Performance Evaluation and Enhancement of Multiple Access Control in IEEE 802.16 Networks

By

Ahmed Doha

A thesis submitted to the
Department of Electrical and Computer Engineering
in conformance with the requirement for
the degree of Master of Science (Engineering)

Queen's University
Kingston, Ontario Canada
April 2005

Your file Votre référence
ISBN: 0-494-01031-2
Our file Notre référence
ISBN: 0-494-01031-2

Abstract

The IEEE 802.16 [1] specifies the air interface of broadband fixed wireless access networks including the multiple access control and physical layers. Owing to its ample QoS provisions, modular bandwidth support, wide radio frequency coverage, robust channel in line of sight and non-line of sight environments, and last but not least its economical deployment compared to underground wire-line technologies, the IEEE 802.16 standard is gaining broad acceptance and consideration for the expanding broadband access markets. The undergoing amendment to the existing IEEE 802.16-2004 standard, supporting broadband mobile access, broadens the usability in different fixed wireless and mobile network scenarios. Therefore, the flexibility and revenue projections of IEEE 802.16 networks suggest increased interest in this standard.

The goal of the thesis is to evaluate and enhance the performance of the multiple access control protocol of the IEEE 802.16 standard.

First, using mathematical modeling, I analyze the contention delay of the reservation requests in the reservation multiple access protocol of the standard in light of the available contention slots in a frame. I obtain insightful observations on the impact of the

various system design parameters on the contention delay among which is the size of the reservation period.

Furthermore, I develop an analytical model to compute the contention delay, data transmission delay and throughput resulting from different reservation period allocations under dynamic traffic conditions. To this end, I instigate different arguments to understand the protocol performance and potentials. I illustrate the performance compromises and remedies with respect to the reservation period allocation.

Based on the analytical model and performance evaluation results, I present a framework of a dynamic reservation period controller to enhance the performance of the protocol. Furthermore, I propose a Markovian optimization model that achieves opportunistic performance improvement, on a per frame basis, over the best case static reservation period. Through simulation, I study the merits of the proposed optimized controller with respect to the framework. I show by illustrative examples and numerical results that the controller successfully fulfills the framework objectives.

Acknowledgements

To my mother,

أماه يا مرج القلب انت مدرستي — علمتني فى البيت كيـف اكون إنسانا

أن أملأ الدنيـا خيـرا منهجـك — أن أستزيد لدروب العيـش إيمـانا

أماه يا سر أمجـادي يا معلمتي — زنة السماء عطـائك أقر عرفـانـا

ربيت طفلك على الهدى فأمسى — يخاصمه النـوم من هـول ما صـار

فليس مالا ولا استكبارا ينقصني — إنما نهضـة الأمـم بعد الزمان انتظارا

والجنة أبواب تـفـاوتت منـازلها — بذل النفيس منسك أعظم الناس مقدارا

والبـذل لـنـآلئ عـقـدك أمـي — فضلا و تسديدا حـانت علي كفـاك

سيـف سـلـلتـه حبـا على كره — فوالله لأصنـعن مـا يروق عينـاك

أضـرب الأرض وأستحـي — ألا أعـود بنصـر إلاك

واليـوم علـم أتـممـت إجازته — والنصـر آت بإذن الله لرضـاك

To my late father,

رحـمـة الله تـرعـاك أبـي — تمنيـت علي اتخـاذ العـلـم سبيـلا

هذا اليـوم مـن فـضـل ربـي — وفضـلك أبـدا حتى هممـت رحيـلا

To my sisters,

أشكـــر أخـــواتـــي داليـــا، دينـــا، شيمـــاء و رؤى على حبـــهـــم وتأيـــيـــدهـــم. أفتقـــد بشـــده جلســـاتنـــا و لعبنـــا.

I am indebted to my intimate friends in Egypt Hossam Mahmoud and Ayman Soliman for their continuous support and lasting friendship.

I am grateful to my thesis advisor, Dr. Hossam Hassanein, for the opportunity to be a member of the Telecommunications Research Lab at Queen's University. His motivation, counseling, and great work ideas have contributed to my academic and professional development. Nevertheless, the blend of his extraordinary personal character, seriousness, and witticism inspired my energies for creativity and made the work more enjoyable.

I am also grateful to Dr. Glen Takahara, my instructor and co-author of a conference paper. His unique teaching style inspired me to broadly utilize stochastic analysis tools in my research. Moreover, his comments and encouragement helped me improve the structure and rigorousness of my technical arguments.

I would like to thank Dr. Saeed Gazor for his valuable support and sincere advices. In addition, it is the generous offer of Dr. Gazor to use his laboratory's computing facilities that expedited the computing work in my thesis.

I would also like to thank Dr. Ahmed Safwat, my instructor and friend, for his continuous support and sincere advices.

I would also like to thank my friend Abd-Elhamid Taha from the TRL lab for putting in his time and resources whenever I needed him.

I enjoyed the company and love of all my TRL lab friends and colleagues especially Quanhong Wang, Kenan Xu, Khaled Ali, Ayman Radwan, Abduladhim Ashitawi, Hassan Ali, Nidal Nasser, and Abdulrahman Hijazi.

I would also like to thank Rashad Sharaf from the Royal Military College for his idea that simplified the simulation design.

I would also like to thank Communications and Information Technology Ontario (CITO) and Queen's University for providing financial support for my research.

Contents

List of Figures

List of Acronyms

BE	Best Effort
BS	Base Station
BW	Bandwidth
BWR	Bandwidth Request
CS	Contention Slot
C-TDMA	Contention TDMA
DL-MAP	Down-Link MAP message
DS	Data Slot
DSL	Digital Subscriber Line
FDD	Frequency Division Duplexing
HC	Head end Controller
HFC	Hybrid Coaxial Fiber
IEEE	Institute of Electrical and Electronics Engineers
MAC	Multiple Access Control
MAN	Metropolitan Area Network

Mbps	Mega bits per second
nrtPS	Non-real time Polling Service
OFDMA	Orthogonal Frequency Division Multiple Access
PHY	Physical layer
PRMA	Packet Reservation Multiple Access
PS	Physical Slot
QoS	Quality of Service
R-Aloha	Reservation-Aloha
R-MAC	Reservation multiple Access Control
rtPS	Real time Polling Service
S-Aloha	Slotted-Aloha
SS	Subscriber Station
TDMA	Time Division Multiple Access
TDD	Time Division Duplexing
UGS	Unsolicited Grant Service
UL-MAP	Up-Link MAP message
VoIP	Voice over Internet Protocol
Wi-Fi	Commercial brand name for the IEEE 802.11 networks
Wi-MAX	Commercial brand name for the IEEE 802.16 networks
WLAN	Wireless Local Area Network
WMAN	Wireless Metropolitan Area Network

List of Symbols

p	Probability of persistent transmission
P_{suc}	Probability of successful contention in a contention slot
P_{Access}	Probability of successful contention during the reservation period
D_{BWR}	Contention delay of bandwidth request
f	Identifier of discrete time points starting at the beginning of uplink frame
T_f	Frame time duration
T_{CS}	Time duration of a contention slot
T_{DS}	Time duration of a data slot
τ	Number of contention slots in the reservation period of a frame
ε	Number of data slots in the service period of a frame
P_a	Probability of transmitting a new bandwidth request

P_r	Probability of retransmitting a backlogged bandwidth request
M	Number of subscriber stations in the R-MAC system
L	Maximum capacity of the service queue at he base station
S	Cumulative number of successful bandwidth requests in a reservation period
B	Number of backlogged subscriber stations or bandwidth requests at the beginning of a frame
W	Number of data packets waiting at the service queue at the beginning pf the frame
π	Steady state probability of a Markov chain
Ψ_i	Identifier of Markov chain i
${}^{\Psi}P$	Indicates the entries of the transition probability matrix of Markov process Ψ
${}_{\Psi_z}P_{u,v}^{(x)}$	Indicates the x-step transition probability from state u to state v according the Markov chain Ψ_z
th	System throughput
R	Reward function
R_D	Total delay related reward function consisting of contention delay and data transmission delay reward functions
R_{th}	Reward function of the frame throughput
R_{Dc}	Reward function of contention delay
R_{Dw}	Reward function of data transmission delay
g_D	Reward weight coefficient of the total delay
g_c	Reward weight coefficient of the contention delay

g_w	Reward weight coefficient of the data transmission delay
g_{th}	Reward weight coefficient of the frame throughput

1 Introduction

The speedy proliferation of the Wireless Local Area Network (WLAN) technology is evidence of a staggering need for broadband mobile access solutions. The IEEE 802.16 standard [1], supporting broadband fixed wireless access networks, is on the verge of redefining the way broadband services are provided. Besides offering higher bandwidth, the IEEE 802.16 standard was designed with ample QoS provisions allowing per-session QoS support. Often cast as economically unviable, extending broadband access to areas of limited demand has never been as feasible as with IEEE 802.16 thanks to its plentiful wireless coverage, reaching 15 km. In addition to the aforementioned strengths, the flexibility of providing wireless broadband service gives the IEEE 802.16 technical and economical advantage over existing DSL and HFC cable wireline technologies. Moreover, the promise of such a QoS-prosperous Metropolitan Area Network (MAN) going mobile presents a challenge to the widely used cellular and Wi-Fi technologies. Currently under development with expectations to be ratified in 2005, the IEEE 802.16e amendment supports mobile broadband access. The use of Orthogonal Frequency Division Multiple Access (OFDMA) maintains a reliable channel at data rates up to 30 Mbps and mobility speeds up to 80 miles/hr [2]. With its support for VoIP traffic, the

IEEE 802.16 standard can provide integrated voice and high speed data access. With bandwidth limitations in 3G technology, and coverage and mobility limitations in Wi-Fi (commercial brand name of the IEEE 802.11 Wireless Local Area Networks) technology, IEEE 802.16 can offer what 3G, WLAN, DSL, and HFC cable technologies can collectively offer, even on a Metropolitan Area Network (MAN) scale. Despite this, in the short term a conversion to mobile WiMAX (commercial brand name of the IEEE 802.16 Wireless Metropolitan Area Networks) is unlikely to occur owing to investment protection and customer loyalty. A more likely scenario is for mobile WiMAX to coexist with existing 3G and Wi-Fi technologies in a heterogeneous platform. Therefore, expediting the emergence of IEEE 802.16 is contingent on the technology's ability to cooperate with existing cellular and Wi-Fi technologies. Of special interest for this thesis is the study of the Multiple Access Control (MAC) of the standard in varying traffic rates. QoS preservation in a heterogeneous traffic environment is desirable. But unpredictable rates of traffic flow arriving to and departing from the network, challenge available resources and QoS preservation.

1.1 Multiple Access Control in the IEEE 802.16 Standard

Reservation Multiple Access (R-MAC) protocols are widely used in the Multiple Access Control (MAC) layer of broadband local and cellular access technologies. The R-MAC protocol is adopted in the IEEE 802.16 standard for its simplicity and efficiency in administering wide bandwidths. The R-MAC protocol organizes Subscriber Stations' (SSs) access to limited bandwidth (BW) Resources. In most of R-MAC protocols, time is organized into frames. A frame comprises an uplink subframe (from subscribers to base

station (BS)) and a downlink subframe (from BS to subscribers) The downlink portion of the frame is used solely by the BS to send a stream of Time Division Multiplexed (TDM) signal to the SSs whereas the uplink portion is shared by all SS's to send their data to the Base Station. The transmission of both portions may be time-separated on the same frequency band in a Time Division Duplex (TDD) fashion or simultaneous in time on separate frequency bands in a Frequency Division Duplex (FDD) setting.

In a centralized manner, media access and BW assignment to SSs are controlled by the head end controller (HC) (i.e. Base Station BS as I call it throughout the thesis). Upon traffic generation, an SS must first send a BW reservation request packet (BWR), through contention, to the BS. The purpose of the BWR packet is to indicate to the BS the amount of BW a SS needs for data transmission. Ultimately, when the BS receives a BW request, it allocates – on the uplink – an equal amount of BW to the SS's request. However, the schedule of BW allocation, which is not necessarily contiguous, is chosen at the BS's own discretion. The diverse priorities of traffic and SSs that are already in the system and also available BW resources are the main factors that shape the BS's chosen allocation schedule.

The uplink subframe is made in the most part of contention and data slots as shown in Figure 1.1. Besides, control and management slots occupy relatively minor portion of the uplink subframe. Since the amount of offered traffic normally exceeds the system's time-unit BW, system delay is inescapable. In reservation MAC protocols, the system delay consists of contention-related and data packets transmission related delays. The delay associated with data packets transmission depends on the number of available data slots in a frame. Since frame size is fixed, simultaneously minimizing both types of delays is a

paradox, though highly desired. An efficient reservation MAC protocol ought to have provisions to competently approach this objective.

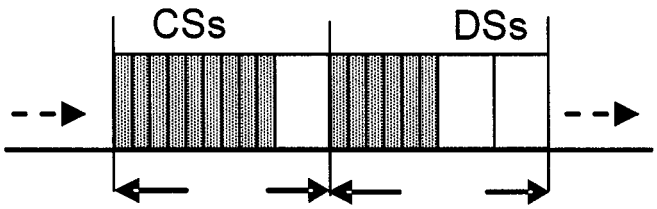

Figure 1.1 Example frame structure of R- MAC system

There have been several proposals of reservation MAC protocols for broadband networks in general. The most popular choice of a multiple access protocol for local access (wired or wireless) and cellular systems [3] supporting voice and data traffic is Reservation Aloha protocol (R-Aloha) [4]. Variations of the R-Aloha protocol are adopted in several local and wide area networks. Generally, the organization of the frame into contention slots (CSs) and data slots (DSs) differs from one protocol to another. In the R-Aloha protocol, proposed in [4], the media is organized into time frames. A frame is divided into a reservation period (I use the terms reservation period and reservation period intermittently to indicate the contention slots period of the frame) and a service period. Any SS must contend using slotted Aloha protocol over a group of CS's to send a BWR to the BS every time it generates traffic. Upon its eventual BWR success, a SS gets assigned the requested BW in the DS period of the frame. A variant of R-Aloha protocol, namely Packet Reservation Multiple Access protocol (PRMA) [5], was proposed by for satellite systems. In PRMA, time slots are grouped into frames. Each slot is recognized as "reserved" or "available" according to the acknowledgement message received from the

BS at the end of the slot. SS's contend to access a slot using slotted Aloha protocol once a data session begins. A slot successfully accessed by a SS is reserved for that SS in future frames till the end of the data session transmission. This represents the main difference from R-Aloha protocol. A PRMA-like protocol was proposed in [6] for microcellular mobile communication systems. It combines features of both R-Aloha and PRMA protocols. In this proposal, uplink and downlink subframes are organized according to Frequency Division Duplexing (FDD). In uplink subframe, a number of uniformly distributed slots are permanently marked as CSs. The remaining uplink slots are marked as DSs. An unused DS is used as an extra CS allowing the contention opportunities in to vary according to the offered data load.

These few proposals discussed above still represent the main categories of today's plentiful variations of R-MAC protocols. The frame structure in this family of protocols is similar in that the uplink subframe is made of CSs and DSs. In general, the allocation of CSs and DSs, in each frame, directly affects the BWR contention delay and system throughput. I define the contention delay suffered by a BWR as the time from the request's first transmission till the time it is successfully transmitted through contention. Increasing the number of CSs in a frame helps increase the probability of successful reservation access during that frame and hence reduces the reservation request's contention delay. Meanwhile, increasing the number of DSs in a frame increases system throughput and hence helps decrease BW allocation schedule-related delay.

1.2 Thesis Objective

A paradox pertinent to the R-MAC protocols is to tune the ratio of the contention and service period resources to serve a better performance. On the one hand, increasing the contention slot (CS) resources is highly desirable at times of excessive arrivals of reservation requests BWRs. The consequent decrease in data slots (DSs) results in an increase in the data transmission delay and a decrease in system throughput since more resources are used for contention. On the other hand in order to remedy the performance, reversing the force by decreasing the contention resources would result in lower number of successful reservations made to the base station resulting in increased contention delay and also decreased throughput. Evidently, network traffic varies over time. Therefore the size of the reservation period is a paramount design parameter for the performance of reservation-based access systems. This control parameter plays a centric role in determining the rate at which BWRs leave the contention phase and also the rate at which data packets are transmitted.

Contention resource allocation has not been specified in any of the aforementioned broadband network technologies. This area was rather left for proprietary solutions and product differentiation. I close in on the importance of the design of the size of the reservation period by reflecting on its impact on the performance of the R-MAC protocol. I first examine the role it plays in determining the contention delay in the system suffered by contending BWRs. I obtain insightful observations on the performance of contention delay under changing network environment (capacity, offered traffic, and resource allocation). The observations attained motivates the study of the underlying reservation

access process adopted in the IEEE 802.16 standard in light of different arrival rates and increased subscriber base, and how system delay and throughput can be impacted. Through analytical modeling and simulation, I show attainable performance improvement, through opportunistic dynamic resources management. This motivates us to design a dynamically responsive method to deal with such dual paradoxical process of contention and service resources allocation under varying traffic load. I aim at enhancing the R-MAC performance opportunistically through dynamic control of the reservation period. I show that a static allocation ratio that is not adaptive to the variations in traffic load or rate may yield poor performance. This thesis is a leading effort to further understand and enhance the performance of the R-MAC control of the IEEE 802.16 networks.

1.3 Thesis Outline

The rest of the thesis is organized as follows

Chapter 2 investigates the characteristics of the reservation based multiple access mechanisms adopted by the IEEE 802.16 standard.

Chapter 3 presents an overview of the literature. It features an overview covering the reservation multiple access protocol proposals, performance evaluation studies, and different approaches of establishing dynamic resources allocation.

Chapter 4 presents an analytical model to study the contention delay suffered in the reservation multiple access protocol of the IEEE 802.16 networks in light of different traffic and network parameters.

Chapter 5 presents a comprehensive performance evaluation of the reservation multiple access protocol of the IEEE 802.16 networks through analytical and simulation techniques.

Chapter 6 presents the results of the numerical experiments run using the analytical model in Chapter 5. In addition it also describes the simulation experiments used to validate the analytical model.

Chapter 7 presents a framework of dynamic reservation period controller as well as a Markovian optimization model to implement the framework.

Chapter 8 presents numerical examples on the optimization model of Chapter 7 under slotted Aloha and p-persistence contention resolution techniques.

Lastly, Chapter 9 presents my conclusions and future work.

2 Related Work

The widespread of access technologies that employ reservation multiple access protocols revived the need to understand their performance in today's network environments. Reservation based multiple access protocols are adopted mainly by broadband access networks such as the Hybrid Fiber Cable (HFC) technology, Digital Subscriber Line (DSL) technology, GPRS technology, and most recently IEEE 802.16 standard supporting both fixed (for the time being) and mobile (projected) broadband access. As will be seen in Chapter 3, a slotted contention protocol is adopted by the IEEE 802.16 standard, largely because of its simplicity in implementation and flexibility in supporting the standard's QoS platform. First, I present an overview of the most prominent reservation multiple access protocols. Then I survey previous studies on performance evaluation and dynamic reservation period allocation policies.

2.1 Reservation Multiple Access Protocols

A number of reservation based multiple access protocols [5]-[14] has been proposed. The most popular reservation multiple access protocol is the one based on slotted Aloha [4]. A closely related reservation multiple access protocol is the packet reservation

multiple access (PRMA) for local wireless communications [5]. In this protocol, a SS with a data session follows a slotted Aloha contention mechanism to access the media. The SS transmits the first packet of the data session by contention to access the media. Once it successfully accesses a slot, it reserves the same slot in future frames until the end of the data session where that slot is released. In PRMA, data packets that remain for long time in contention are discarded, which is not the case in the reservation Aloha protocol.

Another reservation protocol for mobile communications in a microcellular environment is proposed in [6]. An SS maintains a certain transmission probability to transmit a small reservation packet. If the reservation packet is successfully transmitted an information packet is assigned to the SS by the base station until the end of the information session. On the other hand if a collision occurs, an adaptive retransmission probability is adopted in a way to maintain system stability. A deviation from PRMA is Centralized PRMA (C-PRMA) [7] proposed for a microcellular environment. In C-PRMA a group of slots is reserved for contention and another group of slots is reserved for data transmission by polling. The base station sends polling information on a slot by slot basis. C-PRMA adopts a stack algorithm for reservation to avoid slotted Aloha reservation system instability in higher traffic load environments.

Also the contention-TDMA (C-TDMA) protocol [14], a hybrid of R-Aloha and PRMA, was proposed for radio cellular multi-SS systems. In this protocol the SSs contend over a slot if it is free. A list of free slots, updated on a frame by frame basis, is broadcast by the base station to all the SSs. A contention slot of a frame, in which a SS transmitted the first data packet of its data session, being not in the free slot list in the next frame, indicates that the transmission was successful and that the slot is reserved for the specified SS in

consequent frames till the end of the data session. The C-TDMA differs from R-Aloha in that it does not use a broadcast uplink, and it differs from the PRMA protocol in that the slot state is notified by the base station only once per frame, with little overhead.

2.2 Performance Evaluation of R-MAC protocols

The performance of reservation multiple access protocols has been studied [15]-[23]. In [16] a detailed analysis of reservation based access system is featured in a GPRS context. The frame under study consists of a contention period at the beginning of the frame followed by a service period, where the ratio between the size of the two periods is statically chosen throughout the time. An SS may transmit a BWR only once during a reservation period where a contention slot is chosen at random. Success is determined according to a capture model. An SS whose BWR was not successful in a reservation period waits for the next reservation period and randomly chooses a contention sot for its BWR transmission. SSs that successfully transmit their BWRs join a service queue on a First Come First Served (FCFS) basis. The data packets associated with the successful BWRs are served during the service period which has a fixed number of data slots. A Markov renewal process embedded at service departure times is utilized in formulating the equilibrium distribution of the number of customers in the system at arbitrary time instances and at customer arrival times.

In [14] the C-TDMA reservation MAC protocol is studied using a Markovian model. An SS adopts a dynamic permission probability, which is a system parameter, to choose a frame for contention. Once permission is obtained in a frame, a contention slot is chosen at random. An SS may be in one of three states: silent, talking, and backlogged. The bi-

dimensional state of the Markov process consists of the number of backlogged SSs and number of transmitting SSs. As the size of the state space of the Markov process does not allow straightforward manipulation, an equilibrium point analysis is used to analyze the Markov process and investigate the delay and throughput performance. In addition, an optimization technique has been proposed to improve the system resources utilization. Upon visiting a certain state, the optimal permission probability is calculated as a function of the number of SSs, number of slots, traffic characteristics. This optimal permission probability is used to calculate the new values of the equilibrium point. This process is recursively iterated until convergence is reached.

The equilibrium point analysis technique was also used in [6] to evaluate the performance of a proposed R-MAC protocol. The protocol uses a slotted Aloha contention resolution and an adaptive retransmission probability for the purpose of operation stability. The proposed protocol is analytically investigated using a three dimensional Markov process with the state described by the combination of the number of silent SSs, number of backlogged SSs, and number of SSs whether waiting in service queue or in transmission. The throughput and delay performance are calculated using the steady state distribution of the process in equilibrium conditions.

Another performance evaluation study for an integrated reservation TDMA protocol is conducted in [17]. This study utilizes Markov analysis to compute the contention delay experienced by the BWRs. On the other hand an M/G/1 queuing model is utilized to compute the data transmission delay spent in the service queue at the base station.

These are the most prominent performance evaluation studies that were cited in the literature in this field. As shown, a common factor among these studies is the use of discrete-time Markov analysis. Although I also utilize Markov analysis in evaluating the system performance, the approach and objective of this thesis are different than these outlined studies. Unlike others, whose approach is mainly to control the transmission probabilities, this thesis emphasizes the design of the reservation period with an objective to control the delay and throughput performance.

2.3 Reservation Period Allocation Techniques

Currently, the wide spread of access technologies employing reservation based multiple access protocols revived the need for optimal protocol operation. The problem of optimal reservation period allocation can be viewed as the key parameter in controlling both delay and throughput performance of the R-MAC protocols. However, there have been a few proposals [24]-[28] that reflected on this area. Also reference [29] is a comprehensive survey on the contention resolution protocols for the IEEE 802.14 networks. Only recently has the problem of optimal contention slot allocation gained attention in the literature. There have been a few proposals of dynamic contention slot allocation algorithms. In [24], Sriram et al. propose a contention slot allocation algorithm for Hybrid Fiber-Coaxial networks. They observe that a ratio of the number of contention slots to the number of data slots equal to 3:1 would achieve 100% throughput efficiency in a contention slot. This observation is based on 33.3% contention slot throughput efficiency resulting from the use of random binary exponential back-off algorithm for contention resolution. Thus the rational of the proposed contention slot allocation

algorithm is to keep the reservation period throughput efficiency less than or equal to the available number of data slots in a frame. Otherwise the number of contention slots is gradually decreased over the subsequent frames until the ceiling limit of reservation period throughput is restored. The main advantage of this algorithm is preventing the long run overflow of BWRs queue at the BS. However, the main concern that can be shed here is the possible excessive increase in contention delay. In the case of relatively high load of BWRs compared to the frame's contention slot allocation, a scenario that regularly occurs in multiple access networks, high rate of collisions is likely to be experienced. As a consequence, sever contention over the next frames, as a result of the relatively small reservation periods, might lead the contention delay of BWRs beyond acceptable limits of high speed data communications.

In another proposition, Sala et al. introduce in [25] a self-regulating adaptive contention slot allocation mechanism. All unused data slots are initially allocated as contention slots. The self-regulating mechanism dictates that if the number of contention slots is too low, the BWRs will not get to the base station, which automatically triggers additional contention slots allocation. On the other hand, if the number of contention slots is too high, more successful requests will reach the base station and the number of empty slots that can be allocated as contention slots will decrease. Finally, for any number of contention slots allocated in a frame, each SS adopts an optimal transmission probability $p = 1/M_a$, where M_a is the number of active (contending) SSs at the beginning of a contention slot and $M_a \leq M$, where M is the total number of SSs in the system. This approach requires p to be updated at the beginning of each contention slot according to the change in the number of active SSs. Adopting smaller transmission probability by SSs

has the effect of increasing the probability of a BWR success due to decreasing the number of transmitted BWRs. Therefore, knowledge of the number of contending SSs is necessary for calculating p. Because the base station BS can attain knowledge only of successfully received BWRs, a Pseudo Bayesian Estimator is employed to estimate the average number of active SSs in the next contention slot using the feedback information resulting from requests transmission in the previous contention slot. Though this proposal attempts to best utilize the available number of contention slots, it has the effect of enlarging the BWR waiting time in the SS queue. From service delay perspective, it is more advantageous for BWRs to be queued at the base station hoping to gain BW assignment as soon as available resources and service schedules allow.

Another proposal [26] follows a different methodology in dealing with the optimal contention slots allocation problem. They aim at maximizing throughput efficiency in the reservation period in an objective to let, through contention, as many BWRs as possible to get into the service queue. In order to achieve that, they specify that an optimized number of contention slots in a frame should equal to the number of BWRs that will be transmitted in that frame. In order to do that, knowledge of the amount of offered traffic is indispensable. In fact as mentioned earlier, the base station can attain knowledge only about successfully transmitted BWRs. Therefore, two heuristic approaches namely 'time proportionality' and 'most likely number of requests' are proposed to approximately calculate the average number of initial transmissions and backlogged packet retransmissions respectively.

As for performance evaluation studies for the reservation multiple access protocols, the problem of dividing the frame resources between contention and service resources has not

been yet comprehensively studied. As will be seen in Chapter 4 and Chapter 5, the size of the reservation period in a frame is a crucial design parameter for the delay and throughput performance of the reservation multiple access protocols.

Regarding the dynamic reservation period allocation proposals presented in the literature, I note that the techniques used to adjust the reservation period size are essentially driven by throughput performance. So, for certain traffic rates and patterns the delay performance may deteriorate as a result of operating around an optimal throughput point. In telecommunication networks, under limited resources, the delay performance of communications systems retracts after a certain point by increasing the system throughput. Therefore, the delay performance needs to be considered along with throughput performance in allocating the reservation period resources. Such an approach is advantageous for service providers letting them operate their networks according to various objective metrics. In essence, contention slots provide SSs with access opportunities to send their BWRs. Thus, in allocating contention resources, it is crucial to ensure that the level of contention delay does not drive the overall delay (including the data transmission delay) beyond acceptable figures.

The exploitation of possible delay and throughput gains through dynamic resource allocation administered by the base station, at the beginning of each frame, has not been yet explored. In this thesis, besides performance evaluation of the R-MAC protocol of the IEEE 802.16, I propose a novel method for dynamic resources allocation.

2.4 Summary

Reservation multiple access protocols appeared in the past two decades for the emerging broadband technologies. This family of protocols is flexible for integrated voice and data services and differentiated QoS support. Variant contention resolution techniques can be used during the reservation period. Mainly for its simplicity in implementation, Slotted-Aloha is considered by most of the proposed R-MAC protocols.

The performance of reservation multiple access protocols has been previously studied. However, an approach to evaluate the performance with respect to the contention resources allocation has not been comprehensively addressed. The size of the reservation period is a paramount design parameter for controlling the protocol performance. What distinguishes the work in my thesis is that I emphasize the impact of the reservation period size on the protocol performance. Furthermore, I establish a novel optimization method to enhance the performance of the reservation MAC protocol of the IEEE 802.16 networks and broadband networks in general.

Furthermore, I surveyed existing techniques of dynamic reservation period allocation. I noted that the common conceptual design of these proposals is based on adjusting the system parameters to maximize the reservation period throughput. Although this is a desirable aspect of the design, another aspect has to be considered in order to enhance the system throughput; that is to constantly allow more successes of BWRs in order to maximize the resources utilization (i.e. through serving more data packets).

3 Multiple Access Protocol of the IEEE 802.16 Standard: Overview

The IEEE 802.16 standard [1] defines two layers: the multiple access control (MAC) layer, and physical (PHY) layer [30]. In addition, two new MAC sub-layers are defined in the standard: the convergence sub-layer and the security sub-layer as shown in Figure 3.1.

3.1 Physical characteristics of the MAC Protocol

The MAC protocol of the IEEE 802.16 supports both Frequency Division Duplexing (FDD) and Time Division Duplexing (TDD) of the uplink and downlink portions of the frame. In FDD, the uplink and downlink signals are transmitted at the same time on two separate frequency bands. Figure 3.2 shows an FDD frame with the uplink and downlink transmissions occur on two different frequency bands. The DL-MAP message, as shown in Figure 3.2, defines the usage of the downlink intervals which is essential so that SSs tune into listening to their parts of the Time Division Multiplexed (TDM) data stream transmitted by the base station. The UL-MAP message defines the uplink usage in terms of the offset of the burst relative to the start time of the allocation for each SS.

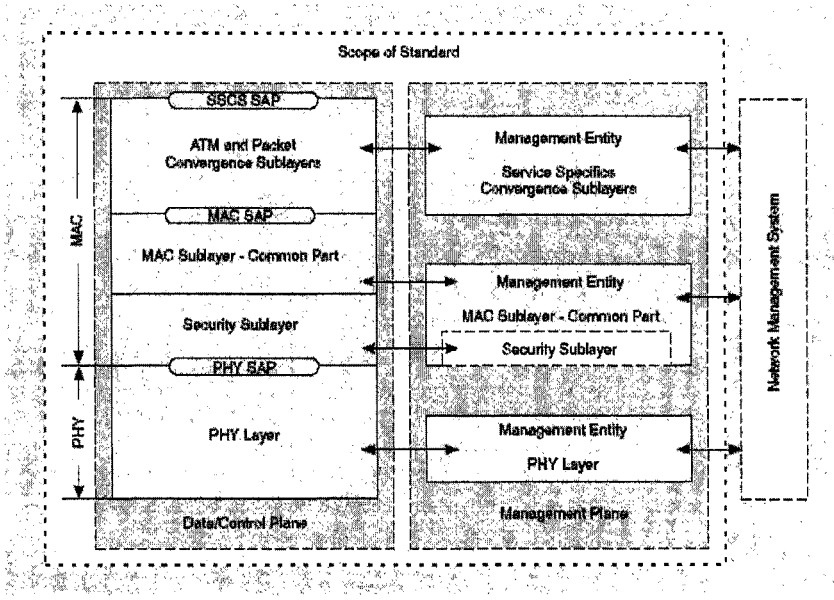

Figure 3.1 IEEE 802.16 network layers

A fixed duration frame is used for both uplink and downlink transmissions in order to

a) facilitate the use of different types of modulation.

b) allow simultaneous use of full-duplex subscriber stations SSs and half-duplex SSs.

c) simplify the bandwidth allocation algorithms.

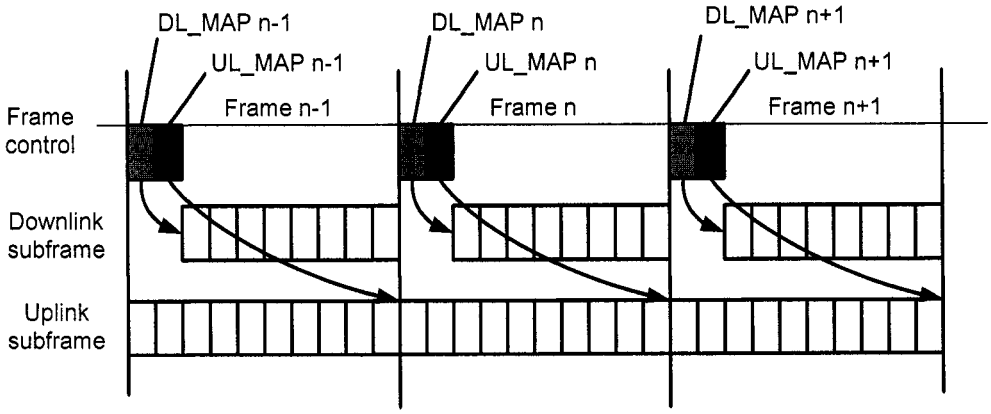

Figure 3.2 Frequency Division Duplex frame organization

In TDD, both uplink and downlink portions of the frame are transmitted on the same frequency bands but separated in time. Figure 3.3 shows a TDD frame where the uplink and downlink portions of the frame are separated in time. The TDD frame is adaptively divided between the downlink and uplink frame. Adaptively sized frame allows the most efficient resources utilization in times when one direction has a heavier traffic load than the other.

From Figure 3.4, it is noteworthy that the downlink subframe comes before the uplink subframe in order. The reason is that the DL-MAP and UL-MAP messages that carry information pertaining to the media assignment on the uplink direction are sent from the base station on the downlink subframe.

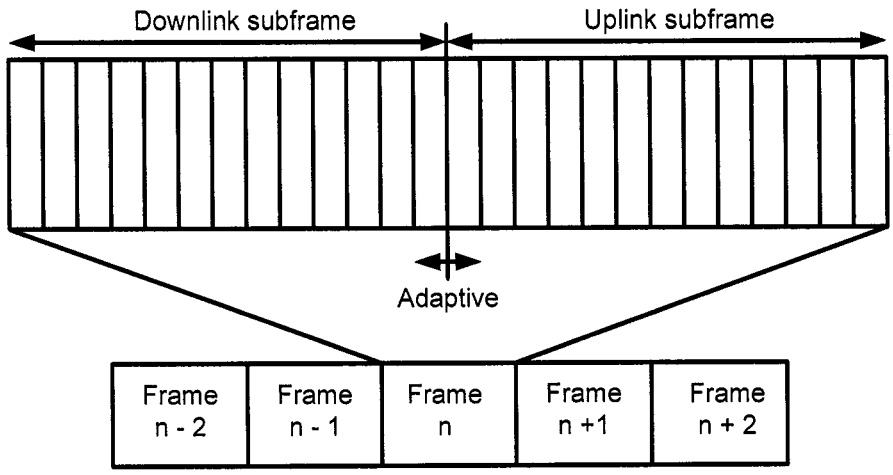

Figure 3.3 Time Division Duplex frame organization

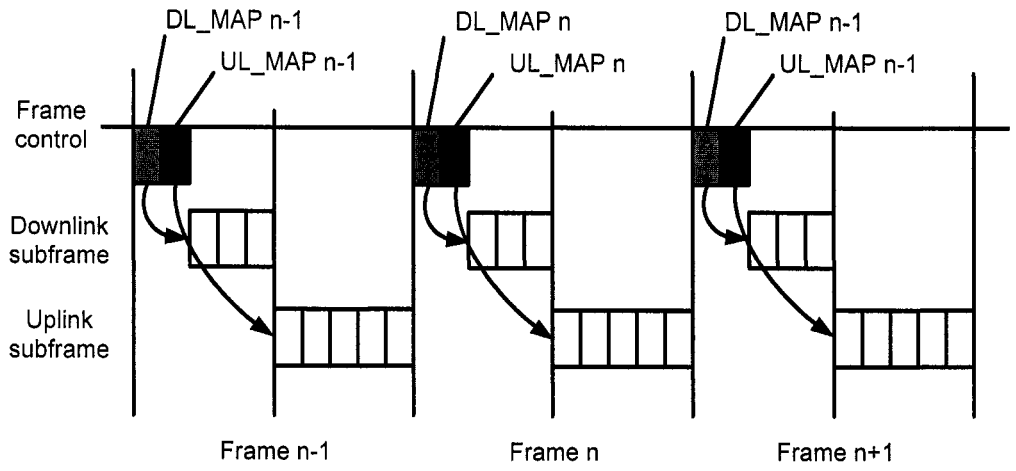

Figure 3.4 Time Division Duplex frame organization

3.1.1 Downlink Broadcast

The downlink, from the base station to the Subscriber Stations (SSs), operates on a point-to-multipoint basis. The IEEE 802.16 wireless link operates with a central base station and sector antenna which is capable of administering multiple independent sectors simultaneously. The base station is the only transmitter operating in this direction, so it transmits without having to coordinate with other stations, except for the overall time division duplexing that may divide time into uplink and downlink transmission periods. The base station sends traffic in a broadcast manner using Time Division Multiplexing (TDM) in the downlink channel. It broadcasts to all stations in the sector (and frequency); stations check the address in the received messages and retain only those addressed to them.

3.1.2 Uplink Multiple Access

Unlike the downlink access, which is a point to multi-point communication link, the uplink is a multi point to one point communication link. SSs share the uplink medium in transmitting their data to the bases station. Therefore organizing the stations access to the media is necessary to efficiently use the media. The IEEE 802.16 standard adopts a reservation based multiple access protocol for the upward communication link in which a combination of Demand Assigned Multiple Access (DAMA) and TDMA technologies are employed. The uplink frame is divided into reservation opportunities and data slots. SSs can transmit their reservation requests through contention only on reservation opportunities (slots) whereas the corresponding data packets to those reservation requests can be transmitted upon assignment only on data slots.

There are four different scheduling services that correspond to the different traffic characteristics,

1) Unsolicited Data Grants (UGS)

UGS is designed to support real-time service flows that generate fixed size data packets on a periodic basis, such as Voice over IP. The service offers fixed size unsolicited data grants (transmission opportunities) on a periodic basis. This eliminates the overhead and latency of requiring the SS to send requests for transmission. In UGS, contention based access is not allowed.

2) Real-Time Polling Service (rtPS) flows

rtPS is designed to support real-time service flows that generate variable size data packets on a periodic basis, such as MPEG video. The service offers periodic unicast request

opportunities, which meet the flow's real-time needs and allow the SS to specify the size of the desired grants. The SS is prohibited from using any contention or piggyback requests.

3) Non-Real-Time Polling Service (nrtPS) flows

nrtPS is designed to support non-real-time service flows that require variable size data grants on a regular basis, such as high bandwidth FTP. The service offers unicast request opportunities (polls) on a periodic basis, but using more spaced intervals than rtPS. This ensures that the flow receives request opportunities even during network congestion. In addition, the SS is allowed to use contention and piggyback request opportunities.

4) Best Effort (BE) Service Flows

The intent of the Best Effort (BE) service is to provide efficient service to best effort traffic. This is maintained by allowing the SSs to use contention request opportunities. This results in the SS using contention request opportunities as well as unicast request opportunities and unsolicited Data Grant Burst Types. SSs access the media through contention. If a SS succeeds in accessing the media it transmits its reservation request to the base station where the base station responds by assigning the required bandwidth in successive frames on a best effort basis.

3.2 Reservation Request and Bandwidth Allocation

An SS having contention-based traffic must first send a reservation request to the base station indicating the amount of bandwidth required. The uplink frame, as shown in Figure 3.4, is divided into Physical Slots (PSs). A number of PSs can be grouped into a

reservation slot (contention slot) or into a data slot. The size of a data slot is usually much larger than that of a reservation slot to maintain high throughput.

Once a contending SS succeeds in winning contention over one of the contention slots, it transmits its requirement of bandwidth to the base station. At this point, it is up to the base station to respond with bandwidth assignment to the requesting SS in the subsequent frames. Scheduling of SSs bandwidth allocation depends on the required service priority whether it is UGS, rtPS, nrtPS, or BE traffic.

3.3 Contention Resolution Mechanism

The base station controls the bandwidth assignments on the uplink channel through the UL-MAP messages and determines which reservation slots are subject to collisions. Collisions occur in contention intervals when more than one SS wishes to access a contention slots to place a BWR. The IEEE 802.16 standard mandates the use of truncated binary exponential backoff method for contention resolution. When a SS experiences collision it backs off its retransmission for a number of contention slots that is randomly selected between a minimum and a maximum backoff window values. This random value indicates the number of contention transmission opportunities that the SS shall defer before transmitting. The minimum and maximum backoff window values are controlled by the base station and distributed to the SSs using a medium access control message transmitted on the downlink.

3.4 Contention and Data Transmission Processes

The Best Effort (BE) class of traffic uses contention based reservation to obtain access to the service period of the frame. Figure 3.5 illustrates the contention and data transmission processes of the Best Effort (BE) traffic.

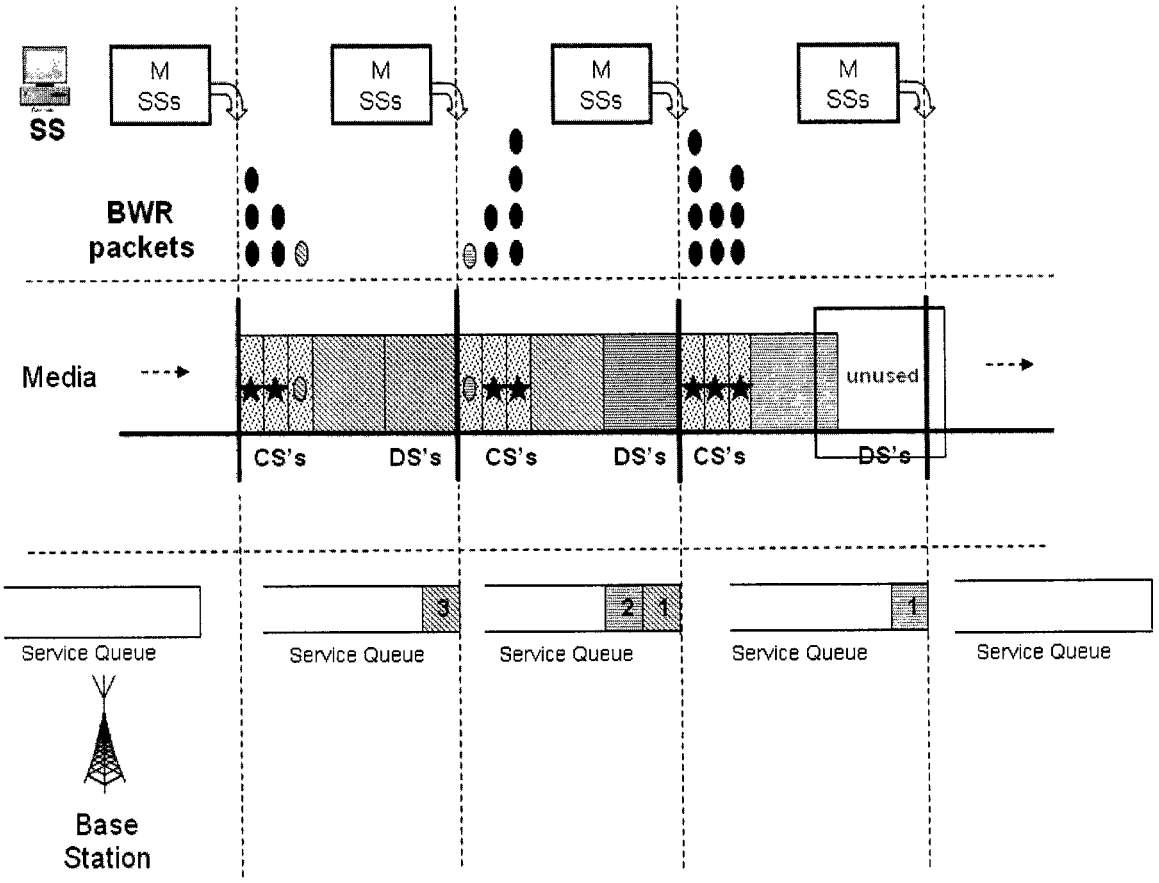

Figure 3.5 Contention and data transmission processes: illustration

As shown in Figure 3.5, multiple BWR transmissions in a contention slot results in a collision. Collided SSs retry transmission of their BWRs in the following contention slots. Once a BWR is successfully transmitted to the BS, it enters the service queue. The BWR carries information about the amount of the bandwidth required by the SS. In the IEEE

802.16 standard, the data session associated wit a successful BWR can start transmission in the following frame. In the illustration in Figure 3.5, for the ease of interpretation, data transmissions are allowed to start in the same frame in which the BWR was successfully transmitted. As shown in the second and third frames, the remainder of a data session in a frame resumes transmission in the service period of the next frame. In the third frame, the number of requested data slots in the service queue (one in this example) is less than the number of available data slots (two sots). The result is that a data slot will be left unused. Accumulation of such unused data slots result in poor bandwidth utilization. I argue that letting more BWRs into the service queue increases the throughput of the system. This can be done by increasing the size of the reservation period, which will create more transmission opportunities for the backlogged BWRs. However, since frame time is constant, the reduced service period may result in increased data transmission delay as the service rate will slow down. The contention and data transmission processes therefore compete for the frame resources. Later in Chapter 7, the an optimization method is presented that make better use of the frame slots.

3.5 Summary

The IEEE 802.16 standard defines the air interface specifications for wireless metropolitan area networks (WMAN). Since the inauguration of this standard there has been growing spread in manufacturing and deploying broadband fixed wireless access networks as the last mile solution to residential complexes and businesses.

Adopting a combination of Demand Assigned Multiple Access (DAMA) and Time Division Multiple Access (TDMA) enables the MAC protocol of the standard to offer

different levels of Quality of Service. The QoS support of the standard can basically be divided into real-time and non- real time service. Delay insensitive non-real time traffic uses contention-based reservation prior to resources assignment whereas real time traffic is assigned bandwidth resources needless of reservation.

The resources allocation between the reservation period and data transmission period has not been specified in the standard. It was rather left for vendor product differentiation according to the operating environment. This thesis broadly studies the resource allocation problem and examines the performance of the MAC protocol, under varying traffic environments.

4 Contention Delay Analysis

In the reservation multiple access protocol of the IEEE 802.16 a SS must first transmit a reservation request packet to the base station through contention. The purpose of the reservation request packet is to inform the base station of the amount of bandwidth required by that SS in order to transmit its data. Although the process of reservation consumes some of the bandwidth, it is considered an efficient way of organizing SSs' access to the media. The long run resources utilization of contention based reservation system is considered to be more efficient than utilization in conflict-free reservation system. This is because the reservation channel of an idle SS in a conflict free reservation system remains unused until that SS starts to transmit into the channel whereas contention based reservation systems do not suffer this problem. In this Chapter I will focus on the delay performance.

4.1 Contention Delay: Definition

When a SS generates data, it transmits a reservation request packet, which will be called Bandwidth Request Packet (BWR) throughout this thesis, to the base station. Once received by the base station, the BWR enters the service queue and waits for data slot

assignment over the data slot period. Therefore, the message delay comprises two parts: contention delay defined as the amount of time from the BWR first transmission attempt until it gets successfully transmitted to the base station, and data transmission delay defined as the amount of time from the BWR contention-based transmission until the data associated with that BWR gets transmitted on the data slots. I focus in this Chapter on studying the behavior of contention delay with respect to the resources allocation on the frame.

The chief two parameters that shape the contention delay are the amount of offered BWRs and the number of contention slots available in a frame. Because I wish to accommodate as much traffic as the network may offer, adequate number of contention slots needs to be allocated so that an acceptable contention delay is maintained from a broadband QoS perspective. In order to do that, I first need to have a mechanism to quantify the impact that the number of contention slots has on contention delay. A contention slot allocation scheme could then be built employing this mechanism to benchmark the calculated contention delay against a threshold level according to the network's QoS requirements. My motivation here is to approach the contention slot allocation problem based on the entailed expected contention delay.

Reservation-based traffic, in the context of reservation multiple access protocols, refers to data packets that need to request bandwidth BW reservations prior to their transmission. In random multiple access networks, where the number of SSs usually exceeds available access channel resources, contention is considered a fair method to provide access to the SSs' random transmissions. As was shown through Chapter 3, there are two traffic categories in the IEEE 802.16 standard that use contention to request reservation prior to

actual data transmission, those are the non-real time Polling Service (nrtPS) traffic and Best Effort (BE) traffic. In this chapter, I study the contention delay of the BWRs in the R-MAC of the IEEE 802.16 standard whose frame is shown in Figure 4.1.

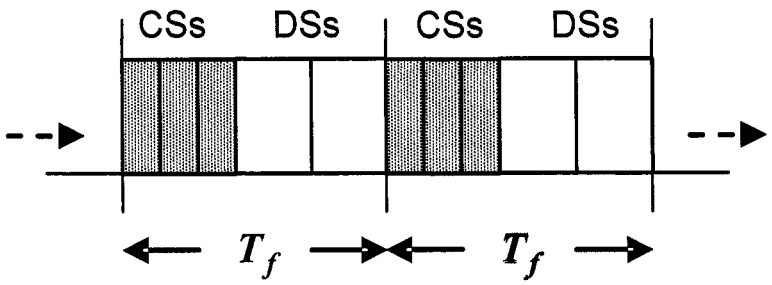

Figure 4.1 IEEE 802.16 reservations MAC frame

Figure 4.1 shows the frame structure of the reservation multiple access protocol of the IEEE 802.16 network. The frame which is periodically reproduced consists of a number of contention slots and data slots. Controlled by the BS, the number of contention slots may vary from one frame to another. In fact most of the multiple access technologies that employ reservation multiple access protocols do not mandate a specific number or schedule of contention slots in the uplink subframe. Contention slots could be scheduled in a collective or sporadic fashion along the uplink subframe. Contention delay improvement resulting from the use of either scheduling fashion is so inconsequential that I choose to disregard its effect in this study. I consider in my analysis collective contention slot allocation in the beginning of the uplink subframe as shown in Figure 4.1. It is worth noting that this setting has been recommended, but not specified, in the IEEE 802.16 standard for the ease of its implementation.

4.2 Contention Delay Analytical Model

A bandwidth request BWR may arrive at an SS at any time. However, as the base station broadcasts to all SSs the schedule of contention and data slots of a frame, the SS attempts transmitting only during the reservation period of the frame. In my analysis, without loss of generality I follow a Frequency Division Duplex FDD frame setting whereby the uplink subframe and downlink subframe are simultaneously established on two separate frequency bands. Upon generation, a BWR packet is transmitted in the beginning of the next available contention slot by carrying out a Bernoulli experiment. The outcome of this Bernoulli experiment is one (successful transmission) with probability P_{suc} and zero (collision) with probability $1-P_{suc}$ [31] P_{suc} is given by

$$P_{suc} = \binom{M}{1} \cdot (1-p)^{M-1} \cdot p = M \cdot p \cdot (1-p)^{M-1} \tag{4.1}$$

Where,

M is the number of SSs in the system.

p is the probability a SS transmits a packet in the beginning of contention slot.

As widely followed in modeling the offered traffic in a slotted Aloha context [24], [25], [32], I model the offered BWRs traffic such that each of the SSs transmits a new BWR or retransmits a backlogged BWR in the beginning of each contention slot persistently with probability p. An increase in p is equivalent to an increase in the offered traffic at each SS. This may sound tricky at first while when I consider only few out of the entire SS

population that actually have an increase in the offered BWR traffic. However, this can be explained with the assumption that an increase in the arrival rate at a SS by $\Delta\lambda$ resembles an increase in the arrival rate at every SS by $\Delta\lambda/M$. This assumption is strongly supported by the fact that a collision, resulting from simultaneous transmissions from multiple SSs, is independent of the SSs identities.

In the uplink subframe, only limited number of contention slots is available. Increasing the reservation period in a frame increases the likelihood of successful transmission during that frame. Nevertheless, with a fixed size uplink subframe, this would decrease the data slots period and hence decreases system throughput in the subject frame. If the contending SS failed to win any contention slot in the current frame, then it waits for the contention slots of the uplink subframe of the next frame in order to resume transmission attempts. Certainly the cost of such failure is an additional contention delay incurred. This process occurs repeatedly over subsequent frames until the SS successfully accesses a contention slot and sends its BWR. I define $P_{Access}(\tau)$ to be the probability that an SS successfully accesses the media in a finite number τ of contention slots. Assuming that a frame contains τ-contention slots, $P_{Access}(\tau)$ implies the probability that an SS successfully transmits its BWR in that frame as mathematically formulated in the following,

$$P_{Access}(\tau) = \sum_{k=1}^{\tau} P_{Access}(k) = \sum_{k=1}^{\tau} (1 - P_{suc})^{k-1} \cdot P_{suc}$$

$$P_{Access}(\tau) = 1 - (1 - P_{suc})^{\tau} \tag{4.2}$$

Denote the number of contention slots in the uplink subframe of frame f by τ_f, where $f \geq 0$. An SS successfully transmits its BWR in frame f with probability $P_{Access}(\tau_f)$ and fails with probability $P'_{Access}(\tau_f) = 1 - P_{Access}(\tau_f)$. If failed, the SS retries transmitting its BWR in subsequent frames. Recall my definition of the contention delay suffered by a BWR as the time from the request's first transmission till the time of its successful transmission. Since the reservation periods are interleaved with data slot periods, the notion of contention delay may comprise both reservation periods and data slot periods. If the BWR was successfully transmitted in one of τ_f contention slots of frame f, then the expected contention delay suffered would be D_f

$$D_f = \frac{\tau_f}{2} \cdot T_{CS} \tag{4.3}$$

where T_{CS} is the time of a contention slot.

If, on the other hand, the BWR could not be successfully transmitted in the reservation period of frame f, the SS will wait for the next frame, frame $f+1$, to resume contention over τ_{f+1} contention slots. In this case, the contention delay encountered in frame f becomes the total length of the uplink subframe T_f. Similarly, the same scenario reoccurs in frame $f+1$ where the average contention delay encountered would equal D_f with probability $P_{Access}(\tau_{f+1})$ and would equal T_f with probability $P'_{Access}(\tau_{j+1})$. The contention process continues on until the BWR eventually gets successfully transmitted in

one of the subsequent frames. I formulate this series of probabilistic events to describe the expected contention delay D_{BWR} as follows

$$D_{BWR} = P_{Access}(\tau_f)\cdot D_f + P'_{Access}(\tau_f)\cdot[P_{Access}(\tau_{f+1})(T_f + D_{f+1})+ \\ P'_{Access}(\tau_{f+1})\cdot[P_{Access}(\tau_{f+2})(2\cdot T_f + D_{f+2})+ P'_{Access}(\tau_{f+2})\cdot[P_{Access}(\tau_{f+3})(3\cdot T_f + D_{f+3})+\cdots\ ]\ ]\ ] \quad (4.4)$$

The formula in (4.4) does not yield a closed form solution for the overall contention delay. This is because the frame in which a successful transmission of the BWR will occur cannot be determined due to the probabilistic nature of the problem. Therefore, I follow a statistical approach to find a closed form solution for the expected contention delay D_{BWR}. If the BWR could not be successfully transmitted in frame f, the delay experienced at the beginning of frame $f+1$ will equal T_f. At that point, and since the contention results in a frame is independent of the contention results in previous frames, the process statistically starts over as mathematically formulated herein,

$$D_{BWR} = D_f + P_{Access}(\tau_f)+(T_f + D_{BWR})(1-P_{Access}(\tau_f))$$

rearranging,

$$D_{BWR} = D_f + T_f \frac{1-P_{Access}(\tau_f)}{P_{Access}(\tau_f)}\cdot\frac{1}{M} \quad (4.5)$$

The expression in (4.5) reveals that the higher $P_{Access}(\tau_f)$ is the lower the second delay component resulting from resuming contention over subsequent frames. Substituting (4.2) into (4.5) to obtain D_{BWR} as a function of τ_f (number contention slots in frame f)

$$D_{BWR} = \frac{\tau_f}{2} \cdot T_{CS} + T_f \frac{(1-P_{suc})^{\tau f}}{1-(1-P_{suc})^{\tau f}} \cdot \frac{1}{M} \tag{4.6}$$

Substituting (4.1) into (4.6), D_{BWR} is obtained as a function of the design parameters M (system capacity), τ_f (no. of frame's contention slot in frame f) and p (SS transmission probability) as follows,

$$D_{BWR} = \frac{\tau_f}{2} \cdot T_{CS} + T_f \frac{\left(1-Mp(1-p)^{M-1}\right)^{\tau f}}{1-\left(1-Mp(1-p)^{M-1}\right)^{\tau f}} \cdot \frac{1}{M} \tag{4.7}$$

The latter Equation (4.7) is important. It describes the impact of system capacity (M) and the probability with which each SS transmits in a contention slot (p) on the expected contention delay. Most importantly for the context of my work, (4.7) also describes the impact of contention slot allocation (τ_f) on the expected contention delay suffered by the BWR

4.3 Numerical Analysis

In order to plot the results from Equation (4.7), the values of T_f and T_{CS} need to be specified. The IEEE 802.16 standard specifies the maximum nominal frame duration to be $200\,ms$, but recommends it to be $1\,ms$ long. I will use the recommended frame duration of $1\,ms$ for T_{frame}. For T_{CS}, the IEEE 802.16 standard specifies that the channel size can be 20 MHz, 25 MHz, or 28 MHz. I consider the 20 MHz channel where the frame is divided into 4000 Physical Slots (PS). Assuming that a contention slot comprises one PS, the contention slot duration would be,

$$T_{CS} = \frac{1\,ms}{4000} = 0.25\mu s$$

I plot Equation (4.7) in Figure 4.2,.which shows the relation between the expected contention delay D_{BWR} and number of contention slots (τ_f) for different values of SS population $(M)$with a fixed value of the per-SS transmission probability(p) given the above values of T_f and T_{CS}. I observe that increasing the number of contention slots, in the immediately following frame, reduces the expected contention delay. In general, the resulting contention delay reduction is intuitive with the increase of the number of contention slots in a slotted Aloha context. What concerns my study in particular however, is making a decision on the number of contention slots to be allocated in the next frame. Therefore I focus the attention on the adequacy of contention slot allocation in the following frame $f+1$ so that the resulting contention delay would be kept within QoS delay threshold D_{QoS}, which could be a service provider proprietary means to establish service differentiation. A frame comprising contention slots and data slots is of a fixed size. Increasing contention slot allocation in a frame improves contention delay but results in reduced data slot period, which reduces system throughput. In Chapter 5 a broader study is, that incorporates the throughput behavior in light of contention slot allocation problem, is presented. Back to our analysis, the result in Figure 4.2 visualizes contention delay sensitivity to the change of the allocated number of contention slots. Contention delay sensitivity for contention slots allocation in the next frame is much higher in the case of low (τ_f) than in the case of high(τ_f). For example, for 120 SSs, allocating extra 50 contention slots on top of low τ_f brings the delay down from about

40 *ms* to about 20 *ms*, while a gain of only about 2 *ms* results from the same additional contention slot allocation in the case of high τ_f. This insightful observation is of chief importance for the study of minimizing reservation-based traffic contention delay and maximizing system throughput through dynamic contention slot allocation over consecutive frames. An increase from 50 contention slots to 100 contention slots allocation shows a drop in the expected average delay by 20 *ms*.

In order to achieve the same amount of contention delay reduction of 20 *ms* where the current allocation is 100 contention slots, the contention slot allocation needs to be increased by over 400 contention slots. Therefore the cost of decreasing contention delay, which is certainly a BW cost, varies significantly according to how severely contention slots are currently under allocated. Accordingly, a variety of decisions could be taken if additional contention slot allocation would not satisfactorily improve contention delay in the next frame.

Maintaining the same number of contention slots in the next frame could be a choice. Another choice could be decreasing contention slot allocation in the next frame. The latter choice would theoretically result in fewer chances for successful access but in the same time would result in better system throughput in the subject frame. Since the predicted contention delay improvement is trivial, improving system throughput in the next frame by allocating some contention slots as data slots would result in better use of system resources. Furthermore, the latter choice would be a more attractive resort if queued BW-requests are starved of BW resources.

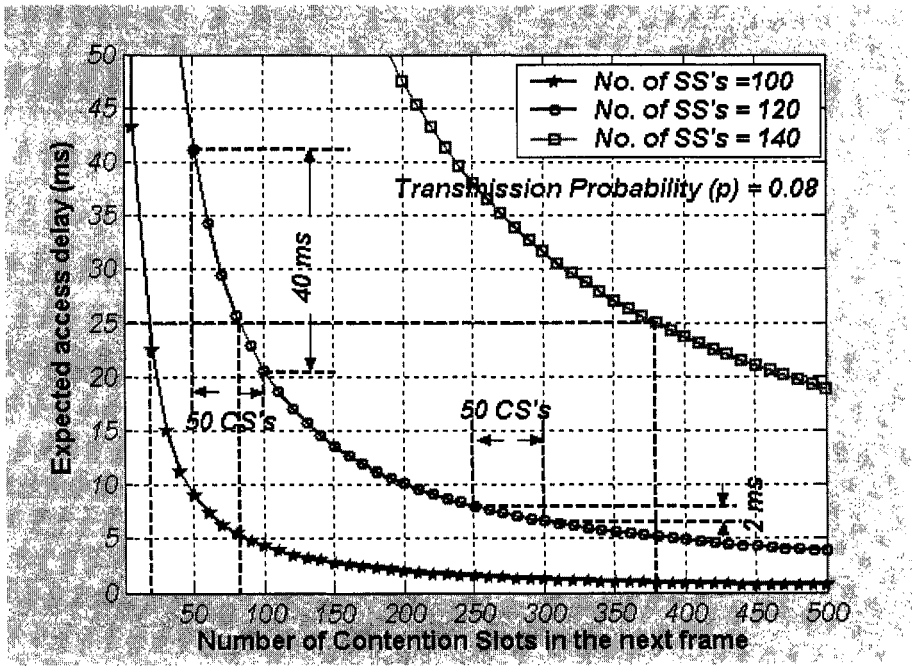

Figure 4.2 Effect of contention slot allocation on expected contention delay

Another observation from the result in Figure 4.2 is that system capacity (number of SSs) can be increased without worsening contention delay by increasing the number of per-frame contention slots. As shown in Figure 4.2, 25 *ms* of contention delay can be maintained with an increase in system capacity from 100 to 120 to 140 by increasing the number of contention slots from about 20 to about 80 to about 370 respectively. This is due to the fact that increasing the number of contention slots balances out the increase in the number of SSs, rendering the density of SSs over available contention slots approximately the same and the probability of successful transmission also approximately the same. However it is worth mentioning that system capacity increase may as well cause demand increase on data slot resources, which in turn will be reduced by the act of widening reservation period.

Figure 4.3 shows the relation between expected contention delay D_{BWR} and number of SSs (M) for different values of contention slot allocation (τ_f) with a fixed value of per-SS transmission probability (p) with the previously indicated values of T_f and T_{CS}. I observe that expected contention delay remains almost unchanged with the increase of contention slot allocation over a range of system capacities. This renders the increase in contention slot allocation, in an attempt to reduce contention delay, ineffective. For example in Figure 4.3, in the range of system capacities up to 70 SSs, contention delay remains approximately unchanged with the increase in contention slot allocation from 100 contention slots to 500 contention slots.

I accordingly define the notion of ineffective contention slot allocation as the increase in contention slot allocation from one frame to another $\Delta^+\tau_{CS} : \Delta^+\tau_{CS} = \tau_{f+1} - \tau_f$ that improves the contention delay by an amount $\Delta D_{BWR}(\Delta^+\tau_{CS}) \leq \partial_{ineffective}$ where $\partial_{ineffective} <<$, and $\Delta D_{BWR}(\Delta^+\tau_{CS}) = D_{BWR}(\tau_{f+1}) - D_{BWR}(\tau_f)$. The ineffective contention slot allocation is therefore characterized by two variables, $\Delta^+\tau_{CS}$ and ΔM. Additional contention slot allocation would be effective only if the resulting improvement in the expected contention delay exceeds a threshold value $\partial_{ineffective}$. In the case of ineffective contention slot allocation, it certainly makes better system resources utilization to allocate the subject $\Delta^+\tau_{CS}$ slots in the data slot period rather than in reservation period (which would not cause improvement in contention delay).

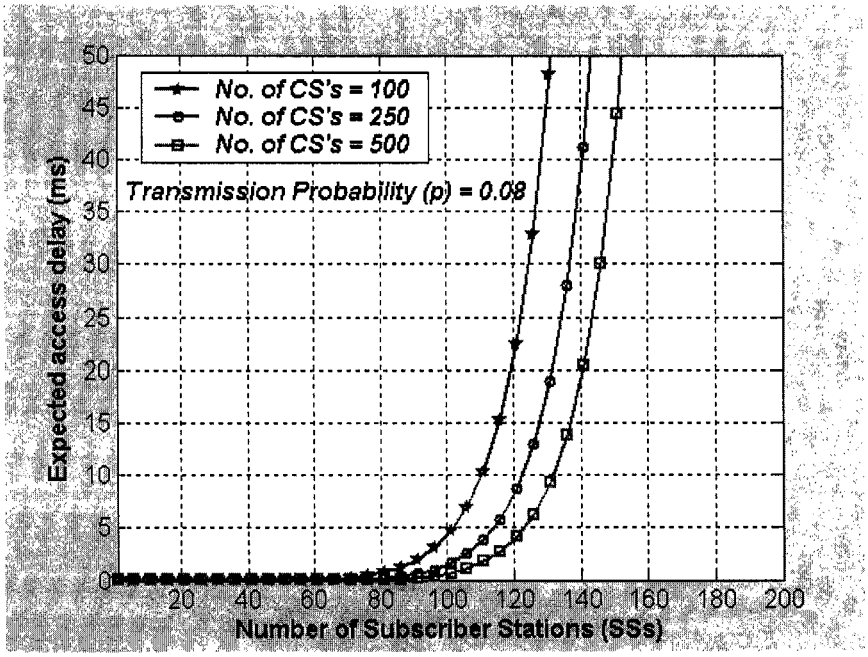

Figure 4.3 Expected contention delay versus the number of subscriber stations

Finally, Figure 4.4 shows the relation between expected contention delay D_{BWR} and per-SS transmission probability(p) for different values of contention slot allocation (τ_f) with fixed number of SSs (M) given the aforementioned values of T_f and T_{CS}. The plots illustrate that the expected contention delay is highly sensitive to slight changes in the per-SS transmission probability(p). As shown in Figure 4.4, when the transmission probability(p) grows beyond 0.1, the resulting expected contention delay grows sharply. I observe that bringing the resulting contention delay increase down requires the assignment of a large number of contention slots, which may affect system throughput. It is noteworthy from Figure 4.3 and Figure 4.4 that the increase of the per-SS transmission probability has a more severe effect on contention delay than the increase in system capacity (number of SSs).

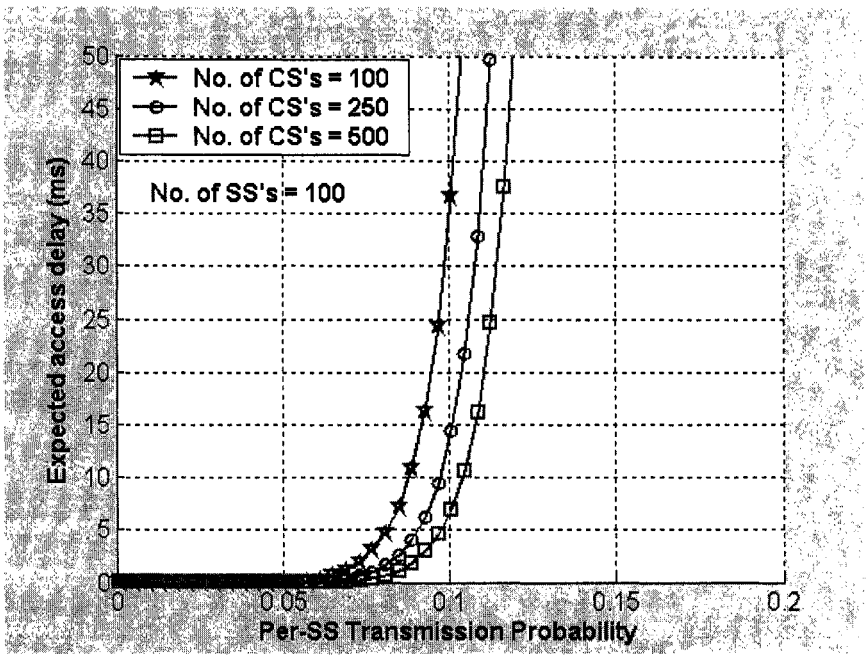

Figure 4.4 Expected contention delay versus per subscriber station transmission probability

4.4 Summary

Variations of reservation MAC protocols are widely used in the MAC layer of local and cellular access technologies that support voice and data communication services. In this chapter I presented a probabilistic analytical model for the contention delay suffered by a reservation request packet in the reservation MAC protocol of the IEEE 802.16 networks. Having formulated the expected contention delay as a function of the number of contention slots allocated in the frame, I drew fundamental observations on their relation. I showed that the contention delay sensitivity to the change in contention slot allocation varies greatly according to the number of contention slots in a frame. I also showed that the increase in system capacity could be counterbalanced by an increase in the number of contention slots in order to maintain the same contention delay. Moreover I

showed that system throughput could be rather improved in situations of imminent ineffective contention slot allocation. Finally, I showed that contention delay is much more sensitive to an increase in the per SS transmission probability, which resembles an increase in the SS's arrival rate, than to an increase in the system capacity.

The study of contention delay is essential especially with the stochastic arrivals and contention process. However, the reservation period allocation cannot be decided upon in isolation from the performance of the data transmission delay and system throughput. Therefore, a thorough study of the reservation multiple access of the IEEE 802.16 requires a more sophisticated analytical model that can reflect on both the contention delay and data transmission delay simultaneously taking into consideration the system throughput.

5 Delay and Throughput Analytical Model

As contention delay and data transmission delay are mutually intertwined, they need to be jointly studied. An analytical model to study the reservation multiple access system of the IEEE 802.16 networks should capture the contention delay side by side with the data transmission delay resulting from different scenarios of reservation period allocation. In other words, the interdependency of the two quantities needs to be incorporated in the computation in order to reach an allocation policy that guarantees optimum performance. I present herein a mathematical model to study the delay and throughput performance of the reservation multiple access protocol of the IEEE 802.16 networks in light of the reservation period size.

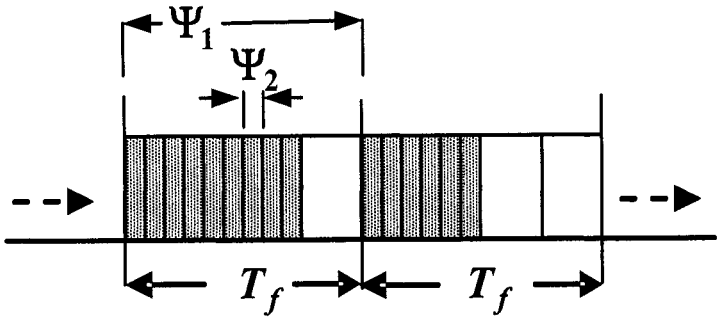

Figure 5.1 IEEE 802.16 reservation MAC frame

Since the frame starts at discrete and fixed time points as shown in Figure 5.1, I would like to examine the Markovian property of the R-MAC system at the time instances in the beginning of the frame. I let B be the number of bandwidth requests (BWRs) that are in contention at the beginning of the frame, and W be the number of waiting data packets at the service queue at the beginning of the frame. The process Ψ_1 with state space given by $\{(B(f), W(f)): \quad f \geq 0\}$ is Markovian where $f = 0, 1, 2, \cdots$ since its future state depends on the past states only through the current state [36]. I use another Markov process denoted Ψ_2 to observe – on a contention slot basis – the arrivals to and departures from the backlog state, as well as the BWR successes that join the service queue. The state transition probabilities of the process Ψ_2 enable the calculation of the transition probability matrix of the process Ψ_1. In what follows, I will derive the transition probability matrices of Ψ_1 after describing the working assumptions.

5.1 Assumptions

Each of M SSs transmits a BWR in a contention slot with probability P_a, and retransmits it after collision with probability P_r. A SS can send only one BWR at a time until successfully transmitted. Therefore, the maximum number of BWRs in contention at any given time equals M. This simplifying assumption used earlier in [31], assists in keeping track of the evolution of number of backlogged BWRs. However, a SS may transmit or retransmit as many times as possible within the same reservation period, unlike the assumption of single transmission per frame in [16]. The service queue at the BS treats BWRs in a First Come First Serve (FCFS) fashion, and is of finite length L,

where overflow is dropped and retransmitted again by contention. The BWR packet size is equal to the size of a contention slot. For simplicity of the model, a BWR packet corresponds to a constant amount of SS data, which equals the size of a data slot. The size of a data slot is greater than that of a contention slot, a design principle in R-MAC protocols that leverages throughput performance.

It is also assumed that data packets associated with a BWR can start transmission in the same frame where the BWR was successfully transmitted., it is assumed data transmission can This assumption, which has negligible effect on the model's accuracy, is used in this Chapter 5and in Chapter 7 to simplify the model calculations.

5.2 Frame Markov Chain

As shown in Figure 5.1, CSs are contiguously organized in the beginning of the frame followed by DSs. Since frame size is constant, knowledge of the number of CSs τ in a frame implies a number of DSs ε in the same frame $\varepsilon = (T_f - \tau \cdot T_{CS})/T_{DS}$ where T_f is the uplink frame time, T_{CS} is the CS time, and T_{DS} is the DS time. The two variables of interest when studying delay and throughput of R-MAC are the number of backlogged BWRs B in contention and the number of waiting BWRs W in the service queue, waiting for data slots assignment. The evolution of B and W is stochastic due to the random nature of both traffic arrivals and contention. I use Ψ_1 to denote a two dimensional Markov process with state space $\{(B,W),\ \ 0 \le B \le M \ \ 0 \le W \le L\}$, where the state is observed in the beginning of a frame, to study delay and throughput performance of the protocol. The time unit of Ψ_1 is the frame time. The choice of the process state to be

(B,W) enables us to cast the underlying process through state transitions along the R-MAC frame. The one-step transition probability matrix of Ψ_1, denoted ${}^{\Psi_1}P$, has an element value ${}^{\Psi_1}P = P\{(B_{f+1},W_{f+1}) | (B_f,W_f)\}$. Computing ${}^{\Psi_1}P$ is not altogether straightforward. It is not a simple counting problem and an approach which calculates binomial probabilities of successes and collisions using B_f, τ, P_a, and P_r becomes complicated quickly. This is due to multiple transmissions and retransmissions in a reservation period and the fact that, a SS may join the backlog state and later depart it in the same frame. Therefore, an enabling mechanism is required to incorporate all combinations of events that may occur during the reservation period.

5.3 Reservation Period Markov Chain

I let Ψ_2 denote a Markov process with state space $\{(B,S),\ 0 \le B \le M\ \ 0 \le S \le \tau\}$, where B is the number of backlogged SSs (BWRs) at the beginning of a CS, and S is the cumulative number of successes since the beginning of the reservation period – be those successes from backlogged or newly transmitted BWRs. The discrete time Markov process Ψ_2 has the contention slot as its one-step time unit. The transition probability matrix of Ψ_2, denoted ${}^{\Psi_2}P$, has an element value given by

$${}^{\Psi_2}P = P\{(B_f + i, S_f + j) | (B_f, S_f)\} \ \forall\ 0 \le i \le M\ ,\ 0 \le j \le \tau\ .$$

In a similar way to the Slotted Aloha analysis in [34], define $Q_a(i,B)$ as the probability that i out of $M - B$ idle SSs simultaneously transmit, each with probability P_a, a new

BWR in a contention slot. Also, define $Q_r(i,B)$ as the probability that i out of B backlogged SSs simultaneously retransmit, each with probability P_r, their backlogged BWR in a contention slot.

$$Q_a(i,B) = \binom{M-B}{i} * {p_a}^i * (1-p_a)^{M-B-i} \tag{5.1}$$

$$Q_r(i,B) = \binom{B}{i} * {p_r}^i * (1-p_r)^{B-i} \tag{5.2}$$

The transition probability matrix Ψ_2 is

$$\Psi_2 P = \begin{cases} Q_a(i,B) & 2 \le i \le M-B \quad j=0 \\ Q_a(1,B)\cdot[1-Q_r(0,B)] & i=1 \quad j=0 \\ Q_a(0,B)\cdot[1-Q_r(0,B)] & i=0 \quad j=0 \\ 0 & i=-1 \quad j=0 \\ 0 & 1 \le i \le M-B \quad j=1 \\ Q_a(1,B)\cdot Q_r(0,B) & i=0 \quad j=1 \\ Q_a(0,B)\cdot Q_r(1,B) & i=-1 \quad j=1 \\ 1 & S=\tau \quad j=0 \end{cases} \tag{5.3}$$

The last line in (5.3) pertains to invalid transitions out of the absorbing states $(B,\tau)\ \forall B$ since only a finite number of CSs, τ, is available in a frame. Now I can obtain the contention results at the end of reservation period through the τ-step transition probability matrix using Markovian properties,

$$\Psi_2 P(\tau) = \left\lfloor \Psi_2 P \right\rfloor^{\tau} \tag{5.4}$$

The τ-step transition matrix is useful in tracking all possible combinations of state evolution over the reservation period which is in effect an alternative approach to

combinatorial techniques in dealing with this problem. Let's now get back to calculating ${}^{\Psi_1}P$ using ${}^{\Psi_2}P$. The state transition in Ψ_1 from (B_f, W_f) to (B_{f+1}, W_{f+1}) entails the following relations

$$W_{f+1} = \begin{cases} W_f + S_f - \varepsilon & \text{if} \quad W_f + S_f > \varepsilon \\ 0 & \text{if} \quad W_f + S_f \le \varepsilon \end{cases} \tag{5.5}$$

From (5.5), knowing W_f, W_{f+1}, and ε, I can determine the S_f value(s) that cause the transition. The entries of ${}^{\Psi_1}P = P\{(B_{f+1}, W_{f+1}) \mid (B_f, W_f)\}$ can be directly obtained from those of ${}^{\Psi_2}P = P\{(B_f + i, S_f = 0) \mid (B_f, S_f)\}$, where I am interested only in elements of ${}^{\Psi_2}P$ where the starting state has $S_f = 0$. Since the system backlog state is identical in both Ψ_1 and Ψ_2, the entry $P\{(B_{f+1}, W_{f+1}) \mid (B_f, W_f)\}$ is given by (5.6). I will frequently use the terms ${}^{\Psi_z}P_{u,v}^{(x)}$ and ${}^{\Psi_z}P_{u,v}$ to denote the x-step and one-step transition probabilities, respectively, from state u to state v of the Markov process Ψ_z.

$${}^{\Psi_1}P = \begin{cases} \sum_{s=0}^{\tau} {}^{\Psi_2}P^{(\tau)}_{(b_f,0),(b_{f+1},s)} & W_{f+1} = 0 \quad \varepsilon \ge L \\ \sum_{s=0}^{\min(\tau, \varepsilon - W_f)} {}^{\Psi_2}P^{(\tau)}_{(b_f,0),(b_{f+1},s)} & W_{f+1} = 0 \quad \varepsilon < L \\ 0 & W_{f+1} > 0 \quad W_f + S_f > L \\ \sum_{s=S_f}^{\tau} {}^{\Psi_2}P^{(\tau)}_{(b_f,0),(b_{f+1},s)} & W_{f+1} > 0 \quad W_f + S_f = L \\ {}^{\Psi_2}P^{(\tau)}_{(b_f,0),(b_{f+1},S_f)} & W_{f+1} > 0 \quad W_f + S_f < L \\ 0 & \text{otherwise} \end{cases} \tag{5.6}$$

where the first line is the case when the sum of W_f and S_f BWRs will all be transmitted in frame f. The second line is similar to the first one but with discarding the possibilities when S_f drives the total number of waiting BWRs beyond ε. The third line is for transitions that entail a total number of waiting BWRs at end of the reservation period exceeding the service queue size L. The fourth line, for the case when the queue is full at the end of the reservation period, considers the possibility of additional successes that were dropped out of the service queue due to its finite length. The fifth line is straightforward.

5.4 Steady State Probabilities of Ψ_1

${}^{\Psi_1}P$ is useful in formulating the contention delay D_c and data transmission delay D_w in terms of τ. For the purpose of this study, τ is considered constant. I solve ${}^{\Psi_1}P$ for the steady state probabilities. $\pi_{(B,W)}$, defined as the discrete time steady state probability of having B backlogged SSs and W waiting BWRs in the beginning of a frame. $\pi_{(B,W)} = \lim_{f\to\infty} {}^{\Psi_1}P\{(b_f, w_f) = (B,W) | (b_0, w_0)\}$ from which $\pi_B = \sum_{w=0}^{L} \pi_{(B,w)}$ and $\pi_W = \sum_{b=0}^{M} \pi_{(b,W)}$.

According to the design of Ψ_1, the state (B,W) of the process is observed at the beginning of each frame. Therefore computing the inflow and outflow of both contending BWRs and successful ones to and from the frame becomes tricky. It is possible that a

BWR starts contention, joins backlog state, and gets successfully transmitted during the same frame. This may provoke a doubt that the steady state probabilities for the number of backlogged BWRs at the beginning of a frame, π_B , is not representative of the all time steady sate probabilities (i.e. steady state probabilities of having B backlogged SSs at any time). Since the design of Ψ_2 considers all the possible scenarios resulting during the reservation period, π_B is reliable to represent the all time steady state probability of having B backlogged SSs in the system. On the other hand, when a BWR gets successfully transmitted during the reservation period of a frame, it joins the service queue immediately. If that BWR reached the head of the queue before the end of the same frame, it will be transmitted before observing the state at the beginning of the next frame. Hence, the definition of π_W, as stated above, does not include the effect of this type of BWRs. This results in the steady state probability of having W BWRs waiting in the service queue at the beginning of a frame π_W to be different than the (all time) steady state probability. So, the (all time) steady state probability of W', denoted $\pi_{W'}$, needs to be calculated.

The transition probability matrix ${}^{\Psi_1}P$ describes the state evolution at fixed and discrete time moments at the beginning of each frame. Conditioning on the state of the process at these moments, the proportion of time out of a frame time that the process is at a certain state can be computed. In other words, I can compute the ratio of the number of slots at the end of which the process is at a certain state, to the total number of slots. Since the time of a contention slot and time of a data slot are not the same, I use time proportionality. Therefore, conditioning on the state at the beginning of a frame, I can

sum all time proportionalities that the process would be in a certain state or $\pi_{W'}$.

If the number of waiting BWRs at the beginning of a frame is W, then W' is the number of waiting BWRs at the end of a slot during the frame. W may be equal to, larger than, or less than W'. Another factor to be taken into account is whether the slot under consideration is a contention slot or a data slot. I will study these cases one by one as follows.

Case 1: $W = W'$ at the end of a contention slot

I would like to calculate the time proportionality of being in state W' at the end of a contention slot given W waiting BWRs in the beginning of the frame where $W = W'$. Since no data transmission occurs during the reservation period, W' can only be larger than or equal to W. In order for a contention slot to witness $W' : W = W'$ at the end of its time, there must have not been any successes in any of the contention slots prior to it. For example, W' will equal W at the end of the second contention slot if and only if there was no successes in either the first or second contention slots. In this example, two contention slots witness $W' : W = W'$ at the end of their respective times. Intuitively, the contention slots that witness $W' : W = W'$ must be lined consecutively, not sporadically at the beginning of the reservation period. Generalizing, for a frame with τ contention slots, a number of k contention slots at the beginning of the reservation period witnesses $W' : W = W'$ with probability equals to the product of the probabilistic events that (a) zero successes occur over the first k contention slots (b) a success occurs in the $(k+1)^{\text{th}}$ slot and (c) any number of successes occur in the last $(\tau - k - 1)$

contention slots, an event that has a probability of one. Therefore, conditioning on the state of the process at the beginning of the frame, the probability of having K contention slots witness W' : $W = W'$ at the end of their respective times is given by,

$$P(K = k | B = b, W = w) = \begin{cases} \sum_{b'=0}^{M} \Psi_2 P^{(k)}_{(b,0),(b',0)} \sum_{b''=b'-1}^{b'} \Psi_2 P_{(b',0),(b'',1)} & 0 \le k < \tau \\ \sum_{b'=0}^{M} \Psi_2 P^{(\tau)}_{(b,0),(b',0)} & k = \tau \end{cases}$$

Unconditioning on B and taking expectation, I find the expected number of contention slots that witness W' : $W = W'$ at the end of the slot time given W at the beginning of the frame

$$E[K | W = w] = \sum_{b=0}^{M} \pi_b \left[\sum_{k=1}^{\tau-1} k \sum_{b'=0}^{M} \Psi_2 P^{(k)}_{(b,0),(b',0)} \sum_{b''=b'-1}^{b'} \Psi_2 P_{(b',0),(b'',1)} + \tau \sum_{b'=0}^{M} \Psi_2 P^{(\tau)}_{(b,0),(b',0)} \right]$$

So I can compute the time proportionality $\gamma_c(W' : W = W')$ of having W' at end of a contention slot given W at the beginning of the frame where $W = W'$ as,

$$\gamma_c(W' : W = W') = \frac{T_{CS}}{T_{frame}} \sum_{b=0}^{M} \pi_b \left[\sum_{k=1}^{\tau-1} k \sum_{b'=0}^{M} \left(\Psi_2 P^{(k)}_{(b,0),(b',0)} \sum_{b''=b'-1}^{b'} \Psi_2 P_{(b',0),(b'',1)} \right) + \tau \sum_{b'=0}^{M} \Psi_2 P^{(\tau)}_{(b,0),(b',0)} \right] \quad (5.7)$$

Case 2: $W = W'$ at the end of a data slot

I would like to calculate the time proportionality of being in state W' at the end of a data slot given W waiting BWRs in the beginning of the frame where $W = W'$. Let the

number of successes over the reservation period of the frame be s. A data slot shall witness $W' : W = W'$, where $W > 0$, at the end of its time if and only if the following conditions apply (a) $s \geq 1$. (b) $\varepsilon \geq s$. If these two conditions apply, then only one data slot shall witness $W' : W = W'$ at the end of its time. The conditional probability of this event is given by,

$$P\left(K = 1 \middle| B = b, W = w\right) = \sum_{s=0}^{\min(\tau,\varepsilon)} \sum_{b'=0}^{M} \Psi_2 P^{(\tau)}_{(b,0),(b',s)} \qquad W > 0$$

On the other side, more than one data slot shall witness $W' : W = W'$ at the end of its time if $W = 0$ and all BWR successes from the reservation period are transmitted before the end of the current frame.

$$P\left(K = \varepsilon - k \middle| B = b, W = w\right) = \sum_{b'=0}^{M} \Psi_2 P^{(\tau)}_{(b,0),(b',k)} \qquad W = 0$$

Unconditioning on B and taking expectation for both cases,

$$E\left[K \middle| W = w\right] = \sum_{b=0}^{M} \pi_b \sum_{s=0}^{\min(\tau,\varepsilon)} \sum_{b'=0}^{M} \Psi_2 P^{(\tau)}_{(b,0),(b',s)} \qquad W > 0$$

$$E\left[K \middle| W = w\right] = \sum_{b=0}^{M} \pi_b \sum_{k=0}^{\min(\tau,\varepsilon)} (\varepsilon - k) \sum_{b'=0}^{M} \Psi_2 P^{(\tau)}_{(b,0),(b',k)} \qquad W = 0$$

Hence, the time proportionality $\gamma_d(W' : W = W')$ of having W' at end of a data slot given W at the beginning of the frame where $W = W'$ is given by

$$\gamma_d(W' : W = W') = \begin{cases} \dfrac{T_{DS}}{T_{frame}} \sum_{b=0}^{M} \pi_b \sum_{k=1}^{\min(\tau,\varepsilon)} \sum_{b'=0}^{M} \Psi_2 P^{(\tau)}_{(b,0),(b',k)} & W > 0 \\ \dfrac{T_{DS}}{T_{frame}} \sum_{b=0}^{M} \pi_b \sum_{k=1}^{\min(\tau,\varepsilon)} (\varepsilon - k) \sum_{b'=0}^{M} \Psi_2 P^{(\tau)}_{(b,0),(b',k)} & W = 0 \end{cases} \quad (5.8)$$

Case 3: $W < W'$ at the end of a contention slot

In this case I would like to calculate the time proportionality of being in state W' at the end of a contention slot given W waiting BWRs in the beginning of the frame where $W < W'$. Let $\theta = W' - W : \theta > 0$. In order to have $W' : W < W'$ at the end of a contention slot, there must have been θ successes somewhere over the reservation period where $\theta < \tau$ or otherwise the transition is invalid. Depending on where there were θ successes along the reservation period, the number of contention slots that witness $W' : W < W'$ at the end of their respective times can be determined. For τ contention slots where $\tau > \theta$, k contention slots, where $k < \tau$, shall witness $W' : W < W'$ at the end of their respective times if and only if the following conditions apply (a) $\tau \geq \theta + k - 1$. (b) the last success of θ successes occurs in the first one of the k contention slots under consideration. (c) no successes occur over the last $(k-1)^{\text{th}}$ contention slots under consideration. (d) a success occurs at the end of the contention slot following the k$^{\text{th}}$ contention slot that witnesses $W' > W$. These k contention slots may be in the middle of or at the end of the reservation period. So the probability of having k contention slots witnessing $\theta = W' - W : \theta > 0$ waiting BWRs at the end of their respective times is given by

$$P\left(K=k \middle| B=b, W=w\right)=\left[\sum_{b'=0}^{M} \Psi_2 P^{(\tau-k)}_{(b,0),(b',\theta-1)} \sum_{b''=0}^{M} \Psi_2 P_{(b',0),(b'',1)} \sum_{b'''=0}^{M} \Psi_2 P^{k-1}_{(b'',0),(b''',0)} + \right.$$
$$\left. \sum_{\beta=\theta-1}^{\tau-k-1} \sum_{b'=0}^{M} \Psi_2 P^{(\beta)}_{(b,0),(b',\theta-1)} \sum_{b''=0}^{M} \Psi_2 P_{(b',0),(b'',1)} \sum_{b'''=0}^{M} \Psi_2 P^{(k-1)}_{(b'',0),(b''',0)} \sum_{b^\circ=0}^{M} \Psi_2 P_{(b''',0),(b^\circ,0)} \right]$$

Unconditioning on B and taking expectations, I get the expected number of contention slots at the end of which there are W' waiting BWRs given W waiting BWRs at the beginning of the frame where $W < W'$.

$$E\left[K \middle| W=w\right]=\sum_{b=0}^{M} \pi_b \sum_{k=1}^{\tau-\theta-1} k \left[\sum_{b'=0}^{M} \Psi_2 P^{(\tau-k)}_{(b,0),(b',\theta-1)} \sum_{b''=0}^{M} \Psi_2 P_{(b',0),(b'',1)} \sum_{b'''=0}^{M} \Psi_2 P^{k-1}_{(b'',0),(b''',0)} + \right.$$
$$\left. \sum_{\beta=\theta-1}^{\tau-k-1} \sum_{b'=0}^{M} \Psi_2 P^{(\beta)}_{(b,0),(b',\theta-1)} \sum_{b''=0}^{M} \Psi_2 P_{(b',0),(b'',1)} \sum_{b'''=0}^{M} \Psi_2 P^{(k-1)}_{(b'',0),(b''',0)} \sum_{b^\circ=0}^{M} \Psi_2 P_{(b''',0),(b^\circ,0)} \right]$$

Hence, the time proportionality $\gamma_c\left(W' : W < W'\right)$ of having W' at end of a contention slot, given W at the beginning of the frame where $W < W'$ is given by

$$\gamma_c\left(W' : W < W'\right)=\frac{T_{CS}}{T_{frame}} \sum_{b=0}^{M} \pi_b \sum_{k=1}^{\tau-\theta-1} k \left[\sum_{b'=0}^{M} \Psi_2 P^{(\tau-k)}_{(b,0),(b',\theta-1)} \sum_{b''=0}^{M} \Psi_2 P_{(b',0),(b'',1)} \sum_{b'''=0}^{M} \Psi_2 P^{k-1}_{(b'',0),(b''',0)} + \right. \quad (5.9)$$
$$\left. \sum_{\beta=\theta-1}^{\tau-k-1} \sum_{b'=0}^{M} \Psi_2 P^{(\beta)}_{(b,0),(b',\theta-1)} \sum_{b''=0}^{M} \Psi_2 P_{(b',0),(b'',1)} \sum_{b'''=0}^{M} \Psi_2 P^{(k-1)}_{(b'',0),(b''',0)} \sum_{b^\circ=0}^{M} \Psi_2 P_{(b''',0),(b^\circ,0)} \right]$$

Case 4: $W < W'$ at the end of a data slot

In this case I would like to calculate the time proportionality of being in state W' at the end of a data slot given W waiting BWRs in the beginning of the frame where $W < W'$. Let $\theta = W' - W : \theta > 0$. Also let the number of successes over the reservation period of the frame be s. Because of the data transmission over every data slot, only one

data slot may have W' : $W < W'$ waiting BWRs at the end of its time. A data slot would have W' : $W < W'$ waiting BWRs at the end of its time starting with W BWRs at the beginning of the frame if and only if the following conditions apply, (a) $\theta \leq \tau - 1$. (b) $\theta \leq s - 1$. (c) $\varepsilon \geq s - \theta$. This multi conditional event occurs with a conditional probability given by

$$P\left(K = 1 \middle| B = b, W = w\right) = \sum_{s=\theta+1}^{\min(\tau,\varepsilon+\theta)} \sum_{b'=0}^{M} \Psi_2 P^{(\tau)}_{(b,0),(b',s)}$$

Unconditioning on B and taking expectations, I get the expected number of data slots at the end of which there are W' waiting BWRs given W waiting BWRs at the beginning of the frame where $W < W'$.

$$E\left[K \middle| W = w\right] = \sum_{b=0}^{M} \pi_b \sum_{s=\theta+1}^{\min(\tau,\varepsilon+\theta)} \sum_{b'=0}^{M} \Psi_2 P^{(\tau)}_{(b,0),(b',s)}$$

Hence, the time proportionality $\gamma_d\left(W' : W < W'\right)$ of having W' at end of a data slot, given W at the beginning of the frame where $W < W'$ is given

$$\gamma_d\left(W' : W < W'\right) = \frac{T_{DS}}{T_{frame}} \sum_{b=0}^{M} \pi_b \sum_{s=\theta+1}^{\min(\tau,\varepsilon+\theta)} \sum_{b'=0}^{M} \Psi_2 P^{(\tau)}_{(b,0),(b',s)} \tag{5.10}$$

Case 5: $W > W'$ at the end of a data slot

In this case I would like to calculate the time proportionality of being in state W' at

the end of a data slot given W waiting BWRs in the beginning of the frame where $W > W'$. Starting with W waiting BWRs at the beginning of a frame, I can observe W' to be less than W only at the end of a data slot. Let $\theta = W - W' : \theta > 0$. Also let the number of successes over the reservation period of the frame be s.

Similar to case 4, only one data slot may have $W' : W > W'$ waiting BWRs at the end of its time. A data slot would have $W' : W > W'$ waiting BWRs at the end of its time starting with W BWRs at the beginning of the frame if and only if the following condition applies (a) $\varepsilon \geq s + \theta$. Therefore, a data slot may witness $W' : W > W'$ where $W' > 0$ with probability

$$P\left(K = 1 \middle| B = b, W = w\right) = \sum_{s=0}^{\min(\tau, \varepsilon-\theta)} \sum_{b'=0}^{M} \Psi_2 P^{(\tau)}_{(b,0),(b',s)} \qquad W' > 0$$

Unconditioning on B and taking expectations to compute the expected number of data slots at the end of which $W' : W > W'$ waiting BWRs are observed in the service queue where $W' > 0$,

$$E\left[K \middle| W = w\right] = \sum_{b=0}^{M} \pi_b \sum_{s=0}^{\min(\tau, \varepsilon-\theta)} \sum_{b'=0}^{M} \Psi_2 P^{(\tau)}_{(b,0),(b',s)} \qquad W' > 0$$

Hence, the time proportionality $\gamma_d\left(W' : W > W'\right)$ of having W' at end of a data slot, given W at the beginning of the frame where $W > W'$ and $W' > 0$ is given by

$$\gamma_d\left(W' : W > W'\right) = \frac{T_{DS}}{T_{frame}} \sum_{b=0}^{M} \pi_b \sum_{s=0}^{\min(\tau, \varepsilon-\theta)} \sum_{b'=0}^{M} \Psi_2 P^{(\tau)}_{(b,0),(b',s)} \qquad W' > 0 \qquad (5.11)$$

On the other hand, if $W' = 0$, then a number k of data slots may witness $W' : W > W'$ and $W' = 0$ if and only if $\varepsilon - k + 1 = s + \theta$. Therefore, the probability that k data slots witness $W' : W > W'$ and $W' = 0$ at the end their respective times is

$$P\left(K = k \middle| B = b, W = w\right) = \sum_{b'=0}^{M} \Psi_2 P^{(\tau)}_{(b,0),(b',\varepsilon-k-\theta+1)} \qquad W' = 0$$

Unconditioning on B and taking expectations to compute the expected number of data slots at the end of which $W' : W > W'$ waiting BWRs are observed in the service queue where c

$$E\left[K \middle| W = w\right] = \sum_{b=0}^{M} \pi_b \sum_{b'=0}^{M} \Psi_2 P^{(\tau)}_{(b,0),(b',\varepsilon-k-\theta+1)}$$

Hence, the time proportionality $\gamma_d\left(W' : W > W'\right)$ of having W' at end of a data slot, given W at the beginning of the frame where $W > W'$ and $W' = 0$ is given by

$$\gamma_d\left(W' : W > W'\right) = \frac{T_{DS}}{T_{frame}} \sum_{b=0}^{M} \pi_b \sum_{b'=0}^{M} \Psi_2 P^{(\tau)}_{(b,0),(b',\varepsilon-k-\theta+1)} \qquad W' = 0 \qquad (5.12)$$

Finally, the (all time) steady state probability of being in a state W' can be calculated as a weighted average of (5.7), (5.8), (5.9), (5.10), (5.11), and (5.12) as follows

$$\pi_{W'} = \pi_W \left[\gamma_c(W' : W = W') + \gamma_d(W' : W = W')\right] + \sum_{W<W'} \pi_W \left[\gamma_c(W' : W < W') + \gamma_d(W' : W < W')\right] + \sum_{W>W'} \pi_W \gamma_d(W' : W > W') \qquad (5.13)$$

5.5 Delay and Throughput Calculation

I employ Little's theorem in calculating the BWR contention delay $D_c = E[B]/\mu_B$, the delay spent in contention before successful transmission. Similarly, the data transmission delay $D_t = E[W']/\mu_{W'}$ is the delay spent in the service queue from the time of successful contention to the time of complete transmission. I have $E[B] = \sum_{b=0}^{M} b \cdot \pi_b$, $E[W'] = \sum_{w'=0}^{L} w' \cdot \pi_{w'}$, and μ_B and $\mu_{W'}$ are the rates at which BWRs depart the backlog state and the rate at which packets depart service queue respectively.

First, I calculate μ_B as the time averaged expected number of BWRs that leave the contention phase after successful transmission over a reservation period of size τ in a frame.

$$\mu_B = E[N_{Bs}(\tau)]/T_f$$

Thus $E[N_{Bs}(\tau)]$ excludes BWRs that were successfully transmitted starting from an idle state. A SS who started out the reservation period in the backlog state may toggle between backlog and idle states several times during the same reservation period. Therefore, a SS may produce several successful BWRs, which were originally backlogged, in the same reservation period. Besides the fact that P_a is different than P_r, computation of μ_B becomes intricate. I follow a simpler approach to compute $E[N_{Bs}(\tau)]$ using Markov chains, compared to the combinatorial approach indicated in [14], [35]. I utilize the

Markov process Ψ_3 with state space $\{B: 0 \le B \le M\}$ used in [34], where the state is observed at the beginning of each CS. Using Equations (5.1) and (5.2). The one-step transition probability matrix of Ψ_3 is given by

$$P_{B,B+i} = \begin{cases} Q_a(i,B) & 2 \le i \le M - B \\ Q_a(1,B)\cdot[1 - Q_r(0,B)] & i = 1 \\ \left(\begin{array}{l} Q_a(1,B)\cdot Q_r(0,B) + \\ Q_a(0,B)[1 - Q_r(1,B)] \end{array}\right) & i = 0 \\ Q_a(0,B)\cdot Q_r(1,B) & i = -1 \\ 0 & \text{otherwise} \end{cases} \tag{5.14}$$

The expected number of successful backlogged SSs (BWRs) in one contention slot conditioned on the number of backlogged SSs in the beginning of the slot, denoted $E[N_{Bs}(1)|b]$ is,

$$E[N_{bs}(1)|b] = 0 * P(N_{Bs}(1)=0) + 1 * P(N_{Bs}(1)=1) = P(N_{Bs}(1)=1)$$

I can interpret $P(N_{bs}(1)=1|b)$ as the transition probability from a state b to state $b-1$ $\forall\, b > 0$ in one contention slot or,

$$P(N_{Bs}(1)=1|b) = {}^{\Psi_3}P_{b,b-1}$$

Moreover, the expected number of successful backlogged SSs (BWRs) in τ CSs conditioned on B at the beginning of a frame is

$$E[N_{Bs}(\tau)|B] = E[N_{Bs}(CS_1)|B_1] + E[N_{Bs}(CS_2)|B_2] + \cdots + E[N_{Bs}(CS_\tau)|B_\tau]$$

where $E\left[N_{Bs}(CS_x)\middle| B_x\right]$ is the expected number of successes in the x^{th} CS conditioned on B at the beginning of that slot. Hence, after unconditioning on B

$$E\left[N_{Bs}(\tau)\right]=\sum_{b=0}^{M}\pi_b\left[{}^{\Psi_3}P_{b,b-1}+\sum_{k=1}^{\tau-1}\sum_{b'=0}^{M}{}^{\Psi_3}P_{b,b'}^{(k)}\cdot{}^{\Psi_3}P_{b',b'-1}\right] \tag{5.15}$$

By direct substitutions

$$\mu_B=\frac{1}{T_f}\sum_{b=0}^{M}\pi_b\left[{}^{\Psi_3}P_{b,b-1}+\sum_{k=1}^{\tau-1}\sum_{b'=0}^{M}{}^{\Psi_3}P_{b,b'}^{(k)}\cdot{}^{\Psi_3}P_{b',b'-1}\right] \tag{5.16}$$

$$D_c=\frac{\sum_{b=0}^{M}b\cdot\pi_b}{\frac{1}{T_f}\sum_{b=0}^{M}\pi_b\left[{}^{\Psi_3}P_{b,b-1}+\sum_{k=1}^{\tau-1}\sum_{b'=0}^{M}{}^{\Psi_3}P_{b,b'}^{(k)}\cdot{}^{\Psi_3}P_{b',b'-1}\right]} \tag{5.17}$$

In (5.17), the time unit of the contention delay calculation during the reservation period is the contention slot, whereas the approximate expression in (4.7) considers the BWR success to occur, in average, in the middle of the reservation period. Therefore, the first expression in (5.7) is considered to be a more accurate estimate of the contention delay, though the difference is marginal.

In a similar manner I can calculate $\mu_{W'}=E\left[N_{Ws}\right]/T_f$, where $E\left[N_{Ws}\right]$ is the expected number of transmitted (served) data packets per frame. Typically, N_{Ws} in a frame f can be expressed as

$$N_{Ws_f} = \begin{cases} W_f + S_f & W_f + S_f < \varepsilon \quad \text{if } W_{f+1} = 0 \\ \varepsilon & W_f + S_f \geq \varepsilon \quad \text{if } W_{f+1} > 0 \end{cases} \tag{5.18}$$

In (5.18), based on my early assumption of fixed and equal data packet size for all SSs, the maximum number of departing data packets in a frame is equal to the number of available data slots in that frame. Otherwise, the number of departing data packets is the sum of waiting packets in the beginning of the frame and the number of successful BWRs that join the service queue) in the same frame. Accordingly, the expected number of served data packets in a frame is computed in (5.19)) where the first and second terms in (5.19)) correspond to the first and second cases of (5.18) respectively.

$$E[N_{Ws}] = \sum_{w'=0}^{L} \pi_{w'} \sum_{b=0}^{M} \pi_b \left[\left(\sum_{n_{ws}=w'}^{\min(w'+\tau,\varepsilon)} n_{ws} \sum_{b'=0}^{M} \Psi_2 P^{(\tau)}_{(b,0),(b',n_{ws}-w')} \right) + \left(\sum_{w''=0}^{L} \varepsilon \sum_{b'=0}^{M} \Psi_1 P_{(b,w'),(b',w'')} \right) \right] \tag{5.19}$$

By direct substitution in $\mu_{W'} = E[N_{Ws}]/T_f$ and $D_t = E[W']/\mu_{W'}$

$$\mu_{W'} = \frac{1}{T_f} \sum_{w'=0}^{L} \pi_{w'} \sum_{b=0}^{M} \pi_b \left[\left(\sum_{n_{ws}=w'}^{\min(w'+\tau,\varepsilon)} n_{ws} \sum_{b'=0}^{M} \Psi_2 P^{(\tau)}_{(b,0),(b',n_{ws}-w')} \right) + \left(\sum_{w''=0}^{L} \varepsilon \sum_{b'=0}^{M} \Psi_1 P_{(b,w'),(b',w'')} \right) \right]$$

$$D_t = \frac{\sum_{w'=0}^{L} w' \cdot \pi_{W'}}{\frac{1}{T_f}\sum_{w'=0}^{L}\pi_{w'}\sum_{b=0}^{M}\pi_b\left[\begin{pmatrix}\sum_{n_{ws}=w'}^{\min(w'+\tau,\varepsilon)} n_{ws}\sum_{b'=0}^{M}\Psi_2 P^{(\tau)}_{(b,0),(b',n_{ws}-w')}\end{pmatrix}+\begin{pmatrix}\sum_{w''=0}^{L}\varepsilon\sum_{b'=0}^{M}\Psi_1 P_{(b,w'),(b',w'')}\end{pmatrix}\right]} \quad (5.20)$$

And the total delay is the sum of contention delay and data transmission delay

$$D = D_c + D_t$$

The throughput of this R-MAC protocol is defined as the effective ratio of the frame used for data transmission. Using (5.19), which formulates the expected number of data packets transmitted in a frame, throughput Th is given by $Th = \left(E[N_{ws}]* T_{DS}\right)/T_f$ or

$$Th = \frac{T_{DS}}{T_f}\sum_{w'=0}^{L}\pi_{w'}\sum_{b=0}^{M}\pi_b\left[\begin{pmatrix}\sum_{n_{ws}=w'}^{\min(w'+\tau,\varepsilon)} n_{ws}\sum_{b'=0}^{M}\Psi_2 P^{(\tau)}_{(b,0),(b',n_{ws}-w')}\end{pmatrix}+\begin{pmatrix}\sum_{w''=0}^{L}\varepsilon\sum_{b'=0}^{M}\Psi_1 P_{(b,w'),(b',w'')}\end{pmatrix}\right] \quad (5.21)$$

5.6 Summary

Through this chapter, I have derived an analytical model to formulate the delay and throughput performance of the multiple access control of the standard. I used a two stage Markov chain to track the evolution of the number of backlogged bandwidth requests and the number of waiting packets in the service queue over the time. After calculating the departure rates of both backlogged bandwidth requests and waiting packets, I employed Little's theorem in calculating the average contention delay and average data transmission

delay, the sum of which represents the average total message delay a data packet experiences in the system, and I also calculated the system throughput.

6 Delay and Throughput Performance Evaluation

I conduct an illustrative analysis on a relatively small size IEEE 802.16 reservation MAC, using the proposed analytical model, to reflect on the comparative influence of design parameters. I study a scaled-down reservation MAC system to draw observations on the performance metrics under different designs of the reservation period.

In addition, I validate the analytical model by implementing a simulator of a real reservation MAC system that resembles the mechanism of the IEEE 802.16 multiple access control.

6.1 Simulation Model

I use Matlab to build a discrete time simulation mimicking the real system mechanisms. Time is divided into frames. Each frame is divided into equal number of slots. System time is the slot time. I use M random-number 0/1 generators representing the SS population. In the beginning of each contention slot, an idle SS generates an event (BWR) with probability P_a while a backlogged SS generates an event (backlog BWR

retransmission) with probability P_r. If more than one event has been generated in a contention slot, a collision occurs. An idle SS involved in the collision will join the backlog list where time of the collision (i.e. start of backlog state) is registered. A backlogged SS involved in the collision will continue in the backlog state.

A successful BWR transmission on the other hand occurs in a contention slot if and only if only one event has been generated by any SS (idle or backlogged) in that slot. At the end of the contention slot, a successful BWR enters a waiting queue (i.e. service queue at the bases station in the actual system). If the successful BWR was originally backlogged, the end time of the successful contention slot marks the end of backlog state and also the start of queue waiting time.

As time goes by, at the beginning of the data transmission period, the BWR at the head of the queue departs when a data slot is available in a First Come First Serve (FCFS) basis. The end of the slot time marks the end of data transmission time. All records pertaining to the BWR generation time, start of backlog state time, end of backlog state time, and start and end of data transmission are recorded throughout the simulation. The simulation is run for a long time. At the end of the run time, the average contention delay and backlog delays (the sum of which is the average total message delay) as well as system throughput are calculated as follows,

$$\text{Average contention delay} = \frac{\sum \text{Contention delay of all backloggd BWRs throughout simulation}}{\text{Number of backlogged packets througout simulation}}$$

$$\text{Average queuing delay} = \frac{\sum \text{Queuing time of all queued BWR's throughout simulation}}{\text{Number of queued BWR's througout simulation}}$$

$$\text{Throughput} = \frac{\sum \text{Data slot time used for data transmission}}{\text{Overall simulation time}}$$

6.2 Numerical Experiments

6.2.1 Experiment 1: Intense BWR Arrival Rate

I assume a system with a number of subscriber stations $M = 15$, maximum service queue size of $L = 30$, frame time duration of $T_f = 1\text{ms}$, and number of slots per frame of $N_{MS} = 33$. A contention slot (CS) comprises one slot whereas a data slot (DS) comprises three slots. Different operating points can be studied using the proposed model. Of special interest is the case of severe contention. To this end, I take $P_a = 0.5$ reflecting intense BWRs arrival, and $P_r = 0.04$ – a relatively small retransmission probability given the size of the SS population. For reservation period size τ varying from 3 to 18 slots where increments are in three slots, identical delay and throughput results are obtained from the proposed analytical model and simulation as shown in Figure 6.1 to Figure 6.4. I study the system under stable operating points. Therefore, I avoid studying points at which $\tau > 18$ where the number of successful BWRs builds up over time and destabilizes the system. Figure 6.1 and Figure 6.2 show that increasing the size of the reservation period, τ, improves the contention delay at the cost of data transmission delay. Both types of delay take part in constituting the total message delay as shown in Figure 6.3. The best operating points from the view point of delay and system throughput vary with the change in BWR arrival rate P_a, retransmission rate P_r, and number of SSs, M. In order to calculate the best operating point from the view point of any of the design

parameters P_a, P_r, M or τ, three of them must be fixed to find the best value of the fourth. For example, for $M = 15$, $L = 30$, $P_a = 0.5$, and $\tau = 9$, message delay is best at $P_r = 0.25$ as shown in Figure 6.5 whereas $P_r = 0.04$ is best for throughput or utilization as shown in Figure 6.6. This is because a value of $P_r = 0.25$ creates more retransmission occurrences than does $P_r = 0.04$, which speeds up the release of backlogged BWRs. However at the same time, the increased contention with new BWRs results in a lower throughput than in the case when $P_r = 0.04$.

From a system capacity perspective, increasing M results in monotonically increasing the total message delay but in the same time results in a retraction in system throughput at a value of $M = 12$ where best throughput is attained as shown in Figure 6.7. Following a constant reservation period size allocation policy, Figure 6.3 shows $\tau = 12$ to achieve the best overall delay performance for the given traffic rate. However, $\tau = 15$ achieves the best throughput performance as shown in Figure 6.4.

The throughput-delay performance in Figure 6.8 shows that up to some point the total message delay improves with throughput increase, after which the delay starts to increase again and throughput to decrease. Since BWR arrival rate $P_a = 0.5$ is relatively high with regard to the reservation period size, the main constituent of the overall message delay is contention delay rather than transmission delay. As the reservation period size increases, contention delay is eased letting more BWRs get into service queue, which increases system throughput. The converse occurs at $\tau \geq 12$ where the DS period gets relatively smaller in size resulting in decreased system throughput. In this range, the increase in message delay is attributed to excessive data transmission delay.

6.2.2 Experiment 2: Relaxed BWR Arrival Rate

In another example, the intensity of BWR arrivals is relaxed by setting $P_a = 0.1$. I notice from Figure 6.3 that the total message delay performance is improved than in the case where $P_a = 0.5$. However contrary to intuition, the best delay performance for $P_a = 0.1$ occurs at a larger reservation period ($\tau = 15$) compared to $\tau = 12$ for $P_a = 0.5$. An explanation is that, while relaxing the intensity of BWR arrivals reduces message delay at all points of τ, less BWRs are produced than in the case of $P_a = 0.5$. A supporting indication can be noticed in Figure 6.4 where system throughput with $P_a = 0.5$ is better than that with $P_a = 0.1$. As a result, most of the message delay with $P_a = 0.1$ comes from contention rather than from queuing. Hence, the best delay performance occurs at a bigger size reservation period with lighter BWRs arrival rate.

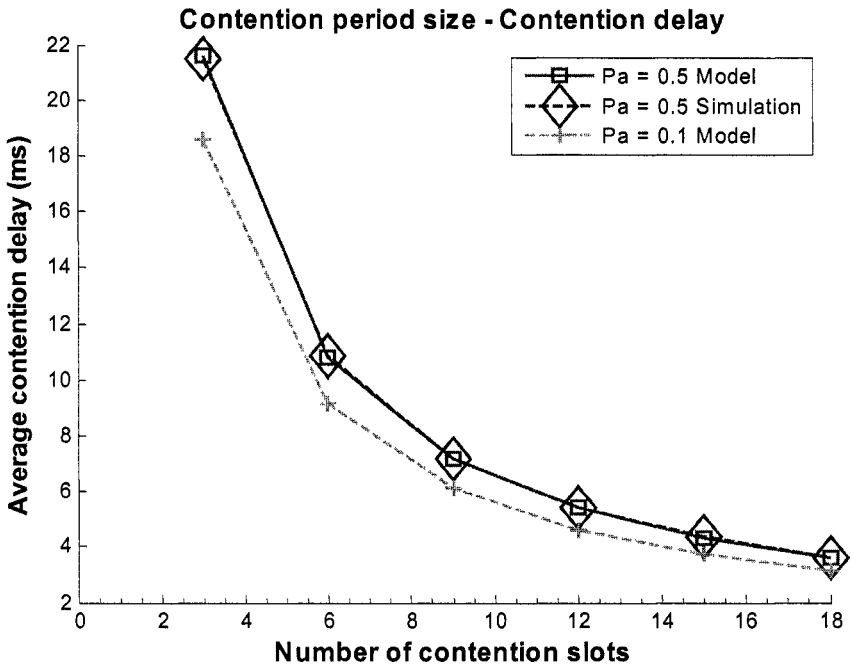

Figure 6.1 Reservation period size vs. contention delay

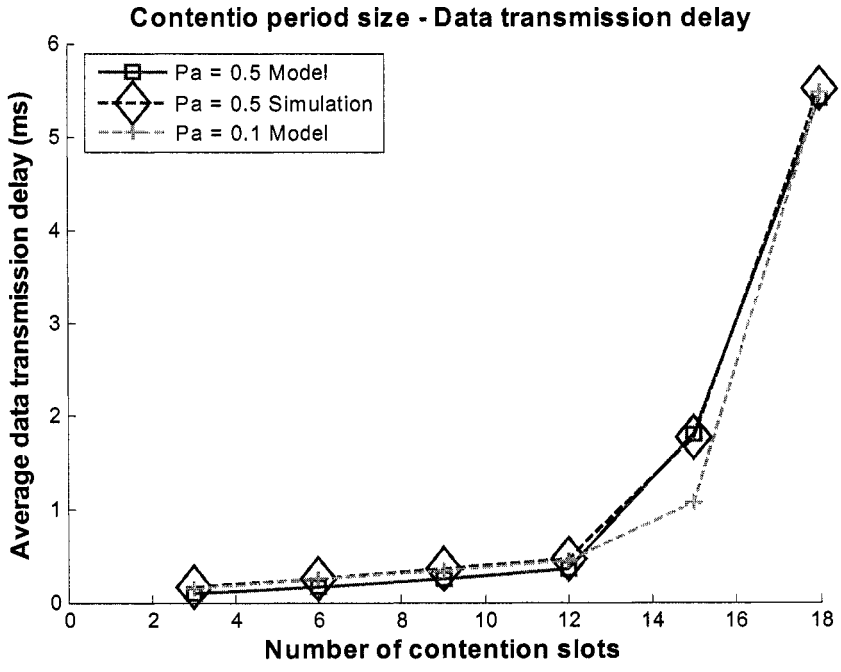

Figure 6.2 Reservation period size vs. data transmission delay

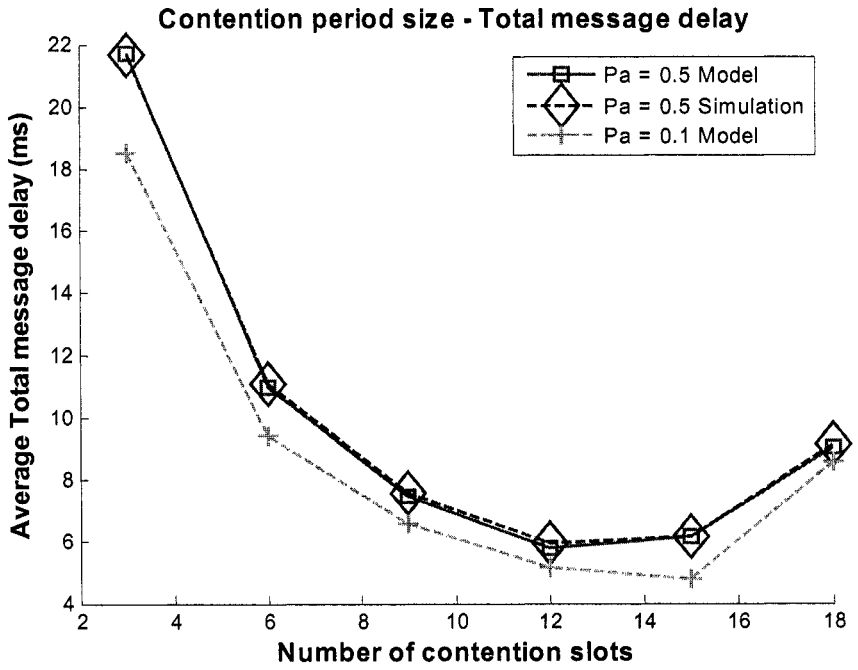

Figure 6.3 Reservation period size vs. Total message delay

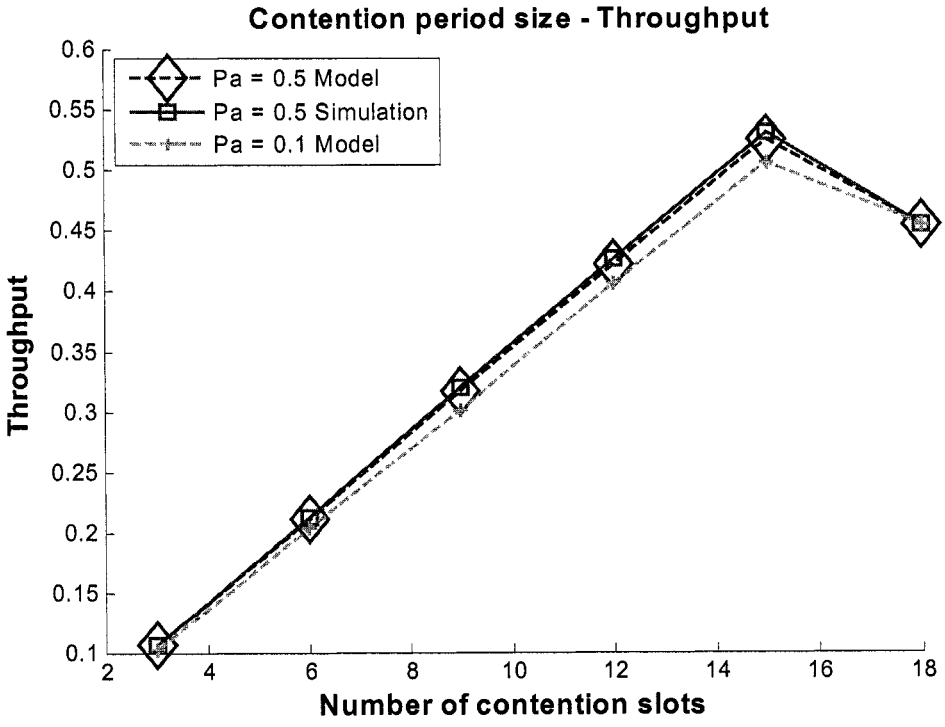

Figure 6.4 Reservation period size vs. system throughput

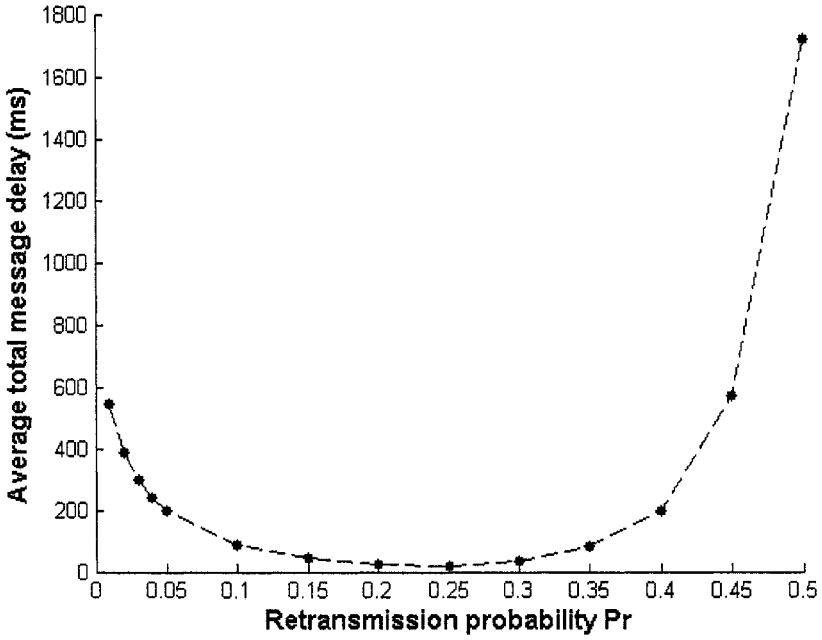

Figure 6.5 Retransmission probability vs. total message delay

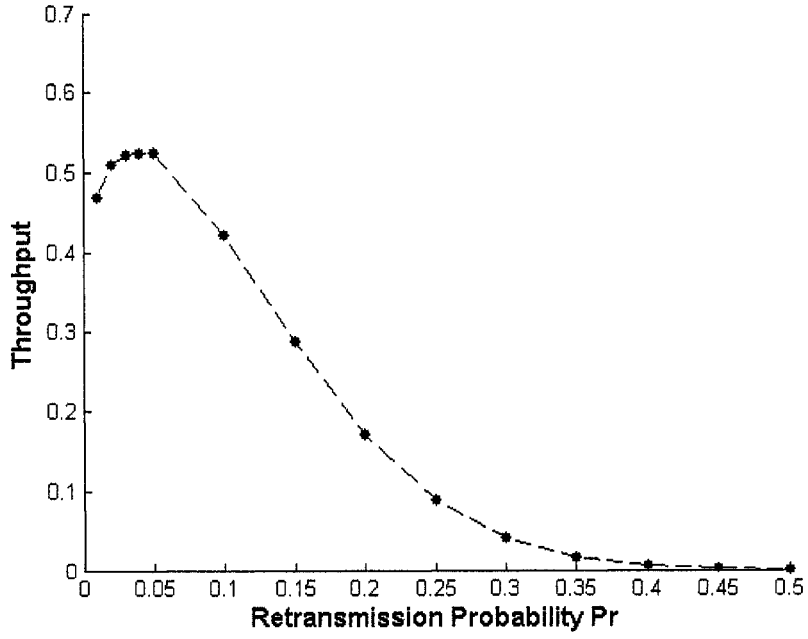

Figure 6.6 Retransmission probability vs. system throughput

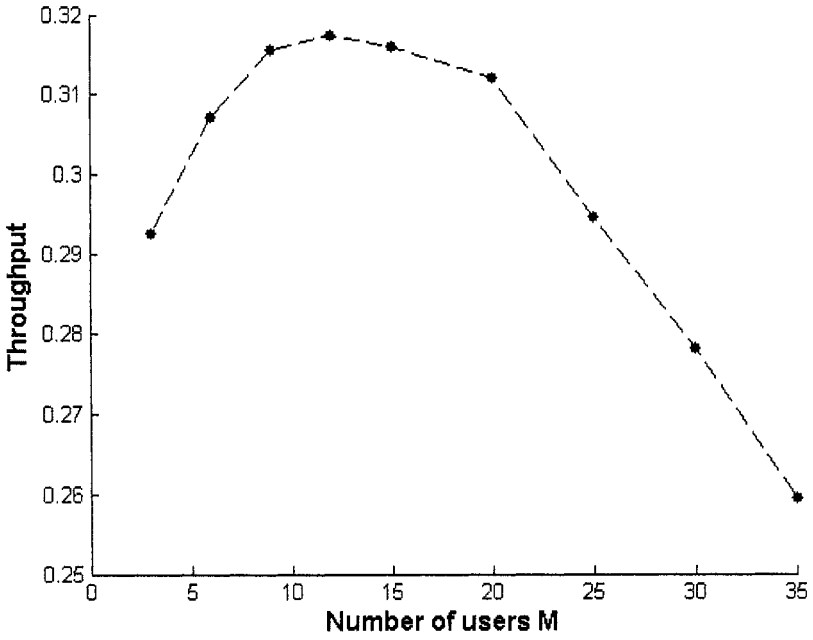

Figure 6.7 Number of SSs vs. system throughput
$P_a = 0.5$ $P_r = 0.04$ $\tau = 9$

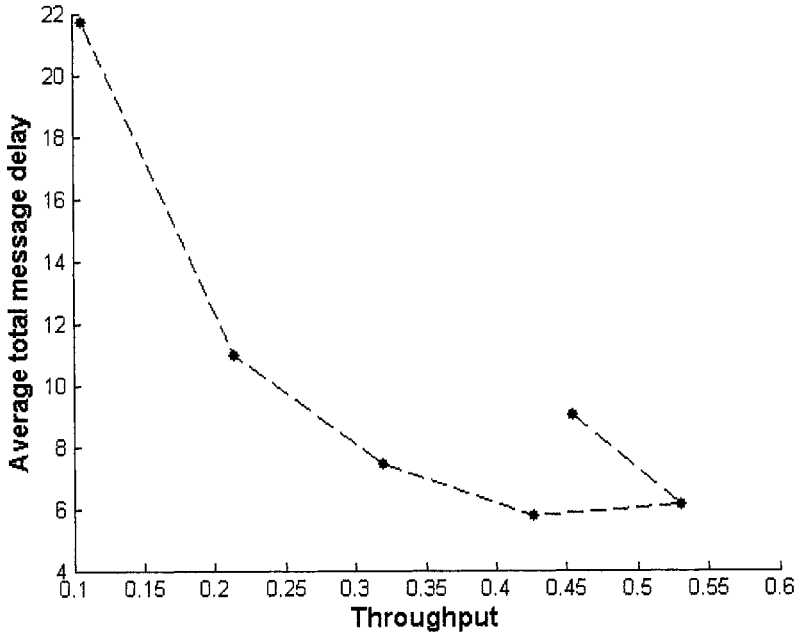

Figure 6.8 Throughput – delay curve

6.3 Discussion

As have been shown through the study of different BWR arrival rates, the reservation period size plays a centric role in tuning the R-MAC to a better performance. Increasing τ has the effect of reducing contention delay, letting contending BWRs into service queue more quickly. This may leverage system throughput on the long run, but in the same time may cause higher queuing delay due to smaller size DS period. Adequately wide reservation period is desirable when P_a is high. On the other hand, as data transmission delay increases, a bigger DS period is desirable to serve waiting packets. To this end, the contention delay may excessively grow due to small size reservation period. With the irregular traffic inherent in wireless and mobile networks, an optimized adaptive CS allocation mechanism is needed to tune the R-MAC to a point where optimal resource utilization and packet delay are attained.

6.4 Summary

Through performance evaluation, important insights have been gained on the tradeoff in contention resources and data transmission resources allocation. Based on the analytical model presented in Chapter 5, I calculated the delay and throughput of the system and showed their behavioral curves.

The analysis of the reservation MAC system in the IEEE 802.16 shows that the delay and throughput performance varies with the change of the reservation request arrival rate. Moreover, I showed that the main control of the protocol performance is largely due to the size of the reservation period. Efficient design of the reservation period requires knowledge of the reservation request arrival rate among other parameters. The change in

the reservation request arrival rate over time requires adaptation of the reservation period size over time to maintain efficient resources utilization and acceptable delay performance. Adapting the reservation period size shall allow for opportunistic delay and throughput improvements.

According to the standard, the base station has the flexibility to allocate resources dynamically at times of the beginning of each frame. Therefore, a dynamically optimized reservation period size needs to be calculated at these times to allow for opportunistic performance evaluation.

7 Markov Decision Process Optimization Model

Most of today's broadband networks, including the recently ratified IEEE 802.16 standard, employ reservation based multiple media access control (R-MAC). The R-MAC frame is divided into two portions; the reservation period, which is divided into contention slots (CSs) and the service period, which is divided into data slots (DSs) as shown in Figure 7.1.

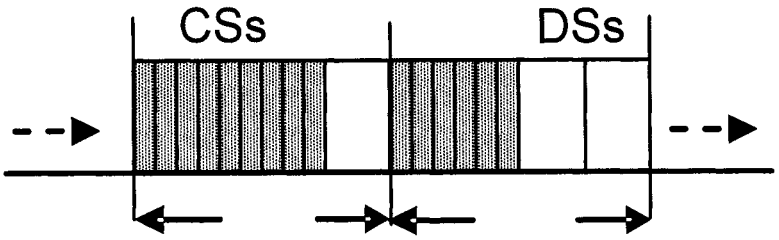

Figure 7.1 Example frame structure of R- MAC system

As discussed earlier, a problem pertinent to reservation MAC protocols is the division of frame slots between the contention and data transmission processes. In most of the reservation MAC protocols, no specific ratio is standardized, leaving proprietary solutions address the local network environment. As both processes are equally important for

maintaining efficient delay and throughput performance, a solution must consider the timely varying traffic load. For example, heterogeneity and cooperation of networks promote access technologies that can sustain waves of increasing traffic load. In this Chapter I start by instituting a framework for efficient allocation of frame resources to the contention and data transmission processes in light of the delay and throughout performance. I then propose a dynamic resource allocation controller based on a Markovian optimization model, where the optimization parameters are tuned according to specific preferential criteria of service providers. I use the transition probabilities and steady state probabilities derived in Chapter 5 in formulating the optimization model. The proposed model achieves opportunistic performance improvements, on a per frame basis, over the best case static allocation. Through simulation, I study the merits of the proposed optimized controller with respect to the proposed framework. I show by illustrative examples and numerical results that the controller successfully fulfills the framework objectives.

7.1 Reservation Period Allocation Controller: Framework

The reservation MAC protocol in most of the broadband technologies, i.e. DSL, HFC cable, GPRS and the IEEE 802.16 standard, is centralized. The base station, which is the central controller, broadcasts the organization of the frames in terms of the start time of the contention opportunities and data slots to the subscriber stations on the downlink portion of the frame. The ratio of the contention and service resources is not necessarily the same in each frame. This ratio is controlled by the base station in the beginning of each frame. Based on input traffic load as well as system performance information, a

controller residing at the base station determines this ratio with an objective to optimize the resources utilization. A block diagram of the proposed reservation period allocation controller is shown in Figure 7.2. In what follows, I describe the ideal input information and characteristics of an optimized resources allocation controller for the R-MAC protocol.

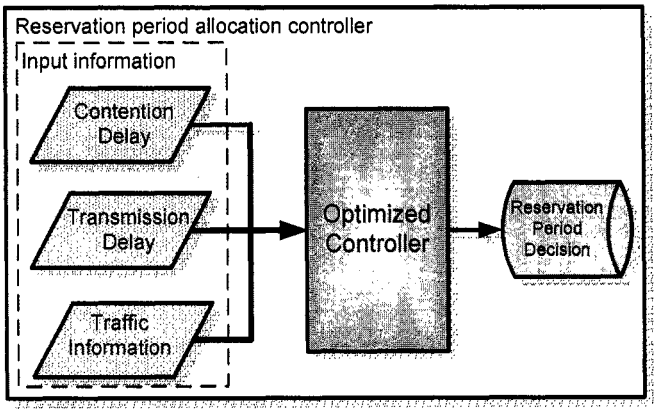

Figure 7.2 Block diagram of the proposed framework of the reservation period controller.

7.1.1 Input Information

The performance metrics of an R-MAC protocol are the packet delay and system throughput. The packet delay is defined as the delay spent from the time of initiating the reservation request until the complete transmission of the associated data session. The system throughput is defined as the percentage of the system time used for data packets transmission. Ultimately, the controller works to optimize the delay and/or throughput of the system. Since the time instances where most of the control information is exchanged between the base station (BS) and subscriber stations (SSs) are at the starting points of the frames, the optimized controller should also be triggered at such points. In other words, at

the beginning of a frame, the controller is triggered to decide on the ratio between the contention and service periods that optimize the throughput and/or delay in that frame. Ideally, this decision requires input knowledge as follows; **a)** the average contention delay of the reservation requests which are currently in contention, **b)** the average data transmission delay of the associated data packets waiting in the service queue and **c)** the reservation requests arrival characteristics (traffic load). This information should influence the decision taken by the optimized controller in a way that rectifies throughput and/or delay performance deterioration.

7.1.2 Optimized Controller Design

The purpose of the controller, as previously stated, is to opportunistically improve the system performance. Any prospective gain in performance would be counterbalanced if the controller's operation presented burdensome constraints to the process. Therefore, the controller must possess a set of characteristics to enhance rather than saddle the performance. I state the desirable characteristics of the optimized controller as follows; **a)** the decision time is infinitesimally small, **b)** the decision is adaptive to performance and traffic conditions and **c)** the controller is tunable to delay and throughput performance improvement.

7.2 Implementation of Reservation Period Allocation Controller

The methodical realization of the framework components, illustrated in Figure 7.2, can have many facets. In my proposal, based on an optimization method, the optimized

controller processes the input information in order to reach a decision (i.e. size of the reservation period), at the beginning of the frame, which leads to opportunistically achieve performance gains. I am focused on investigating the merits of Markov Decision Processes (MDP) as the core component of the optimized controller. MDP is a useful dynamic optimization tool that is based on Markov processes and is often used in control theory and more generally in decision theory. Reference [37] is a compact reference on MDP optimization models. In this Section I will start by setting up the Markov process of the R-MAC system shown in Figure 7.1. Using the state transition probabilities of the frame Markov process described in Chapter 5, I formulate the MDP model. By solving the optimization problem, I reach the optimal decisions (set of reservation period sizes). In Section 7.2.1 I shall briefly reflect on the realization of the input information components shown in Figure 7.2, and in Section 7.2.2 I will describe the proposed optimization model.

7.2.1 Input Information Realization

In this Section I investigate the enabling mechanisms to collect contention delay, data transmission delay and traffic information.

7.2.1.1 Contention Delay Information

In the R-MAC protocols employed in DSL, HFC, GPRS, and IEEE 802.16 standard, the base station has no means to evaluate the contention delay suffered by the BWRs that are contending for the media. I use the number of BWRs that are in contention, denoted B, as a measure of the contention delay. References [25] and [26] apply estimation

techniques to compute an average value of B since the BS is incapable of directly recognizing the contending SSs.

7.2.1.2 Data Transmission Delay

Accurate information about the number of and delay experienced by data packets waiting in the BS's service queue is available at the BS. In a similar manner to the contention delay information, I use the number of waiting data packets, denoted W, as a measure of the data transmission delay.

7.2.1.3 Traffic Information

I characterize the BWR traffic by the probability that an SS transmits an BWR in an CS, and I denote it by P_a. Since the base station is incapable of realizing the traffic characteristics of BWR arrivals, average estimation techniques like pseudo-Bayesian estimator [34] can be employed in order to estimate P_a at the beginning of the frame.

7.2.2 MDP Optimization Model

As outlined in Section 7.1.1, the input information necessary for the operation of the optimized controller at the beginning of the frame includes the contention delay, data transmission delay, and traffic information. I utilize Markov Decision Processes (MDP) in implementing the proposed optimized controller. In MDP, assuming a Markov process with a countable state space S, after observing the state of the process a decision must be chosen. Assume A, which is finite, be the set of all possible decisions. If the process is in

state i, where $i \in S$, at time f, where $f = 0, 1, 2, \cdots$ and decision a, where $a \in A$, is chosen, then the following occur:

a) A reward $R(i,a)$ is gained.

b) The next state will be chosen according to the transition probabilities $P_{ij}(a)$ where $j \in S$.

Thus, both the rewards and the transition probabilities are functions only of the current state and the subsequently chosen decision. I examine the Markovian property of the R-MAC system shown in Figure 7.1. In Chapter 5 I have formulated the Markov process Ψ_1 with state space $\{(B(f),W(f)): \quad f \geq 0\}$, which describes the operation of the R-MAC system through the system state (B,W). Moreover, since the rate of BWRs success and data packets service depend primarily on the size of the reservation period denoted τ, then the state transition in Ψ_1 is a function of the current state (B,W) and size of the reservation period τ. Furthermore, as the frame has a finite number of slots, the value of τ has a limited domain. Based on these characteristics, the operation of the R-MAC system resembles to a great extent the dynamics of a Markov Decision Process model. I am now ready to formulate the MDP optimization model.

7.2.2.1 MDP Optimization Problem Formulation

A policy determines the rule used to choose a decision i.e. size of the reservation period, should the system be in a specific state [37]. Policies may differ according to the local environment and objectives of service providers. Truly it is the reason why a

reservation period allocation policy has not been endorsed by any of the standards employing R-MAC protocols. If I let X_f denote the state of the process at time f (at the beginning of frame f) and τ_f be the decision (size of reservation period) chosen at time f, then the transition probability from state $u=(B_f,W_f)$ to state $v=(B_{f+1},W_{f+1})$ is given by

$$P\{X_{f+1}=v \mid X_0,\tau_0,X_1,\tau_1,\cdots\cdots,X_f=u,\tau_f=\tau\}=P_{u,v}(\tau)$$

For any policy β, the limiting probability π_u^τ that the process will be in state u and decision τ will be chosen is given by

$$\pi_u^\tau=\lim_{f\to\infty}P_\beta\{X_f=u,\tau_f=\tau\} \tag{7.1}$$

where π_u^τ must satisfy the following

a) $\pi_u^\tau \geq 0 \quad \forall\, u,\tau$

b) $\sum_u\sum_\tau \pi_u^\tau = 1$

c) $\sum_\tau \pi_v^\tau = \sum_u\sum_\tau \pi_u^\tau P_{u,v} \quad \forall\, v$

where $P_{u,v}$ is taken from ${}^{\Psi_1}P$ derived in Chapter 5. When the system is in state $u=(B_f,W_f)$ and decision τ_f is taken, a reward $R(u,\tau_f)$ is achieved. Using π_u^τ, I can calculate the steady state expected reward as follows

$$E[R] = \lim_{f\to\infty} E\left[R(u_f,\tau_f)\,\middle|\,u_f,\tau_f\right] = \sum_u \sum_\tau \pi_u^\tau\ R(u,\tau) \tag{7.2}$$

Consequently, the optimal policy that maximizes the expected average reward with respect to τ is

$$\underset{\pi=\pi_u^\tau}{\text{maximize}} \sum_u \sum_a \pi_u^\tau R(u,\tau) \tag{7.3}$$

subject to the following conditions,

$$\sum_a \pi_u^\tau = \sum_v \sum_\tau \pi_v^\tau P_{v,u}(\tau)$$

$$\sum_u \sum_\tau \pi_u^\tau = 1$$

$$\pi_u^\tau \geq 0$$

where the policy P_u^τ (the probability of taking decision τ when the process is in state u) is computed from the following equalities

$$\pi_u^\tau = \pi_u P_u^\tau$$

$$\sum_\tau \pi_u^\tau = \pi_u$$

The maximum average reward can be achieved by a nonrandomized policy [36], i.e. the decision that must be taken when in a state u is a deterministic function of u. I now need to design a reward function such that controlled state transitions yield performance enhancement.

7.2.2.2 Reward Function

The reward function $R(u,\tau)$ is a function of the current state $u=(B,W)$ and the subsequently chosen size of the reservation period τ. The reward function should be designed such that desirable state transitions result in higher rewards than undesirable transitions. After observing the state at the beginning of a frame to be $u=(B,W)$, the reward function shall be calculated for all possible values of the decision – which corresponds to the size of the reservation period. Finally, the chosen decision is the one that achieves the highest return of the reward function. According to the framework in Section 7.1, the choice of a reservation period should consider improving the performance opportunistically. In order to comply with this requirement, and since throughput performance is inseparable from the delay performance, I use a multi-objective reward function structure. I let $R_D(u,\tau)$ denote the delay objective function and $R_{th}(u,\tau)$ denote the throughout objective function. Ultimately, $R(u,\tau)$ is a parametric function of $R_D(u,\tau)$ and $R_{th}(u,\tau)$. For this multi-objective optimization, I choose the aggregation function technique [38] to aggregate all the objectives into a single function using a form of weighted sum as follows

$$R(u,\tau)=g_D\cdot R_D(u,\tau)+g_{th}\cdot R_{th}(u,\tau) \tag{7.4}$$

where $0\le g_D\le 1$ and $0\le g_{th}\le 1$ are the weighting coefficients representing the relative influence of the delay and throughput components.

7.2.2.2.1 Delay Objective Function

Two parts comprise the delay performance; contention delay and data transmission delay. As I have mentioned, improving one part usually results in deterioration of the other part. Therefore, the two delay components must be separated and represented in the reward function. Thus, Equation (7.4) is slightly expanded as follows,

$$R(u,\tau) = g_c \cdot R_{D_c}(u,\tau) + g_w \cdot R_{D_w}(u,\tau) + g_{th} \cdot R_{th}(u,\tau) \tag{7.5}$$

where $R_{D_c}(u,\tau)$ and $R_{D_w}(u,\tau)$ pertain to contention and data transmission delays respectively, with $0 \le g_c \le 1$ and $0 \le g_w \le 1$ as their respective weight coefficients.

7.2.2.2.1.1 Contention Delay Objective Function

My objective is to reduce the number of BWRs in contention at the beginning of the next frame. Intuitively, the bigger the size of the reservation period is the higher the reduction in B. Meanwhile, the reward function should consider the relative value of B in deciding on the value of τ. Accordingly, I propose the following experimental form of contention delay objective function

$$R_{D_c}(B,\tau) = 1 - \exp\left(\frac{-\tau}{B}\right) \tag{7.6}$$

The proposed function in (7.6) has good characteristics of an objective reward function. For numerical illustrations, I take $B \le 15$, $W \le 30$, and $\tau = \{3,6,9,12,15,18,21,24,27\}$ is the domain of reservation period size. As shown in

Figure 7.3, the proposed objective function establishes desirable reward differentiation according to the combination of B and τ. For a certain value of B, the reward proportionaly icreases with the increase in the size of the reservation period τ.

7.2.2.2.1.2 Data Transmission Delay Objective Function

Similar to the contention delay reward function, the data transmission delay objective reward function is

$$R_{D_w}(W,\tau) = 1 - \exp\left(\frac{-\tau}{W}\right) \tag{7.7}$$

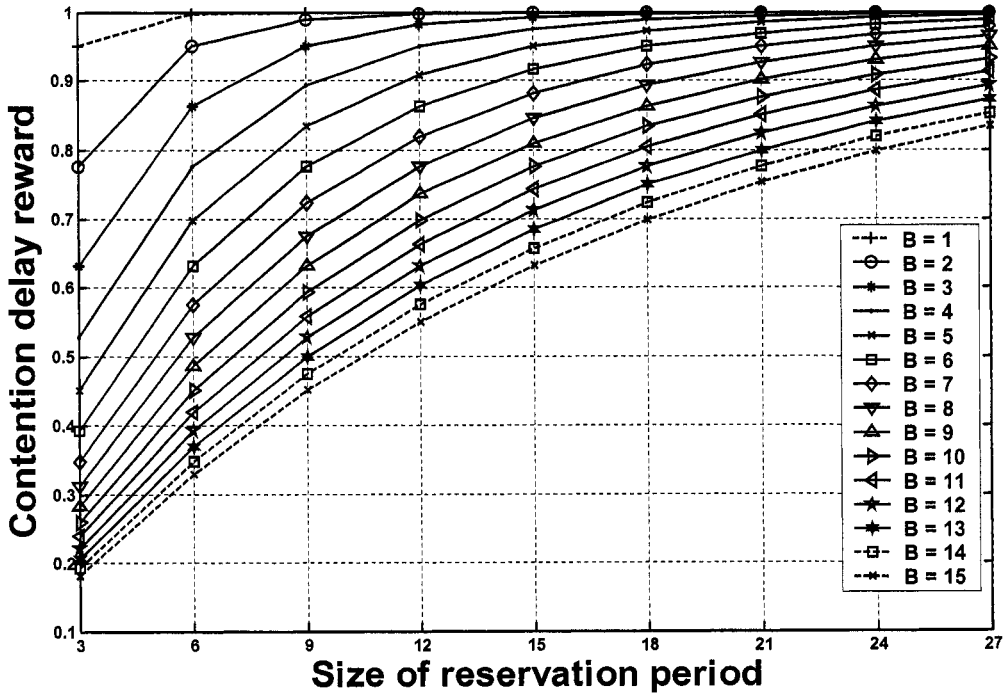

Figure 7.3 Contention delay reward differentiation with different combinations of B and τ

7.2.2.2.2 Throughput Objective Function

Given the state of the process $u=(B,W)$ at the beginning of a frame, I need to calculate the throughput of that frame for a certain size of the reservation period τ. I let $E\left[N_{Ws_f} \mid B,W\right]$ denote the expected number of served data packets in frame f. The frame throughput is defined as the effective fraction of the frame time utilized in data packets transmission. Therefore, the average conditional throughput of frame f, th_f, is given by

$$E\left[th_f \mid B,W\right]=\frac{T_{DS}}{T_f}E\left[N_{Ws_f} \mid B,W\right] \tag{7.8}$$

Typically, N_{Ws_f} is expressed as

$$N_{Ws_f}=\begin{cases} W_f+S_f & W_f+S_f<\varepsilon \\ \varepsilon & W_f+S_f\geq\varepsilon \end{cases} \tag{7.9}$$

In (7.9), based on my early assumption of fixed and equal size of data packets, the maximum number of served data packets in a frame is equal to the number of available data slots in that frame. Otherwise, the number of served data packets is the sum of waiting data packets in the beginning of the frame and the number of data a packets that join the service queue in the same frame i.e. number of successful BWRs. Accordingly, the conditional expected number of served data packets in a frame is computed as

$$E\left[N_{Ws_f} \mid B=b, W=w\right]=\begin{cases} w+\sum_{b'=0}^{M}\sum_{s=0}^{\varepsilon-w-1} s\ \Psi_2 P^{(\tau)}_{(b,0),(b',s)} & w<\varepsilon \\ \varepsilon & w\geq\varepsilon \end{cases} \tag{7.10}$$

where the first and second terms in (7.10) correspond to the first and second cases of (7.9) respectively. By direct substitution in (7.8), the conditional frame throughput is

$$E\left[th_f \mid B=b, W=w\right]=R_{th}(u,\tau)=\begin{cases} \frac{T_{DS}}{T_f}\left[w+\sum_{b'=0}^{M}\sum_{s=0}^{\varepsilon-w-1} s\ \Psi_2 P^{(\tau)}_{(b,0),(b',s)}\right] & w<\varepsilon \\ \frac{T_{DS}}{T_f}\varepsilon & w\geq\varepsilon \end{cases} \tag{7.11}$$

The frame throughput function in (7.11) has appropriate characteristics of an objective reward function because; **a)** Its maximum value is one, **b)** It is a function of the current state $u=(B,W)$ and chosen decision τ and **c)** It appropriately establishes reward differentiation by emphasizing the desirable throughput increase over undesirable throughput decrease associated with different state transitions. Hence, I directly use it as an objective function. As shown in Figure 7.4, I plot the throughput objective function in for W =12, and P_a =0.3 to show sample reward differentiation policy. I note that as B increases, the reward is highest at a median value of τ. This is desirable to maintain efficient contention and data transmission processes. I also note that as B decreases, the reward tends to be highest at lower values of τ, which is also desirable in order to efficiently serve the waiting data packets in the service queue.

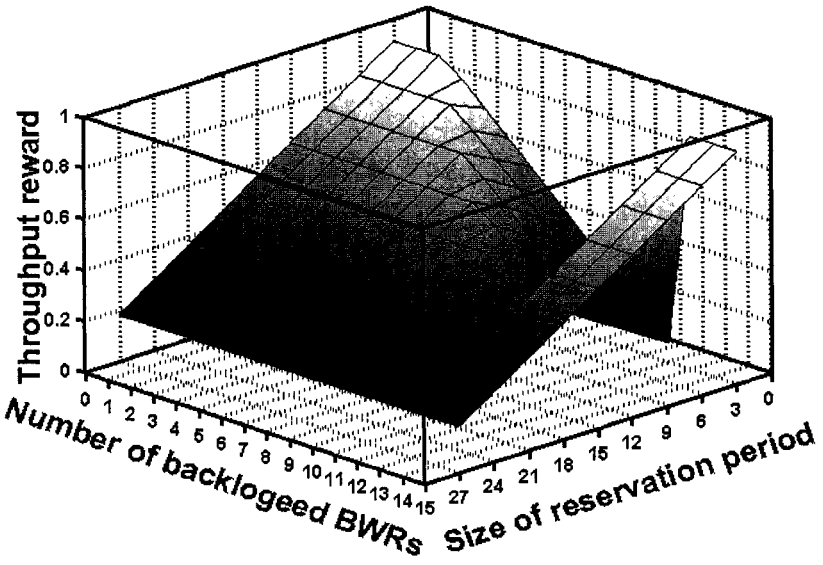

Figure 7.4 Throughput reward differentiation for different combinations of B and τ with $W = 12$

Plugging (7.6), (7.7), and (7.11) in (7.5), I obtain the reward function

$$R(u,\tau) = g_c \cdot \left[1 - \exp\left(\frac{-\tau}{B}\right)\right] + g_w \left[1 - \exp\left(\frac{-\tau}{W}\right)\right] + g_{th} \cdot \begin{cases} \frac{T_{DS}}{T_f}\left[w + \sum_{b'=0}^{M} \sum_{s=0}^{\varepsilon-w-1} s \, \Psi_2 P^{(\tau)}_{(b,0),(b',s)}\right] & w < \varepsilon \\ \frac{T_{DS}}{T_f}\varepsilon & w \geq \varepsilon \end{cases} \quad (7.12)$$

I am now ready to solve for the optimal decision that should be chosen should the process be in state $u = (B, W)$. I use an MDP Matlab toolbox [39] to solve the numerical optimization examples I present in Chapter 8.

7.2.3 Operation of the Optimized Controller

At the beginning of a frame, the base station collects the values of B, W, P_a, and P_r (note: P_r can be a system parameter in slotted Aloha or follow the change of P_a in

p-persistence contention resolution algorithms). According to the combination of B, W, P_a, and P_r, the controller determines the optimum value of τ, at the beginning of the frame, from the entries of the pre-calculated lookup table administered hosted at the base station.

7.2.4 Implementation Complexity

Markov decision processes are known for their computational complexity with large-scale state space [40]. For large scale Markov processes, the state space grows significantly. To alleviate burdensome computational complexities at the beginning of each frame, I propose to perform offline calculations of the optimized decision (τ) for all possible combinations of B, W, and P_a. The resulting optimized decisions will be the entries of a lookup table administered by the base station. After observing a state (B,W) and estimating P_a, the controller retrieves the corresponding size of the reservation period for the current frame. This method results in considerable performance enhancements as will be seen in the numerical examples in Chapter 8

7.3 Summary

In this Chapter I established a framework of an optimized controller to dynamically tune the size of the reservation period in the beginning of each frame. Based on the Markovian behavior of the R-MAC system under study, I formulated a Markov Decision Process optimization model that resembles that dynamics of the R-MAC system. My objective here is establishing a sort of intelligence at the base station in order to

opportunistically improve the performance on a frame level according to the local performance and traffic information. At the beginning of each frame, the base station will collect the delay and traffic information. Using the proposed MDP model, I established a method to dynamically calculate the optimized size of the reservation period at the beginning of each frame given the traffic information and state of the system. The proposed optimization method uses an offline-calculated lookup table, which alleviates the MDP computational complexities, and hence adds no delay to the R-MAC performance. In Chapter 8 I examine the efficiency of the proposed optimization model under the slotted Aloha and p-persistence contention resolution mechanisms.

8 Performance Evaluation of MDP Optimization Model

I conduct an illustrative analysis for a relatively small size R-MAC system. In the simulation, the base station deterministically collects the values of B, W, P_a, and P_r at the beginning of each frame. Using the lookup table, the controller determines the value of τ to be used in that frame. I consider $M = 15$, $L = 30$, and T_f to be 1 ms resembling the IEEE 802.16 recommendation. However, I consider a downscaled number of slots per frame, 33 slots / frame. I let the size of a reservation request packet be one slot, and I let the size of a data packet be three slots (although a fixed size data packet is considered, the proposed model can be slightly modified to accommodate variable size data packets). Thus, a frame can have a maximum of 33 contention slots or alternatively a maximum of 10 data slots. The size of the reservation period will have the domain $\tau = \{3, 6, 9, 12, 15, 18, 21, 24, 27\}$. I implement the proposed optimization model in two contention resolution environments; the slotted Aloha and p-persistence mechanisms and obtain performance results as follows.

8.1 Slotted Aloha Contention Resolution

The objective here is to study the performance enhancements resulting from the optimization model under constant and variable values of P_a. I implement the BWR traffic intensity by varying the value of P_a. However, P_r remains unchanged considering it to be a system parameter. In this example I show the transient R-MAC performance degradation as a result of the increase in P_a. I let the simulation run enough time to reach the steady state. Taking $P_r = 0.1$, I let $P_a = 0.15$ up till the 10^{th} second. I find that under static reservation period allocation, τ=15 achieves the best throughput performance among all values of static allocations. In Figure 8.1, up till the 10^{th} second, the best throughput performance under static allocation approaches 44.7% which is the same as the throughput bound derived in [25]. During the same period, the optimized controller achieves 49 % throughput which represents about 10% throughput improvement. Meanwhile, the throughput enhancement results in increased packet delay as shown in Figure 8.2. At the 10^{th} second, I increase P_a to 0.3 until the end of the simulation time. As a result, the system throughput under both static and dynamically optimized allocations is slightly affected; however, the optimized allocation maintains the same 10% throughput improvement. The reason the performance is only marginally affected by doubly increasing P_a is twofold: first, because the unchanged retransmission probability partially absorbs the increase in P_a and second, because by assumption B is not allowed to grow indefinitely, by restricting the generation of new BWRs to those SSs in idle state. In Figure 8.4, instead of alternating P_a only once, I continuously alternate between

$P_a = 0.15$ and $P_a = 0.3$ by tossing a coin with equal probability of returning heads or tails. Once one value is chosen, it remains active for a geometrically distributed number of frames, with an average of 50 frames. As shown, the performance reaches steady state at around the fifth second of the simulation time. At this time and on, the optimized controller consistently outperforms the throughput of the best case static allocation by about 12%. The corresponding delay increase is shown in Figure 8.5. Figure 8.3 and Figure 8.6 show the dynamically optimized allocation of the reservation period, in an interval of 200 frames around the 10^{th} second of the simulation time, under the two traffic scenarios. In Figure 8.3, in response to the double increase of P_a from 0.15 to 0.3 at the 10^{th} second, the optimized controller adapts quickly by increasing the reservation period to τ=21 more frequently. In Figure 8.6, the more frequent alternation of P_a between 0.15 and 0.3 causes deeper dynamicity of the adaptive response and more visits to the decision τ=21.

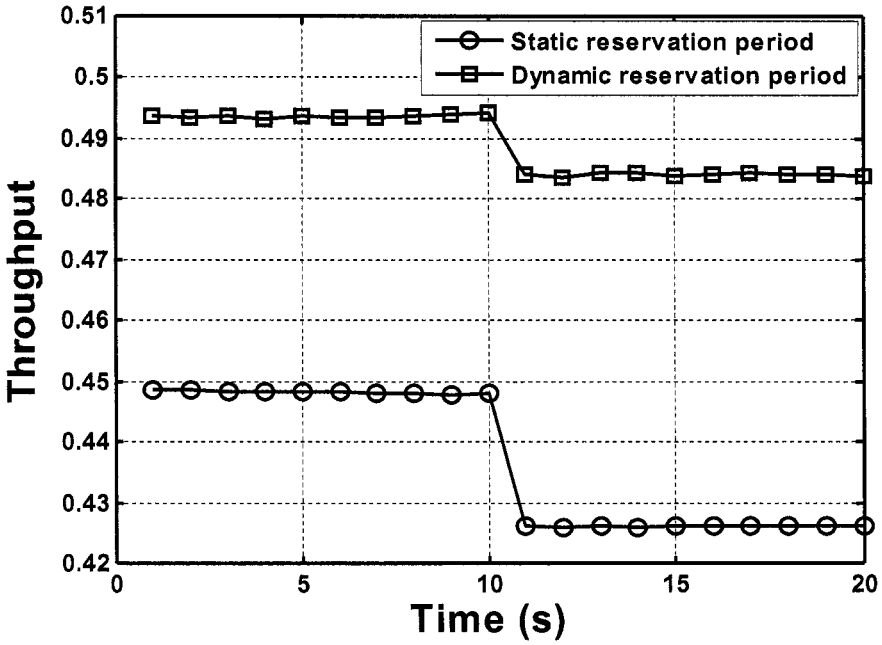

Figure 8.1 Slotted-Aloha throughput with increasing P_a at the 10th second

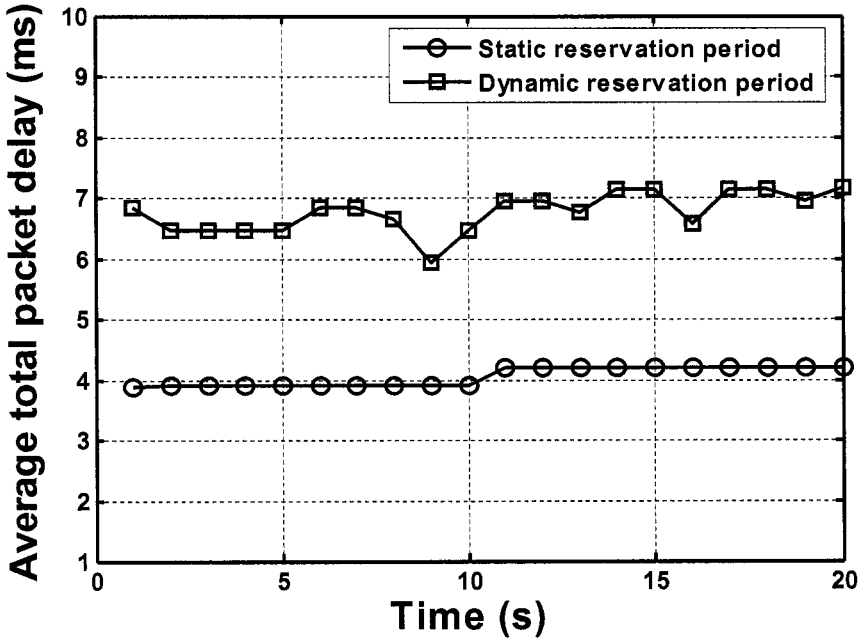

Figure 8.2 Slotted-Aloha packet delay with P_a increasing at the 10th second

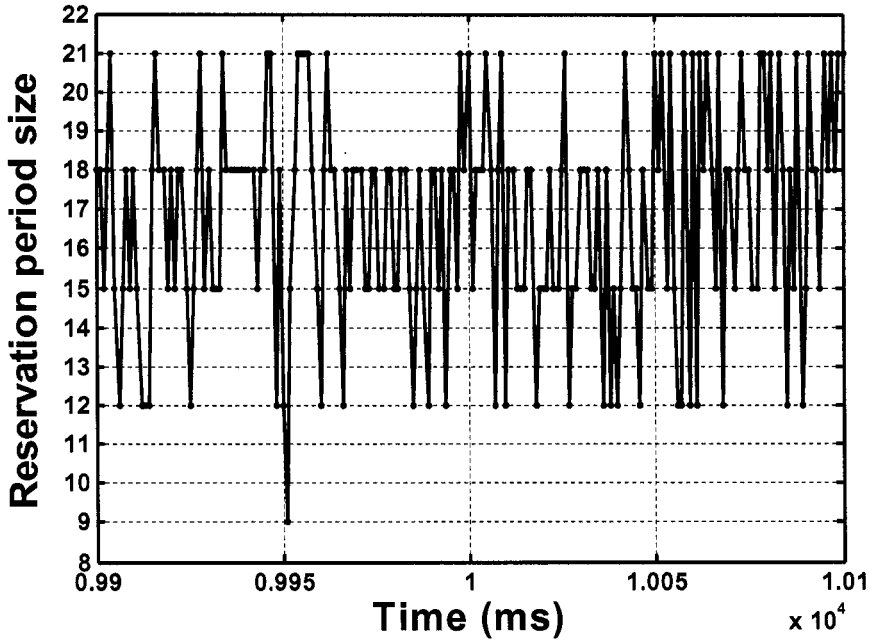

Figure 8.3 Optimized controller transient response around the 10^{th} second

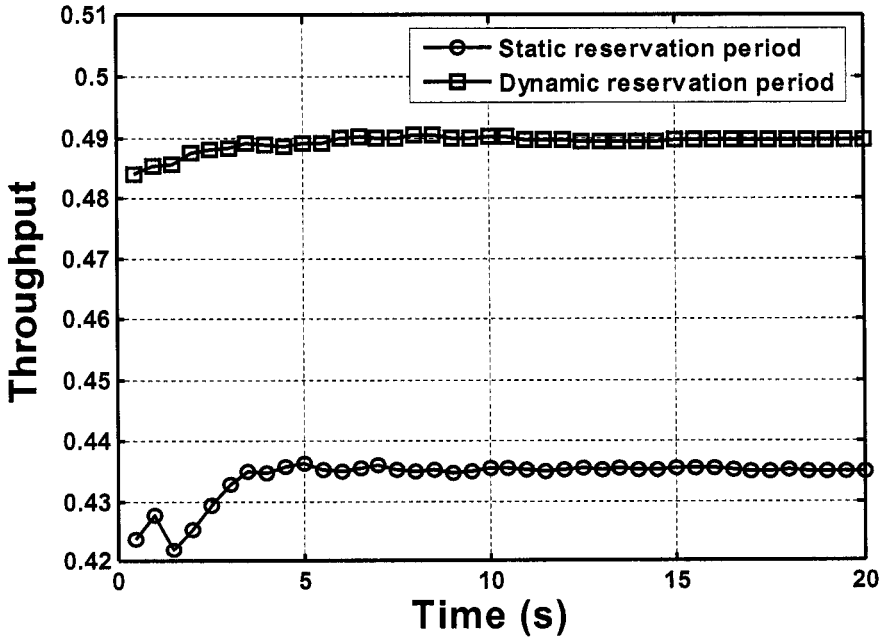

Figure 8.4 Slotted-Aloha throughput with P_a frequent oscillation.

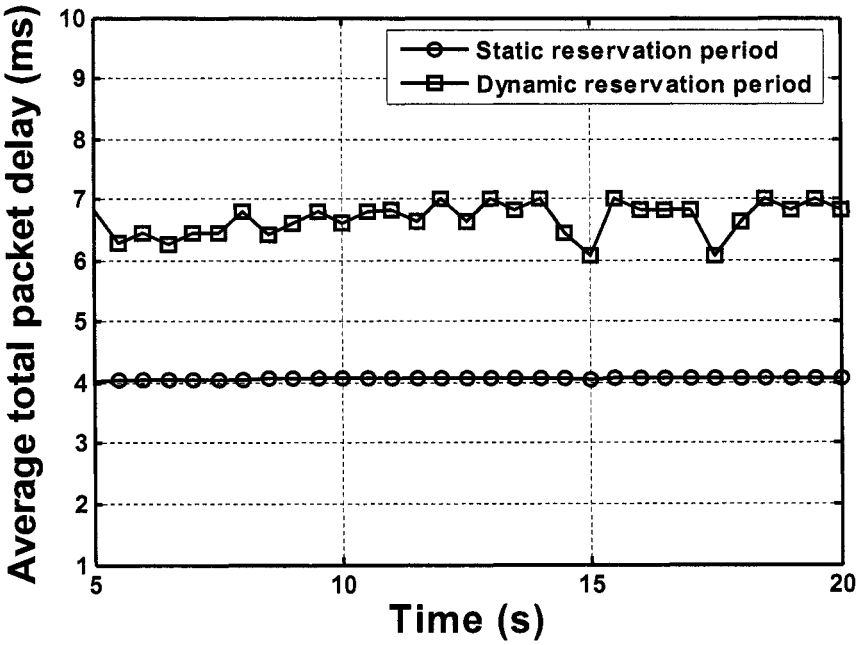

Figure 8.5 Slotted-Aloha packet delay with P_a frequent oscillation

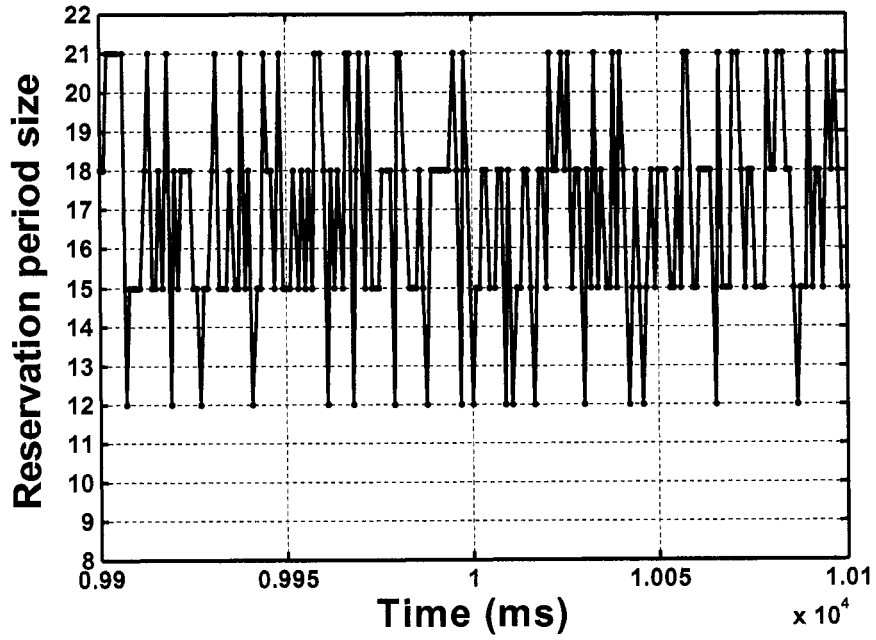

Figure 8.6 Optimized controller response to P_a frequent oscillation

8.2 p-persistence Contention Resolution

With a similar traffic scenario to the slotted Aloha case, I investigate the performance improvement achievable through the optimized controller. In the p-persistence mechanism, each SS persistently attempts to transmit a BWR with the same probability p in every contention slot. The first traffic scenario is designed to observe the transient behavior associated with the increase in p from 0.15 to 0.3 at the 10^{th} second of the simulation time. As shown in Figure 8.7, under both optimized dynamic and best case static reservation period allocation, the increase in the BWR traffic results in sudden deterioration in both allocation policies. However, the dynamically optimized performance is better by about 30% than the best case static allocation τ=15, with the delay increase shown in Figure 8.8. Moreover, when I alternate between $p = 0.15$ and $p = 0.3$ throughout the simulation time, in the same way described in Section 8.1, the dynamically optimized performance is about 44% better than the best case static allocation τ=15 from a throughput perspective as shown in Figure 8.10 (Note that the steady state performance is reached around the fifth second of the simulation time). In the same time, the corresponding delay increase is shown in Figure 8.11. In Figure 8.9, the transient response of the optimized controller to the steep performance degradation, in an interval of 200 frames around the 10^{th} second of the simulation time, is shown. The controller adapts to the traffic increase by consistently maintaining a wide reservation period. Also Figure 8.12 shows the change activity of the controller decisions, in an interval of 200 frames around the 10^{th} second of the simulation time, where the widest reservation period τ=27, from the set of possible decisions, is taken more often.

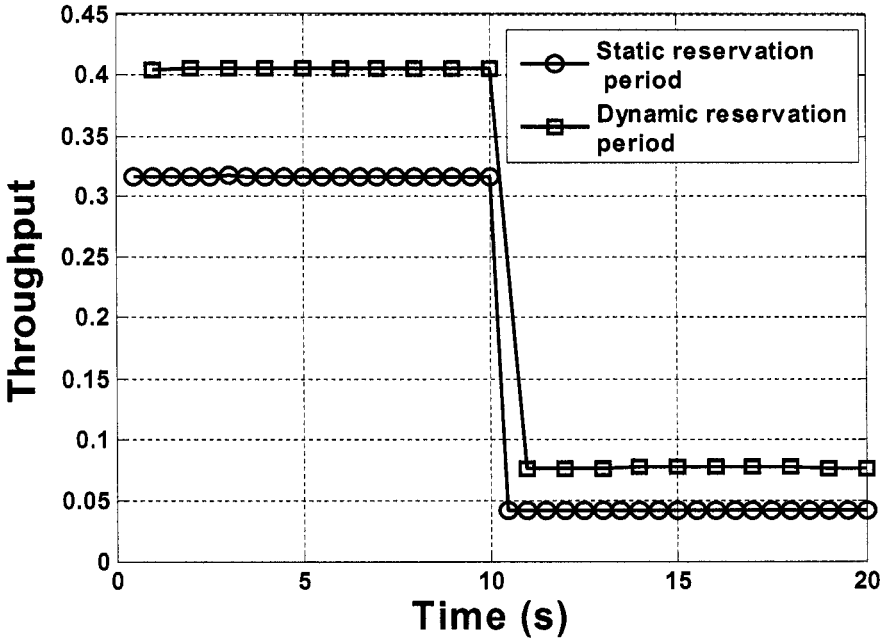

Figure 8.7 *p*-persistence throughput with increasing *p* at the 10th second

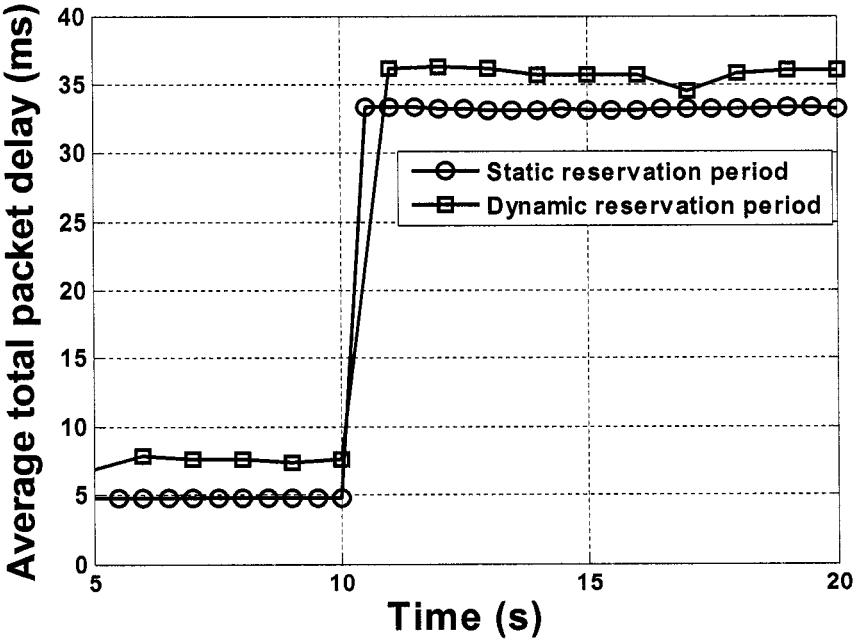

Figure 8.8 *p*-persistence packet delay with increasing *p* at the 10th second

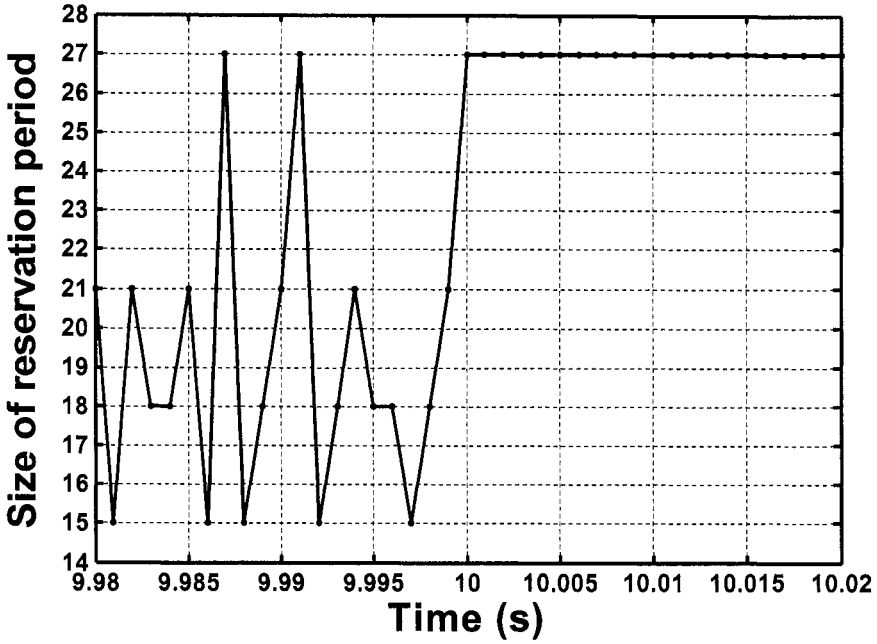

Figure 8.9 Optimized controller transient response around the 10^{th} second

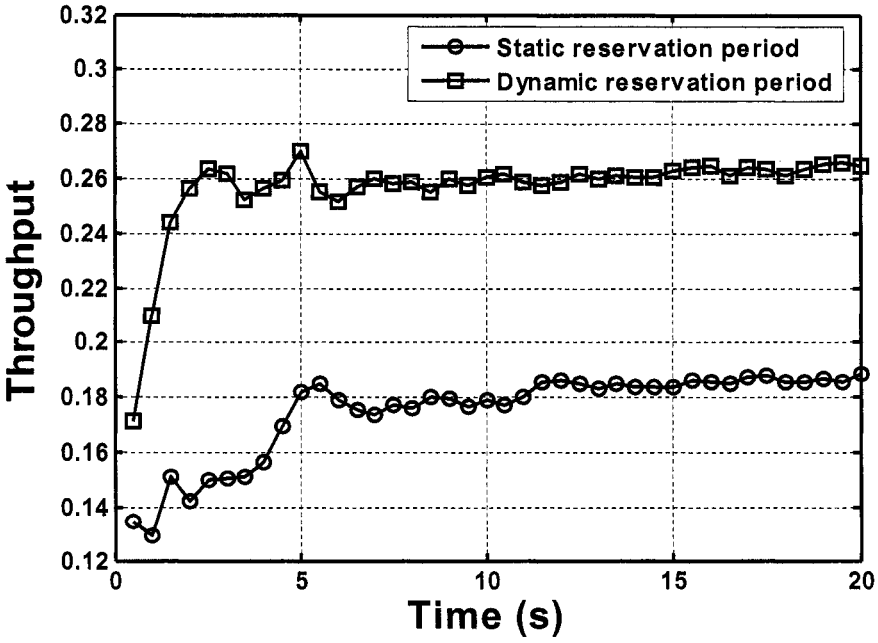

Figure 8.10 p -persistence throughput with p frequent oscillation

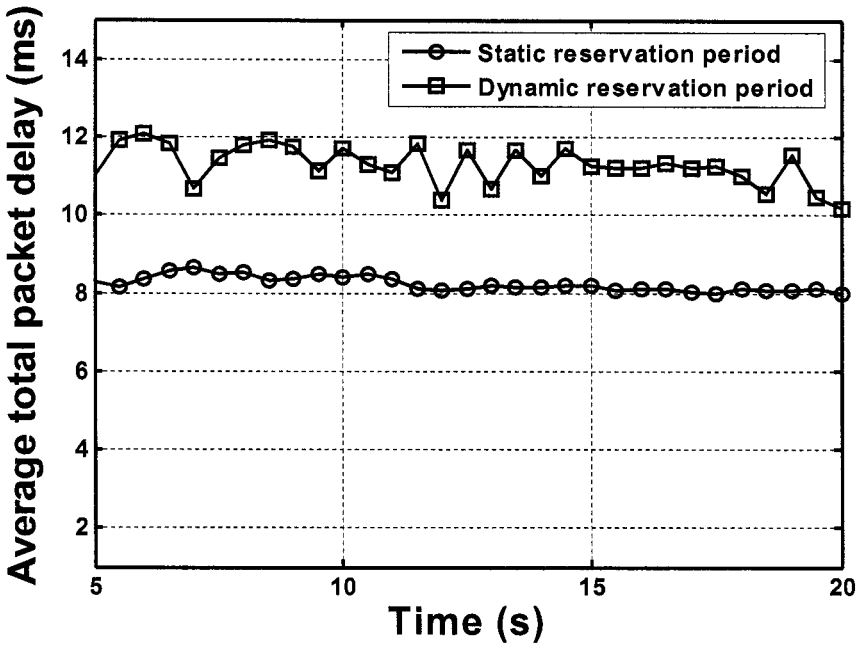

Figure 8.11 p -persistence packet delay with p frequent oscillation

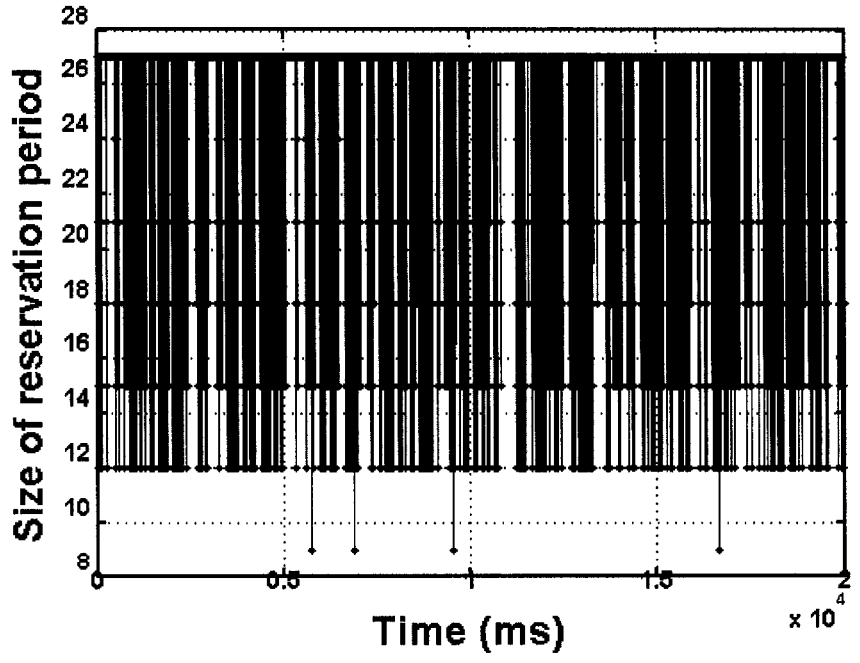

Figure 8.12 Optimized controller response to p frequent oscillation

8.3 Summary

The dynamic characteristics of today's traffic may degrade the R-MAC performance. A paramount design parameter in R-MAC systems is the size of the reservation period. Through proper control, tuning the size of the reservation period according to the traffic and operations dynamics leads to considerable performance enhancement. To illustrate the merits of proposed dynamic method, I examined the proposed optimization model under the slotted Aloha and p-persistence contention resolution mechanisms. With both techniques I obtained considerable throughput enhancement over the static reservation period policy. Furthermore, the proposed method meets the proposed framework's operational characteristics. The results I have obtained indicate a potential enhancement in the R-MAC performance over previously derived bounds in reference [25]. Therefore, there is a need to investigate a bound on the throughput obtainable through the proposed dynamic optimization model.

9 Conclusions

With extensive QoS provisions, the IEEE 802.16 standard is evolving as a cost-effective platform offering higher bandwidth for fixed and mobile communications. A swift materialization requires the IEEE 802.16 networks to sustain and embrace the inherent uneven traffic characteristics. The emerging interest in broadband networks that employ reservation multiple access protocols revived the need to understand and improve the performance metrics of this family of protocols.

In this thesis I have studied the multiple access control protocol of the IEEE 802.16 standard. I have identified areas that have not been specified in the standard. Those areas were left so that vendor product differentiation would present alternative solutions that suit different objectives of operation. One of these areas is the ratio between contention resources and data transmission resources in the multiple access frames.

Through illustrative analysis, by means of analytical modeling and simulation, the performance of the IEEE 802.16 multiple access protocol was found to respond differently to different ratios of resources allocation. On the one hand, increasing the size of reservation period decreases the contention delay of bandwidth requests and hence

increases the number of bandwidth requests that can be processed on the subsequent service period of the frame. I have shown that letting the bandwidth requests reach the service queue at the base station more quickly, through increasing the size of the reservation period, is advantageous from a resources utilization perspective than spending longer time in contention. However on the other hand, the increase in the size of the reservation period also has the effect of decreasing the size of the service period in the frame, which results in longer data transmission delay over the time and also results in decreasing the system throughput.

The ultimate goal of the multiple access protocol performance is to achieve the best total delay and throughout performance, where both must be within the acceptable bounds. For a constant rate of bandwidth arrivals, I showed the ratio of the resources allocation that can achieve best delay and throughput performance. I also showed that obtaining the best delay and throughput performance may require two different ratios. The experiments that led to these conclusions were run under fixed ratio of reservation and service periods in all frames.

In a typical network environment, a time variable rate of bandwidth request arrivals is more probable than the fixed rate. Running the same experiments with different rates of bandwidth request arrivals, I found that the optimal points of operation with respect to delay and throughput performance have changed.

In order to maintain certain delay and throughput performance, the resources allocation ratio should be adaptive to the change in traffic conditions (i.e. bandwidth requests arrival rate). I found, through the use of dynamic optimization of the reservation period, that

improved performance can be obtained under varying intensities of bandwidth request arrivals. In particular, I obtained improved performance under the slotted Aloha and *p*-persistence contention resolution mechanisms using a Markov Decision Process dynamic optimization model. Depending on the state of the Markov process and the intensity of the bandwidth requests arrivals, the proposed optimized controller chooses the size of the reservation period that would opportunistically improve the frame performance metrics.

The proposed method does not require additional standardization mandates and can be directly implemented in the R-MAC protocol of the DSL, HFC Cable, and GPRS technologies in addition to the IEEE 802.16 networks.

References

[1] IEEE 802.16-2004, "IEEE Standard for Local and Metropolitan Area Networks — Part 16: Air Interface for Fixed Broadband Wireless Access Systems".

[2] A. Ghosh, D. R. Wolter, J. G. Andrews, and R. Chen "Broadband wireless access with WiMax/8O2.16: current performance benchmarks and future potential," IEEE Communications Magazine, Feb. 2005 pp: 129- 136.

[3] A. Brand, and H. Aghvami, "Multiple access protocols for mobile communications" John Wiley & Sons Ltd. 2002.

[4] W. Crowther, R. Rettberg, D. Walden, S. Ornstein and F. Heart, "A system for broadcast communication: Reservation-Aloha," Proc. 6th Hawaii Int. Conf. Syst. Sci., 1973, pp 596-603.

[5] D. J. Goodman, R. A. Valenzuela, K. T. Gayliard, and B. Ramamurthi, "Packet reservation multiple access for local wireless communications," Communications, IEEE Transactions on, Vol. 37, Issue: 8, Aug. 1989 pp: 885 – 890.

[6] N.M. Mitrou, T.D. Orinos, and E.N. Protonotarios, "A reservation multiple access protocol for microcellular mobile-communication systems," IEEE Transactions on

Vehicular Technology, Vol. 39, Issue: 4, Nov. 1990 pp: 340 – 351.

[7] G. Bianchi, F. Borgonovo, L. Fratta, L. Musumeci, and M. Zorzi, "C-PRMA: A centralized packet reservation multiple access for local wireless communications," IEEE Transactions on Vehicular Technology, Vol. 46, Issue: 2, May 1997 pp: 422 – 436.

[8] D. J. Goodman, "Cellular Packet Communications", IEEE Transactions on Vehicular Technology. 1990 pp: 1272-1280.

[9] D. Tsai, and J.-F Chang, "A hybrid contention based TDMA technique for data transmission," IEEE Transactions on communications, Vol. 36, issue 2, Feb. 1988 pp: 225 - 228.

[10] C. G. Kang, C. W. Ahn, K. H. Jang, and W. S. Kang, "Contention-free distributed dynamic Reservation MAC protocol with deterministic scheduling (C-FD3R MAC) for Wireless ATM networks" IEEE journal on selected areas in communications, Vol. 18, no. 9, Sep. 2000 pp: 1623 - 1635.

[11] A. Sugihiara, K. Enomoto and I. Sasai, "Hybrid contention/reservation channel-access schemes for integrated voice/data wireless networks," Seventh IEEE International Symposium on Personal, Indoor and Mobile Radio Communications PIMRC'96, Vol. 2 Oct. 1996 pp: 638 – 642.

[12] J.-F. Frigon, V. C. M. Leung, and H. C. B. Chan, "Dynamic Reservation TDMA protocol for Wireless ATM networks," IEEE journal on selected areas in communications, Vol. 19, no. 2, Feb. 2001 pp: 370 – 383.

[13] H. C. B. Chan, J. Zhang, and H. Chen, "A dynamic reservation protocol for LEO mobile satellite systems," IEEE journal on selected areas in communication, Vol.

22, no. 3, Apr. 2004 pp: 559 – 573.

[14] G. Pierobon, A. Zanella, and A. Salloum. "Contention-TDMA protocol: performance analysis," IEEE transactions on vehicular technology, Vol. 51, issue: 4, July 2002 pp: 781 - 788.

[15] S. Tasaka, "Stability and performance of the R-Aloha packet broadcast system" IEEE transactions on Computers, Vol. C-32, Aug. 1983, pp 717-726.

[16] J.N. Daigle, and M.N. Magalhaes, "Analysis of packet networks having contention-based reservation with application to GPRS," IEEE/ACM Transactions on networking, Vol. 11, Issue: 4, Aug. 2003 pp: 602 – 615.

[17] Y. Cao, H.R. Sun, and K.S. Trivedi, "Performance analysis of reservation media-access protocol with access and serving queues under bursty traffic in GPRS/EGPRS," IEEE transactions on vehicular technology, Vol. 52, issue: 6, Nov. 2003 pp: 1627 – 1641.

[18] R. Fantacci, and S. Nannicini, "Performance evaluation of a reservation TDMA protocol for voice/data transmission in microcellular systems," IEEE Journal on Selected areas in Communications, Vol. 18, no. 11, Nov. 2000 pp: 2404 – 2416.

[19] K. Crisler and M. Needham. "Throughput analysis of reservation Aloha multiple access," Electronics Letters, Vol. 31, issue: 2, 19 Jan. 1995 pp: 87 - 89.

[20] D. G. Jeong and W. S Jeon. "Performance of an exponential backoff scheme for slotted Aloha protocol in local wireless environment," IEEE transactions on vehicular technology, Vol. 44, issue 3, Aug. 1995 pp: 470 - 479.

[21] T. K. Liu, J. A. Sivester, and A. Ploydoros, "Performance evaluation of R-Aloha in distributed packets radio networks with hard real-time communications," IEEE 45th

Vehicular Technology Conference, 1995, Vol. 2 , July 1995 pp: 554 – 558.

[22] F. L. Presti and V. Grassi, "Markov analysis of the PRMA protocol for local wireless networks", J.C. Baltzer AG, Science Publishers.

[23] C. S. Wu and G.-K. Ma, "Performance of packet reservation MAC protocols for wireless networks," IEEE 48th Vehicular Technology Conference, Vol. 3, May 1998 pp: 2537 – 2541.

[24] K. Sriram and P. D. Magil, "Enhanced throughput efficiency by use of dynamically variable request slots in MAC protocols for HFC and wireless access networks," Kluwer academic publishers, Telecommunication Systems Journal, Vol. 9, issue 3, 1998 pp. 315-333.

[25] D. Sala, J. Limb and S. Khaunte, "Adaptive Control Mechanism for Cable Modems MAC Protocols," Proc. Infocom 1998, March 1998.

[26] W. M. Yin, and Y. D. Lin;" Statistically optimized slot allocation for initial and collision resolution in hybrid fiber coaxial networks," IEEE Journal on Selected Areas in Communications, Vol. 18, issue: 9, Sept. 2000 pp: 1764 – 1773.

[27] G.-H. Hwang and D.-H. Cho, "Dynamic Random channel allocation scheme in HiperLAN Type 2," IEEE International Conference on Communications, 2002. ICC 2002, Vol. 4 pp: 2253 – 2257.

[28] Z. J. Haas, and J. Deng, "On optimizing the backoff interval for random access schemes," IEEE Transactions on Communications, no. 12, Dec 2003 pp. 2081-2090.

[29] N. Golmie, Y. Saintillan, and D. Su, "A Review of Contention Resolution Algorithms for IEEE 802.14 Networks," IEEE Communications Surveys, 1999.

[30] C. Eklund, R. B. Marks, K. L. Stanwood and S. Wang "IEEE 802.16 standard: a

technical overview of the Wireless MAN air interface for broadband wireless access," IEEE Communications Magazine, June 2002.

[31] R. Rom and M. Sidi, Multiple Access Protocols: Performance and Analysis, Springer Verlag, 1990.

[32] N. Golmie, F. Mouveaux, D.H. Su, "Differentiated services over cable networks," IEEE GLOBECOM 1999, Vol. 2 pp: 1109 – 1115.

[33] L. Kleinrock, and S. Lam, "Packet switching in a multiaccess broadcast channel: performance evaluation," IEEE Transactions on Communications [legacy, pre - 1988], vol. 23, issue: 4, Apr 1975 pp: 410 – 423.

[34] D. Bertsekas, and R. Gallager, Data Networks, 2nd ed. Englewood Cliffs, NJ: Prentice Hall, 1992.

[35] J. Riordan, An introduction to combinatorial analysis. Princeton, NJ: Princeton Univ. Press, 1978

[36] S.M. Ross. Introduction to Probability Models (8th Ed.), Academic Press, 2003.

[37] S. M. Ross, "Applied probability models with optimization applications", Dover Publications, 1970.

[38] C. Coello, A. Carlos; V. Veldhuizen, A. D. Lamont, and B. Gary, "Evolutionary algorithms for solving multi-objective problems," Kluwer Academic Publishers, Boston, 2002.

[39] Institute National de la Recherche Agronomique, http://www.inra.fr/bia/produits /logiciels/Page_home.php/

[40] C. H. Papadimitriou, and J. N. Tsitsiklis, "The complexity of Markov decision processes", Mathematics of Operations Research, 1978.