers the successfully transmitted bits but not the received bits. As a result, stations receive packets with higher rates and the station's transmitted throughput

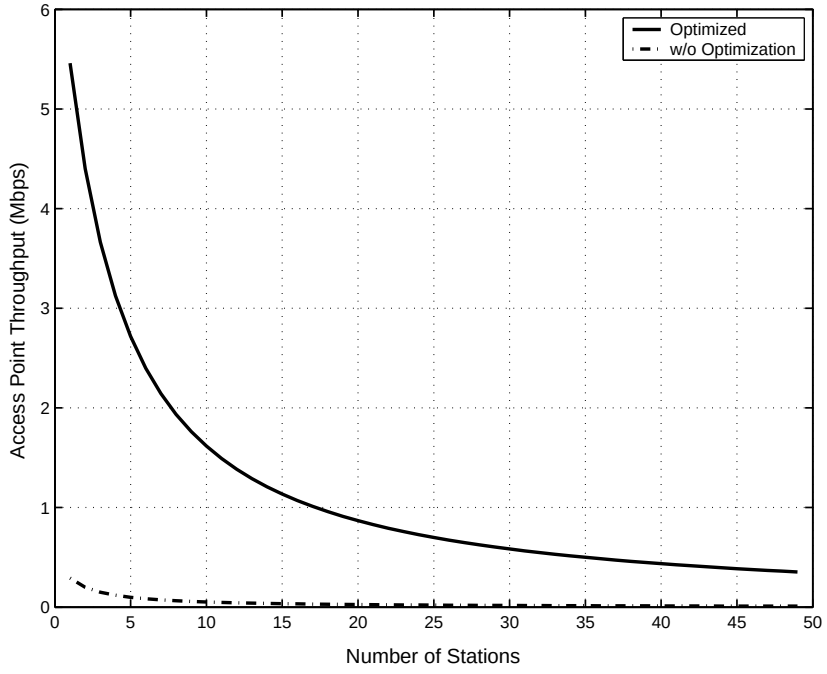


Figure 9.12: Access point throughput for frame aggregation (w/o RTS/CTS)

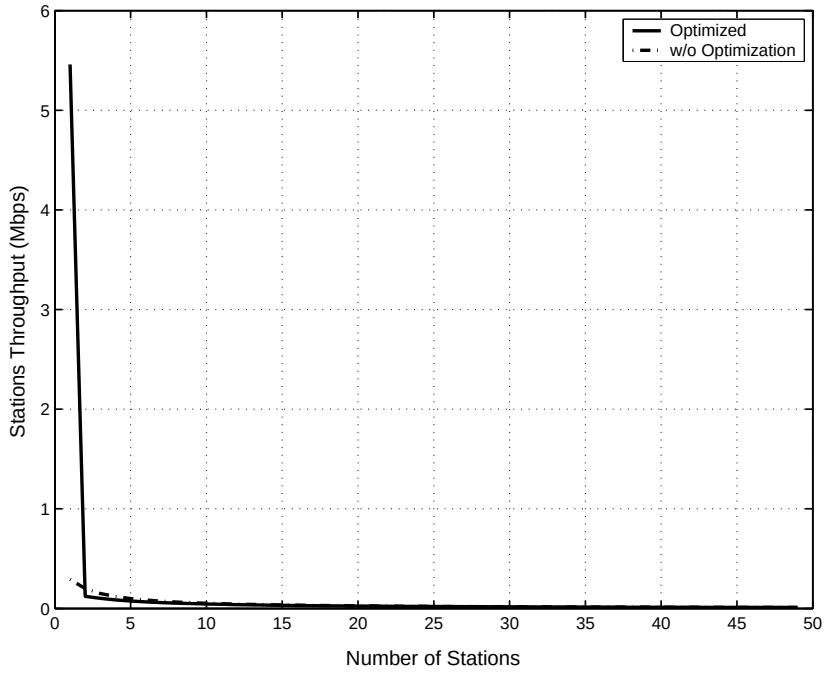


Figure 9.13: Station throughput for frame aggregation (w/o RTS/CTS)

S_S is found with the formula below:

$$S_S = \frac{1}{n} \frac{P_s P(2)}{(1 - P_{tr})\sigma + \bar{T}_s + \bar{T}_c} \tag{9.3}$$

Figure 9.10 depicts the comparison. As expected, a station's throughput S_S with VoIP algorithm should be lower than that without the algorithm. This is because higher packet size of the access point increases the average duration values. One can easily see that the following equation should hold:

$$S_T = S_A + (n-1)S_S$$

Frame Aggregation

From Figure 8.1 we can infer that sending bigger size packets increases the throughput. We can easily conclude that each station can aggregate the small size packets into a big one and then send it. If we look at a scenario when only the access point aggregates the packets into a 2024 bytes size packets ($P(1)$) we obtain a performance increase which gradually decreases as a new station pops up.Figure 9.11 shows the total throughput and Figures 9.12 and 9.13 show the access point throughput and stations throughput, respectively. In frame aggregation traffic direction could be one destination or multiple destinations. If it is sent with multicast, then there is no need for an ACK packet. Otherwise, for one destination only, one ACK is needed and for multiple destination each station may send an ACK packet one by one with respect to their order in the received packet. This topic is currently being considered in IEEE 802.11n working group.

Chapter 10

Adaptive Bit Loading and Subcarrier Allocation

10.1 Introduction

Broadband Wireless Access (BWA) is an appealing system for providing flexible and easy deployment solution to high-speed communications. It is an alternative to wireline broadband access techniques such as copper line, coaxial cable, xDSL and cable modem [56, 57]. Visionaries predict a big market because of its distributed installation and semi-ad-hoc routing protocol that reduces the need for an infrastructure [1, 3, 58].

Vector Orthogonal Frequency Division Multiplexing (VOFDM) is considered as a base setting for BWA systems by the Broadband Wireless Internet Forum (BWIF), one of the programs of the IEEE Industry Standards and Technology Organization (IEEE-ISTO) [56]. Some vendors offer BWA system with the existing wireless LAN technologies such as IEEE 802.11a and IEEE 802.11b. IEEE 802.16 group aims to unify the BWA

solutions [59]. IEEE 802.16 group issued standards in the 10-66 GHz bands and IEEE 802.16a group was formed to develop standards to operate in the 2-11 GHz bands in which channel impairments, multipath fading and path loss become more significant with the increase in the number of subscribers. Improved and flexible multiple access methods are needed to cope with these impairments. Orthogonal Frequency Division Multiple Access (OFDMA) is a promising multiple access scheme that has attracted interest. OFDMA is based on OFDM and inherits its immunity to inter-symbol interference and frequency selective fading [60, 61].

Achieving high transmission rates depends on the ability of BWA system to provide efficient and flexible resource allocation. Recent studies [57, 62, 63, 64, 65, 66, 67] on resource allocation demonstrate that significant performance gains can be obtained if frequency hopping and adaptive modulation are used in subcarrier allocation, assuming knowledge of the channel gain in the transmitter. The frequency hopping strategy resembles the interference cancelation of CDMA [68].

In a multiuser environment, a good resource allocation scheme leverages multiuser diversity and channel fading [69]. It was shown in [70] that the optimal solution is to schedule the user with the best channel at each time. Although in this case, the entire bandwidth is used by the scheduled user, this idea can also be applied to OFDMA system. Here, the channel is shared by the users, each owing a mutually disjoint set of subcarriers, by scheduling the subcarrier to a user with the best channel among others. Of course, the procedure is not simple since the best subcarrier of the user may also be the best subcarrier of another user who may not have any other good subcarriers. The overall strategy is to use the peaks of the channel resulting from channel fading. Unlike in the traditional view

where the channel fading is considered to be an impairment, here it acts as a channel randomizer and increases multiuser diversity [69].

The resource allocation problem has been recently considered in many studies. Almost all of them define the problem as a real-time resource allocation problem in which Quality of Service (QoS) requirements are fixed by the application. QoS requirement is defined as achieving a specified data transmission rate and bit error rate (BER) of each user in each transmission. In this regard, the problem differs from the water-pouring schemes wherein the aim is to achieve Shannon capacity under the power constraint [62].

We introduce an iterative multiuser bit and power allocation scheme so that the QoS requirements of users are fulfilled. Our objective is to minimize the total transmit power by allocating subcarriers to the users and then to determine the number of bits transmitted on each subcarrier. Variable transmittable bits (i.e adaptive modulation) was considered in [62, 63]. Our scheme is simple and sufficiently fair to meet real-time applications criteria in which a quick scheme is needed to allocate subcarriers before the channel changes and a fair scheme is needed to treat each user.

We also consider a continuous allocation scheme where the allocator uses the previous channel information per user for the current allocation. We try to extend the point-to-point version of proportional fair scheduling (as in [69]) to a point-to-multipoint version. In this scheme there is no fixed requirements per symbol, the aim is to maximize capacity.

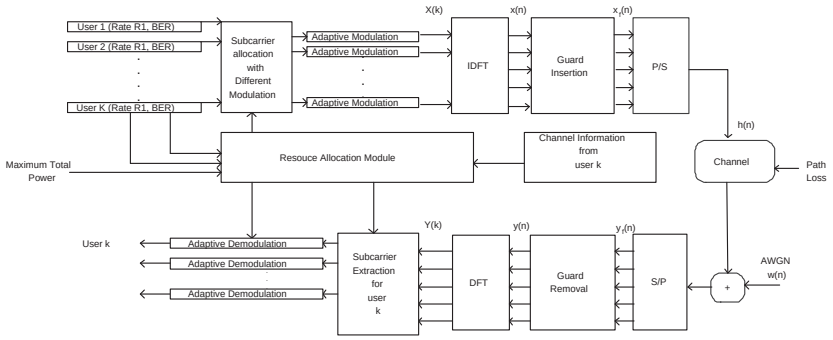


Figure 10.1: Orthogonal Frequency Division Multiple Access System

10.2 Orthogonal Frequency Division Multiple Access

This section outlines the OFDMA system and states the resource allocation problem. Unlike in an OFDM system [60], K users are involved in the OFDMA system to share N subcarriers. The difference arises in the forming and an deforming of FFT block. The rest is the same as an OFDM system as seen in Fig. 10.1. Each user allocates non-overlapping set of subcarrriers S_k where the number of subcarriers per user is $J(k)$ following the notation in [71]. We denote by $X_k(l)$ the l^{th} subcarrier of the FFT block belonging to k^{th} user. $X_k(l)$ is obtained by coding the assigned bits c with the corresponding modulation scheme. In the downlink the $X_k(l)$ are multiplexed to form the OFDM symbol[1] of length ($N + L$) with the appended guard prefix L in order to eliminate ISI. At the uplink, the OFDM symbol is formed in the base station with a synchronization error as follows:

$$x(l) = \sum_{k=0}^{K-1} \sum_{n=0}^{J(k)-1} X_k(n) e^{j\frac{2\pi}{N}(I_k(n))l}, \tag{10.1}$$

[1] A OFDMA symbol is defined as one OFDM FFT block.

with $n = -L, ..., N-1$, where $I_k(n)$ denotes the subcarrier assigned to the k^{th} user. A resource allocation problem comes into the picture when associating the set of subcarriers to the users with different bits loaded into them. The received signal from the j^{th} user is:

$$y_j(l) = x(l) \bigotimes h_j(l) + w(l), \tag{10.2}$$

where $h_j(t)$ is the baseband impulse response of the channel between base station (BS) and j^{th} user. Equation (10.2) is the received signal $y(t)$ sampled at rate $1/T$. The first L samples are discarded and the N-point FFT is computed. The data of the j^{th} user is:

$$Y_j(n) = \begin{cases} X_j(n)H_j(i_j(n)) + W(n), & if\, i_j(n) \in S_j \\ 0, & otherwise, \end{cases} \tag{10.3}$$

where $H_j(n) := \sum_i h_j(i) exp(j2\frac{\pi}{N}ni)$ is the frequency response of the channel of k^{th} user.

In a perfectly synchronized system, the allocation module of the transmitter assigns subcarriers to each user according to some QoS criteria. QoS metrics in the system are rate and bit error rate (BER). Each user's bit stream is transmitted using the assigned subcarriers and adaptively modulated for the number of bits assigned to the subcarrier. The power level of the modulation is adjusted to overcome the fading of the channel. The transmission power for AWGN channel can be predicted. In addition the channel gain of subcarrier n to the corresponding user k should be known. The channel gain of the subcarrier is defined as:

$$\alpha_{k,n} = H_k(n) * PL_k, \tag{10.4}$$

where PL is the path loss, defined by:

$$PL_k = PL(d_o) + 10\alpha log_{10}(d_k/d_o) + X_\sigma,$$

where d_o is the reference distance, d_k is the distance between transmitter and receiver, α is the path loss component and X_σ is a Gaussian random variable for shadowing with a standard deviation σ [64]. An example of channel gain can be seen in Fig. 10.2.

The problem above is called resource allocation in the OFDMA literature. The channel information is assumed to be known at transmitter and receiver [57, 62, 63, 64, 65, 66, 67]. The channel is assumed to be reciprocal; BS is able to estimate the channel of all BS-to-mobile links based on the received uplink transmission as long as the channel variation is slow [65]. As a result, the resource allocation should be done within the coherence time [62].

With the channel information, the objective of resource allocation problem can be defined as maximizing the throughput subject to a given total power constraint regarding the user's QoS requirements. As we clarify further, BER_k of the transmission should not be higher than the required BER_k and data rate of every user should be equal to the requirement R_k.

Let's define $\gamma_{k,n}$ as the indicator of allocating the n^{th} subcarrier to the k^{th} user. The transmission power allocated to the n^{th} subcarrier of k^{th} user is expressed as:

$$P_{k,n} = \frac{f_k(c_{k,n}, BER_k)}{\alpha_{k,n}^2}, \tag{10.5}$$

where $f_k(c_k, n)$ is the required received power with unity channel gain for reliable reception of $c_{k,n}$ bits per symbol [63]. We can formulate the resource allocation problem with an imposed power constraint as:

$$\max_{c_{k,n},\gamma_{k,n}} R_k = \sum_{n=1}^{N} c_{k,n}\gamma_{k,n} \; for\, all\, k$$

$$subject\, to\, P_T = \sum_{k=1}^{K}\sum_{n=1}^{N} \frac{f_k(c_{k,n},BER_k)}{\alpha_{k,n}^2}\gamma_{k,n} \leq P_{Max}$$

where the limit on the total transmission power is expressed as P_{Max} for all $n \in \{1, ..., N\}$, $k \in \{1, ..., K\}$ and $c_{k,n} \in \{1, ..., M\}$.

If there is no power constraint, equation (10.6) is changed in order to minimize P_T subject to allocating R_k bits for all k (i.e problem is to find the values of the $\gamma_{k,n}$ and the corresponding $c_{k,n}$ while minimizing P_T) [62, 63, 66]. As it can be seen the cost function in our system is the power consumption matrix in equation (10.5). Rather than using $\alpha_{k,n}^2$ as in [62, 63, 66], we adopt $P_{k,n}$ from [64] since in this case, modulation type and BER are involved in the decision process.

10.3 Optimal Solution

In a multiuser environment with multiple modulation techniques, the solution to the problem is complicated since the optimal solution needs to pick the subcarriers in balance. We can classify the problem according to each set of bits assigned to a subcarrier. For a user k, $f_k(c_{k,n}) \in \{f_k(1, BER_k), ..., f_k(M, BER_k)\}$. We can construct M times $[K * N]$ power

matrices$\{P^c\}$ for each c. For a constant c, $\{f(c)\}$ can be computed and the transmission power requirement can be found with equation (10.5). The dimension of the indicator function is incremented and represented by $\gamma_{k,n,c}$ and defined as follows [63]:

$$\gamma_{k,n,c} = \begin{cases} 1, & if\, c_{k,n} = c \\ 0, & otherwise \end{cases} \tag{10.6}$$

The above problem can be solved with Integer Programming (IP). We refer to the IP approach as the optimal solution to the resource allocation problem. As stated in [63], the non linear approximation in [62, 67] requires more computation than the IP.

There are $K * N * M$ indicator variables and M power matrices where the entries of each matrix for a given c can be found from:

$$P^c_{k,n} = \frac{f_k(c, BER_k)}{\alpha^2_{k,n}}. \tag{10.7}$$

Using equation (10.7) as an input, the cost function now can be written as:

$$P_T = \sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{c=1}^{M} P^c_{k,n} \gamma_{k,n,c} \tag{10.8}$$

and the description of the IP problem is:

$$\min_{\gamma_{k,n,c}} P_T, \;\; for\, \gamma_{k,n,c} \in \{0, 1\} \tag{10.9}$$

subject to:

$$R_k = \sum_{n=1}^{N} \sum_{c=1}^{M} c.\gamma_{k,n,c}, \ \ for\, all\, k,$$

and

$$0 \leq \sum_{k=1}^{K} \sum_{c=1}^{M} \gamma_{k,n,c} \leq 1, \ \ for\, all\, n.$$

Although the optimal solution gives the exact results, from an implementation point of view, it is not preferred since in a time varying channel, in order to allocate the subcarriers within the coherence time, the allocation algorithm should be fast and the IP complexity increases exponentially with the number of constraints. This real-time requirement leads to searching suboptimal solutions that are fast and close to the optimal solution. Several suboptimal allocation schemes are proposed for different settings in the literature [62, 63, 64, 65, 66, 72]. Up to now suboptimal solutions differ in the modulation type. There are a few suboptimal schemes that use adaptive modulation; the rest assume fixed modulation, i.e. same number of bits are assigned to each subcarrier. We will describe current solutions and compare them with our iterative solution.

10.4 Suboptimal Solutions

In most attempts to simplify the resource allocation problem, the problem is decomposed into two procedures: a subcarrier allocation with fixed modulation, and bit loading. Subcarrier allocation with fixed modulation deals with one P^c matrix with fixed c and then by using bit loading scheme, the number of bits is incremented.

10.4.1 Subcarrier Allocation

We know that $f_k(x, y)$ is a convex function [62, 63]. We can start with $P^1_{k,n}$ and we can define new $\bar{R}_k$ with $\sum_{k=1}^{K} \bar{R}_k \leq N$ which can be obtained by decrementing R_k properly. Then the solution to this problem can be solved with Linear Programming or mapping to the Hungarian problem. Although the Hungarian algorithm is proposed as an optimal solution for resource allocation with a fixed modulation in [64, 66], we consider it as a suboptimal solution for adaptive modulation.

Linear Programming

Linear programming is investigated in [63]. For comparison purposes, we briefly restate the problem description:

$$P_T = \min \sum_{k=1}^{K} \sum_{n=1}^{N} P^1_{k,n} \rho_{k,n} \quad \rho_{k,n} \in [0, 1], \tag{10.10}$$

subject to:

$$\sum_{n=1}^{N} \rho_{k,n} = \bar{R}_k \quad \forall\, k \in \{1, ..., K\}$$

$$\sum_{k=1}^{K} \rho_{k,n} = 1 \qquad \forall\, n \in \{1, ..., N\}.$$

After linear programming, the $[K * N]$ allocation matrix has entries ranging between 0 and 1. The entries are converted to integers by selecting the highest $\bar{R}_k$ non-zero values from N columns for each k and assigning them to the k^{th} user.

Hungarian Algorithm

The problem described above can also be solved by an assignment method such as Hungarian algorithm [73]. The Hungarian algorithm works with square matrices. Entries of the square matrix can be formed by adding $\bar{R}_k$ times the row of each k. The problem formulation is:

$$P_T = \min \sum_{k=1}^{N} \sum_{n=1}^{N} P_{k,n}^{1} \rho_{k,n} \;\; \rho_{k,n} \in \{0,1\} \tag{10.11}$$

and the constraints become:

$$\sum_{n=1}^{N} \rho_{k,n} = 1 \quad \forall\, n \in \{1,...,N\}$$

$$\sum_{k=1}^{N} \rho_{k,n} = 1 \quad \forall\, k \in \{1,...,N\}$$

as stated in [66]. Although the Hungarian method has computation complexity $O(n^4)$ in the allocation problem with fixed modulation, it may serve as a base for adaptive modulation.

10.4.2 Bit Loading Algorithm

The bit loading algorithm (BLA) appears after the subcarriers are assigned to users that have at least $\bar{R}_k$ bits assigned. Bit loading procedure is as simple as incrementing bits of the assigned subcarriers of the users until $P_T \leq P_{Max}$. Following the notation of [62, 63], define $\Delta P_{k,n}(c)$ as the additional power needed to increment one bit of the n^{th} subcarrier of k^{th} user as represented in equation (10.12),

$$\Delta P_{k,n}(c_{k,n}) = \frac{[f(c_{k,n}+1) - f(c_{k,n})]}{\alpha_{k,n}^2} \tag{10.12}$$

The bit loading algorithm assigns one bit at a time with a greedy approach to the subcarrier. Representation is as follows: $\{\arg\min_{k,n} \Delta P_{k,n}(c_{k,n})\}$.

BL Algorithm

STEP 1: For all n, Set $c_{k,n} = 0$, $\Delta P_{k,n}(c_{k,n})$, and $P_T = 0$;

STEP 2: Select $\bar{n} = \arg\min_n \Delta P_{k,n}(0)$;

STEP 3: Set $c_{k,\bar{n}} = c_{k,\bar{n}} + 1$ and $P_T = P_T + \Delta P_{k,n}(c_{k,n})$;

STEP 4: Set $\Delta P_{k,n}(c_{k,\bar{n}})$;

STEP 5: Check $P_T \leq P_{Max}$ and R_k for $\forall k$, if not satisfied GOTO STEP 2.

STEP 6: Finish.

It is a simple algorithm. Bits on the subcarriers are incremented one by one. If there is no power constraint, procedure runs for $\sum_{k=1}^{K} R_k$ times. This algorithm enable us to convert the fixed modulation schemes into adaptive modulation ones.

The Hungarian approach and LP approach with bit loading appear as two different suboptimal solutions to the resource allocation with adaptive modulation. We use these schemes as a reference in our simulations and call them GreedyHungarian and GreedyLP, respectively, in our simulations.

10.5 Iterative Solution

The GreedyLP and GreedyHungarian methods both first determine the subcarriers and then increment the number of bits on them according to the rate requirements of users.

This may not be a good schedule in some certain cases. For instance, consider a user with only one good subcarrier and a low rate requirement. The best solution for that user is allocating its good carrier with high number of bits. But if GreedyLP or GreedyHungarian is used, user may have allocated more than one subcarrier with lower number of bits and in some cases, its good subcarrier is never selected. Consider another scenario where a user does not have any good subcarrier (i.e. it may have a bad channel or be at the edge of the cell). In this case, rather than pushing more bits and allocating less subcarriers as in GreedyLP and GreedyHungarian, the opposite strategy is preferred since fewer bits in higher number of subcarriers give better result. Another difficulty arises in providing fairness. Since GreedyLP and GreedyHungarian are based on greedy approach, the user in the worst condition usually suffers. In any event, these are complex schemes and simpler schemes are needed to finish the allocation within the coherence time. To cope with these challenges, we introduce a simple, efficient and fair subcarrier allocation scheme with iterative improvement.

Our scheme is composed of two modules named “scheduling” and “improvement” modules. In the scheduling section, bits and subcarriers are distributed to the users and passed to the improvement module where the allocation is improved iteratively by bit swapping and subcarrier swapping algorithms.

10.5.1 Fair Scheduling Algorithm

We introduce a simple and mixed allocation scheme that considers fair allocation among users with adaptive modulation. The allocation procedure starts with the highest level of

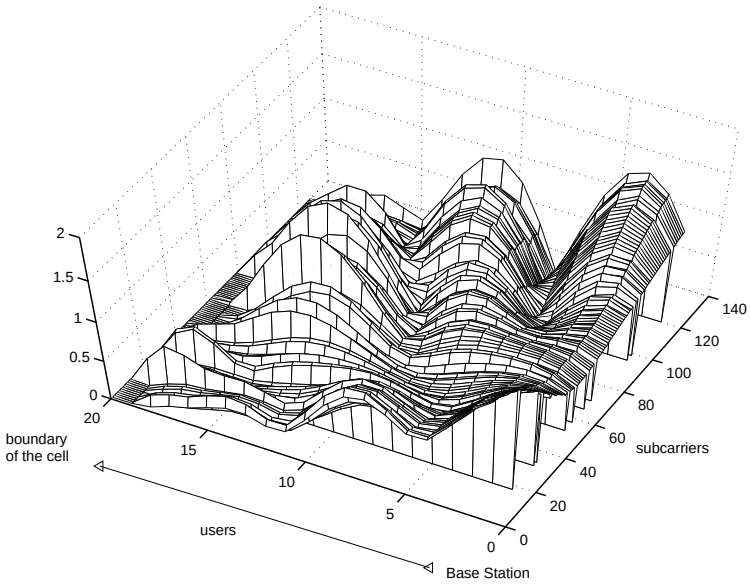


Figure 10.2: An example of channel gain

modulation scheme. In this way, it tries to find the best subcarrier of a user to allocate the highest number of bits. We can describe the strategy by an analogy: “The best strategy to fill a case with stone, pebble and sand is as follows. First filling the case with the stones and then filling the gap left from the stones with pebbles and in the same way, filling the gap left from pebbles with sand. Since filling in opposite direction may leave the stones or pebbles outside”. With this strategy more bits can be allocated and the scheme becomes immune to uneven QoS requirements. The fair scheduling algorithm (FSA) runs a greedy release algorithm (GRA) if there are non-allocated subcarriers after the lowest modulation turn and the rate requirement is not satisfied. GRA decrements one bit of a subcarrier to gain power reduction, which is used to assign higher number of bits to the users on the whole. FSA is described as follows:

FS Algorithm

STEP 1: Set $c = M$, Select a k, and $P_T = 0$;

STEP 2: Find $\bar{n} = arg\min_n P^c_{k,n}$;

STEP 3: Set $R_k = R_k - c$ and $\rho_{k,\bar{n}} = 1$, Update P_T, Shift to the next k;

STEP 4: If $P_T > P_{Max}$, Step Out and Set $c = c - 1$, GOTO STEP 2.

STEP 5: If $\forall k$, $R_k < c$, Set $c = c - 1$, GOTO STEP 2.

STEP 6: If $\{c == 1\}$,$\sum_{k=1}^{K}\sum_{n=1}^{N}\rho_{k,n} < N$, $P_T > P_{Max}$, Run "Greedy Release" and GOTO STEP 2.

STEP 7: Finish.

10.5.2 Greedy Releasing Algorithm

The GRA tends to fill the unallocated subcarriers. It releases one of the bits of the most expensive subcarrier to gain power reduction in order to drive the process. GRA works in the opposite direction of BLA. GRA is described as follows:

GR Algorithm

STEP 1: Find $\{\bar{k}, \bar{n}, \bar{c}_{\bar{k},\bar{n}}\} = \arg\max_{k,n,c} P^c_{k,n}\rho_{k,n}\ \forall c$;

STEP 2: Set $\bar{c}_{k,n} = \bar{c}_{k,n} - 1$,$P_T = P_T - \Delta P_{\bar{k},\bar{n}}(c_{\bar{k},\bar{n}})$;

STEP 3: Set $c = c_{\bar{k},\bar{n}} - 1$;

STEP 4: Finish.

10.5.3 Horizontal Swapping Algorithm

The horizontal swapping algorithm (HSA) aims to smooth the bit distribution of a user. When the subcarriers are distributed, the bit weight per subcarrier can be adjusted to reduce power. One bit of a subcarrier may be shifted to the other subcarrier of the same user if there is a power reduction gain. Therefore, variation of the power allocation per subcarrier is reduced and a smoother transmission is performed. HSA is described as follows:

HS Algorithm

STEP 1: Set P_C=∞

STEP 1a: Find $\{\bar{k}, \bar{n}, \bar{c}_{\bar{k},\bar{n}}\} = \arg\max_{k,n,c}(P^c_{k,n}\rho_{k,n}) < P_C \; \forall c$;

STEP 2: Define $n \in S_k$, where $\{\rho_{k,n} == 1\}$ for $\forall n$;

STEP 3: Set $\Delta_{\dot{n}}$=$\max_n \Delta P_{\bar{k},\bar{n}}(c_{\bar{k},\bar{n}} - 1)$-$\Delta P_{\bar{k},\dot{n}}(c_{\bar{k},\dot{n}})$, $\dot{n} \in S_k$;

STEP 4: Set $P_C = P^{\bar{c}}_{\bar{k},\bar{n}}$;

STEP 4a: if $\Delta_{\dot{n}} > 0$, Set $P_T = P_T - \Delta_{\dot{n}}$

STEP 4b: Set $c_{\bar{k},\bar{n}} = c_{\bar{k},\bar{n}} - 1$, $c_{\bar{k},\dot{n}} = c_{\bar{k},\dot{n}} + 1$ GOTO Step 1a;

STEP 5: if $\{P_C == \min_{k,n,c}(P^c_{k,n}\rho_{k,n})\}$, Finish.

10.5.4 Vertical Swapping Algorithm

Vertical swapping is done for every pair of users. In each iteration, users try to swap their subcarriers such that the power allocation is reduced. There are different types of vertical swapping. For instance, in triple swapping, user i gives its subcarrier to user j and in the same way user j to user k and user k to user i. Pairwise swapping for fixed modulation is described in [64, 66] with a slight difference: the former uses power and the latter uses channel gain as a decision metric. We modified pairwise swapping to cope with adaptive modulation case. In this case, there is more than one class where each class is defined with its modulation (i.e number of bits loaded to a subcarrier) and swapping is only within the class. Each pair of users swap their subcarriers that belong to the same class if there is a power reduction. In this way, adjustment of subcarrier is done across users, to try to approximate the optimal solution. VSA is described as follows:

VS Algorithm

STEP 1: $\forall$ pair of user $\{i, j\}$;

STEP 1a: Find $\partial P_{i,j}(n) = P^{\dot{c}}_{i,n} - P^{\dot{c}}_{j,n}$ and $\Delta^{\hat{n}} P_{i,j} = \max \partial P_{i,j}(n), \forall n \in S_i$;

STEP 1b: Find $\partial P_{j,i}(n) = P^{\dot{c}}_{j,n} - P^{\dot{c}}_{i,n}$ $\Delta^{\check{n}} P_{j,i} = \max \partial P_{j,i}(n), \forall n \in S_j$;

STEP 1c: Set $\Omega^{\hat{n},\check{n}} P_{i,j}$=$\Delta^{\hat{n}} P_{i,j}$+$\Delta^{\check{n}} P_{j,i}$;

STEP 1d: Add $\Omega^{\hat{n},\check{n}} P_{i,j}$ to the $\{\Lambda\}$ list;

STEP 2: Select Ω=$\max_{(i,\hat{n}),(j,\check{n})} \Lambda$;

STEP 3: if $\Omega > 0$, Switch subcarriers and $P_T = P_T - \Omega$ GOTO STEP 1a;

STEP 4: if $\Omega \leq 0$, Finish.

10.6 Resource Allocation Regarding Capacity

In the previous sections, we considered the problem where the QoS requirements per symbol is fixed. Another way to approach resource allocation is in terms of capacity [62]. Suppose there is no fixed requirements per symbol and the aim is to maximize capacity.

It has been shown in [69] that for point-to-point links, a fair allocation strategy maximizes total capacity and the throughput of each user in the long run, when the user's channel statistics are the same. This idea underlying the proposed fair scheduling algorithm exploits the multiuser diversity gain.

With a slight modification, we can extend the fair scheduling algorithm for point-to-point communication to an algorithm for point-to-multipoint communication. Suppose time-varying data rate requirement $R_k(t)$ is sent by the user to the base station as feedback of the channel condition. We treat symbol time as the time slot, so t is discrete, representing the number of symbols. We keep track of average throughput $t_{k,n}$ of each user for a subcarrier in a past window of length t_c. The scheduling algorithm will schedule a subcarrier $\bar{n}$ to a user $\bar{k}$ according to the following criterion:

$$\{\bar{k}, \bar{n}\} = \arg\max_{k,n} \frac{r_{k,n}}{t_{k,n}} \tag{10.13}$$

where $t_{k,n}$ can be updated using an exponentially weighted low-pass filter described in [69]. Here, we are confronted with determining the $r_{k,n}$ values. We can set $r_{k,n}$ to

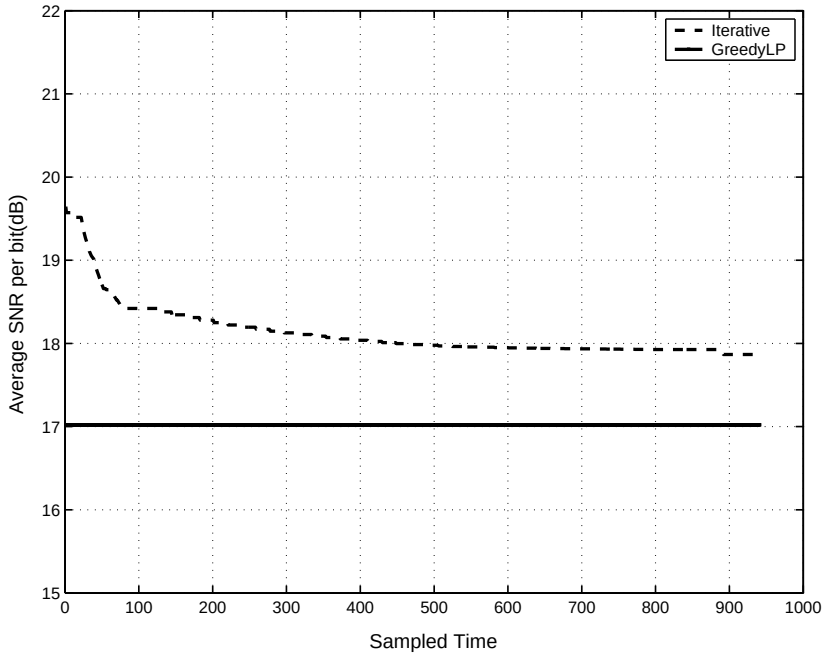


Figure 10.3: Comparison of convergence of the iterative approach to the GreedyLP one

R_k/N, where N is the number of carriers. With this setting, the peaks of the channel for a given subcarrier can be tracked. The algorithm schedules a user to a subcarrier when the channel quality in that subcarrier is high relative to its average condition in that subcarrier over the time scale t_c. When we consider all subcarriers the fairness criterion match with the point-to-point case as follows:

$$\bar{k} = \max_k \frac{R_k}{T_k}, \tag{10.14}$$

where $T_k = \sum_{n=1}^{N} t_{k,n}$. The theoretical analysis of fairness property of (10.14) for point-to-point communication is derived in [69]. We can apply these derivations for point-to-multipoint communication.

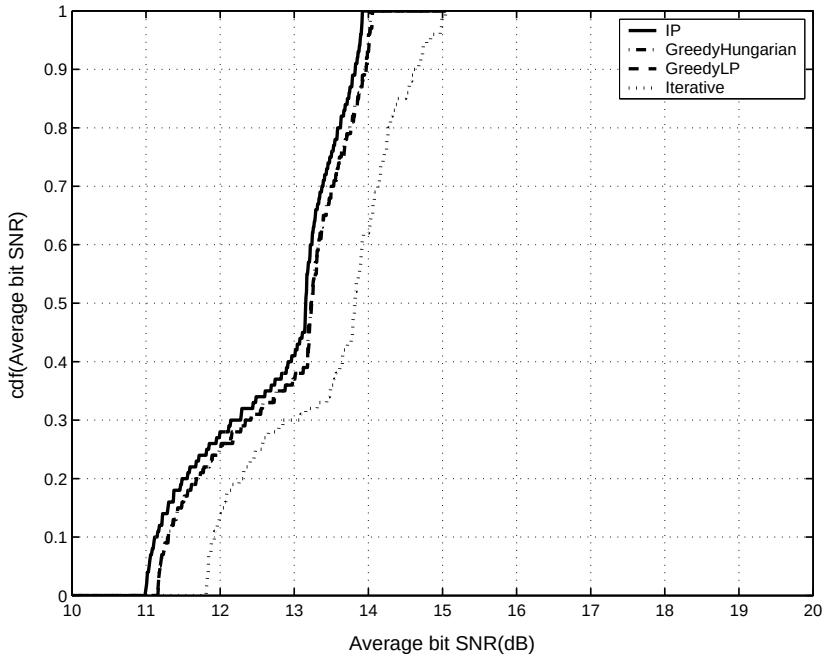


Figure 10.4: Comparison of the cumulative distribution function of the average bit SNR (w/o Power Constraint)

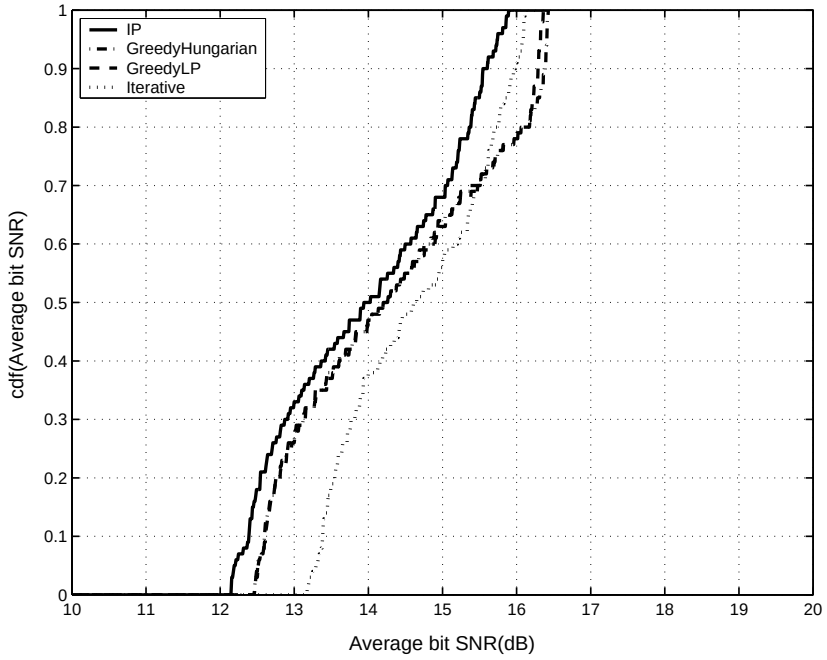


Figure 10.5: Comparison of the cumulative distribution function of the average bit SNR (w/o Power Constraint)

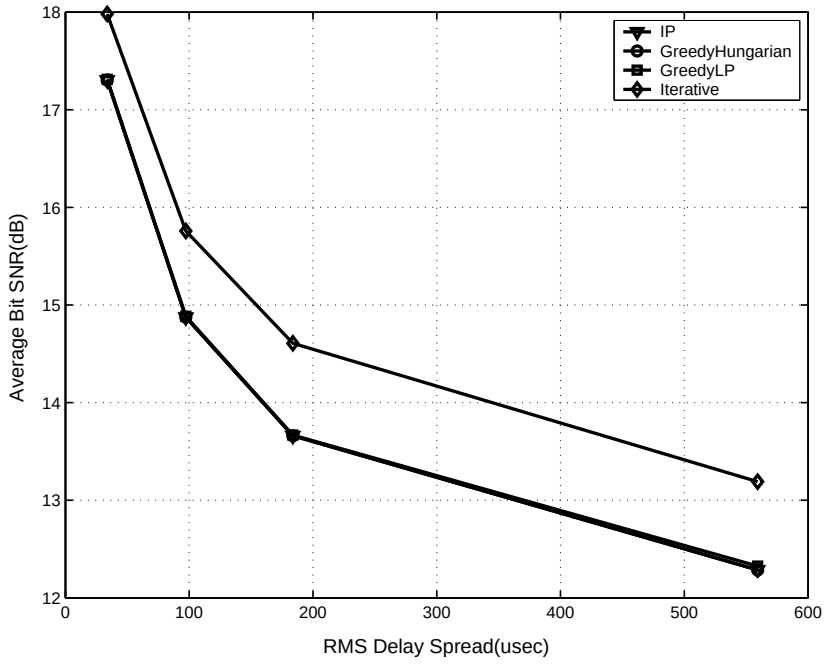


Figure 10.6: Average bit SNR versus channel fading and multiuser diversity: Average vs Delay Spread

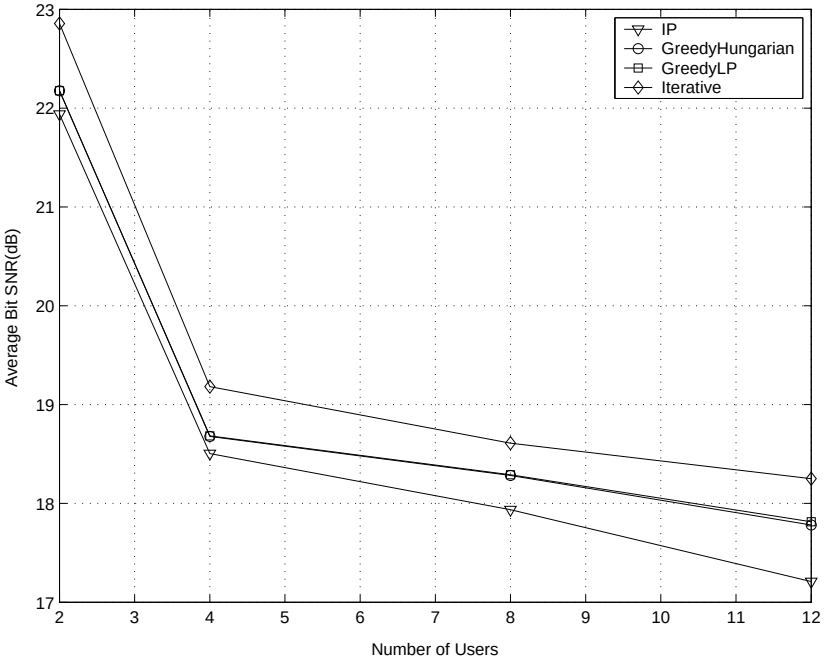


Figure 10.7: Average bit SNR versus channel fading and multiuser diversity: Average bit SNR vs Number of Users

10.7 Performance Analysis

We compare the performance of our iterative algorithm with the proposed suboptimal GreedyHungarian and GreedyLP schemes and optimal IP scheme. We adopt the M-ary quadrature amplitude modulation of 4-QAM, 16-QAM, and 64-QAM which are used to carry two, four, or six bits/subcarrier [62]. Required transmission power for c bits/subcarrier at a given BER with unity channel gain is:

$$f(c, BER) = \frac{No}{3}[Q^{-1}(\frac{BER}{4})]^2(2^c - 1) \tag{10.15}$$

where $Q^{-1}(x)$ is the inverse function of:

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2} dt.$$

We evaluate our scheme in Rayleigh fading channels [74, 75]. The power spectral density level N_o is equal to unity, and gain of Rayleigh channel $E|H_k(n)|^2$ is also equal to unity. We use 128 subcarriers with a total transmission rate between 480 bits/symbol and 768 bits/symbol. BER requirement of users is selected from the list $\wp = \{1e-2, 1e-4\}$. In the simulations, depending on the constraint, either the rate requirements are fulfilled when the transmit power is minimized or the power constraint is fulfilled when rates are maximized. BER requirement, on the other hand, is satisfied in both situations. We distinguish the settings by naming them with or without power constraint.

Figure 10.3 shows the convergence of our scheme with iterative betterment. In each iterative step, the power is reduced keeping the total number of bits constant. The steepest

decrease is observed in the HSA step since the power reduction in bit swapping is higher than the one in subcarrier swapping because of the exponential growth of the $f(x, y)$ function. It can be seen from the figure that the iterative solution approximates the GreedyLP with time.

Figures 10.4 and 10.5 present the cumulative distribution functions of the average bit SNR for the cases with and without power constraints. There are four users in two sets of BER requirement and each user has rate requirement of 120bits/symbol. It can be seen from the Figure 10.4 that the iterative approach approximates the optimal solution up to $0.9dB$ when there is no power constraint. When there is a power constraint, as seen in Figure 10.5, the iterative approach outperforms the GreedyHungarian and GreedyLP approach and is close to the IP solution within $0.3dB$. The reason why iterative solution gives better performance than the suboptimal solution is its tight power control scheme, which allows transmission of higher number of bits. GRA is very important since it decreases the variance of average bit SNR and makes the iterative approach perform better at the end by exchanging one high cost bit with more than one low cost bit i.e. lower level modulation.

Figures 10.6 and 10.7 shows the performance of the schemes in various channel fading and multiuser diversity situations [69]. Figure 10.6 presents the average bit SNR as a function of root mean square (RMS) delay spread for different resource allocation schemes. As RMS delay spread increases, the fading variation increases, so higher gains are obtained by adaptive allocation. We find that iterative approach is never more than $0.9dB$ above the IP approach. Figure 10.7 shows the average bit SNR versus the number of users where each has the same BER requirement of $1e-4$ and RMS value of

30usec. As the number of users increases, the probability of obtaining a good channel in the subcarriers increases. The iterative approach follows the lower bound within $0.9dB$ and follows the GreedyHungarian and GreedyLP schemes within $0.5dB$.

Figure 10.8 and 10.9 show the standard deviation of the bit allocation of the users for different power constraints or different number of users. Each user has a BER requirement of $1e-4$ and total transmit rate is 480bits/symbol which is equally distributed to each user. Each user has a 180usec RMS delay spread. Figure 10.8 presents the standard deviation of bit distribution among users under the total power constraint. It can be observed from the graph that iterative approach outperforms the GreedyHungarian and GreedyLP and is close to the Integer Programming (IP). The FSA distributes the bits fairly compared to the greedy approach. The fairness property is an important metric for real-time data if there is tight power control. The iterative solution maintains fairness. As the total transmit power increases, the significance of a power control scheme decreases as can be inferred from the graph. In Figure 10.9 fairness is tested under varying number of users. The iterative approach again outperforms the GreedyHungarian and GreedyLP and closely follows the Integer Programming (IP).

Figure 10.10 shows the average data rates per subcarrier versus total power constraint when there are four users. Each user has a rate requirement of 192bits/symbol (maximum rate) and BER requirement of $1e-4$. The performance of the iterative approach is close to that of the optimal and the difference between suboptimal and iterative approaches decreases as the total transmit power increases.

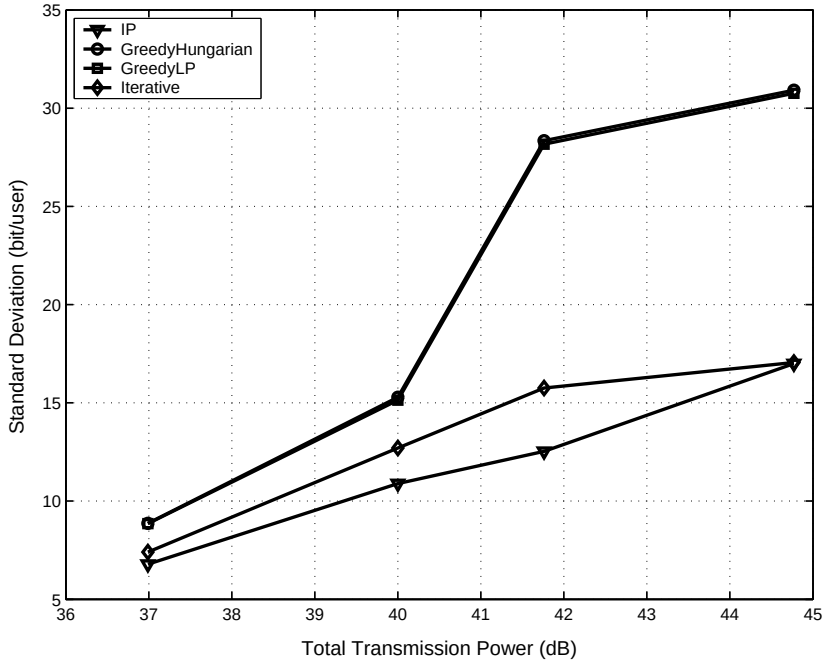


Figure 10.8: Spectral efficiency vs Total power

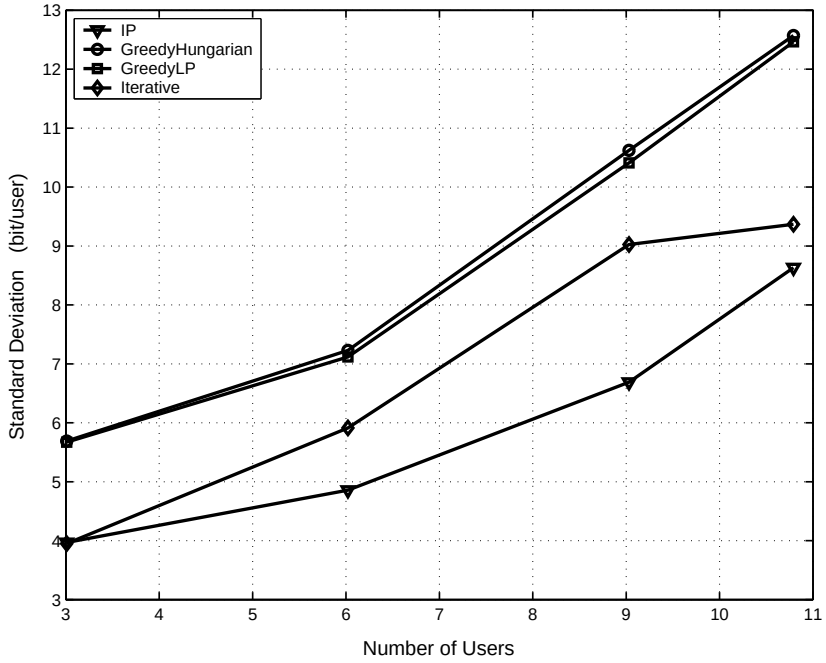


Figure 10.9: Standard deviation of Bits/User vs Number User

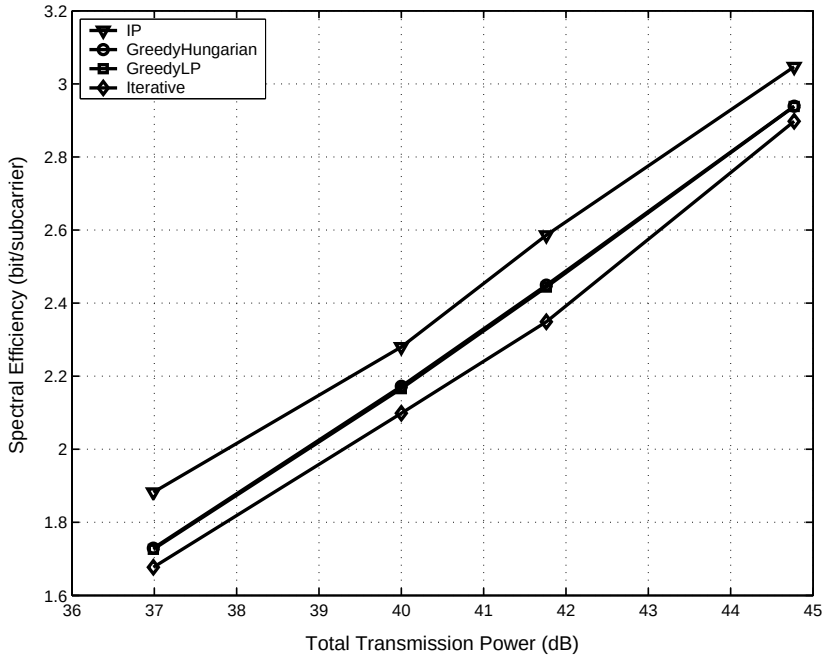


Figure 10.10: Spectral efficiency versus total transmission power

Remark

This chapter is published as a journal article titled “QoS Aware Adaptive Resource Allocation Techniques for Fair Scheduling in OFDMA Based Broadband Wireless Access Systems in IEEE Transaction on Broadcasting, vol. 49, December 2003. Authors are Mustafa Ergen, Sinem Coleri and Pravin Varaiya.

Chapter 11

Wireless LAN with Adaptive Antennas

11.1 Introduction

Wireless communication systems demand a higher system capacity to satisfy the increasing needs in voice over IP, multimedia communication. Existing bandwidth must be utilized in a much more efficient way. This can be achieved by increasing the bit error rate performance. Adaptive antenna systems creatively form a directional beam with infinite number of patterns that are adjusted in real-time. These antenna systems configure themselves with the detected signal strength and create a heightened sensitivity in particular directions which increases the signal-to-noise ratio and decrease the bit error rate [76].

Adaptive antenna technology takes advantage of its ability to locate and track various types of signals to maximize signal reception with minimized interference [77]. Figure 11.1 shows a representative of a coverage where the main lobe is extended toward a user with a null directed toward a co-channel interferer. In some cases, having a directional antenna can reduce the number of APs needed within a network but the APs require

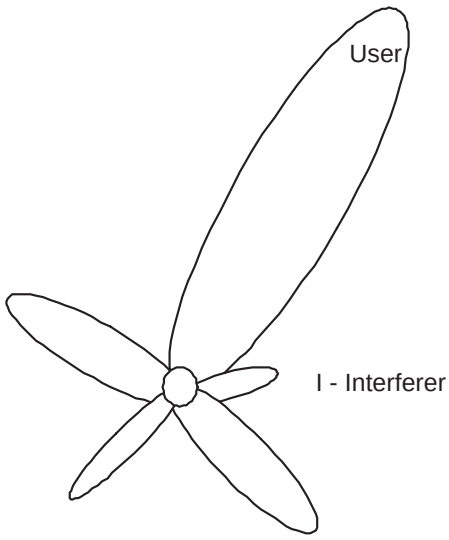


Figure 11.1: Adaptive array coverage

more power.

WLAN with directional antennas is also considered for Broadband Wireless Access systems. High-gain antennas work best for covering large distances in narrow areas or for supporting point-to-point links between buildings [2].

One of the central motivations for this chapter comes from investigating the throughput performance with the implication of the adaptive antenna system. The system throughput for a packet network with a smart antenna system may be increasing substantially with still low packet delay [78, 79]. We build a smart antenna system in OPNET simulation platform to investigate the overall network throughput in various configuration. Furthermore, the chapter presents strategies to use the adaptive antenna system in order to leverage its ability (partially stated in [80]).

11.2 Configuration of MAC for Adaptive Antenna System

There are several configurations that can be applied to the adaptive antenna system. Configurations can be for an ad-hoc network that is for station-to-station communication or for an access point network that is access point-to-station network or for broadband network which is access point-to-service provider. The network can be adjusted according to these networks by sending the packets in directional or omni-directional [81].

11.2.1 Directed transmission of Data frame

When only data frame is transmitted with directional antennas, then the coverage remains the same but the access point can reach the boundary with higher data rate. This scheme is very immune to the hidden terminal and exposed terminal problem since every station is aware of the end of a transmission by receiving an ACK frame, which is sent in omni-direction.

11.2.2 Directed transmission of ACK frame

If the ACK frame is also sent with a directional antenna then each sender-receiver pair constructs a high power tube which could result in isolated islands of connected graphs. This kind of configuration is very suitable for fixed mesh networks which suffer less from hidden and exposed terminal problems and the range can be very large. For an ad-hoc network this configuration is also a good option since non-uniform traffic distribution

could be leveraged and isolated islands can be created with less number of nodes. As a result, the throughput is expected to be higher but this kind of system can suffer from the hidden and exposed terminal problem more than the others. This is because the station can move arbitrarily and can be in the vicinity of a station whose ongoing transmission could not be heard due to the narrow beam width.

11.3 Performance

In this section we look into the performance of the system. We first explain the simulation architecture and then explain the results [4].

11.3.1 Rate Adaptation

The system has a rate adaptation mechanism which decreases its rate every time it experiences a collision. It brings the data rate back to the highest value after a successful transmission, as seen in Figure 11.2. The rate adaptation mechanism is crucial since the station adapts itself to the wireless channel condition.

11.3.2 Antenna Architecture

In an adaptive antenna system, the power strengths are determined in each antenna depending on the reception and those values are considered in the transmission. In OPNET there are four directional antennas and one isotropic control antenna. The reception is always in the isotropic antenna. The transmission of the control packets are sent via the control antenna and data packets are sent with one of the directional antennas. This is

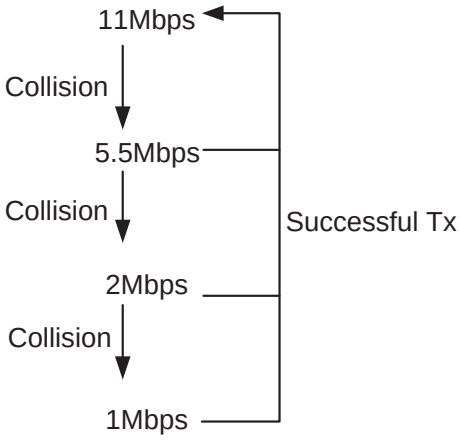


Figure 11.2: Rate adaptation mechanism

reasonable in a free space propagation environment since in free space, there is line of sight and the physical direction matches with the adaptive antenna wave direction. An illustration of the first antenna pointing to 0-90 degrees is shown in Figure 11.3 with 10dB in direction and -20dB out of direction, which corresponds to 10 mW and 0.01 mW, respectively. The isotropic antenna seen in Figure 11.4 on the other hand has 0 dB gain in all degrees which has a power level of 1mW.

We considered a multi antenna system where the beam width could change between 45^o, 90^o, 180^o, 360^o and the power level changes from 1 mW to 10 mW in the main lobe and 0.01 mW in the side lobes.

11.3.3 Node Architecture

OPNET uses source streams as the transmission pipeline. Each source stream has an id number and a packet is sent to the id number. Each transmitter and receiver has four source streams each for a data rate chosen from the set (1, 2, 5.5, 11 Mbps). The sim-

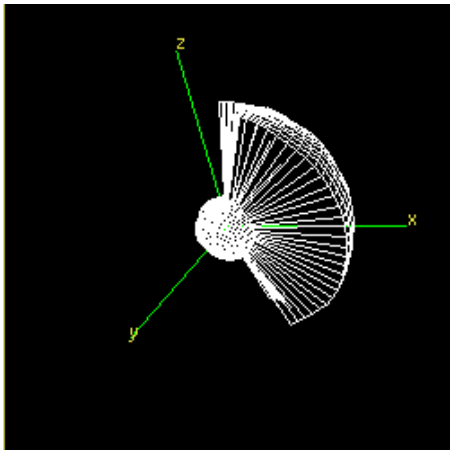


Figure 11.3: Antenna pattern when the beam width is $\pi/2$

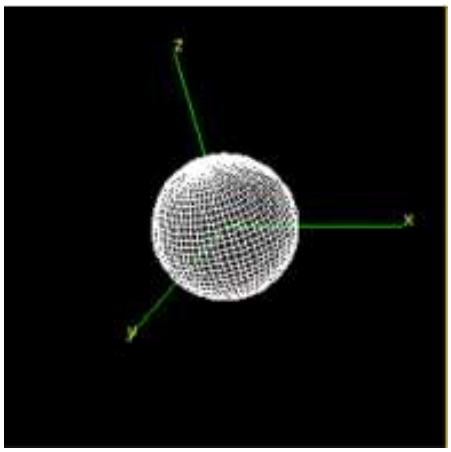


Figure 11.4: Isotropic antenna pattern

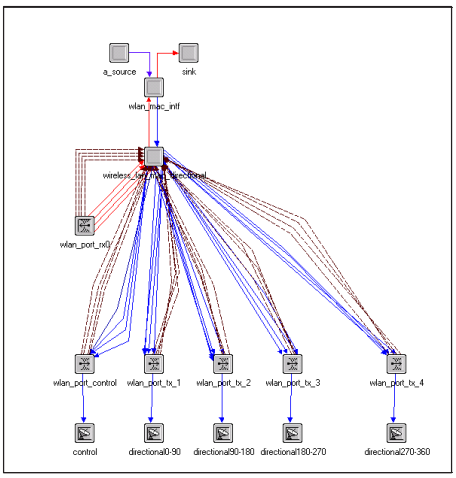


Figure 11.5: Node model for the adaptive antenna when the beam width is $\pi/2$

ulation can be configured in a way that the packets either can use the omni-directional antenna or one of the directional antennas. Position information is saved into a list and in each packet transmission, the source and destination pairs are extracted from the packet header. The antenna is determined according to the position of the transmitter and receiver. The model of a node is shown in Figure 11.5. In this model we focus only MAC performance. There is a direct source and sink attached to the MAC module which does not consider the network and TCP layer overhead.

11.3.4 Infrastructure Mode

Access-Point-Only configuration is where only the access point has an adaptive antenna and mobile stations have isotropic antennas. In this configuration, the AP reaches to the farthest point with higher data rate. Figure 11.9 shows the possible range configuration. Mobile stations operating with the lowest data rate have a higher range because a lower data rate is more tolerant to noise. Another key point is that multiple antennas consume

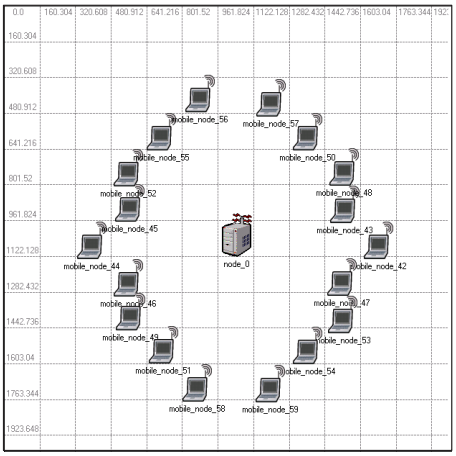


Figure 11.6: Configuration for the network where only access point has adaptive antenna

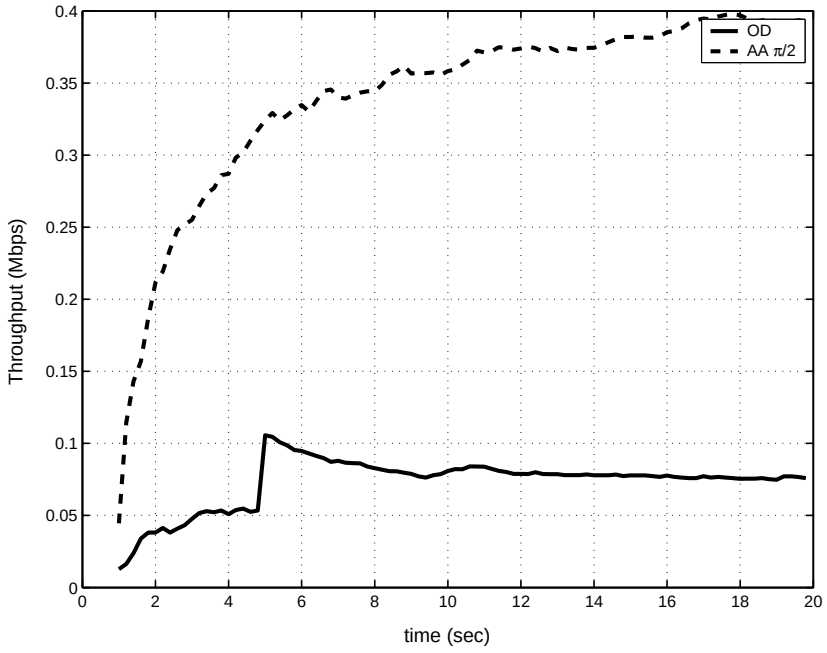


Figure 11.7: Throughput for the network where only access point has adaptive antenna

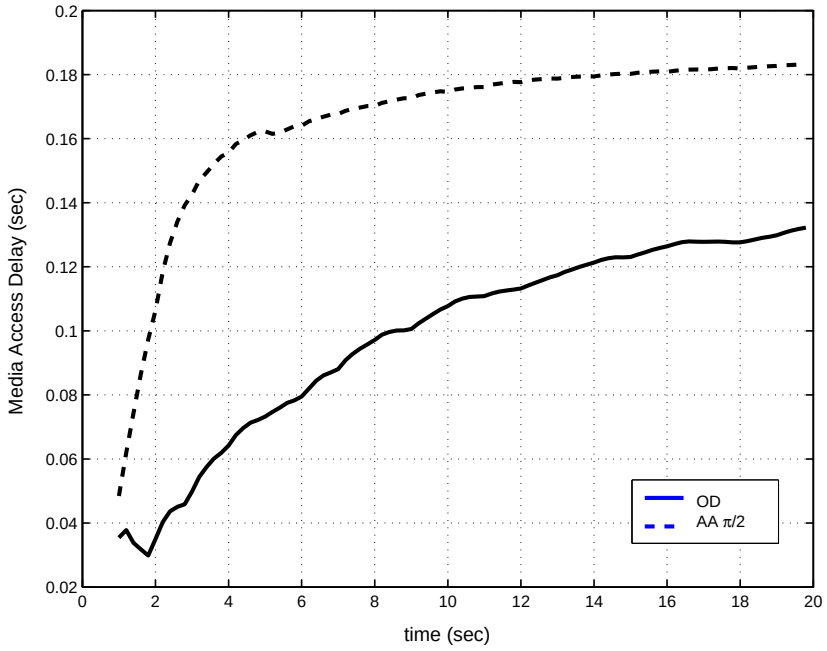


Figure 11.8: Media Access Delay for the network where only access point has adaptive antenna

more energy and it is not preferable for certain cases in mobiles. However, this is not applicable in this Access Point Only configuration since, normally, APs are wired.

Figure 11.6 shows the scenario for this mode. Mobile stations are located around the lowest data rate region and their destination is the AP. It is configured to see the performance increase with the adaptive antenna. Figure 11.7 shows the time-averaged throughput comparison with and without the adaptive antenna in the access point. As seen from the figure, an adaptive antenna impacts performance greatly. We can also see this in Figure 11.8 where time-averaged media access delay is presented. The performance increase comes from the condition that packets transmitted with higher data rate have shorter time of flight.

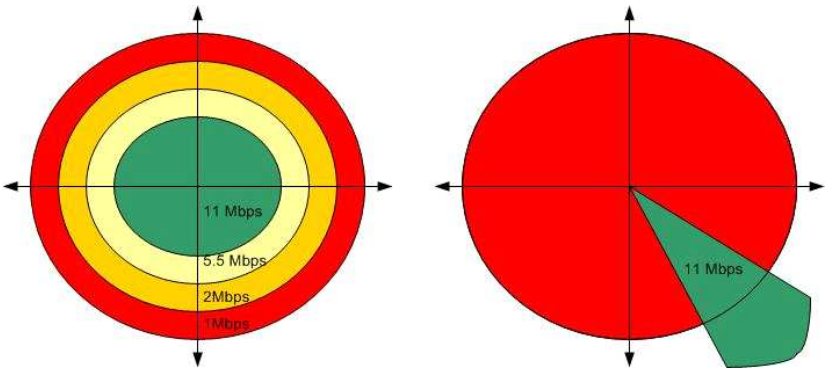


Figure 11.9: Adaptive antenna system illustration in OPNET

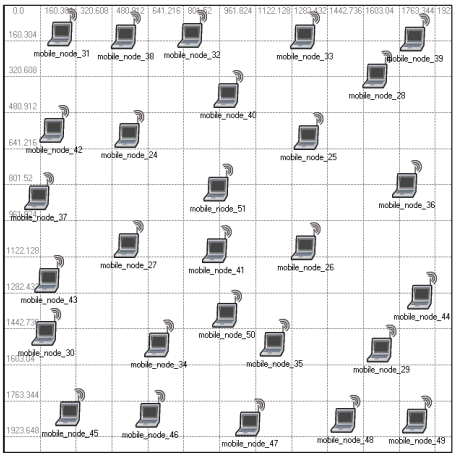


Figure 11.10: Configuration for the network where all stations have an adaptive antenna

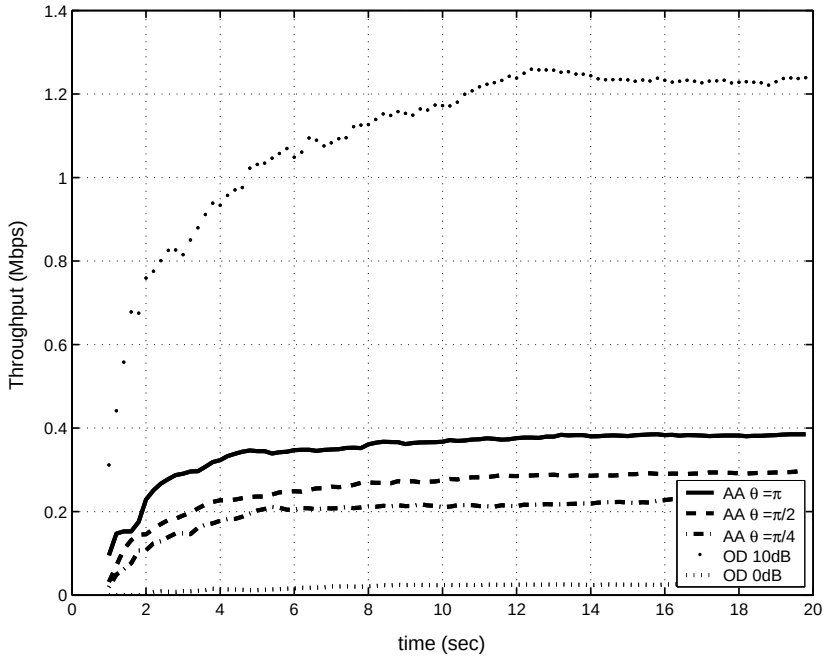


Figure 11.11: Throughput for the network where all stations have adaptive antenna

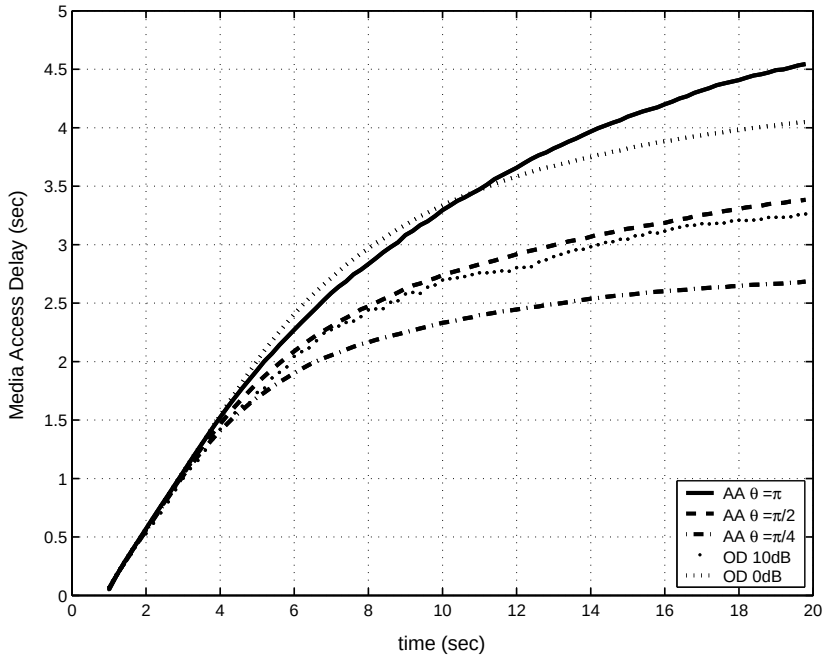


Figure 11.12: Media access delay for the network where all stations have adaptive antenna

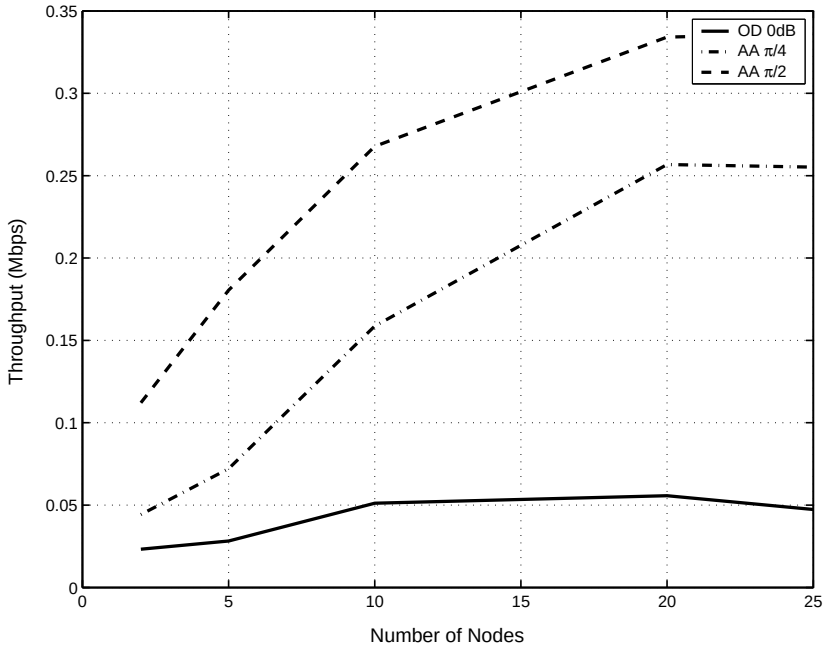


Figure 11.13: Throughput for the network where all stations have adaptive antenna (w/o RTS/CTS)

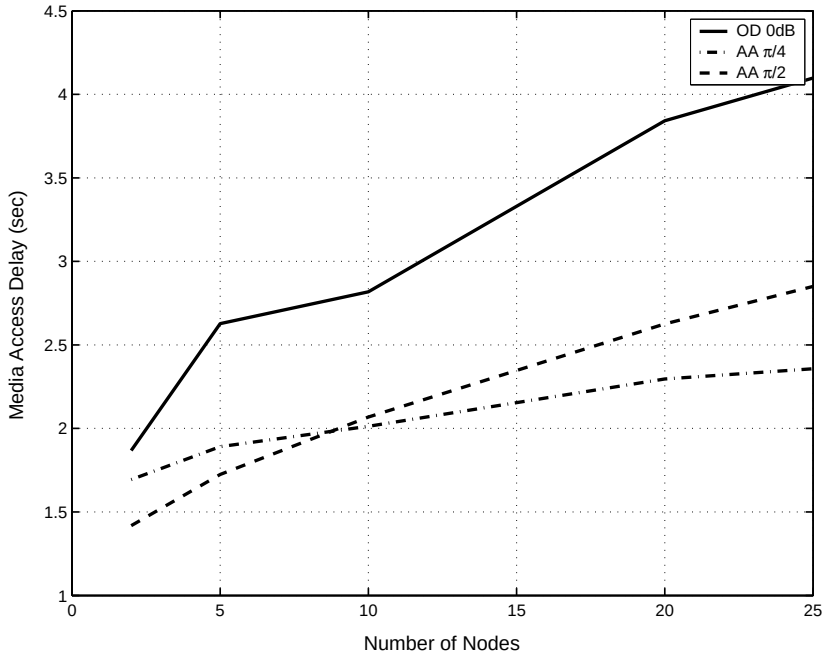


Figure 11.14: Media access delay for the network where all stations have adaptive antenna (w/o RTS/CTS)

11.3.5 Independent Mode

In independent mode, all stations have an adaptive antenna and there is no access point. Traffic destination is random for all stations and there are two cases considered: with or without RTS/CTS. Figure 11.10 shows the ad-hoc network.

Figure 11.11 shows the throughput comparison for adaptive antenna (AA) and omni-directional (OD) cases. As shown, when the RTS/CTS is applied in direction, OD 10 dB which is 10 mW, all directions give the highest throughput. Performances of 180^o, 90^o, and 45^o lie within a close vicinity and are substantially higher than OD 0dB. We can infer from the graph that with OD 10 dB, the hidden and exposed terminal problems are reduced as much as possible. However, the degradation due to these problems becomes increasingly active as the beam width decreases. Figure 11.12 also brings an interesting case where the AA 45^o case has the lowest media access delay.

It is interesting to investigate the throughput performance when all packets are sent in directional including control packets which could create isolated networks within the connected network as well as could create more hidden terminals and exposed terminals. When we increment the active stations in the network step by step we obtain the Figure 11.13 for throughput. As can be inferred from the graph in Figure 11.14, the throughput increases with the beam width and media access delay decreases with the beam width. We can conclude that as the beam width increases, there is less chance to transmit since there are more stations with suspended transmission which introduces delay.

Chapter 12

Conclusion

Next-generation wireless networks should be open to control and re-configuration, which requires estimation of the performance a priori. Next-generation networks should also support multiple access to provide ubiquitous sessions.

In this dissertation, we have presented an evaluation of existing problems in IEEE 802.11 systems that prevent scalability and intelligent network management required for future wireless networks. We also introduced a multiple access mechanism for OFDMA.

Our introductory chapter gave an extensive overview of IEEE 802.11 systems, with an emphasis on medium access layer and physical layer. We explained Distributed Coordination Function and OFDM PHY layer in detail.

Then we proposed a new Markov model for the distributed coordination function (DCF) of IEEE 802.11. The model incorporates consecutive transmission, non-saturated traffic and signal-to-noise ratio for both basic and RTS/CTS access mechanisms. We compared our model with the existing models [5] and found that our model gives better performance. We also investigated the throughput characteristic under constant load, in-

creasing load, and increasing number of stations. The throughput increases as the number of stations increases until the network gets congested, after which the throughput starts to decrease.

A novel throughput formulation was introduced for the case where the stations in the network operate with mixed data rates. The formulation changes the Markov model and duration values considerably. Performance analysis shows that lower data rate station degrades the performance of the system but has the same individual throughput with higher data rate stations. We then used our model to find the delay characteristics of the time interval between two successful transmissions from one station. We first calculated the average delay and then the probability mass function in order to determine the characteristics of the delay as a function of the number of stations. Analysis of the model showed that the throughput first increases and subsequently decreases with the number of active stations, suggesting the need for an admission control mechanism. We introduced such a mechanism, which tries to maximize the throughput while maintaining a fair allocation. The maximum achievable throughput is tracked by the mechanism as the number of active stations increases. An extensive performance analysis showed that the mechanism provides significant improvements.

This admission control mechanism provides a controllable access to DCF, which is random access. If a single station is allowed to be in the access list then the admission control scheduler always selects one station at a time, converting DCF to PCF, which is centralized access. As a result, the admission control mechanism brings a tuning mechanism that can shift the performance from DCF to PCF.

Wireless LANs are being widely deployed today. Managing large-scale wireless

LANs is a difficult and time-consuming process. The individual throughput of stations within the coverage area of an access point depends heavily on interference from other access points and stations operating on the same channel. We focused on a network deployed indoors typically bordered on a building; radio propagation, therefore, is influenced by the structure of the building. We then extended our analysis to determine and optimize the performance of an extended service set in an IEEE 802.11 network in which there is more than one access point. The method leverages indoor radio propagation models and RF prediction. We used our Markov model to support non-saturated traffic and different data rates. Indoor radio propagation simulation was used to determine the coverage area of each access point and signal-to-noise ratio of each station. The signal-to-noise ratio was used to set the data rate of the station. The number of associated stations to an access point was determined by the number of stations whose location fell within the coverage of the access point.

We also introduced a real-time network management model. This model predicts the throughput of a station and adjusts dynamically the power levels of access points to satisfy quality of service.

Based on our observation of a network with mixed data rate connections, we concluded that the CSMA/CA scheme does not provide a fair scheduling in terms of giving equal amount of channel usage to each station, but it does provide an equal chance of a channel access. We analyzed the case where there is a low-rate station in the medium and concluded that throughput could be increased by turning that station off. A fair allocation, however, should give equal amounts of channel usage per station. In this regard, we introduced a packet-size adaptation with respect to data rate. As a result, the slow station

increases its packet size to occupy the channel for as much time as the fast station, which does not make the fast station suffer.

Wireless LANs are being utilized for voice communication in terms of their widely deployed condition and cheap, high-speed communication. A voice communication requires guaranteed bandwidth. We introduced the monitoring scheme of a network that can distinguish the lowest throughput level and take action to increase it with a network management tool. Another modification can be introduced in the access point from which all voice packets are distributed. The access point can aggregate the voice packets and send the packets in a multicast packet. Each station can identify its packet from the payload by its identifier, which is acquired through subscription to a multicast group. The results show that the throughput increases significantly. We also observed that as the packet size increases, so does the throughput of a system. We conclude that if stations aggregate the packets and send them in a big frame, the overall throughput increases.

The next-generation stations will use multiple access, which could be modified depending on the network to which they subscribe. They can connect to a WLAN or to a broadband wireless access (BWA) network, or WLAN and BWA can bring an access together. OFDM is considered as a PHY layer for the last mile access. A configurable radio can support WLAN and BWA PHY layer together. A system based on orthogonal frequency division multiple access (OFDMA) has been developed to deliver mobile broadband data service at data rates comparable to those of wired services, such as DSL and cable modems. We considered the problem of resource allocation for adaptive modulation in OFDMA systems. The problem was considered in two different approaches: one maximizes the capacity, and the other satisfies fixed QoS criteria (i.e. the rate and bit error

rate requirements) in each symbol. Recent work has focused on developing algorithms to meet the QoS criteria [57, 62, 63, 64, 65, 66, 67].

In an OFDMA system, subcarriers are distributed among users and the number of bits transmitted in each subcarrier is adjusted according to the rate requirements of users to minimize total transmit power. It has been shown that resource allocation can be optimized by Integer Programming [63]. However, the optimal solution can not be implemented in real-time. We proposed a simple suboptimal solution that fairly allocates and efficiently converges close to optimal meeting the QoS criteria per symbol. The algorithm achieves good performance in terms of tight power control, iterative betterment and fair scheduling among users when compared with the optimal solution and previously proposed suboptimal schemes. The proposed solution can also be applied to the uplink when there is perfect synchronization. We also considered a possible resource allocation scheme when the objective is to maximize capacity, based on proportional fair scheduling algorithm for point-to-point communication introduced in [69].

Instead of shaping the directional antenna pattern with the metallic properties and physical design of a single element (like a sectorized antenna), adaptive beam systems combine the outputs of multiple antennas in such a way as to form finely sectorized (directional) beams with more spatial selectivity than can be achieved with conventional, single-element approaches. We investigated qualitatively the performance improvement in WLAN when smart antenna systems are used for communication. The simulations studied the effect of the beam width and various combinations of adaptive antenna usage with control and data packets. This configuration could be implemented within an AP subnet or AP-to-AP or AP-to-service provider to satisfy the connection for the last-mile

access. We looked at the performance of two different networks: access point only, in which only the access point has an adaptive antenna system, and ad-hoc configuration, in which all stations have the adaptive antenna system. We note that the performance of the system is affected by some conditions. For example, directed high power increases the signal-to-noise ratio; on the other hand, hidden terminal and exposed terminal problems become more effective, which causes interference and produces a negative effect on the signal-to-noise ratio. Therefore a balance must be reached with these dynamics.

Overall, we coordinate possible components of a wireless network that make the system more intelligent and can be implemented in next-generation wireless networks.

Bibliography

[1] M. Ergen, S. Coleri, P. Varaiya. *QoS Aware Adaptive Resource Allocation Techniques for Fair Scheduling in OFDMA Based Broadband Wireless Access Systems.* IEEE Transaction on Broadcasting, vol.49, December 2003.

[2] *Understanding Wi-Fi and WiMAX as Metro-Access Solutions.* Intel White Paper. Online: `http://www.intel.com`

[3] M. Ergen, A. Puri. *MEWLANA-Mobile IP Enriched Wireless Local Area Network Architecture.* Proceedings of IEEE VTC, vol.4, Fall 2002.

[4] OPNET. *Optimum Network Performance Simulation Tool for Communication Networks.* Online: `http://www.opnet.com`

[5] G. Bianchi. *Performance Analysis of the IEEE 802.11 Distributed Coordination Function.* IEEE Journal on Selected Areas in Communications, vol.18, March 2000.

[6] A. Bahai, B. Saltzberg, M. Ergen. *Multi-Carrier Digital Communications: Theory and Applications of OFDM.* Springer-Verlag, New York, September 2004.

[7] B. O'Hara, A. Petrick. *The IEEE 802.11 Handbook: A Designer's Companion.* IEEE Press, 1999.

[8] IEEE 802.11. *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications.* IEEE 802.11 Standard, August 1999.

[9] IEEE 802.11b. *Part 11: Wireless LAN, Medium Access Control (MAC) and Physical Layer (PHY) Specifications: High-Speed Physical Layer Extension in the 2.4GHz Band.* Supplement to IEEE 802.11 Standard, September 1999.

[10] IEEE 802.11a. *Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: High-Speed Physical Layer Extension in the 2.4GHz Band.* Supplement to IEEE 802.11 Standard, September 1999.

[11] J. Geier. *802.11 Alphabet Soup.* Wi-Fi Planet, August, 2002. Online: `http://www.wi-fiplanet.com/.`

[12] H. J. Yeh, J. C. Chen, C. C. Lee. *WLAN standards.* IEEE Potentials, Oct/Nov 2003.

[13] IEEE 802.11e. *Medium Access Control (MAC) Enhancements for Quality of Service (QoS).* IEEE Standard 802.11e/D4.4, June 2003.

[14] L. Bisdounis. *HiPERLAN/2 vs IEEE 802.11a: Physical Layer.* Intracom. Online: `http://easy.intranet.gr/paper_10.pdf.`

[15] M. Ergen. *IEEE 802.11 Tutorial.* UC Berkeley, June 2002. Online:`http://www.eecs.berkeley.edu/~ergen/docs/ieee.pdf/.`

[16] C. Perkins, *Ad Hoc Networking.* Addision-Wesley Professional, 2001.

[17] H. Li, D. Yu. *Comparison of ad hoc and centralized multihop routing.* Proceedings of The 5th International Symposium on Wireless Personal Multimedia Communications, vol.2, October 2002.

[18] N. Moghim, F. Hendessi, N. Movehhedinia. *An improvement on ad-hoc wireless network routing based on AODV.* Proceedings of ICCS 2002. vol. 2, November 2002.

[19] E. M. Royer, C. E. Perkins. *An implementation study of the AODV routing protocol.* Proceedings of IEEE WCNC, vol.3, September 2000.

[20] K. Paul, D. Westhoff. *Context aware detection of selfish nodes in DSR based ad-hoc networks.* Proceedings of IEEE GLOBECOM, vol.1, November 2002.

[21] C. Perkins. *Mobile IP: Design Principles and Practices.* Prentice Hall, 1998.

[22] M. Ergen, S. Coleri, B. Dundar, A. Puri, J. Walrand, P. Varaiya. *Position Leverage Smooth Handover Algorithm For Mobile IP.* Proceedings of IEEE ICN, August, 2002.

[23] T. M. Cover, J. A. Thomas. *Elements of Information Theory.* Wiley-Interscience, August 1991.

[24] A. J. Viterbi. *CDMA, Principles of Spread Spectrum Communication.* Reading, MA: Addison Wesley, 1995.

[25] E. Bogenfeld, et al. *Influence of Nonlinear HPA on Trellis-Coded OFDM for Terrestrial Broadcasting of Digital.* Proceedings of IEEE GLOBECOM, 1993.

[26] B. Come, R. Ness, L. V. Perre, W. Eberle, P. Wambacq, M. Engels, I. Bolsens. *Impact of front-end non-idealities on Bit Error Rate performances of WLAN-OFDM transceivers.* IMEC. Online: `http://www.imec.be/.`

[27] IMEC. *Interuniversity MicroElectronics Center.* Online: `http://www.imec.be/.`

[28] J. Tubbax, B. Come, L. V. Perre, L. Deneire, S. Donnay, M. Engels. *Compensation of IQ imbalance in OFDM systems.* White Paper IMEC, 2003.

[29] OFDM White Paper. *OFDM (Orthogonal Frequency Division Multiplexing).* Nova Engineering. Online: `http://www.nova-eng.com/.`

[30] R. Tesi, M. Codreanu, I. Oppermann. *Interference Effects of UWB Transmission in OFDM Communication Systems.* Centre for Wireless Communications, University of Oulu, Finland.

[31] J. Bellorado, S. S. Ghassemzadeh, L. J. Greenstein, T. Sveinsson, V. Tarokh. *Coexistence of UltraWideband Systems with IEEE-802.11a Wireless LANs.* Proceedings of IEEE GLOBECOM, 2003.

[32] D. Qiao, S. Choi, K. G. Shin. *Goodput Analysis and Link Adaptation for IEEE 802.11a Wireless LANs.* IEEE Transactions on Mobile Computing, vol.1, October-December 2002.

[33] Y. Xiao. *A Simple and Effective Priority Scheme for IEEE 802.11.* IEEE Communication Letters, vol. 7, February 2003.

[34] H. Wu, Y. Peng, K. Long, S. Cheng, J. Ma. *Performance of Reliable Transport Protocol over IEEE 802.11 Wireless LAN: Analysis and Enhancement.* Proceedings of IEEE INFOCOM, 2002.

[35] G. R. Cantieni, Q. Ni, C. Barakat, T. Turletti. *Performance Analysis under Finite Load and Improvements for Multirate 802.11.* preprint submitted to Elsevier Science.

[36] C. Wang, B. Li, B. Li, K. Sohraby. *An effective collision resolution mechanism for wireless LAN.* Proceedings of ICCNMC, October 2003.

[37] A. C. Boucouvalas, V. Vitsas. *Influence of channel BER on IEEE 802.11 DCF Chatzimisios.* Electronics Letters, November 2003.

[38] Y. Peng, H. Wu, S. Cheng, K. Long. *A new self-adapt DCF algorithm.* Proceedings of IEEE GLOBECOM, vol.1, November 2002.

[39] M. Ergen, D. Lee, R. Sengupta, P. Varaiya. *WTRP- Wireless Token Ring Protocol.* IEEE Transaction on Vehicular Technology, November 2004.

[40] Z. Kong, D. H. K. Tsang, B. Bensau, D. Gao. *Performance Analysis of IEEE 802.11e Contention-Based Channel Access.* IEEE Journal on Selected Areas in Communications, vol.22, December 2004.

[41] M. Ergen, P. Varaiya. *Individual Throughput with Different Data Rates from Markov Model of IEEE 802.11.* Proceedings of IEEE GLOBECOM-CAMAD, December 2004.

[42] M. Heusse, F. Rousseau, G. Berger-Sabbatel, A. Duda. *Performance Anomaly of 802.11b.* Proceedings of IEEE INFOCOM 2003.

[43] Z. Hadzi-Velkov, B. Spasenovski. *Saturation throughput - delay analysis of IEEE 802.11 DCF in fading channel.* Proceedings of IEEE ICC, vol.1, May 2003.

[44] G. Wang, Y. Shu, L. Zhang, O. W. W. Yang. *Delay analysis of the IEEE 802.11 DCF.* Proceedings of IEEE PIMRC, September 2003.

[45] P. Chatzimisios, V. Vitsas, A. C. Boucouvalas. *Throughput and delay analysis of IEEE 802.11 protocol.* Proceedings of IEEE 5th International Workshop on Networked Appliances, October 2002.

[46] M. Ergen, P. Varaiya. *Admission Control and Throughput Analysis in IEEE 802.11.* appear on ACM-Kluwer MONET Special Issue on WLAN Optimization.

[47] M. L. Lahti. *IEEE 802.11 Wireless LAN.* Helsinki University of Technology, May 2000.

[48] J. Liesenborgs. Voice over IP in networked virtual elements. Ph.D dissertation, University of Maastricht, 2000.

[49] S. Garp, M. Kappes. *Can I add a VoIP call?* Proceedings of IEEE ICC 2003, May 2003.

[50] D. Awoniyi, F. A. Tobagi. *Effect of fading on the performance of VoIP in IEEE 802.11a WLANs.* Proceedings of IEEE ICC, June 2004.

[51] W. Wang, S. C. Liew, and V. O. K. Li. *Solutions to Performance Problems in VoIP over 802.11 Wireless LAN.* to appear in IEEE Trans. On Vehicular Technology. Online:`http://www.ie.cuhk.edu.hk/fileadmin/staff_upload/soung/.`

[52] P. C. Ng, S. C. Liew, and W. Wang. *Voice over WLAN via Wireless Distribution System.* Proceedings of IEEE VTC, September 2004.

[53] H. Tounsi, L. Toutain, F.Kamaoun. *Small packets aggregation in an IP domain.* Proceedings of Sixth IEEE Symposium on Computers and Commuincations, July 2001.

[54] W. Wang, S. C. Liew, Q. X. Pang, and V. O.K. Li. *A Multiplex-Multicast Scheme that Improves System Capacity of Voice-over-IP on Wireless LAN by 100%.* Proceedings of The Ninth IEEE Symposium on Computers and Communications, June 2004.

[55] W. Wang, S. C. Liew, and J. Y. B. Lee. *ABRC: An End-to-End Rate Adaptation Scheme for Multimedia Streaming over Wireless LAN.* Proceedings of IEEE WCNC, March 2004.

[56] I. Koffman, V. Roman. *Broadband Wireless Access Solutions Based on OFDM Access in IEEE 802.16.* IEEE Communications Magazine, April 2002.

[57] I. Koutsopoulos, L. Tassiulas. *Channel state-Adaptive techniques for Throughput Enhancement in Wireless Broadband Networks.* Proceedings of IEEE INFOCOM. vol.2, 2001.

[58] M. S. Corson, R. Laroia, A. O'Neill, V. Park, G. Tsirtsis. *A New Paradigm for IP-Based Cellular Networks.* Proceedings of IEEE INFOCOM, vol.3, 2001.

[59] IEEE 802.16. *IEEE Std. 802.16-2001 IEEE Standard for Local and Metropolitan area networks Part 16: Air Interface.* IEEE 802.16-2001 Standard, 2002.

[60] S. Coleri, M. Ergen, A. Puri, A. Bahai. *Channel Estimation Techniques Based on Pilot Arrangement in OFDM Systems.* IEEE Transactions on Broadcasting, vol.48, September 2002.

[61] R. Nogueroles, M. Bossert, A. Donder, and V. Zyablov. *Performance of a Random OFDMA System for Mobile Communications.* International Zurich Seminar on Accessing, Transmission, Networking. Proceedings, 1998.

[62] C. Y. Wong, R. S. Cheng, K. B. Letaief, R. D. Murch. *Multiuser OFDM with Adaptive Subcarrier, Bit, and Power Allocation.* IEEE Journal on Selected Areas in Communications, vol.17, October 1999.

[63] I. Kim, H. L. Lee, B. Kim, Y. H. Lee. *On the Use of Linear Programming for Dynamic Subchannel and Bit Allocation in Multiuser OFDM.* Proceedings of IEEE GLOBECOM, vol.6, 2001.

[64] S. Pietrzyk, G. J. M. Janssen. *Multiuser Subcarrier Allocation for QoS Provision in the OFDMA Systems.* Proceedings of IEEE VTC, vol.2, Fall, 2002.

[65] Y. Zhang, K. B. Letaief. *Multiuser Subcarrier and Bit Allocation along with Adaptive Cell Selection for OFDM Transmission.* Proceedings of IEEE ICC, vol.2, 2002.

[66] C. Y. Wong, C. Y. Tsui, R. S. Cheng, K. B. Letaief. *A Real-time Sub-carrier Allocation Scheme for Multiple Access Downlink OFDM Transmission.* Proceedings of IEEE VTC, vol.2, 1999.

[67] W.Rhee, J. M. Cioffi. *Increase in Capacity of Multiuser OFDM System Using Dynamic Subchannel Allocation.* Proceedings of IEEE VTC, vol.2, 2000.

[68] C. Ibars, Y. Bar-Ness. *Comparing the Performance of Coded Multiuser OFDM and Coded MC-CDMA over Fading Channels.* Proceedings of IEEE GLOBECOM. vol.2, 2001.

[69] P. Viswanath, D. N. C. Tse, R. Laroia. *Opportunistic Beamforming Using Dumb Antennas.* IEEE Transaction on Information Theory, vol.48, June 2002.

[70] R. Knopp, P. Humblet. *Information capacity and power control in single cell multiuser communications.* Proceedings of IEEE ICC, vol.1, 1995.

[71] S. Barbarossa, M. Pompili, G. B. Giannakis. *Time and Frequency Synchronization of Orthogonal Frequency Division Multiple Access Systems.* Proceedings of IEEE ICC 2001, vol.6, 2001.

[72] D. Kivanc, H. Liu. *Subcarrier Allocation and Power Control for OFDMA.* Conference Record of the Thirty-Fourth Asilomar Conference on Signals, Systems and Computers, vol.1, 2000.

[73] H. W. Khun. *The Hungarian Method for the Assignment Problem.* Naval Research Logistics Quarterly 2, pp. 83-97, 1955.

[74] *Digital video broadcasting (DVB): Framing, channel coding and modulation for digital terrestrial television.* Draft ETSI EN300 744 V1.3.1 (2000-08).

[75] A. Bohdanowicz, G. J. M. Janssen, S. Pietrzyk. *Wideband Indoor and Outdoor Multipath Channel Measurements at 17 GHz.* Proceedings of IEEE VTS, vol.4, 1999.

[76] P. H. Lehne, M. Pettersen *An overview of smart antenna technology for mobile communications systems.* IEEE Communications Surveys, Fourth Quarter 1999, vol.2 no.4.

[77] *Smart Antenna Systems.* Internaitonal Engineering Consortium.
Online: `http://www.iec.org/online/tutorials/smart_ant/.`

[78] A. Nasipuri, S. Ye, J. You, R.E. Hiromoto. *A MAC protocol for mobile ad hoc networks using directional antennas.* Proceedings of IEEE WCNC, vol.3, September 2000.

[79] T. Ren, I. Koutsopoulos, L. Tassiulas. *Efficient media access protocols for wireless LANs with smart antennas.* Proceedings of IEEE WCNC, vol.2, March 2003.

[80] O. Gurbuz. *Wireless LANs with Smart Antennas.*
Online: `http://www.telecom.sabanciuniv.edu/.`

[81] C. Sakr, T. D. Todd. *Carrier-sense protocols for packet-switched smart antenna base stations.* Proceedings of International Conference on Network Protocols, October 1997.