PACKET TRANSMISSION SCHEDULING FOR SUPPORTING REAL-TIME TRAFFIC IN WIRELESS MESH NETWORKS

PACKET TRANSMISSION SCHEDULING FOR SUPPORTING REAL-TIME TRAFFIC IN WIRELESS MESH NETWORKS

BY

JUN ZOU, M. ENG

A THESIS

SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

McMaster University

Doctor of Philosophy (2008) McMaster University

(Electrical and Computer Engineering) Hamilton, Ontario

TITLE: Packet Transmission Scheduling for Supporting Real-time Traffic in Wireless Mesh Networks

AUTHOR: Jun Zou, M. ENG (Tianjin University, Tianjin, China)

SUPERVISOR: Dr. Dongmei Zhao

NUMBER OF PAGES: xxii, 152

Abstract

Packet transmission scheduling plays a key role in Quality of Service (QoS) support for real-time traffic and efficient radio resource utilization in a wireless mesh network (WMN). It is a highly complicated problem due to the fact that any scheduling decision at one mesh access point (AP) may affect the scheduling decisions in the entire network. The strict delay requirement of real-time applications makes the scheduling problem even more challenging.

In this thesis, the packet transmission scheduling problem for real-time constant-bit-rate (CBR) traffic in a WMN is first formulated as a standard integer linear programming problem, which takes into consideration both the multihop packet transmission delay and timeline coordinations of the mesh APs. The objective is to efficiently utilize the radio resources, subject to available bandwidth of the mesh APs, co-channel interference, and packet transmission latency requirement.

Two heuristic schemes, namely AP-based scheduling (ABS) and connection-based

scheduling (CBS) schemes, are then proposed to support real-time CBR traffic. ABS makes scheduling decisions on a per-AP basis. Scheduling decisions at APs with a higher traffic load are determined before those at APs with a lower traffic load. ABS achieves close-to-optimum capacity but may go through multiple iterations before reaching a feasible solution. CBS makes scheduling decisions on a connection-by-connection basis. It gives a higher priority to connections with more hops. In CBS, connections with a lower priority can only use resources remaining from serving all higher priority connections. CBS requires much lower complexity than ABS while achieving capacity performance slightly lower than ABS.

We extend the proposed ABS and CBS scheduling schemes for supporting real-time variable bit rate (VBR) traffic in a WMN. By combining the concept of effective bandwidth and the proposed scheduling schemes, both delay and packet loss performance of the VBR traffic can be effectively satisfied. The scheduling schemes are further extended for supporting real-time traffic in a WMN with multi-radio APs.

All the scheduling decisions are done at the time when new connection requests arrive and the results are used to make admission control decisions. In this sense, the work in this thesis is for both packet transmission scheduling and admission control for real-time traffic in WMNs.

Acknowledgements

First of all, I am deeply grateful to my supervisor Dr. Dongmei Zhao for her invaluable guidance, continuous encouragement, tremendous time and efforts she has put throughout my Ph.D. study. She is always available, supportive, nice and patient. From her, I can always get valuable advice when facing difficulties. From her, I have learned how to do research, write papers and make presentations. Without her, this work would not have been possible.

I would like to thank Dr. Terence Todd, Dr. Steve Hranilovic and Dr. Douglas Down for being members of my supervisory committee and their valuable and helpful comments on my research and thesis. I would like to thank Dr. Tim Davidson for his advice on the mathematical optimizations. I also would like to thank Dr. Narayanaswamy Balakrishnan, Dr. Zdislav V. Kovarik and Dr.Rong Zhu for being my thesis examining committee members and providing valuable comments.

I would like to thank Cheryl Gies, Helen Jachna, Cosmin Coroiu, and Terry Greenlay for their administrative and technical support.

I would like to acknowledge the Ontario Centers of Excellence (OCE) for its support for my conference travel and research. I would like to acknowledge the Value-Added Personnel (VAP) program offered by OCE where I expanded my horizon and learnt business skills. I also would like to acknowledge the Ontario Graduate Scholarship and McMaster University for supporting me in completing my PhD study.

My appreciation is also to the team members of Wireless Networking Laboratory at McMaster University. I have learned a lot from them through discussions and group meetings and had a good time working with them.

To my wife, thanks for your love, support and taking care of me in my daily life.

To my parents and sister, thank you for your endless love and sacrifice.

List of Abbreviations

ABS	AP-Based Scheduling
ACK	ACKnowledgement frame
AP	Access Point
BE	Best Effort
BRS	Bottleneck Region Size
BS	Base Station
BSS	Basic Service Set
CBR	Constant-Bit Rate
CBS	Connection-Based Scheduling
CFP	Contention-Free Period
CP	Contention Period
CSMA/CA	Carrier Sense Multiple Access with Collision Avoidance
CTS	Clear-To-Send
DCF	Distributed Coordination Function

DIFS	Distributed InterFrame Space
EB	Effective Bandwidth
EDCA	Enhanced Distributed Channel Access
EDF	Earliest Deadline First
ertPS	extended real-time Polling Service
FTP	File Transfer Protocol
GPS	Generalized Processor Sharing
HC	Hybrid Coordinator
HCCA	Hybrid coordinator function Controlled Channel Access
IEEE	Institute of Electrical and Electronics Engineers
ILP	Integer Linear Programming
IP	Internet Protocol
JS	Job Shop
MAC	Medium Access Control
MANET	Mobile Ad-hoc Network
MEDF	Modified Earliest Deadline First
MS	Mobile Station
NP	Non-deterministic Polynomial time
nrtPS	non-real-time Polling Service
OPT	Global Optimum Scheduling

OPT-CB	Connection-Based Optimum Scheduling
PCF	Point Coordination Function
PHY	Physical layer
PIFS	PCF InterFrame Space
PMP	Point-to-MultiPoint
QoS	Quality of Service
rtPS	real-time Polling Service
RTS	Request-To-send
SI	Scheduling Interval
SIFS	Short InterFrame Space
SS	Subscriber Station
TDM	Time Division Multiplexing
TDMA	Time Division Multiple Access
UGS	Unsolicited Grant Service
VBR	Variable-Bit Rate
VC	Virtual Clock
VOIP	Voice Over IP
WDS	Wireless Distribution System
WFQ	Weighted Fair Queuing
WLAN	Wireless Local Area Network

WMAN	Wireless Metropolitan Area Network
WMN	Wireless Mesh Network

List of Notations

α	transition probability from ON state to OFF state
β	transition probability from OFF state to ON state
$\lambda_{i,m,x}$	binary variable indicating whether packet i is transmitted by AP m at time x
$\tau_{i,h}$	time when packet i is transmitted at hop h
$\tau_{i,m}$	time when packet i is transmitted by AP m
A	offered traffic load
$A_{m,t}$	binary variable indicating availability of AP m at time t
b	buffer size
C	physical channel transmission rate
C_e	equivalent bandwidth of a VBR connection
$\mathcal{C}_{Rm}$	a set of connections that AP m receives
$\mathcal{C}_{Tm}$	a set of connections that AP m transmits
c	number of frequency channels
$\mathcal{D}_{i,h}$	candidate set of timeslots for connection i at hop h

d_h	delay budget for home hop
d_h^*	optimum delay budget for home hop
d_i	delay budget for packet i
d_m	delay budget for mesh hops
H_i	number of hops of connection i
$\mathbf{I}_{m,f}$	a set of APs that transmit at frequency channel f and are within interference range of AP m
L_m	traffic load of AP m
M	total number of APs
$\mathcal{M}$	a set of all APs
m_i^-	the immediate upstream AP of AP m along route of connection i
$m_{i,h}$	the transmitting AP of the h-th hop of connection i
N	total number of connections
N_p	number of ON-OFF mini-sources for a VBR connection
$\mathcal{N}$	a set of all connections
$\mathcal{N}_m$	a set of connections processed by AP m
p	maximum packet loss rate of VBR traffic
p_c	packet loss rate due to channel impairment
p_d	packet loss rate due to long delay
$q_{i,h1,h2}$	reserved time for the hops between hops h_1 and h_2

R	packet generation rate of a CBR connection
R_p	peak packet generation rate of an ON-OFF source
$\mathcal{R}_i$	a set of the APs along the route of connection i
$T_{rt,m}$	real-time portion of AP m
$T_{rt,\max}$	maximum real-time portion of all APs
T_{SI}	Duration of one scheduling interval
w_i	transmission delay of packet i
$w_{\max}$	maximum packet transmission delay of all packets
Z_m	number of radios of AP m

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

Wireless Mesh Networks (WMNs) have emerged as a key technology for next generation wireless networking. A WMN consists of mesh access points (APs) and mobile stations (MSs). Mesh APs form a wireless backbone via multiple wireless hops and serve MSs in their coverage [1]. WMNs bring attractive advantages such as low installation cost and extended service coverage that traditional wireless networks do not have.

Wireless local area networks (WLANs) can also extend the limited coverage of individual APs by deploying a large number of APs. As each AP has a cabled connection to the wireline network, a WLAN can have a highly expensive infrastructure, which poses a heavy burden to service providers in their deployment in cities where digging tunnels and laying down network cables are very expensive and time-consuming. Even in rural areas, the cost of connecting remote sites via wired networks can be high. In contrast, the mesh APs of WMNs connect to each other wirelessly and only a few APs need to have wired access to the wireline network. Hence, WMNs are able to extend the service coverage rapidly and at a lower cost and become an economic alternative for the high-speed Internet access network.

Similar to WMNs, mobile ad hoc networks (MANETs) [1] are also multihop wireless networks that can provide extended service coverage to mobile users. In a MANET, MSs are connected to each other via an ad hoc topology [4]. Because of the high mobility of MSs, the network topology can change dynamically [5], which makes it difficult to provide guaranteed quality of service (QoS) and prevents the wide deployment and commercial applications of MANETs. In contrast to MANETs, WMNs have a relatively stable network infrastructure as mesh APs are usually static. Compared to MANET, the wireless infrastructure in WMNs leads to a more reliable topology, which assures service availability and exhibits better manageability.

Because of the advantages over other wireless networks, wireless mesh networks

[1] In some literature, for example [2] and [3], MANETs are also referred to as (infrastructureless) mesh networks, where MSs communicate with each other directly. In contrast, the WMNs that are studied in this thesis are infrastructure-based, where the infrastructure is formed by mesh APs.

are experiencing rapid progress and numerous applications such as high-speed Internet access networks, community networks, vehicle networks, high-speed metropolitan area networks, and emergency networks [1] [6]. Meanwhile, many issues in the WMN remain to be solved before its full potential is realized. For example, radio techniques, network management and security are ongoing research topics in WMNs. QoS provisioning is an essential requirement and a key challenge, since the envisioned scenario of WMNs is to provide broadband Internet access and support real-time applications such as voice over Internet protocol (VOIP) [7]. However, wireless links are subject to limited bandwidth, hostile physical propagation channels, interference and noise [8]. The multi-hop and meshed scenario in WMNs makes it more difficult to satisfy various QoS requirements such as throughput, delay, and packet loss rate. In order to support more traffic with stringent QoS requirements, carefully arranging operations of the mesh APs and scheduling packet transmissions are necessary. In this thesis, we focus on scheduling problems for supporting real-time traffic in a wireless mesh network.

1.2 Motivations and Challenges

A WMN is expected to support a high volume of traffic. Due to the limited bandwidth of the wireless medium, an efficient wireless resource management mechanism is necessary to make "wise" resource allocations to the connections while maintaining their QoS performance. Scheduling is concerned with allocations of scarce wireless resources with the objective of optimizing performance measures such as system resource utilization, while meeting specific QoS requirements posed by applications. In a WMN, the scheduler should operate across different APs, reserve and assign radio time to connections in order to guarantee their QoS requirements such as throughput, delay and packet loss requirements [9]. Effective and efficient packet transmission scheduling is very important for both network providers and users' applications. Although extensive work has been done for traffic scheduling in wireless networks, there has been very limited work on scheduling in WMNs. The motivation behind this thesis is the need of packet transmission scheduling solutions to coordinate the radio resource allocation of mesh APs in order to effectively support real-time traffic and efficiently utilize the radio resources in WMNs.

The unique characteristics of WMNs pose some special problems that do not exist in scheduling in other types of networks and make the scheduling in WMNs a highly challenging problem.

- The scheduling problem in WMNs is similar to a job shop (JS) scheduling problem. A job shop consists of a set of machines that perform operations on jobs while each WMN consists of a set of APs that serve packet transmissions. The APs and packets in a WMN are similar to the machines and jobs in the

JS problem, respectively. Each job/packet has a specified processing order through the machines/APs and each machine/AP can handle only one job/packet at a time. The JS/scheduling problem is to find the processing time of the jobs/packets on the machines/APs with an objective to optimize system performance, for example, to minimize the completion times of all operations [10]. The JS problem belongs to the most intractable problems in machine scheduling [11]. For example, a $K \times M$ size problem, where K and M are the total number of jobs and machines, respectively, has an upper bound of $K!^M$ combinations, thus a 20×10 problem may have up to 7.27×10^{183} combinations. The JS problem with three machines and unit processing time has been proved to be strongly NP-hard [12]. The JS problem with three jobs is NP-hard [13].

- Each packet transmission requires simultaneous availability of both the receiver and the transmitter, and the radio time allocations at these APs can affect other APs that directly or indirectly communicate with them and further affect the scheduling of the entire network. Thus the radio resource allocations at different APs in a WMN are not independent. Careful coordination of the radio resource allocations among different APs is necessary for successful packet transmissions and efficient radio resource utilization. This makes the scheduling problem in WMNs even more difficult than the JS problem.

- The strict latency requirement for supporting real-time traffic in WMN can be another challenging issue. When coordinating the timeline allocations of the mesh APs, minimizing the multihop transmission delay of individual packets and efficiently utilizing the radio time resources can be two contradictory objectives. The problem becomes more complicated when co-channel interference exists and the scheduler is obliged to avoid strong interference and ensure successful packet transmissions.

1.3 Objectives

The objective of this thesis is to both find optimum solutions and design practical heuristic methods for effective and efficient packet transmission scheduling in order to support real-time traffic in WMNs. Specifically we will take the following performance measures into consideration:

- **Packet transmission delay**: Scheduling is expected to provide a delay bound for individual real-time connections based on their respective delay budgets.

- **Packet loss rate**: Scheduling is expected to provide zero packet loss for CBR traffic and a bounded packet loss rate for VBR traffic.

- **Radio resource utilization**: Scheduling should be able to achieve efficient utilization of the limited radio resources. That is, to use the minimum amount of radio resources for supporting the real-time traffic.

1.4 Organization of the thesis

The remainder of this thesis is organized as follows. In Chapter 2, we provide the background on which this thesis is based and review different scheduling problems and their respective solutions in wireline networks and in different types of wireless networks. In Chapter 3, the scheduling problem for supporting real-time CBR traffic in WMNs is formulated as a standard integer linear programming problem. The solution to the optimization problem provides the global optimum scheduling. In Chapter 4, we propose an AP-based scheduling scheme and show that its performance is close to the global optimum scheduler. A connection-based heuristic scheduling scheme is proposed in Chapter 5 which aims to achieve good performance with low complexity. A connection-based optimum scheduling scheme is also proposed in the same Chapter. In Chapter 6, we extend the proposed schemes to schedule real-time VBR traffic and in a WMN with multi-radio APs. Chapter 7 summarizes the main contributions of the thesis and discusses possible future work.

Chapter 2

Background

This chapter introduces the wireless mesh network architecture and two IEEE standards for WMNs, i.e., IEEE 802.11s and IEEE 802.16a. A literature review on various scheduling problems in contrast and comparison to scheduling in WMNs is provided.

2.1 Wireless mesh networks architecture

In WMNs, multiple mesh APs are interconnected to each other wirelessly to form an infrastructure. An example topology of a WMN is shown in Fig. 2.1. The WMN has two tiers: mesh APs at the top tier and mobile stations at the bottom tier. Generally, an AP has dual functions: serving the mobile stations and relaying the traffic to and from other APs. The APs with wired access to the wireline network are referred to as root APs. One of the root APs serves as a gateway and enables the integration of the WMN with other types of networks such as Ethernet or Internet. MSs are associated with mesh APs and transmit packets to their associated APs which then forward the packets to their destination APs or Internet through one or multiple hops [2] [3].

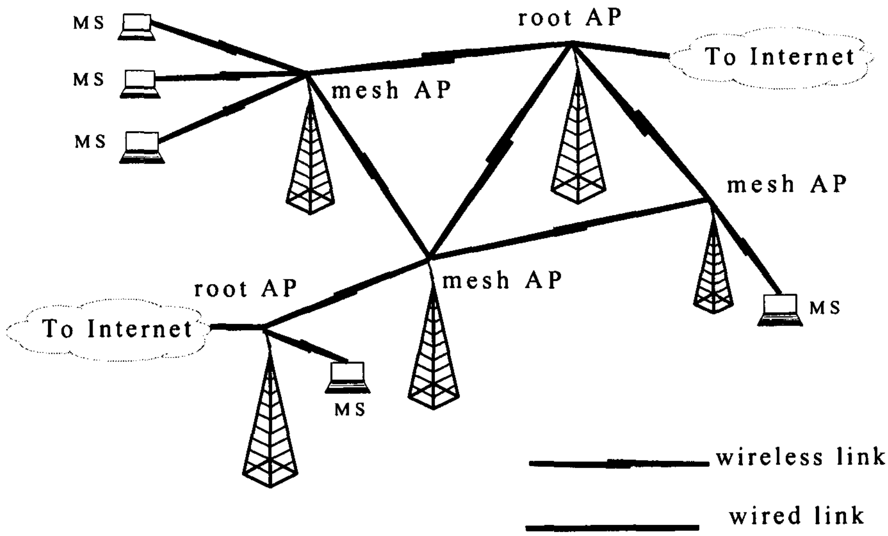

Figure 2.1: Infrastructure-based wireless mesh network

2.2 IEEE 802.11s

The IEEE has been playing a key role in the development of standards for WMNs. Two main IEEE standards for WMNs are IEEE 802.11s and IEEE 802.16a which are used in the WLAN and WMAN (wireless metropolitan area network) mesh networks respectively [14]. We briefly describe IEEE 802.11s for supporting mesh networking

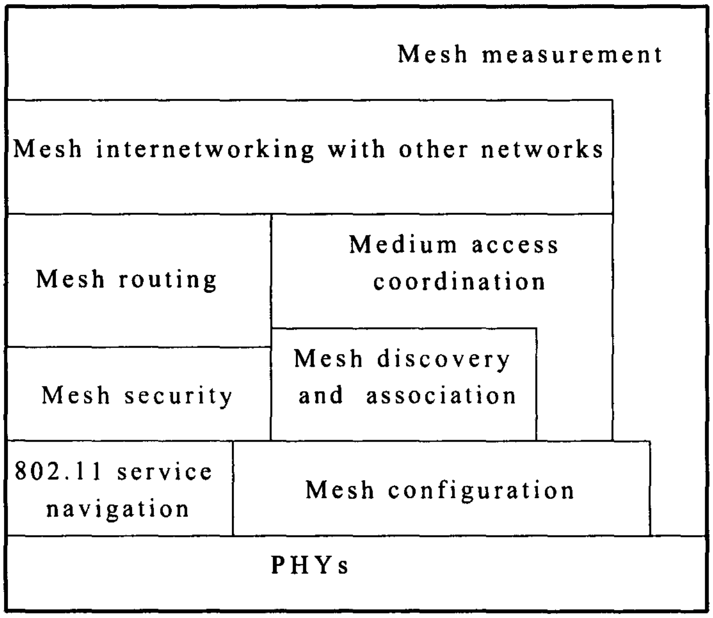

Figure 2.2: Architecture for the IEEE 802.11s standard

in this section and IEEE 802.16a in the next section.

2.2.1 Overview

The IEEE 802.11 family is currently the most successful wireless networking standard for WLANs. Aiming at building a wireless multi-hop infrastructure, IEEE 802.11s was formed in 2004 to address the need for the wireless mesh in WLANs [14] [15]. In the IEEE 802.11-based WMN, the basic building block is a basic service set (BSS) which consists of one AP and multiple MSs. Multiple BSSs may be interconnected through a wireless distribution system (WDS) to form a wireless infrastructure that supports MAC (Medium Access Control)-layer broadcast, multicast and unicast transmissions. The WLAN mesh builds on top of existing physical (PHY) layers of IEEE 802.11 a/b/g operating in the unlicensed spectrum of 2.4 and 5GHz frequency bands. IEEE 802.11s supports mesh APs that are able to work at multiple frequency channels and equipped with multiple radios. This standard also defines the installation, configuration and operation of wireless mesh networks. The major components of the IEEE 802.11s standard are shown in Fig. 2.2 [14]. The internetworking of the WLAN mesh with other types of networks is based on the IEEE 802.1D standard. The root APs which incorporate the functionality of IEEE 802.1D serve as the gateway to other networks. The security of WLAN mesh is based on the IEEE 802.11i standard, which addresses the wireless security issues for all WLAN networks [14]. In the following, we only describe the functionalities related to mesh networking [16].

2.2.2 Mesh discovery and association

There are two types of association in the mesh network: an MS associating with an AP, and an AP associating with a neighboring AP in the WMN. The association of an MS and an AP is performed in the traditional IEEE 802.11 manner. When an MS moves into the coverage area of an AP, it scans for beacons. After receiving a beacon, it associates with the AP. Packets from the MS are always routed to the AP. WDS can use the association information to determine which AP to use in order to deliver the packets to a certain MS. For the association with a neighboring AP, an AP needs to discover the network by either actively sending probe messages or passively listening to beacons. This discovery phase results in basic connectivity among mesh APs in the network. The WDS updates the association record to reflect the change of the mesh topology. After the initial discovery phase, beacon messages remain to be transmitted periodically for topology maintenance [14] [17].

2.2.3 Routing

In contrast to traditional single-hop WLANs, packet transmissions between mesh APs in WMNs rely on routing protocols. Mesh APs are pure layer-2 devices and layer-2 routing protocols based on the MAC address are used to handle packet transmissions in WMNs. There are mainly three kinds of routing protocols: proactive routing (table-driven), reactive routing (on-demand) and hybrid routing. In reactive routing protocols, nodes do not need a priori knowledge of network topology and need to find a route before transmitting any packet. The advantage of reactive routing is that it can adapt to the rapidly changing topology and avoid invalid routing [18]. However, a discovery delay is incurred when an application requires a route to a destination and the discovery process is required every time a source node needs to transmit data packets. In contrast, session establishment time in a proactive routing protocol is greatly reduced from an application's point of review, because routes to all destinations are already maintained by each node. Also, the proactive protocol is not affected by an increase in the number of active nodes due to the proactive route discovery mechanism [19]. The main disadvantage of proactive routing is the large overhead incurred in maintaining the routes. In WMNs, APs are fixed in the WMN and the connectivity among them is relatively static. APs generally handle a high volume of traffic and they can afford the overheads of maintaining the routing tables in the proactive routing. Thus, a proactive routing protocol is more suitable for WMNs than a reactive one. An alternative for routing in WMNs is the hybrid routing which incorporates both proactive and reactive routing. For example, a hybrid routing in [20] is based on the ad hoc on-demand distance vector [21] and the optimized link state routing [22].

2.2.4 Medium access control mechanisms

The medium access control provides channel access control mechanisms that make it possible for multiple users to communicate via a shared medium. Two categories of MAC protocols employed in IEEE 802.11-based WMNs are: i) Contention-based channel access, which includes DCF (distributed coordination function) and its enhancement EDCA (enhanced distributed channel access), and ii) Contention-free channel access, which includes PCF (point coordination function) and its enhancement HCCA (hybrid coordinator function controlled channel access) [23] [24].

Contention-based MAC

DCF is basically a carrier sense multiple access with collision avoidance (CSMA/CA) mechanism. A station wishing to transmit senses the medium. If the medium is busy, it defers its access to the medium for a random backoff period. If the medium is free for a specified time, called the distributed interframe space (DIFS) in the standard, the station is allowed to transmit [15]. DCF also has an optional virtual carrier sense mechanism that exchanges short request-to-send (RTS) and clear-to-send (CTS) frames in order to reduce frame collisions. DCF includes a positive acknowledgement scheme [25] [26].

EDCA introduces priorities to different types of traffic. High priority traffic has a shorter waiting time in contentions and thus has a higher chance of being sent than low priority traffic. In addition, each priority level is assigned a bounded time interval during which a station can send as many frames as possible which alleviates the problem of low rate stations gaining an inordinate amount of channel time in the legacy 802.11 DCF [27].

Contention-free MAC

PCF is a centralized and contention-free MAC. PCF relies on the central node AP to communicate with MSs. The AP polls the stations listed on its polling list one-by-one in a round-robin fashion. A waiting time period, PIFS (PCF interframe space), is at the beginning of a PCF transmission period. After that, the AP begins to poll the first station on its polling list. After receiving the polling packet, the polled station replies with an ACK packet after a SIFS (short interframe space) waiting time interval. If the polled station has data packets for the AP, a data packet is allowed to be piggybacked with the ACK packet. After sending out the polling packet, the AP waits for, at most, a PIFS time interval. If no packet is received from the polled station during the PIFS period, the AP removes the polled station from its polling list and polls the next station immediately. At the end of the PCF period, the AP issues a contention-free-end packet to terminate the PCF period. In DCF, the overhead comes from DIFS, random backoff, SIFS, RTS/CTS and ACK packets. In PCF, the overhead comes from PIFS, SIFS and Polling/ACK packets. PCF can also solve the hidden terminal problem inside the WLAN, since all the stations are centrally controlled by an AP [28].

HCCA works similar to PCF. During the contention-free period (CFP), the hybrid coordinator (HC) function controls the access to the medium while all stations work in EDCA during the contention period (CP). Traffic class and traffic stream are defined in HCCA. This means that HC can provide per-connection service and coordinate the connections in any fashion, not just the round-robin in PCF. Moreover, MSs are able to give traffic load information on a connection basis and request specific QoS requirements such as throughput and delay. HC can use the information to adjust its scheduling mechanism, for example, to give priority to one station/connection over another. Hence, HCCA is able to allow various real-time applications like VOIP to work more efficiently [29] [30].

Comparison of the two MACs

Now we discuss the advantages and disadvantages of the two MACs in IEEE 802.11-based WMNs. The main advantage of DCF/EDCA is its simplicity. But as a stochastic contention-based technique, it does not guarantee QoS like delay and bandwidth requirements of real-time applications. Moreover, CSMA/CA does not function well in a wireless multi-hop environment due to the instable throughput and unfairness caused by the hidden terminal problem, the exposed terminal problem, and the binary exponential backoff [31] [32]. It is shown that the capacity of CSMA/CA-based WMNs degrades significantly with the increase of traffic load and network size [33]. Therefore, the contention-based DCF/EDCA MAC is not a good choice for WMNs, especially for supporting real-time applications.

On the other hand, TDMA (time division multiple access)-based PCF/HCCA is a centralized protocol which is able to allocate wireless resources based on the various needs of connections and provide QoS support for them. In addition, the overhead of PCF/HCCA is lower compared to DCF/EDCA, resulting in more efficient recourses usage and higher capacity. Therefore, the contention-free PCF/HCCA is more suitable for WMNs for supporting real-time traffic.

2.3 IEEE 802.16a

The IEEE 802.16 standard aims to build WMANs. This standard initially defines the point-to-multipoint (PMP) mode and incorporates the mesh mode in IEEE 802.16a, which allows non-line-of-sight communications [34] [35]. A key difference between the PMP mode and mesh mode is the ability of the latter to enable multihop communications. While the PMP mode requires each subscriber station (SS) to be connected to a base station (BS), SSs can directly communicate with each other in the mesh mode. Hence, an SS in the mesh mode has the function of relaying traffic among the SSs and the BS [36] [37]. IEEE 802.16a defines some similar mesh concepts and mechanisms such as routing and security as in IEEE 802.11s. In the following, we only describe the major services and functions specific to the IEEE 802.16a-based mesh network.

2.3.1 Mesh network entry mechanism

IEEE 802.16a specifies a network entry mechanism to find sponsor mesh nodes and establish connectivity with neighboring nodes [38]. Active nodes in a mesh network periodically broadcast the mesh network configuration messages (MSH-NCFG) in order to maintain the network and allow new nodes to synchronize with the existing network. Upon entering the mesh network, a new SS actively scans for the MSH-NCFG message. After hearing the message, the SS acquires synchronization and initiates a network entry process. It sends a mesh network entry message (MSH-NENT) with a request to join the mesh network to the candidate sponsor node. If the request is accepted by the sponsor node, the SS will be assigned a channel to register with the BS. After the new SS is authorized to enter the mesh network by the BS, the SS will receive a 16-bit node identifier which uniquely identifies it within the mesh network. Then the SS can request bandwidth from the BS and establish connectivity with SSs other than the sponsor node [36] [38].

2.3.2 Medium access control

IEEE 802.16a supports a frame-based transmission. Each frame is divided into timeslots for the purpose of bandwidth allocation and packet transmissions [36]. Another key feature of the MAC is that it is connection-oriented where services are mapped into connections. The SS cannot transmit data until it has been allocated a timeslot and channel by the BS. This allows IEEE 802.16 to provide strong support for QoS [39] and also requires a scheduling mechanism which enables the BS to control QoS parameters by balancing the timeslot assignments among connections of SSs [40] [41].

2.3.3 Scheduling services

The IEEE 802.16 standard specifies five different scheduling services: unsolicited grant service (UGS), real-time polling service (rtPS), extended real-time polling service (ertPS), non-real-time polling service (nrtPS), and best effort (BE) service. In the UGS service, the BS offers a fixed size burst in timeslots to an SS periodically, and the SS does not have to make any explicit bandwidth requests. This is suitable for real-time CBR traffic. For the rtPS service, the BS periodically polls the SSs, which make bandwidth requests at specified uplink timeslots, and bandwidth grants are then broadcast from the BS through downlink transmissions. The rtPS service is designed for real-time services with variable packet generation rates. The ertPS service is a new addition in IEEE 802.16e. In this service, the BS keeps offering the same amount of bandwidth to the SS unless explicitly requested by the SS. The nrtPS service is designed to support non-real-time data which require variable bandwidth on a regular basis such as FTP (file transfer protocol). BE is to provide efficient services to best effort traffic.

IEEE 802.16a does not specify specific scheduling algorithms but leaves the venders to differentiate their equipment. The standard defines two types of scheduling: a centralized scheduling and a distributed scheduling. In the centralized scheduling, the BS is responsible for coordinating the packet transmissions in the entire mesh network. The BS collects bandwidth requests from the SSs and computes the uplink and downlink bandwidth grants to each SS in the network. Determination of the flow assignments for each SS, however, is not specified in the standard. The BS also dictates the number of timeslots dedicated to the centralized scheduling data which is recorded in a parameter called MSH-CSCH-DATA -FRACTION. The BS then sends this bandwidth grants information via the mesh schedule assignment (MSH-CSCH) message to SSs directly communicating with the BS. These SSs then forward the MSH-CSCH message to the SSs that are further away from the BS [38] [42].

In the coordinated distributed scheduling, the nodes (BS and SSs) are peers in coordinating their transmissions. The nodes express their bandwidth requirements and grants and confirm the grants in the MSH-DSCH-NUM portion of a control subframe [42]. The coordinations of packet transmissions are restricted in the two-hop neighborhood of each node, and do not rely on the operation of BS. In contrast, the uncoordinated distributed scheduling is able to build schedules in ad-hoc mode and on a link-by-link basis [38].

Although the distributed scheduling is more flexible and scalable, it is not efficient in QoS guarantee. In addition, the centralized scheduling ensures the collision-free packet transmission in the multi-hop network [38]. As a result, better QoS support and more efficient wireless resources utilization can be achieved in the centralized scheduling. Therefore, the centralized scheduling is more suitable for real-time applications in WMNs.

2.4 Related work on packet transmission scheduling

In this section, we present a review on scheduling approaches and schemes in various network environments in contrast and comparison to the scheduling problem in WMNs for supporting real-time applications.

2.4.1 Scheduling in packet-switched wireline networks

In packet-switched networks, scheduling plays an important role in enabling sharing resources such as buffers and link bandwidths. It determines which packets to serve and at what time in order to satisfy QoS of the traffic and fully utilize the network resources. Scheduling algorithms can be classified as either work-conserving where a server is always busy whenever there is a packet to send, or non-work-conserving where each packet is assigned an eligible time and will not be transmitted unless it is eligible, no matter whether the server is idle or busy [43].

Many work-conserving scheduling algorithms have been proposed and widely used in packet-switched networks [43]. Virtual clock (VC) [44] scheduling aims to emulate the time division multiplexing (TDM) system. A virtual transmission time is the time at which the packet would have been transmitted if the server was actually doing TDM. Packets are assigned virtual transmission time and transmitted in the increasing order of the virtual time. In earliest deadline first scheduling (EDF) [45] [46], each real-time packet has a deadline requirement. If delivered after the deadline, the packet is dropped. Packets are scheduled in the order of increasing deadlines. Weighted fair queueing scheduling [47] (WFQ) is a packet-based approximation of generalized processor sharing (GPS), an optimal but impractical scheduling discipline used as a scheduling performance benchmark. In WFQ, when the server is ready to transmit the next packet, it picks the packet that would complete the transmission first in the corresponding GPS system among all packets which are ready for transmission. These work-conserving scheduling algorithms provide basic scheduling approaches, but they assume a single-server environment. Hence, they cannot be applied to the scheduling in multi-hop WMNs.

Some non-work-conserving scheduling algorithms address multi-server network environment. Two examples are jitter earliest-due-date [48] and stop-and-go [49]. In jitter earliest-due-date, each packet is stamped with the time difference between its required deadline and the actual finishing time. The next server holds the packet for a period before the packet is eligible in order to ensure the bound on the delay jitter of the packet. In the stop-and-go scheduling, time is divided into frames. The packet that arrives during a frame time is postponed to be transmitted until the beginning of the next frame. Packets on the same frame at the source stay in the same frame throughout the network [43]. Non-work-conserving schedulers achieve higher average packet transmission delays but better delay jitter performance than the work-conserving counterparts [9] [43]. The basic mechanism for these algorithms to handle the multi-server scheduling is that a server recodes the traffic status at one hop and servers at later hops make use of the traffic status to improve network performance such as throughput and delay jitter. Servers schedule packets based on the recorded packet status, and do not jointly coordinate their scheduling decisions. So these scheduling schemes are not suitable to WMNs where coordinations among multiple APs are essential to ensure the successful packet transmissions between a pair of transmitter and receiver. In addition, these scheduling algorithms are designed for the wireline network and not obliged to consider such characteristics in the wireless network as wireless channel interference and multiple access situations and thus are not suitable for scheduling in WMNs.

2.4.2 Scheduling in single-server wireless networks

The design of scheduling algorithms is especially challenging given the interference and limited link bandwidth in a wireless network. A typical single-server wireless network consists of one BS and a number of MSs. Scheduling is implemented at the BS which communicates with all MSs. Most existing scheduling schemes assume that the BS has the channel states and packet queue status of its MSs [9].

Scheduling schemes for TDMA-based wireless networks have been extensively studied in the literature. Channel state dependent packet scheduling, for example [50] and [51], monitors channel states and defers transmissions when the channel is in bad state to avoid wasting wireless resources. One merit of this approach is that it is able to employ different service disciplines such as round robin and EDF for scheduling decisions. Idealized wireless fair queueing [52] incorporates a compensation mechanism for lagging sessions. Other scheduling algorithms for TDMA-based wireless networks can be found in review papers such as [9].

Scheduling algorithms for IEEE 802.11-based WLANs and IEEE 802.16-based WMANs in the PMP mode have also been proposed. Fair scheduling [53] [54] is implemented at APs and aims to achieve fairness among MSs. A scheduling scheme which combines the earliest deadline and weighted fair queue approaches is proposed in [55]. In [56], the scheduler at the BS decreases the amount of granted bandwidth for a voice connection by half when the connection is silent and changes back to the full amount of granted bandwidth after the connection is active. This mechanism is able to improve the network efficiency. In [57], priorities are given to different categories of connections and resources are allocated to satisfy the QoS of higher priority services first.

In addition to the above heuristic schemes, mathematical optimization is also an approach to solving the scheduling problem in single-server wireless networks. Fair scheduling in wireless networks under TDMA-based MAC framework [58] is formulated as an assignment problem. Dynamic channel-sensitive scheduling to optimize the wireless data throughput is presented in [59]. Packet scheduling with smart antennas is formulated in [60] as combinatorial optimization problems. In [61], a resource allocation problem is formulated and solved using linear programming relaxation.

The algorithms mentioned above focus on fairness, throughput and efficiency of wireless channels. In the single-server environment, the BS and MSs work in the simple client-server mechanism. All MSs are assumed to be available to the BS, and the BS makes scheduling decisions which all MSs will follow. Thus the coordination between the BS and MSs is simple. The above algorithms cannot be applied to WMNs as APs can communicate with multiple other APs and coordinating their timelines can be much more complicated.

2.4.3 Scheduling in MANETs

Scheduling in MANETs has been an ongoing research topic in both academia and industry. Most of the scheduling schemes, for example [62] and [63], work distributedly, where link activations are scheduled to achieve efficient channel utilization (high spatial reuse) and avoid collisions.

Clustering is employed in quasi-centralized schemes, for example [64] [65] [66]. Network nodes are grouped into clusters, each of which has a cluster head. The cluster head is responsible for scheduling within the cluster and forwarding packets among clusters. Because of the double roles of a cluster head, coordinating its functions within the cluster and among the cluster heads is an important and difficult issue for scheduling performance. In addition, the size of clusters affects the performance of the scheduling schemes and allocating nodes to different clusters and selecting the cluster heads are separate research topics.

Different TDMA-based scheduling schemes [67] [68] have been designed for MANETs. In [67], a topology transparent scheduling algorithm for a multi-channel TDMA system can provide maximal guaranteed throughput to each network node. This algorithm requires that multiple receivers are equipped at each station. In [68], a topology-independent scheduling is proposed using TDMA-based transmission scheduling vector. It focuses on topology transparency, fairness and high mobility of wireless nodes. It only guarantees that one timeslot is free of contention at each frame and does not consider timeslot coordination in the TDMA-based network. Based on a preset service priority, a polling-based 2-layer integrated multihop scheduling algorithm is proposed in [69] for both intra-cluster and inter-cluster scheduling. The polling mechanism employed is unable to give a higher priority to the packets with more strict delay requirement. As a result, the packet loss rate can be high.

Overall, because of the dynamic and distributed network environment in MANETs, most scheduling schemes for MANETs focus on throughput fairness/maximization, topology transparency, and other issues for supporting non-real-time traffic and cannot be applied directly to schedule real-time traffic in WMNs.

2.4.4 Scheduling in WMNs

There has been little work done on scheduling in multi-hop wireless mesh networks. Most of them, for example [70] and [71], use contention-based MACs where wireless nodes compete for medium access. The contention-based scheduling is more suitable for supporting non-real-time data traffic than real-time traffic as it does not guarantee the delay requirement. Furthermore, transmission collisions can significantly reduce the wireless resource utilization and network capacity.

There has been very limited work on real-time packet transmission scheduling in

WMNs. The spatial-temporal scheduling scheme proposed in [72] for IEEE 802.16-based mesh networks uses a station's moving direction and speed to select repeaters for forwarding its data packets. This is for a special application in the high-speed railway scenario. The routing and centralized scheduling scheme proposed in [73] provides QoS in IEEE 802.16 mesh networks. It reduces the network to a tree and presents a scheduling approach for interactive data applications. This work focuses on minimizing total number of timeslots for a given traffic load in a tree structure. A simple even-odd link activation framework is proposed in [74], where the odd and even nodes send packets alternatively in different time in order to avoid the conflict of simultaneous transmissions and receptions at each node. This scheme requires multiple radios at each AP for sending to or receiving from multiple nodes simultaneously. A relay strategy in [75] is designed for mesh nodes in a transmission tree by taking channel utilization and transmission delay into consideration. This strategy is based on a single-channel environment and schedules on a timeslot basis. In most cases, the scheme allows only one transmission link in each timeslot, and the channel utilization is rather low.

2.5 Summary

Compared to the contention-based channel access, a TDMA-based MAC protocol is better for supporting real-time traffic in WMNs due to several reasons: i) QoS of traffic can be satisfied more easily; ii) the channel resources can be utilized more efficiently when traffic load is relatively high; and iii) it is more convenient to use centralized control in TDMA.

Although extensive work has been done on traffic scheduling, existing scheduling schemes are not suitable for scheduling real-time traffic in WMNs. Careful timeline coordinations are important for satisfactory delay performance as well as efficient radio resource utilization in scheduling real-time traffic in WMNs.

Chapter 3

Problem Formulation and Optimum Scheduling

In this chapter we first describe the scheduling problem in WMNs and then formulate the scheduling for real-time CBR traffic as an optimization problem. The objective of the optimization is to minimize the total amount of radio time for supporting the real-time CBR traffic, given the number of the connections and their respective routes in the WMN. This is subject to the delay requirement of the packets, available radio resources, and co-channel interference condition. System performance, including network capacity and packet transmission delay, is then studied based on the optimum scheduling solution.

3.1 System description

We consider an infrastructure-based WMN where mesh APs are interconnected through one or more hops. Each AP is equipped with one radio for communicating with its associated MSs as well as with other APs. A more general case where each AP may be equipped with multiple radios is considered in Section 6.2. A number of MSs are associated to each AP which is referred to as the home AP of the MSs. We consider that mesh APs are fixed and MSs are relatively static so that mobility is not a problem.

Each AP stores a routing table which records the next hop AP for specific destinations. We assume that the routing tables are relatively static. Dynamically changing the routing tables is possible, but may trigger re-scheduling, which introduces extra overhead.

A connection may carry internal or access traffic. The internal traffic flows between end stations within the same WMN and the access traffic is between an MS in the WMN and a station in the wireline network. We assume that the wireline network always has sufficient resources to support a connection as long as the WMN can accept it. We further assume that the transmission delay introduced by the wireline network can be neglected, compared to that in the wireless network. Both of the assumptions on the wireline network are true in most practical networks. In the remaining part of the thesis, we only consider packet transmission scheduling in the WMN.

Each hop along the route of a connection has an index number. For access traffic, the first hop in the uplink (from the MS to the wireline network) is the hop from the MS to its home AP, and the first hop in the downlink (from the wireline network to the MS) is the hop from the root AP to the next downstream AP. For internal traffic, the first hop is the hop from the source MS to its home AP and the last hop is the hop from the home AP of the destination MS to the destination MS. The index number of the hops increases along the direction where packets are forwarded.

Channel time is divided into equal size timeslots so that one timeslot is exactly for one packet transmission, including the data frame, ACK, and necessary channel idle time as the MAC protocol specifies, such as the interframe spaces in IEEE 802.11. Below all time is normalized to the duration of one timeslot. Note that the condition of equal-size timeslots is only for simplifying the presentation but is not necessary for the proposed scheduling solutions to work. TDMA is adopted for serving real-time traffic. Compared to contention-based MACs such as CSMA/CA, TDMA allows centralized timeslot assignments and is more efficient for providing strict latency requirement.

A root AP, which is directly connected to the wireline backbone, is a central station responsible for admission control and resource allocation for all QoS traffic in the network. A WMN may have multiple root APs, one of which then serves as the central station for resources management and admission control. We assume that a certain amount of resources in each AP is reserved for control signaling exchanges for admission control and other network management purposes, and there is no transmission error in the control signal transmissions.

In addition to signaling exchanges, the root AP needs a way to gather the network and traffic information. A new connection with strict QoS requirements should pass admission control before entering the system. When making a QoS connection request to its home AP, the MS specifies the traffic characteristics and QoS requirements of the connection. Upon receiving the connection request, the home AP forwards the request, together with the parameters of the connection and its QoS requirements, to the root AP through one or more hops. An admission decision is made at the root AP and forwarded back to the home AP hop by hop. Therefore, the root AP has information about the traffic load in the network, including the number of connections associated to each AP, their traffic parameters and QoS requirements. The root AP also has information about resource availability at each AP. The admission decision should guarantee that all APs along the route of the new connection have sufficient resources, and accepting the new connection will still guarantee the QoS of all existing connections. Details about the admission control in a WMN can be found in [76] [77].

A limited number of frequency channels are available for the network. Co-channel interference is present among APs sharing the same frequency channel. A packet cannot be received successfully if more than one AP within the interference range of the considered receiver is transmitting at the same frequency channel. We consider that a frequency channel assignment scheme is in place and the frequency assignments are relatively static for the duration of the real-time connections. Frequency channel assignment in a WMN is another important and challenging research topic and is beyond the scope of this thesis. Several frequency assignment schemes are proposed in [78] [79] [80].

We consider homogeneous real-time CBR connections. (Scheduling for real-time VBR traffic will be performed in Section 6.1.) The WMN may support other types of traffic, but the real-time service is given the highest priority and its performance is not affected by other traffic. We define a scheduling interval (SI) so that scheduling decisions are repeated in different SIs unless traffic load is changed. Each SI consists of two parts: a real-time portion and a non-real-time portion, for processing real-time and non-real-time data traffic, respectively. For example, in legacy IEEE 802.11 protocol, the SI can be a superframe, and the real-time portion and non-real-time portion, respectively, correspond to the CFP and CP in the same superframe. Similarly, the SI can be a superframe in IEEE 802.16, which also includes contention-free timeslots for real-time services (e.g., UGS and rtPS services) and contention-based timeslots for non-real-time services (e.g., nrtPS and BE services) [40]. Each CBR source generates a bit sequence at a constant rate. The generated bit sequence is packed into fixed-size packets which arrive at the system periodically. We set the duration of an SI to be equal to the packet inter-arrival time of a CBR connection. That is, one packet is generated from each CBR connection in every SI. To provide a constant service rate to each CBR connection, one packet should be served for the connection in every SI. In this way, it is guaranteed that packets generated in one SI are delivered to the next hop before new packets from the same connection arrive at the same hop in the next SI. Therefore, we only need to describe the scheduling of each AP in a typical SI. All other SIs of the AP repeat the same schedule. Below we use T_{SI} to denote the normalized duration of one SI with respect to a timeslot duration, i.e., T_{SI} is the number of timeslots in each SI.

Comment: For heterogeneous CBR traffic, different connections may have different packet inter-arrival times. In this case, the duration of the SI can be set to be the least common multiple of all the packet inter-arrival times (in number of timeslots). Then one or multiple packets (depending on the ratio of the least common multiple of all connections to the SI of a particular connection) should be served for the connection during one SI. The proposed scheduling schemes can be easily extended to this case.

3.2 Timeline coordination

The task of the scheduler is to coordinate timelines of APs in the WMN for packet transmissions in order to avoid any conflict in their timeline operations and unacceptable co-channel interference, while achieving satisfactory end-to-end transmission delay for real-time packets.

Let $\mathcal{M}$ be a set of all APs, indexed by $m = 1, 2, \ldots, M$, and $\mathcal{N}$ be a set of CBR connections, indexed by $i = 1, 2, \ldots, N$, where M and N are the total number of APs and CBR connections, respectively. Since each CBR connection generates one packet in each scheduling interval, we use "packet i" to denote the packet to be scheduled for connection i in the considered SI. We consider that $M > 1$ and $N > 1$, i.e., there is more than one AP and more than one connection in the WMN.

Each connection is processed by a set of APs in a predefined order based on the routing protocol. Let $\mathcal{R}_i$ be a set of the APs along the route of connection i. For $m \in \mathcal{R}_i$, we use m_i^- to denote the immediate upstream AP of AP m. For a special case, if AP m is the first AP along the route of the connection, AP m_i^- does not exist. Let $\tau_{i,m}$ be the time when packet i is transmitted by AP m ($m \in \mathcal{R}_i$), then $[\tau_{i,m}]$ is an $N \times M$ matrix (although we are only interested in values of $\tau_{i,m}$'s for $i \in \mathcal{N}$ and $m \in \mathcal{R}_i$). Based on the above notations, we can use τ_{i,m_i^-} to denote the time when packet i is received by AP m.

The transmission time for a packet at an upstream hop must be smaller than that at a downstream hop. That is

$$\tau_{i,m_i^-} < \tau_{i,m}, \quad m_i^-, m \in \mathcal{R}_i. \tag{3.1}$$

The one hop transmission delay for connection i is $\tau_{i,m} - \tau_{i,m_i^-}$, and the total multihop delay should be less than d_i, the normalized maximum delay budget. That is,

$$\sum_{m_i^-, m \in \mathcal{R}_i} (\tau_{i,m} - \tau_{i,m_i^-}) + 1 \leq d_i, \tag{3.2}$$

where the extra one timeslot is for the packet transmission between the active MS and its home AP.

Different packets may traverse the same AP, but each AP can only process (either receive or transmit) one packet at a time. Let $\mathcal{N}_m$ be the set of connections processed by AP m and $i, j \in \mathcal{N}_m$. We have (i) $\tau_{i,m} \neq \tau_{j,m}$ for $i \neq j$, meaning that AP m cannot transmit the two packets at the same time; (ii) $\tau_{i,m_i^-} \neq \tau_{j,m_j^-}$ for $i \neq j$, that is, AP m cannot receive the two packets at the same time; and (iii) $\tau_{i,m} \neq \tau_{j,m_j^-}$, i.e., AP m cannot receive and transmit at the same time.

The real-time portion of AP m is the maximum time difference between the times when AP m processes (transmits or receives) the real-time packets, i.e.,

$$T_{rt,m} = \max_{i,j \in \mathcal{N}_m} (|\tau_{i,m} - \tau_{j,m_j^-}| + 1, |\tau_{i,m} - \tau_{j,m_j}| + 1, |\tau_{i,m_i^-} - \tau_{j,m_j^-}| + 1). \tag{3.3}$$

The transmitting time is always later than the receiving time for a given packet at a given hop, i.e., $\tau_{j,m_j^-} < \tau_{j,m}$ and $\tau_{i,m_i^-} < \tau_{i,m}$. Then (3.3) is reduced to

$$T_{rt,m} = \max_{i,j \in \mathcal{N}_m} |\tau_{i,m} - \tau_{j,m_j^-}| + 1. \tag{3.4}$$

Since the real-time portion is defined within one SI, we have

$$T_{rt,m} \leq T_{SI}, \quad m \in \mathcal{M}. \tag{3.5}$$

Co-channel interference also affects the timeline coordination. We define $\mathbf{I}_{m,f}$ as a set of APs that are located within the interference range of AP m and assigned frequency channel f. When "an AP is assigned to frequency channel f", the AP uses the channel to transmit, and the corresponding receiver is tuned to the same channel. AP m is able to receive a packet successfully at frequency channel f only if AP m_i^- is the only AP in $\mathbf{I}_{m,f}$ to transmit at the same time, i.e.,

$$|\mathbf{I}_{m,\tau_{i,m_i^-},f}| \leq 1, \; m_i^-, m \in \mathcal{M} \tag{3.6}$$

where $\mathbf{I}_{m,t,f} \subseteq \mathbf{I}_{m,f}$ is a set of APs which belong to set $\mathbf{I}_{m,f}$ and are scheduled to transmit at time t, and $|\cdot|$ denotes the cardinality of a set. Note that the above condition in general also guarantees that AP m_i^- can transmit at the time only if it is not in the interference range of any AP that has been scheduled to receive in the same frequency channel and at the same time.

3.3 Objective functions

For a given number of real-time connections, it is desirable that the real-time portion to serve these connections is minimized in order to leave more time for serving non-real-time traffic. For example, in legacy IEEE 802.11 network this is equivalent to minimize the CFP so as to leave more time for serving contention-based data traffic. For this purpose the objective is to minimize the maximum of $T_{rt,m}$'s over all APs in the WMN, i.e.,

$$\min \max_{m \in \mathcal{M}} T_{rt,m}. \tag{3.7}$$

Note that an AP may not be busy at all timeslots within the real-time portion, as its corresponding AP may be busy in communicating with another AP or the interference condition prevents it from transmitting or receiving at the time. Because of the mesh topology, each AP usually needs to forward packets to or from multiple other APs, and the timeline arrangments of all the APs can be dependent on each other. Therefore, completely removing such idle timeslots from the real-time portion of the APs is impossible in some networks. However, the number of "idle" timeslots within the real-time portion can be minimized with the objective of minimizing the real-time portion.

On the other hand, minimizing the maximum real-time portion does not necessarily result in the optimum delay performance. An alternative objective function is to minimize the maximum packet transmission delay. That is

$$\min \max_{i \in \mathcal{N}} w_i, \tag{3.8}$$

where $w_i = \sum_{m_i^-, m \in \mathcal{R}_i} (\tau_{i,m} - \tau_{i,m_i^-}) + 1$. The optimum scheduling problem formulation and approach to solving the optimization problem for the two objective functions are very similar, and below we only present details with (3.7) as the objective function.

3.4 Optimization problem

Based on the above discussions, the scheduling problem for supporting CBR real-time traffic in a WMN can be formulated as:

Problem 3.1

$$\min_{[\tau_{i,m}],\,[T_{rt,m}]} \max_{m\in\mathcal{M}} T_{rt,m} \tag{3.9}$$

$$\text{s.t.}\quad \tau_{i,m_i^-} < \tau_{i,m},\;\; m_i^-, m \in \mathcal{R}_i,\; i \in \mathcal{N} \tag{3.10}$$

$$\sum_{m_i^-, m\in\mathcal{R}_i} (\tau_{i,m} - \tau_{i,m_i^-}) + 1 \le d_i,\; i \in \mathcal{N} \tag{3.11}$$

$$T_{rt,m} = \max_{i,j\in\mathcal{N}_m} |\tau_{i,m} - \tau_{j,m_j^-}| + 1,\; m_j^-, m \in \mathcal{M} \tag{3.12}$$

$$T_{rt,m} \le T_{SI},\;\; m \in \mathcal{M} \tag{3.13}$$

$$\tau_{i,m} \ne \tau_{j,m},\;\; m \in \mathcal{R}_i \cap \mathcal{R}_j,\; i,j \in \mathcal{N},\; i \ne j \tag{3.14}$$

$$\tau_{i,m_i^-} \ne \tau_{j,m_j^-},\;\; m_j^-, m_i^- \in \mathcal{R}_i \cap \mathcal{R}_j,\; i,j \in \mathcal{N},\; i \ne j \tag{3.15}$$

$$\tau_{i,m} \ne \tau_{j,m_j^-},\;\; m_j^-, m \in \mathcal{R}_i \cap \mathcal{R}_j,\; i,j \in \mathcal{N} \tag{3.16}$$

$$|\mathbf{I}_{m,\tau_{i,m_i^-},f}| \le 1,\; m_i^-, m \in \mathcal{M},\; i \in \mathcal{N} \tag{3.17}$$

$$\tau_{i,m} \in Z^{++},\; i \in \mathcal{N},\; m \in \mathcal{M} \tag{3.18}$$

where $[T_{rt,m}]$ denotes the vector with $T_{rt,m}$ ($m \in \mathcal{M}$) as its entry.

Problem 3.1 is not a standard optimization problem and should be reformulated. First of all, let $T_{rt,\max} = \max_{m\in\mathcal{M}} T_{rt,m}$. The objective function is then converted to

$$\min_{[\tau_{i,m}],\, T_{rt,m}} T_{rt,\max}. \tag{3.19}$$

This introduces an additional constraint: $T_{rt,\max} = \max_{m\in\mathcal{M}} T_{rt,m}$, which is equivalent to

$$T_{rt,m} \le T_{rt,\max}, \quad m \in \mathcal{M}. \tag{3.20}$$

Constraint in (3.12) is equivalent to the following inequality constraint

$$|\tau_{i,m} - \tau_{j,m_j^-}| + 1 \le T_{rt,m}, \ m_j^-, m \in \mathcal{M}, \ i,j \in \mathcal{N}_m. \tag{3.21}$$

Based on the definition of the absolute value function, we have

$$\tau_{i,m} - \tau_{j,m_j^-} + 1 \le T_{rt,m}, \ m_j^-, m \in \mathcal{M}, \ i,j \in \mathcal{N}_m, \tag{3.22}$$

and

$$\tau_{j,m_j^-} - \tau_{i,m} + 1 \le T_{rt,m}, \ m_j^-, m \in \mathcal{M}, \ i,j \in \mathcal{N}_m. \tag{3.23}$$

Together with (3.20), we have

$$\tau_{i,m} - \tau_{j,m_j^-} + 1 \le T_{rt,\max}, \ m_j^-, m \in \mathcal{M}, \ i,j \in \mathcal{N}_m, \tag{3.24}$$

and

$$\tau_{j,m_j^-} - \tau_{i,m} + 1 \le T_{rt,\max}, \ m_j^-, m \in \mathcal{M}, \ i,j \in \mathcal{N}_m. \tag{3.25}$$

Now we look at the constraints in (3.14)-(3.16). Since these three constraints have the same inequality format, we only take (3.14) as an example. The straightforward approach is to rewrite it as "$\tau_{i,m} < \tau_{j,m}$ or $\tau_{i,m} > \tau_{j,m}$". Then for each inequality specified by (3.14), the problem is divided into two sub-problems, and the optimal result of $T_{rt,\max}$ is the smallest among all the sub-problem solutions. However, the total number of such sub-problems will be very large with the increase of the number of connections and APs in the WMN, which leads to an unmanageable situation. Below we propose a different approach to reformulating the optimization problem by introducing a set of binary variables $\lambda_{i,m,x}$'s as

$$\lambda_{i,m,x} = \begin{cases} 1, & \text{if packet } i \text{ is transmitted by AP } m \text{ at time } x \\ 0, & \text{otherwise} \end{cases} \tag{3.26}$$

Based on the definition of m_i^- and (3.26), $\lambda_{i,m_i^-,x} = 1$ means that AP m is receiving packet i at time x. At the same time, we introduce one additional constraint

$$\sum_{x=1}^{d_i} \lambda_{i,m,x} = 1, \;\; i \in \mathcal{N}, m \in \mathcal{R}_i, \tag{3.27}$$

because a packet only needs to be transmitted once by each AP along its route.

With (3.26) and (3.27), (3.14)-(3.16) together can be rewritten in terms of $\lambda_{i,m,x}$'s as

$$\sum_{i \in \mathcal{N}_m} (\lambda_{i,m,x} + \lambda_{i,m_i^-,x}) \leq 1, \;\; m_i^-, m \in \mathcal{R}_i, \;\; x = 1, ..., d_i. \tag{3.28}$$

Given values of i and m, $\lambda_{i,m,x} = 1$ only when AP m is along the route of connection i and $\tau_{i,m} = x$. Therefore, $\tau_{i,m}$ can be expressed in terms of $\lambda_{i,m,x}$'s as

$$\tau_{i,m} = \sum_{x=1}^{d_i} x \cdot \lambda_{i,m,x}. \tag{3.29}$$

Then $\tau_{i,m}$ in (3.10)-(3.12) can be expressed in terms of $\lambda_{i,m,x}$'s accordingly.

The interference constraint in (3.17) is rewritten as

$$\sum_{i \in \mathcal{N}} \sum_{m' \in \mathbf{I}_{m,f}} \lambda_{i,m',x} \leq 1, \;\; m \in \mathcal{M}, \tag{3.30}$$

which means that at any time x, at most one AP using frequency f and in the interference range of AP m is scheduled to transmit a packet to AP m.

Now Problem 3.1 can be reformulated as the following standard integer linear programming (ILP) problem:

Problem 3.2

$$\min_{[\lambda_{i,m,x}],\, T_{rt,\max}} T_{rt,\max} \tag{3.31}$$

$$\text{s. t.} \quad \sum_{x=1}^{d_i} x \cdot \lambda_{i,m_i^-,x} < \sum_{x=1}^{d_i} x \cdot \lambda_{i,m,x}, \quad m_i^-, m \in \mathcal{R}_i, \quad i \in \mathcal{N} \tag{3.32}$$

$$\sum_{m_i^-, m \in \mathcal{R}_i} \left(\sum_{x=1}^{d_i} x \cdot \lambda_{i,m,x} - \sum_{x=1}^{d_i} x \cdot \lambda_{i,m_i^-,x}\right) + 1 \le d_i, \; i \in \mathcal{N} \tag{3.33}$$

$$\sum_{x=1}^{d_i} x \cdot \lambda_{i,m,x} - \sum_{x=1}^{d_i} x \cdot \lambda_{j,m_j^-,x} + 1 \le T_{rt,\max}, \quad i,j \in \mathcal{N}_m, \quad m_j^-, m \in \mathcal{M} \tag{3.34}$$

$$\sum_{x=1}^{d_i} x \cdot \lambda_{j,m_j^-,x} - \sum_{x=1}^{d_i} x \cdot \lambda_{i,m,x} + 1 \le T_{rt,\max}, \quad i,j \in \mathcal{N}_m, \quad m_j^-, m \in \mathcal{M} \tag{3.35}$$

$$T_{rt,\max} \le T_{SI} \tag{3.36}$$

$$\sum_{i \in \mathcal{N}_m} (\lambda_{i,m,x} + \lambda_{i,m_i^-,x}) \le 1, \quad m_i^-, m \in \mathcal{R}_i, \; x = 1, ..., d_i \tag{3.37}$$

$$\sum_{x=1}^{d_i} \lambda_{i,m,x} = 1, \quad i \in \mathcal{N}, m \in \mathcal{R}_i \tag{3.38}$$

$$\sum_{i \in \mathcal{N}} \sum_{m' \in \mathrm{I}_{m,f}} \lambda_{i,m',x} \le 1, \; m \in \mathcal{M}, \; x = 1, ..., d_i \tag{3.39}$$

$$\lambda_{i,m,x} \in \{0,1\}, i \in \mathcal{N}, m \in \mathcal{M}, \; x = 1, ..., d_i \tag{3.40}$$

where $[\lambda_{i,m,x}]$ denotes the matrix with $\lambda_{i,m,x}$ as its entry.

The branch and bound algorithm [81] [82] is used for solving the ILP problem. The algorithm could potentially search all $2^{\sum_{i=1}^{N}(H_i \cdot d_i)}$ binary integer vectors to find the optimal point, where H_i is the number of hops of connection i. Therefore, with the increase of number of connections and their number of hops, the computation time to solve the optimization problem increases exponentially. Hence, solving the above optimization problem is not a practical solution to the scheduling problem in the WMN, and heuristic scheduling schemes should be proposed. Nevertheless, the solution to the optimization problem provides the global optimal scheduling, which is an important benchmark for designing practical scheduling schemes. In the remaining part of this chapter we look at performance using the optimum scheduling solution.

3.5 Numerical results and discussions

3.5.1 Simulation system setting

We consider the WMN as shown in Fig. 3.1 where 9 APs form a 3×3 grid[1]. AP 1 is the root AP and the solid line represents the wired link. The transmission range and interference range of each AP is 150m and 250m, respectively. MSs are uniformly distributed in the network coverage area, and each MS is associated to the AP closest

[1]Due to the high complexity of solving the optimization problems, obtaining optimal scheduling solutions for larger WMNs is beyond the capability of our computing facilities. Scheduling in larger WMNs will be tested using the proposed heuristic schemes in later chapters.

to it. All results are for access traffic unless explicitly indicated for internal traffic. We use constant rate voice traffic as an example of the real-time CBR traffic. Each voice connection includes a two-way communication in both the uplink and downlink directions. One voice packet is generated in every 20ms from each voice connection in each direction, and the maximum delay tolerance of the packets is 60ms. Each connection selects the route with the minimum number of hops to the destination. Each MS can have at most one connection. The default channel transmission rate is 2Mbps. Based on the parameters for a typical IEEE 802.11 mesh network in [83], the time for transmitting one packet is 1.39ms, which includes the transmission time for the voice packet, ACK and the interframe space. Then the duration of one timeslot is 1.39ms, and each SI includes 14 timeslots. Let c denote the total number of frequency channels available in the network. When $c = 1$, all APs share the same frequency channel, which represents a case with strong co-channel interference. When $c = 3$, each of the following three APs share one frequency channel, APs 1, 6, and 8, APs 2, 4, and 9, and APs 3, 5, and 7. This represents a case with moderate co-channel interference. When $c = 9$, each AP has its own frequency channel and there is no co-channel interference.

In the remaining part of this section we first do snapshot experiments and collect values of $T_{rt,\max}$ and $w_{\max}$ by solving the optimization problems with (3.7) and (3.8) as objective functions, respectively. We then consider a WMN where new connection

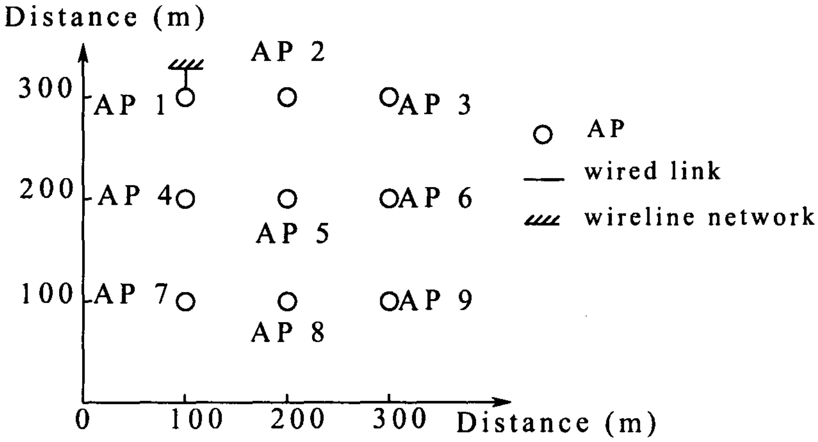

Figure 3.1: A WMN with 3×3 topology

requests arrive as a Poisson process and collect the connection level performance by solving the optimization problem with (3.7) as the objective function.

3.5.2 $T_{rt,\max}$ and $w_{\max}$ vs. number of connections

In this experiment we fix the number of hops of all connections, i.e., $H_i = H$ for all $i = 1, 2, \dots, N$, and vary N, the total number of connections. We set $H = 3$, so that MSs with active traffic can only be associated with five APs (APs 3, 6, 7, 8, and 9). When the total number of connections is an integer multiple of five, an equal number of active MSs are associated to each of the five APs; otherwise we distribute the number of connections so that the difference between the numbers of active MSs associated to any two of the five APs is at most 1. The purpose of this setting is to minimize the bottleneck effect (This will be discussed in details in Section 6.2).

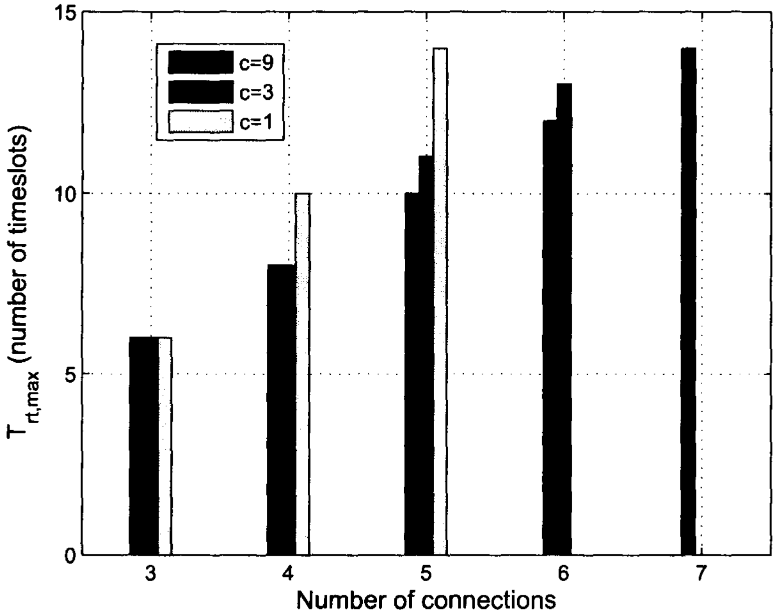

Figure 3.2: Maximum real-time portion vs. number of connections ($H = 3$)

Fig. 3.2 shows that $T_{rt,\max}$ monotonically increases with the number of connections due to the fact that more timeslots are consumed at the APs when more connections traverse them. We also observe from the figure that for a given number of connections, $T_{rt,\max}$ can be larger if a fewer number of frequency channels are available. For example, for $N = 5$, $T_{rt,\max}$ is 10, 11 and 14, respectively, when $c = 9$, 3 and 1. This is because when more APs share one frequency channel, co-channel interference among them results in more timeslots unavailable for multiple APs to transmit simultaneously. It is also shown that co-channel interference does not affect $T_{rt,\max}$ when the number of connections is relatively small, for example, $N = 3$. This is because the pressure of timeline coordination in a network with light traffic load

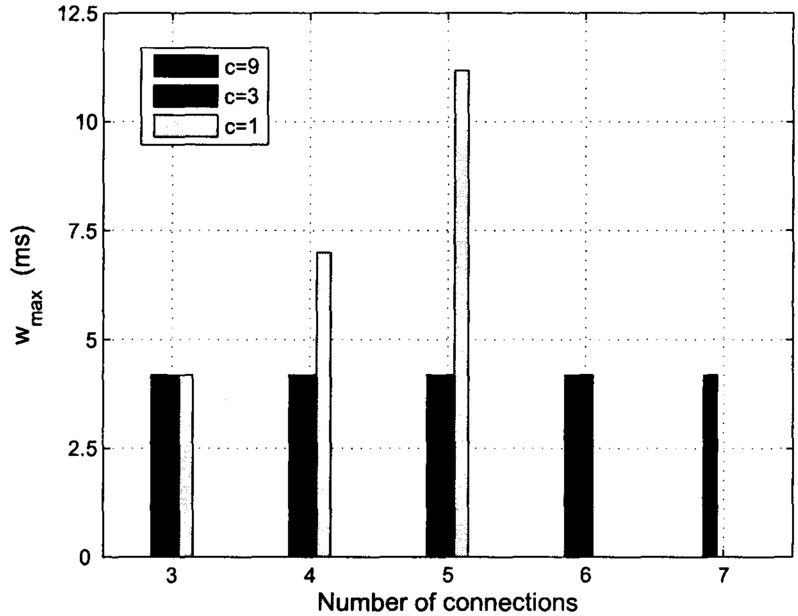

Figure 3.3: Maximum packet transmission delay vs. number of connections ($H = 3$)

is much less—it is easier to find a real-time portion within an SI without any "idle" timeslots in between. The effect of co-channel interference is also reflected by the system capacity. When $c = 9$, at most 7 3-hop connections can be supported, while only 6 and 5 connections can be supported for $c = 3$ and 1, respectively.

Fig. 3.3 shows the maximum packet transmission delay w_{max} vs. the number of connections. We observe that when $c = 9$ and 3, w_{max} remains the same as N is increased, which indicates that the packet transmission delay is independent of traffic load. This is because with $c = 9$ and 3, it is possible to schedule the connections so that there is no delay caused by buffering packets in intermediate APs. However, when $c = 1$, the maximum packet transmission delay increases with the number of connections. This results from more timeslots being unavailable due to high interference, causing packets to experience longer wait time at APs. The effect of interference is reflected in the increase in delay when c is decreased from 3 to 1 for a given number of connections. For instance, $w_{\max}$ is increased from 4.2ms to 11.2ms when $N = 5$.

Both bandwidth availability at the APs and delay budget of the traffic can affect the feasibility of the scheduling problem. It is shown from Figs. 3.2 and 3.3 that compared to the delay budget, the available bandwidth is the dominant factor that determines how many connections can be supported in the considered WMN. When $T_{rt,m} = T_{rt,\max} = 14$, all available bandwidth of AP m is occupied and the system reaches its capacity, while the maximum packet transmission delay is much lower than the delay budget. For example, when $c = 9$ and $N = 7$, $T_{rt,\max}$ is 14 timeslots which is equal to the total number of timeslots in one SI, while $w_{\max}$ is 4.2ms which is much less than the delay budget of 60ms.

We also perform the experiments by further setting the total number of connections to 7 for $c = 1$ and 8 for $c = 3$, respectively, and the optimization problem becomes infeasible.

3.5.3 $T_{rt,\max}$ and $w_{\max}$ vs. number of hops

Next we fix the total number of connections to $N = 5$ and vary H, the number of hops of the connections. All connections still have the same number of hops. When $H = 1$, all active MSs are associated to the root AP; when $H = 2$, active MSs can be associated to three APs (APs 2, 4, and 5); and when $H = 3$, active MSs can be associated to five APs (APs 3, 6, 7, 8, and 9). For $H = 2$ and 3, the active MSs are as evenly distributed among the respective APs as possible, i.e., the difference between numbers of active MSs associated to any two of the three (for $H = 2$) or five (for $H = 3$) APs is at most 1. Note that for $H = 1$, 2 and 3, all connections go through the root AP, but not necessarily the other APs. Figs. 3.4 and 3.5 show the $T_{rt,\max}$ and $w_{\max}$ values, respectively, in this experiment.

It is shown in Fig. 3.4 that when $c = 9$, $T_{rt,\max}$ remains the same when the number of hops is increased from 1 to 3. Since the network bottleneck is at the root AP in the considered network (the bottleneck effect will be discussed in Chapter 6), $T_{rt,\max}$ is determined by the load of the root AP. Without co-channel interference ($c = 9$), extra hops of the connections only increases the traffic load of non-root APs but not the root AP, and therefore $T_{rt,\max}$ is not affected. However, this is not the case when co-channel interference exists, i.e., for $c = 3$ or 1. When $c = 3$, although $T_{rt,\max}$ is not affected by increasing H from 1 to 2, it is increased when H is increased from 2 to 3. When $c = 1$, $T_{rt,\max}$ is increased when H is increased from 1 to 2 and from 2 to 3. The

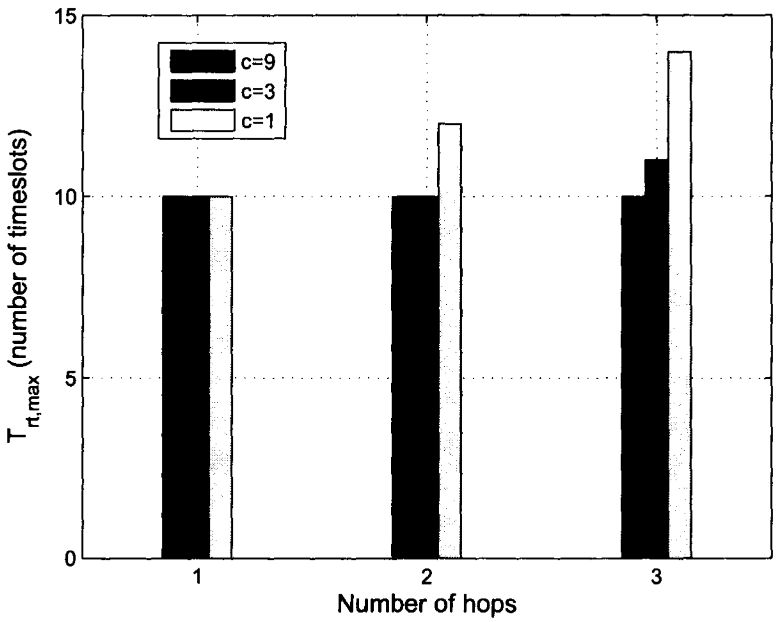

Figure 3.4: Maximum real-time portion vs. number of hops ($N = 5$)

reason that $T_{rt,\max}$ is affected by the number of hops of the connections is because of the co-channel interference which causes more timeslots unavailable for more than one AP to transmit simultaneously. As a result, there have to be some "idle timeslots" within the real-time portion of the APs. The number of such "idle timeslots" increases with the level of co-channel interference and traffic loads of individual APs.

Fig. 3.5 shows that $w_{\max}$ monotonically increases with the number of hops due to the fact that each extra hop introduces more transmission time to the delay. We can further find from the figure that $w_{\max}$ increases much faster when $c = 1$, compared to the cases when $c = 3$ and 9. This is due to the increased co-channel interference which makes packets to be buffered for a longer time in order to wait for the next

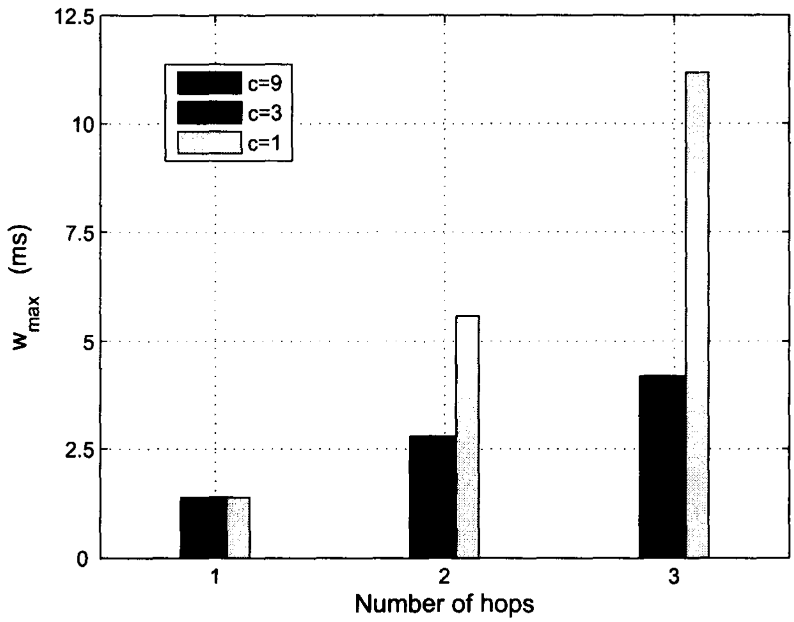

Figure 3.5: Maximum packet transmission delay vs. number of hops ($N = 5$)

hop transmission.

3.5.4 Poisson traffic

We then consider a dynamic traffic scenario and look at the connection blocking performance. CBR connections arrive at the network according to a Poisson process. Active MSs are selected so that they are associated to each of the APs in the network with equal probability. Upon receiving a new connection request, the root AP solves the optimization problem by assuming that the new connection is already in the system. If a feasible schedule is found, the new connection is accepted. Otherwise, it is rejected. The connection blocking rate is a ratio of the number of rejected

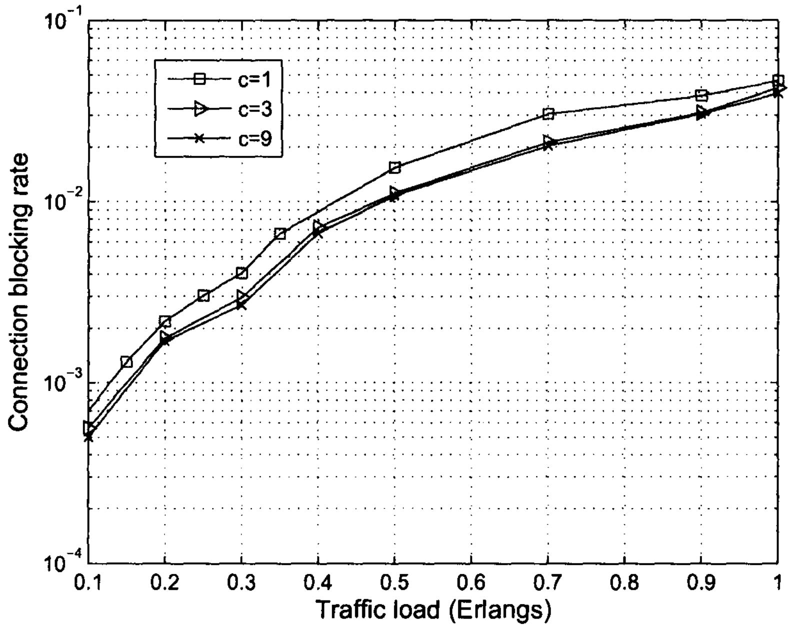

Figure 3.6: Connection blocking rate: optimum scheduling (C =1Mbps)

connections to the total number of connection requests. The duration of accepted connections follows an exponential distribution. We define the WMN capacity as the maximum offered traffic load in Erlangs so that the connection blocking rate is below a predefined value. In the following simulations, we set the average connection duration to 60s. We also set the physical channel transmission rate (C) to 1Mbps in order to reduce the running time of solving the optimization problem.

Fig. 3.6 shows that the connection blocking rate increases with the offered traffic load in Erlangs. For a given offered traffic load, the connection blocking rate is lower if more frequency channels are available in the network due to reduced co-channel interference. However, it is also seen from Fig. 3.6 that there is only very minor performance improvement by increasing the total number of available frequency channels from 1 to 9. This is because the system capacity is mainly limited by the APs with heaviest traffic load, or the bottleneck APs, which are usually the root AP or the APs closest to the root AP [84], and increasing the total number of frequency channels in the network does not affect the capacity of the bottleneck APs.

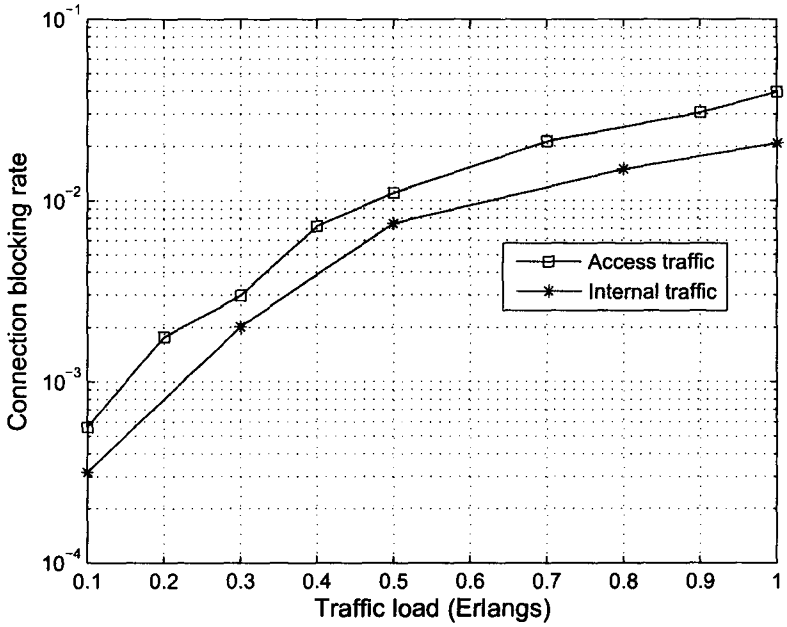

Figure 3.7: Connection blocking rate: access and internal traffic (C =1Mbps)

Although all previous experiments are for access traffic, the optimum scheduling can also be applied for internal traffic. Fig. 3.7 shows the connection blocking rate of both the internal and access traffic when $c = 3$. When generating the internal traffic, two MSs are randomly selected to be the source and destination stations for each internal connection. A shortest route between the two home APs is chosen for the connection. For a given offered traffic load, the internal traffic has a lower connection blocking rate than the access traffic. This is because the internal traffic is more evenly distributed in the WMN than the access traffic. The system capacity for internal traffic is less restricted by a single or a small number of APs than the capacity for access traffic. For a given blocking rate of 1%, the system capacity for internal and access traffic is 0.48 and 0.62 Erlangs, respectively. Because of the dependence of scheduling and admission control performance on the distribution of the connections in the network, routing can strongly affect the performance of both scheduling and admission control for internal traffic, while joint routing and scheduling for internal traffic is a much more difficult issue than scheduling for internal traffic and beyond the scope of this thesis. Because of this, most experiments in this thesis are for access traffic only, although the scheduling schemes proposed in later chapters can also be applied to internal traffic.

3.6 Summary

In this chapter, the scheduling problem in a WMN is formulated as a standard integer linear programming problem. The complexity of solving the optimization problem can be high. Therefore, heuristic schemes should be designed. Based on the numerical results we find that with optimum scheduling, the performance for supporting real-time traffic in a small size WMN is mainly limited by the available bandwidth of the bottleneck APs. We also find that with optimum scheduling, increasing the number of frequency channels does not increase the system capacity very much. We will propose an AP-based scheduling scheme in Chapter 4 and a connection-based scheduling scheme in Chapter 5.

Chapter 4

AP-based Scheduling

Although the computational complexity of the optimum scheduling solution proposed in Chapter 3 may prevent it from being implemented in practice, results based on the optimum scheduling have shown that the system capacity is limited by the APs that carry the heaviest traffic load. Based on this idea, we propose a heuristic scheme in this chapter, namely AP-based scheduling (ABS), where scheduling decisions at APs with higher traffic loads are determined before those at APs with lower traffic loads. At each AP, scheduling decisions for the real-time packets with more hops to their destinations are made first. Numerical results show that the proposed ABS scheme achieves the capacity performance close to the optimum one while obtaining satisfactory delay performance for real-time CBR traffic.

4.1 Candidate set

For an easy presentation we use a slightly different notation as in Chapter 3 to denote the packet transmission times. Let $\tau_{i,h}$ be the time when connection i is transmitted at hop h along its route. The scheduling scheme is to find $\tau_{i,h}$'s for all connections at all their hops, i.e., $i = 1, 2, \ldots, N$ and $h = 1, 2, \ldots, H_i$. Before making scheduling decisions for the h-th hop of connection i, some timeslots of the transmitting and receiving APs may have been scheduled for serving other connections and thus are not available for connection i. Define a candidate set $\mathcal{D}_{i,h}$ as a set of possible timeslots that $\tau_{i,h}$ can take. Note that each hop involves two APs, one transmitting and the other receiving, and determination of $\tau_{i,h}$ requires the availability of both the APs. Let m and n, respectively, represent the transmitting and receiving APs of the h-th hop of connection i. A timeslot is in $\mathcal{D}_{i,h}$ if both the transmitting and receiving APs have not been scheduled for processing other packets at the time, and the interference condition does not prevent the APs from transmitting and receiving at the time.

In order to find $\mathcal{D}_{i,h}$, we define a set of binary variables $A_{m,t}$'s with $A_{m,t} = 1$ representing that AP m has not been scheduled for serving real-time traffic at time t and $A_{m,t} = 0$ otherwise. These variables are updated during the process of making scheduling decisions. If $t \in \mathcal{D}_{i,h}$, then $A_{m,t} = 1$ and $A_{n,t} = 1$, i.e., both the transmitting and receiving APs are available at the time. Co-channel interference is another factor that affects $\mathcal{D}_{i,h}$. Let f represent the frequency channel assigned to

AP m. If $t \in \mathcal{D}_{i,h}$, then except for AP m, no other AP that is assigned frequency channel f and is within the interference range of AP n has been scheduled to transmit at time t, and at the same time, AP m is not in the interference range of any AP that has been scheduled to receive at frequency channel f at time t.

$\mathcal{D}_{i,h}$ can be found for each hop of connection i right before making the scheduling decision for $\tau_{i,h}$. The task of the scheduling is to find $\tau_{i,h}$'s to satisfy the latency requirement of all connections, while keeping the real-time portions as short as possible for all APs. The real-time portion in one typical SI of AP m is defined as

$$T_{rt,m} = t_{m,\max} - t_{m,\min} + 1, \tag{4.1}$$

where $t_{m,\max} = \text{argmax}_t\{A_{m,t} = 0\}$, $t_{m,\min} = \text{argmin}_t\{A_{m,t} = 0\}$, and t is in the considered typical SI of AP m. The value of $T_{rt,m}$ is updated every time a new timeslot of AP m is allocated for serving the real-time traffic.

4.2 Outline of ABS

In this section we describe the outline of the ABS scheme for supporting real-time CBR traffic. The scheme is performed when a new connection request arrives. ABS first assumes that the new connection is already in the network and makes scheduling decisions. At the end of the scheduling process, if a feasible solution is found, i.e., all connections can be supported with the required delay performance, the new connection is accepted and all connections are transmitted based on the new schedule; otherwise, the new connection is blocked, and existing connections keep the current transmission schedule. In this sense, ABS is not only a scheduling scheme but also can be used for admission control.

Comment: The admission control and scheduling can also be performed in a slightly different way. When a new connection arrives, the scheduler first tries to find a timeslot for each hop along its route using the remaining resources after serving all existing connections. If a feasible schedule is found for the new connection, the new connection is accepted. Otherwise, the complete process of ABS as described above is performed. This method makes no change to the service schedules of existing connections, but the drawback is that it may not allocate the AP resources efficiently.

The basic idea of the proposed ABS scheme is that, scheduling decisions of transmissions/receptions at APs with higher traffic loads are made earlier, since their timeline allocations are less flexible and packets are more likely to experience longer delay in these APs. Time allocations of the APs with lower traffic loads can be done one by one while avoiding any conflicts with the time allocations of scheduled APs. This results in that, for a given packet, the order of making scheduling decisions for each of its hops is not necessarily the same as the order in which the packet is forwarded along these hops. Therefore, a sufficient amount of resources, i.e., a certain number of timeslots, should be reserved for any intermediate hops before both their upstream and downstream hops are scheduled. Within each AP, scheduling decisions for packets with more hops to their destinations are made first.

We define the traffic load of AP m as the total number of real-time CBR packets that AP m should receive and transmit in one scheduling interval, i.e.,

$$L_m = |\mathcal{C}_{Rm}| + |\mathcal{C}_{Tm}| \tag{4.2}$$

where $\mathcal{C}_{Rm}$ and $\mathcal{C}_{Tm}$ are sets of real-time CBR packets that AP m should receive and transmit, respectively, in one scheduling interval.

Pseudocode 4.1 gives the outline of ABS, where $\mathcal{M}$ is a set of the APs whose timelines need to be scheduled, and $\mathcal{N}_m$ is a set of the connections to be processed by AP m. Initially, $\mathcal{M}$ is set to all the APs in the network, and $\mathcal{N}_m$ is set to all the connections traversing AP m. $\mathcal{M}$ will be updated every time the timeline of a certain AP is determined, and $\mathcal{N}_m$ will be updated every time the transmission decision for a connection at AP m is completed. q_{i,h_1,h_2} is the number of timeslots reserved for packet transmissions between hops h_1 and h_2. It is initialized to zero and will be updated in Line 8 (or Pseudocode 4.2 in Section 4.3) when necessary. ENDM is a binary variable, which is set to zero (Line 3) when the scheme starts scheduling for a new AP. To schedule individual packets at a given AP, the scheme first finds $\mathcal{D}_{i,h}$ before calling Pseudocode 4.2. If Pseudocode 4.2 finds that the current values of

q_{i,h_1,h_2} need to be increased, ENDM is set to one. When this occurs (in Lines 12-14), $\mathcal{M}$ and $\mathcal{N}_m$ are re-initialized, and the scheduling starts over again. This will be detailed in Section 4.3. ENDALL is also a binary variable, which is initialized to zero and set to one if Pseudocode 4.2 finds that no feasible schedule is available. After all connections have been scheduled, if $\tau_{i,H_i} - \tau_{i,1} + 1 > d_i$, i.e., the packet transmission delay of connection i exceeds its delay tolerance, then ENDALL is set to 1 and ABS stops with no feasible solution.

Pseudocode 4.1: Outline of ABS

1: Let $\mathcal{M} = \{1, 2, \ldots, M\}$, $\mathcal{N}_m = \{i | m \in \mathcal{R}_i\}$ for $m = 1, 2, \ldots, M$, $q_{i,h_1,h_2} = 0$ for $i = 1, 2, \ldots, N$ and $h_1, h_2 = 1, 2, ..., H_i$, and ENDALL=0

2: **while** $\mathcal{M}$ is not empty and ENDALL=0 **do**

3: Let ENDM $= 0$

4: Find AP m, which is the AP with the highest traffic load in set $\mathcal{M}$

5: **while** ENDM=0 and $\mathcal{N}_m$ is not empty **do**

6: Find connection i, which is the one with the largest number of hops in set $\mathcal{N}_m$

7: Find h, which is the index number of the hop with AP m as the transmitter along the route of connection i

8: Find $\mathcal{D}_{i,h}$ and then call Pseudocode 4.2 for finding $\tau_{i,h}$

9: Let $\mathcal{N}_m = \mathcal{N}_m - \{i\}$

end while

Let $\mathcal{M} = \mathcal{M} - \{m\}$

if ENDM=1 and ENDALL=0 **then**

Let $\mathcal{M} = \{1, 2, \ldots, M\}$ and $\mathcal{N}_m = \{i | m \in \mathcal{R}_i\}$

end if

end while

if $\tau_{i,H_i} - \tau_{i,1} + 1 > d_i$ for any connection i **then**

Let ENDALL=1

end if

4.3 Timeline coordination

We then describe how to find $\tau_{i,h}$ for a typical hop h of connection i with AP m and AP n as the transmitter and receiver, respectively.

There are four different cases to be considered, depending on which hops have been scheduled along the route of the connection at the time of making scheduling decision for $\tau_{i,h}$. Case 1: no hop has been scheduled for connection i; Case 2: at least one upstream hop but no downstream hop has been scheduled for the connection; Case 3: at least one downstream hop but no upstream hop has been scheduled for the connection; and Case 4: at least one upstream hop and one downstream hop have been scheduled for the connection. Pseudocode 4.2 gives details of the scheduling, where Lines 4-11 are for Case 1, Lines 13-23 for Case 2, Lines 25-35 for Case 3, and Lines 37-46 for Case 4. When a feasible solution is found, Line 49 updates the real-time portion of the transmitting and receiving APs; otherwise the remaining part of the code either updates q_{i,h_1,h_2} or sets binary variables for stopping the scheduling process. Below we describe details of the code.

The Pseudocode specifies an upper limit, $T_{\max}$, and a lower limit, $T_{\min}$, in Line 2 so that all values of t in Lines 4, 14, 26 and 38 (corresponding to Cases 1, 2, 3, and 4, respectively) can ensure $\tilde{T}_{rt,m} \leq T_{SI}$, where $\tilde{T}_{rt,m}$ is the real-time portion of AP m, assuming $\tau_{i,h} = t$. The real-time portion of AP n is guaranteed in Lines 6, 17, 29 and 40, respectively, for Cases 1, 2, 3, and 4.

In Case 1 (Lines 4-11), $\tau_{i,h}$ can take any timeslot in $\mathcal{D}_{i,h}$ as long as both $T_{rt,m}$ and $T_{rt,n}$ are not larger than T_{SI}. In Lines 7-9, the Pseudocode selects the timeslot which minimizes the real-time portion of AP m among all feasible timeslots. In Case 2 (Lines 13-23), hop h' is the closest upstream hop of hop h among all hops that have been scheduled. The time to schedule packet i at hop h should be later than that at hop h' by at least $(h - h')$ timeslots, in order to reserve at least one timeslot for each intermediate hop between hops h' and h. That is, $\tau_{i,h}$ should be larger than $\tau_{i,h'}$ by at least $(h - h')$ timeslots. The value of $q_{i,h',h}$ in Line 15 is an additional number of reserved timeslots for the intermediate hops. $q_{i,h',h}$ is set to zero initially (in Pseudocode 4.1), but a higher value may be required, since the packet may have to be buffered in intermediate APs. When $q_{i,h',h} = 0$, one timeslot is reserved for each intermediate hop between hop h and h'. That is, the service time of each intermediate hop between the two hops has been determined after $\tau_{i,h}$ is determined. However, that time may not be in the candidate set of the hop. In this case, the value of $q_{i,h',h}$ will be increased (in Case 4 of Pseudocode 4.2). The scheduling in Case 3 (Lines 25-35) is similar to that in Case 2 except that $\tau_{i,h}$ should be less than $\tau_{i,h''} - (h'' - h) - q_{i,h,h''}$, where hop h'' is the hop closest to hop h among all the scheduled downstream hops. In case 4 (Lines 37-46), besides the same considerations as in Case 1, a timeslot is feasible only if $t > \tau_{i,h'} + (h - h') + q_{i,h',h}$ and $t < \tau_{i,h''} - (h'' - h) - q_{i,h,h''}$.

Finally, if a feasible timeslot for $\tau_{i,h}$ is found, i.e., $\tau_{i,h} < \infty$, the real-time portions of APs m and n are updated in Line 49. Otherwise, the scheme treats Cases 1-3 and Case 4 differently. For Cases 1-3, $\tau_{i,h} = \infty$ means that no feasible schedule exists, and ENDALL is set to 1 (Line 52). When this value is returned back to Pseudocode 4.1, the scheduling scheme stops running. Note that ENDM is also set to 1 in Line 52 so as to stop the inner **while**-loop in Pseudocode 4.1, i.e., Pseudocode 4.2 is not called repeatedly. If $\tau_{i,h} = \infty$ in Case 4, then either there is no feasible solution or q_{i,h_1,h_2}'s are not sufficiently large for finding a feasible solution. In this case, q_{i,h_1,h_2} is updated as $q_{i,h_1,h_2} = q_{i,h_1,h_2} + 1$ for all $h_1 \leq h'$ and $h_2 \geq h''$. These updates will probably affect some or all other scheduled packets. Therefore, when these values are returned to Pseudocode 4.1, both $\mathcal{M}$ and $\mathcal{N}_m$ are re-initialized (Line 13 in Pseudocode 4.1) and the scheme starts over again. During this process, if $q_{i,h_1,h_2} > d_i$ for any hop between hops $h^{'}$ and $h^{''}$, the scheduling process should stop, because the multihop transmission delay has exceeded the delay budget. This is done by setting both ENDALL and ENDM to 1 in Pseudocode 4.2 (Line 56), and Pseudocode 4.1 stops running when these values are returned.

Pseudocode 4.2: Find $\tau_{i,h}$

1: $\tau_{i,h} = \infty$ and $T_{temp} = \infty$ % T_{temp} is a temporary variable

2: $T_{\min} = t_{m,\max} - T_{SI}$ and $T_{\max} = t_{m,\min} + T_{SI}$

3: **if** Case 1 **then**

4:     **for** $t = T_{\min}$: 1: $T_{\max}$ **do**

5:         Find $\tilde{T}_{rt,m}$ and $\tilde{T}_{rt,n}$

6:         **if** $\tilde{T}_{rt,n} \leq T_{SI}$ and $t \in \mathcal{D}_{i,h}$ **then**

7:             **if** $\tilde{T}_{rt,m} < T_{temp}$ **then**

8:                 Let $T_{temp} = \tilde{T}_{rt,m}$ and $\tau_{i,h} = t$

9:             **end if**

10:         **end if**

11:     **end for**

12: **else if** Case 2 **then**

13:     Find hop $h^{'}$, the hop closest to hop h among all the scheduled hops in the upstream of hop h

for $t = T_{\min}$: 1: $T_{\max}$ **do**
  if $t > \tau_{i,h'} + (h - h') + q_{i,h',h}$ **then**
    Find $\tilde{T}_{rt,m}$ and $\tilde{T}_{rt,n}$
    if $\tilde{T}_{rt,n} \leq T_{SI}$ and $t \in \mathcal{D}_{i,h}$ **then**
      if $\tilde{T}_{rt,m} < T_{temp}$ **then**
        Let $T_{temp} = \tilde{T}_{rt,m}$ and $\tau_{i,h} = t$
      end if
    end if
  end if
end for
else if Case 3 **then**
  Find hop h'', the hop closest to hop h among all the scheduled hops in the downstream of hop h
  for $t = T_{\min}$: 1: $T_{\max}$ **do**
    if $t < \tau_{i,h''} - (h'' - h) - q_{i,h,h''}$ **then**
      Find $\tilde{T}_{rt,m}$ and $\tilde{T}_{rt,n}$
      if $\tilde{T}_{rt,n} \leq T_{SI}$ and $t \in \mathcal{D}_{i,h}$ **then**
        if $\tilde{T}_{rt,m} < T_{temp}$ **then**
          Let $T_{temp} = \tilde{T}_{rt,m}$ and $\tau_{i,h} = t$
        end if

end if

end if

end for

else

Find hops h' and h'', the ones closest to hop h among all the scheduled hops in the upstream and downstream of hop h, respectively

for $t = T_{\min}: 1: T_{\max}$ **do**

if $\tau_{i,h'} + (h - h') + q_{i,h',h} < t < \tau_{i,h''} - (h'' - h) - q_{i,h,h''}$ **then**

if $\tilde{T}_{rt,n} \leq T_{SI}$ and $t \in \mathcal{D}_{i,h}$ **then**

if $\tilde{T}_{rt,m} < T_{temp}$ **then**

Let $T_{temp} = \tilde{T}_{rt,m}$ and $\tau_{i,h} = t$

end if

end if

end if

end for

end if

if $\tau_{i,h} \neq \infty$ **then**

Let $T_{rt,m} = \tilde{T}_{rt,m}$ and $T_{rt,n} = \tilde{T}_{rt,n}$

else

if Cases 1, 2, or 3 **then**

    Let ENDALL=1 and ENDM=1
  else
    Let $q_{i,h_1,h_2} = q_{i,h_1,h_2} + 1$ for all $h_1 \leq h'$ and $h_2 \geq h''$
    if any $q_{i,h_1,h_2} > d_i$ **then**
      Let ENDALL=1 and ENDM=1
    else
      Let ENDALL=0 and ENDM=1
    end if
  end if
end if

Now we discuss the complexity of the ABS scheme. When scheduling a particular connection at an AP, ABS checks at most all timeslots in one SI before it can find a feasible timeslot or conclude that no feasible timeslot exists. Thus the running time of finding a feasible timeslot for one connection at AP m is at most $a_{ABS} \cdot T_{SI}$, where a_{ABS} is the running time of checking feasibility of one timeslot and assumed to be a constant. Then the running time of scheduling all connections at AP m is at most $a_{ABS} \cdot N_m \cdot T_{SI}$. If no packet is buffered in intermediate APs and no-rescheduling is triggered, the running time of scheduling all APs is at most $a_{ABS} \cdot T_{SI} \cdot M \cdot N_m$. However, if the number of reserved timeslots is not sufficient, the scheduling process needs to be run again after q_{i,h_1,h_2}'s are increased. In order to minimize the reserved

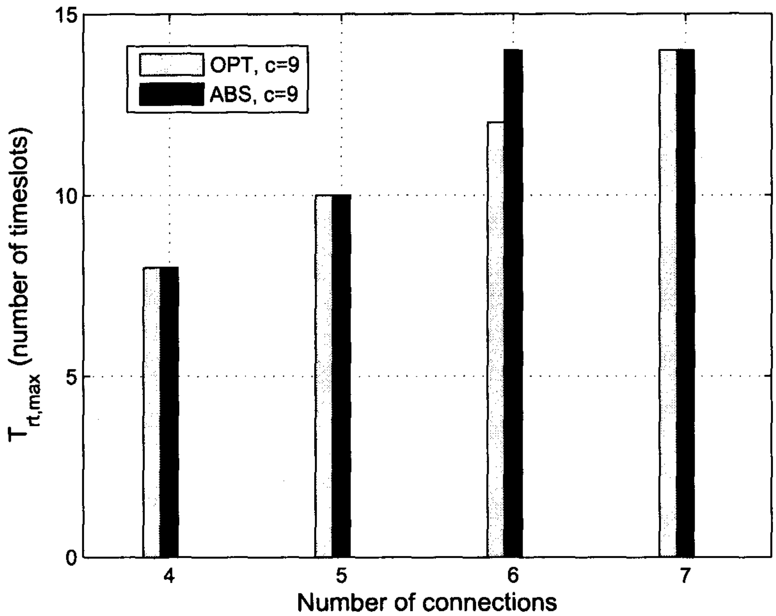

Figure 4.1: Maximum real-time portion: ABS and OPT (c=9)

timeslots, we conservatively increase q_{i,h_1,h_2} by one timeslot each time until a feasible timeslot is found or a conclusion is reached that no feasible schedule exists. Therefore, the overall running time of the scheduling depends on the final values of q_{i,h_1,h_2}'s, which, however, are very difficult to find as they depend on the network topology, routing, traffic distribution, and delay tolerance within the network. The complexity of ABS will be tested based on actual CPU time in computer simulations.

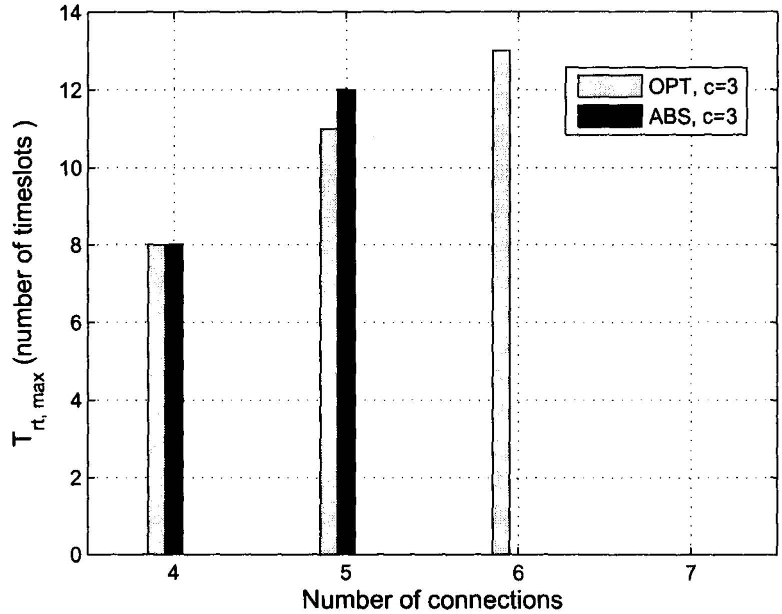

Figure 4.2: Maximum real-time portion: ABS and OPT (c=3)

4.4 Numerical results

In this section we present numerical results of the ABS scheme in comparison with the optimum scheduling (denoted as OPT) proposed in Chapter 3. We first consider the same WMN topology as in Fig. 3.1, adopt the same parameter settings as described in Section 3.5.1, and do snapshot experiments similar to those in Section 3.5.2. We then simulate Poisson traffic arrivals and collect connection level performance using ABS.

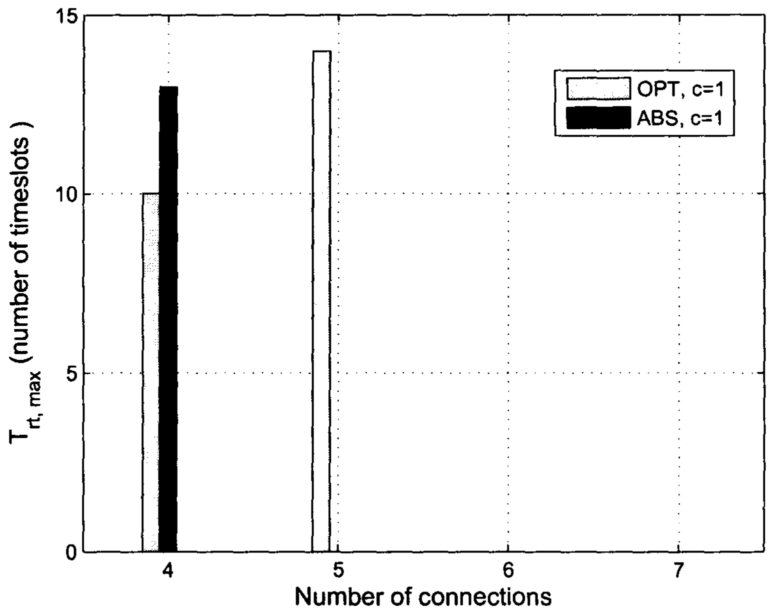

Figure 4.3: Maximum real-time portion: ABS and OPT (c=1)

4.4.1 Snapshot experiments: $T_{rt,\max}$ and $w_{\max}$

We first fix the number of hops of each connection to 3 and vary the number of connections as described in Section 3.5.2. Figs. 4.1, 4.2 and 4.3 compare the $T_{rt,\max}$ using ABS and OPT in the WMN where 9, 3 and 1 frequency channels are available, respectively. We observe that using ABS results in the same $T_{rt,\max}$ as using OPT at some instances. This indicates that ABS is able to efficiently utilize wireless resources as OPT does at these instances. We also observe that ABS does result in slightly larger $T_{rt,\max}$ than OPT in other cases, because ABS schedules one AP at a time and does not search all possible schedules globally as OPT does. Overall, these results indicate that the basic idea of giving a higher priority to APs with a higher traffic load in ABS works well in efficiently allocating the radio resources.

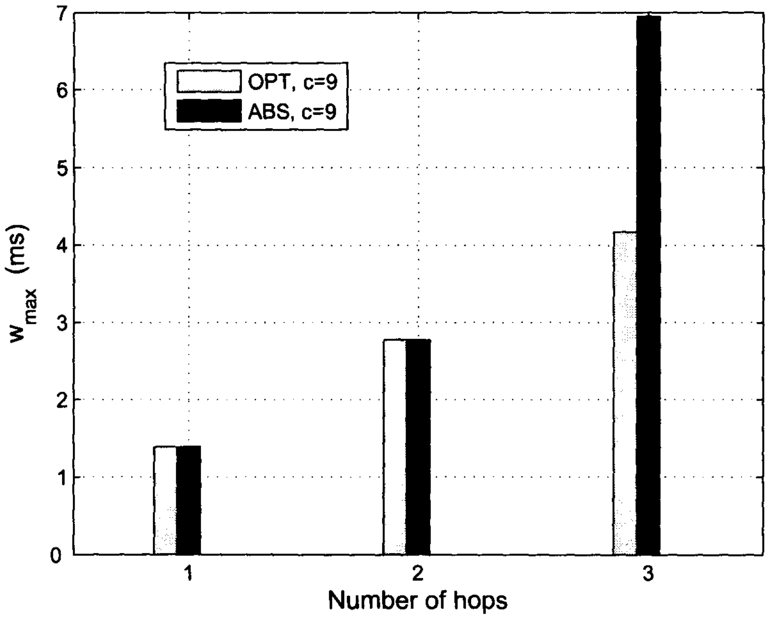

Figure 4.4: Maximum packet transmission delay: ABS and OPT (c=9)

Then we fix the number of connections to 5 and vary the number of hops of each connection as described in Section 3.5.3. Figs. 4.4, 4.5 and 4.6 compare w_{max} using ABS and OPT in the WMN where 9, 3 and 1 frequency channels are available, respectively. It is shown that the maximum delay given by ABS is lower than the delay budget of 60 ms, although ABS may result in longer delay than OPT at some instances. These results indicate that ABS can satisfy the delay requirement of the real-time traffic.

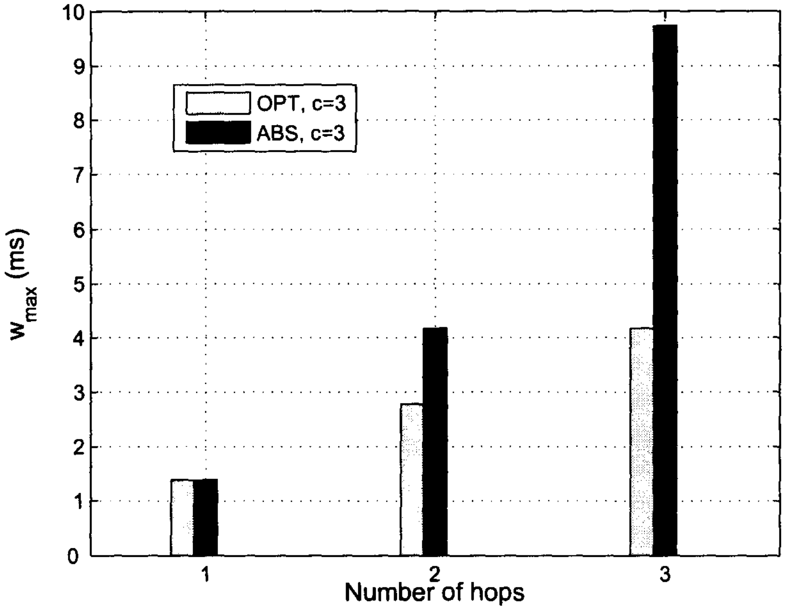

Figure 4.5: Maximum packet transmission delay: ABS and OPT (c=3)

4.4.2 Poisson traffic

We then change the simulation scenario to Poisson arrival connections with exponential connection duration as described in Section 3.5.4.

Figs. 4.7, 4.8 and 4.9 compare the connection blocking rate of ABS and OPT in the 3×3 WMN where 9, 3, and 1 frequency channels are available, respectively. The blocking rate of ABS is very close to that of OPT, and this implies that the capacity using ABS is close to that using OPT. For instance, for a given blocking rate of 1%, the capacity given by ABS and OPT is 0.42 and 0.49 Erlangs (c=9), respectively. That is, ABS can achieve around 85% of the optimal capacity in this case. It is also

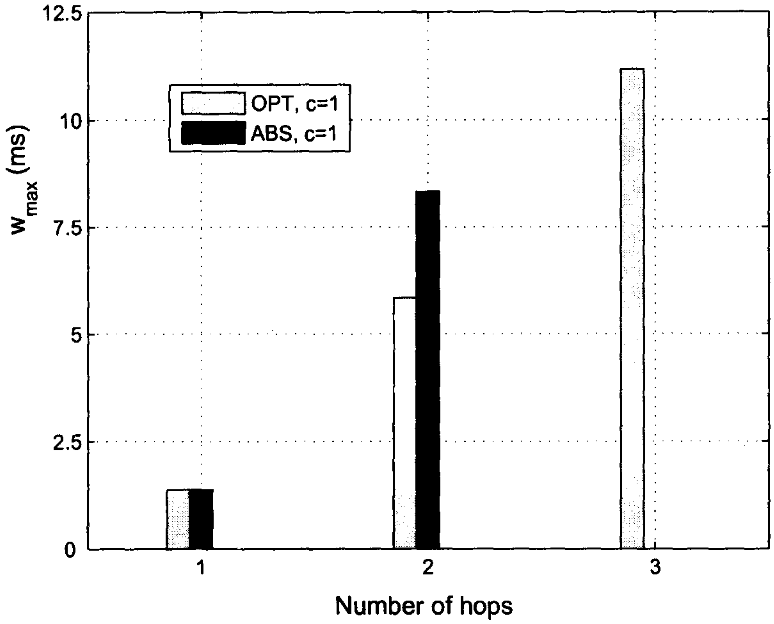

Figure 4.6: Maximum packet transmission delay: ABS and OPT (c=1)

shown that with fewer frequency channels, the performance gap between ABS and OPT increases slightly. For example, when $c = 1$, the capacity given by ABS and OPT, respectively, is 0.33 and 0.42 Erlangs, meaning that ABS can achieve around 80% of the optimal capacity in this case. With stronger co-channel interference, the scheduler should search more timeslots in order to find a feasible solution. However, ABS does not search all possible schedules, therefore results in increased performance gap from the optimum scheduling.

We further simulate the connection blocking performance of both ABS and OPT in the $1 + 2 \times 4$ and $1 + 4 \times 2$ topologies shown in Fig. 4.10. The transmission range and interference range of each AP are 150m and 250m, respectively. The home AP

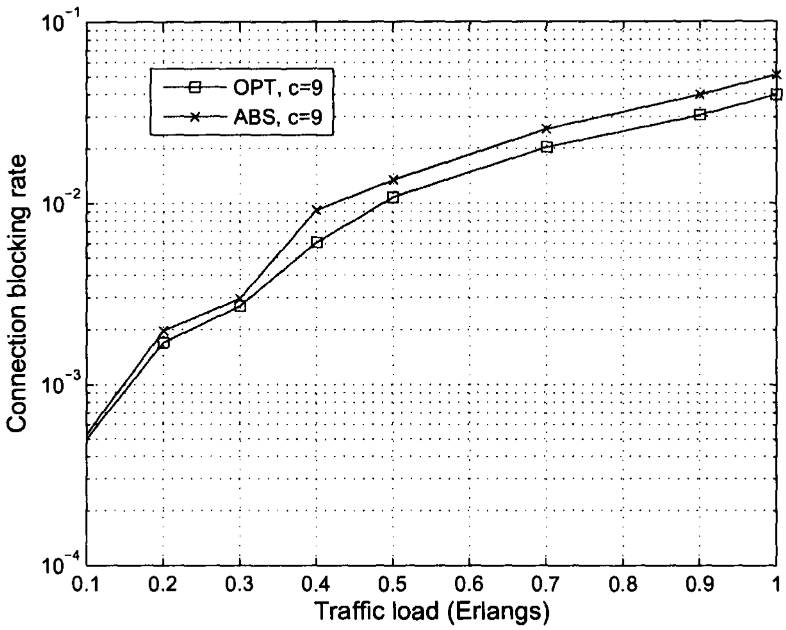

Figure 4.7: Connection blocking rate: ABS and OPT (c=9)

with an active MS is randomly selected from all the APs in the network with equal probability. The difference in the topologies is the number of mesh APs that the root AP directly communicates with and the maximum number of hops of mesh APs to the root AP. Fig. 4.11 shows that the connection blocking rate of ABS is close to that of OPT in both topologies. It is also seen from the figure that the blocking rate in the $1 + 4 \times 2$ topology is higher than that in the $1 + 2 \times 4$ one. Comparing the two topologies in Fig. 4.10 we find that i) in the $1 + 2 \times 4$ topology, connections may experience shorter delay as on average each connection traverses a fewer number of hops; and ii) in the $1 + 2 \times 4$ topology, traffic can be distributed in four different "branches" before (for the uplink) or after (for downlink) the root AP, but only two

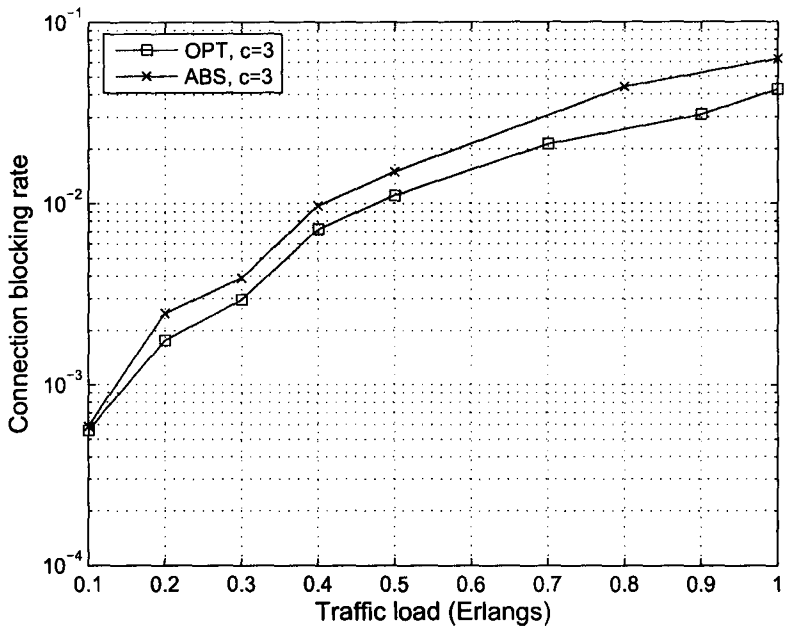

Figure 4.8: Connection blocking rate: ABS and OPT (c=3)

"branches" in the $1 + 4 \times 2$ topology.

4.4.3 Running time

Now we compare the running time of ABS and OPT based on snapshot experiments. We first consider the 3×3 topology with 9 frequency channels and the system setting as described at Section 3.5.1. We set the channel transmission rate to 2Mbps and vary the number of connections. Both ABS and OPT are run at an IBM server with dual 2.4 GHz Pentium 4 Xeon processors with 1 GB of memory. Fig. 4.12 shows that the running time of ABS is several magnitudes lower than that of OPT, and the difference between the two curves increases with traffic load. The running time of OPT increases

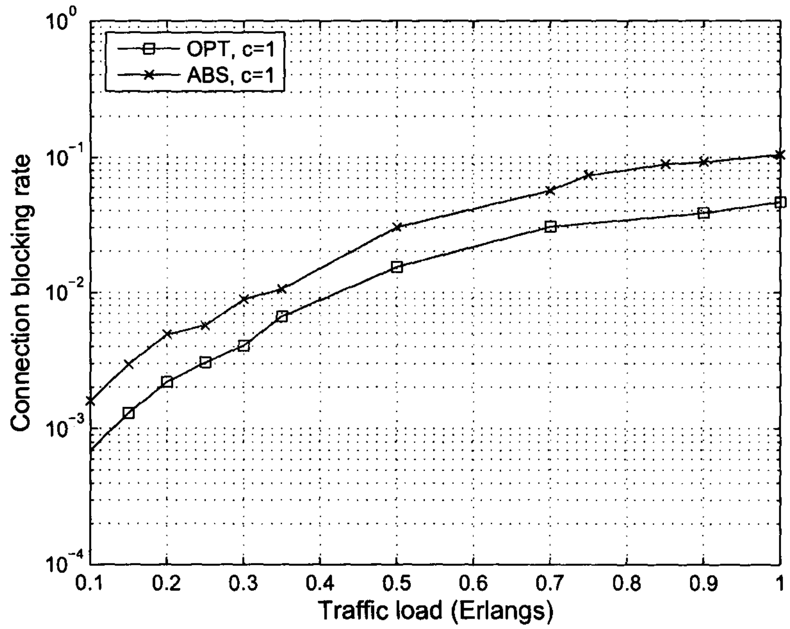

Figure 4.9: Connection blocking rate: ABS and OPT (c=1)

exponentially with the number of connections as explained in Section 3.4, while the running time of ABS increases much slower with the number of connections. This indicates that ABS is a more practical method than OPT for scheduling in a WMN with moderate or high traffic load.

We then run ABS in the $1 + 2 \times 4$, 3×3 and $1 + 4 \times 2$ topologies and set the channel transmission rate to 11Mbps. Fig. 4.13 shows the running time of ABS in the three topologies. The running time of ABS in $1 + 4 \times 2$ topology is the largest in the three topologies, because the connections in the $1 + 4 \times 2$ topology generally have more hops to the root AP than the connections in the other two topologies. It is also shown from the figure that the running time of ABS can be very large when the

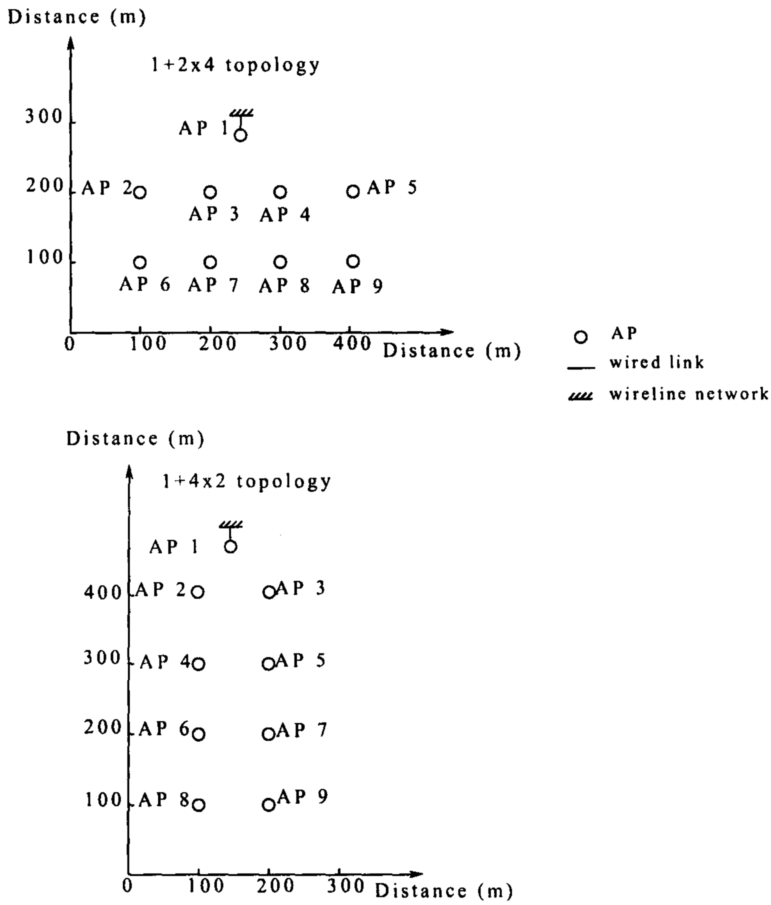

Figure 4.10: WMNs with $1 + 2 \times 4$ and $1 + 4 \times 2$ topologies

traffic load is relatively high and connections have a long route (e.g., 17 connections in $1 + 4 \times 2$ topology). This is because once the reserved timeslots for any hop of any connection are not sufficient, the ABS scheme has to restart and may need multiple iterations before finding a feasible schedule. Thus the complexity of ABS can be high for supporting real-time applications in a WMN with a high traffic load, and a scheduling scheme with lower complexity is desired.

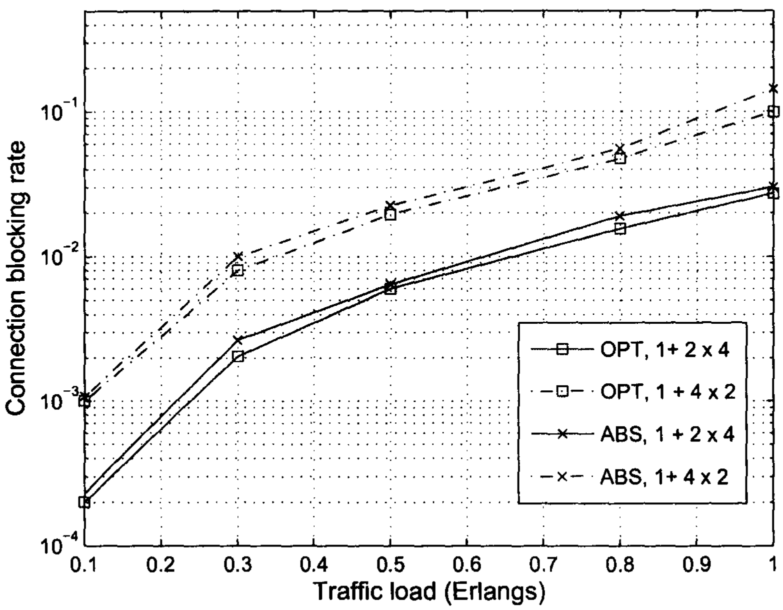

Figure 4.11: Connection blocking rate of ABS and OPT: different topologies ($c = 3$)

4.5 Summary

In this chapter we have proposed an AP-based scheduling (ABS) scheme for supporting real-time CBR traffic in WMNs. In ABS, APs with higher traffic loads are given a higher priority and their scheduling decisions are made before those with lower traffic loads. It is shown that ABS achieves capacity performance close to that of the optimum scheduler but with lower complexity. On the other hand, the complexity of ABS can still be high (even much lower than OPT) and performing ABS can take a long time in a WMN supporting a relatively large number of connections with long routes. In the next chapter we will propose a

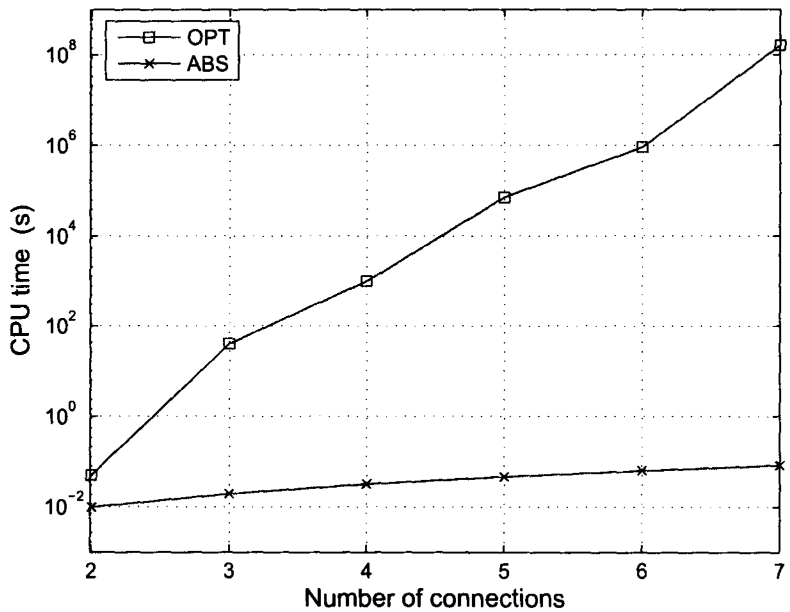

Figure 4.12: Running time of ABS and OPT (C=2Mbps, $c = 3$)

connection-based scheduling scheme to further reduce the complexity of the scheduling.

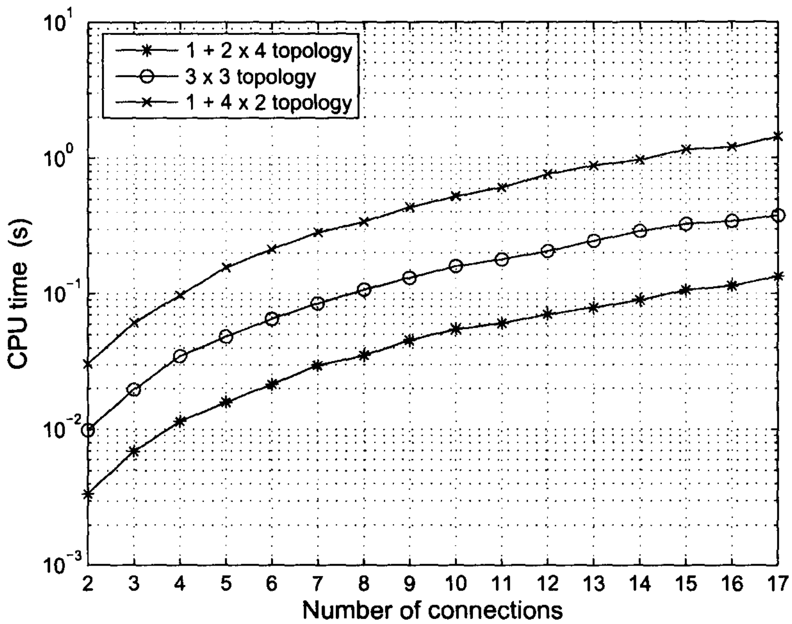

Figure 4.13: Running time of ABS: different topologies (C=11Mbps, $c = 3$)

Chapter 5

Connection-based Scheduling

The ABS scheme proposed in Chapter 4 achieves capacity performance very close to that of the global optimal scheduling. However, its complexity can also be high. In this chapter we propose a connection-based scheduling (CBS) scheme for supporting real-time traffic in a WMN. The intention of the CBS scheme is to reduce the complexity of the scheduling while supporting as many connections in the WMN as possible. In the CBS scheme, connections with a larger number of hops are given a higher priority and their scheduling decisions are made first. Connections traversing a fewer number of hops use resources remaining from serving connections with a higher priority. The scheduling decisions for higher priority connections are made independent of those for lower priority ones. This simplifies the scheduling process, as compared with ABS. For each packet, CBS first determines the transmission time at the AP with the highest traffic load along its route. In order to evaluate the performance of the CBS, a connection-based optimum scheduling is then developed, where scheduling decisions for multiple hops of each connection are jointly optimized with an objective to minimize the total amount of required radio resources, subject to the delay requirement of each connection.

5.1 Connection-based scheduling

We consider the same system as described in Section 3.1. As in Chapter 4, the CBS scheme is performed when a new connection request arrives. When making scheduling decisions, the scheduler assumes that the new connection is already in the network. At the end of the process, if a feasible solution can be found, the new connection is accepted and all connections are transmitted based on the new schedule; otherwise, the new connection is blocked and existing connections keep the current transmission schedule.

In general, connections with a larger number of hops are more likely to experience a longer delay. Thus their scheduling decisions are made earlier, and the scheduling decisions of connections with a fewer number of hops are made using the remaining resources. For a given packet/connection i, let h^* be the hop whose transmitter is the AP with the highest traffic load along the route of the connection, where traffic load of an AP is defined the same as in Section 4.2. The value of τ_{i,h^*} is determined first, and transmitting times at other hops along the route of the packet are then determined one by one from hop $h^* - 1$ to the first hop, and then from hop $h^* + 1$ to hop H_i.

An outline of the CBS scheduling is given in Pseudocode 5.1, where Lines 3-14 give steps for scheduling all hops of one connection. Line 5 is to call Pseudocode 5.2 for finding the transmission time at the h^*th hop, Lines 6-10 are for finding the transmission time for all the upstream hops, and Lines 11-15 are for finding the transmission time for all the downstream hops. ENDALL is a binary variable initialized to zero. It is set to 1 either because the latency requirement of the packet cannot be satisfied (Lines 17-19), or no sufficient radio resource available (as will be described in Pseudocode 5.2). The scheduling process stops when a feasible solution is found for all packets, or ENDALL is equal to 1. In the latter case, no feasible schedule is found and the new connection request is rejected.

Pseudocode 5.1: Outline of CBS

1: Let $\mathcal{N} = \{1, 2, \ldots, N\}$ and ENDALL=0

2: **while** $\mathcal{N}$ is not empty and ENDALL=0 **do**

3: Find connection i, which is the one with the largest number of hops in set $\mathcal{N}$

4: Find h^*, which is the index number of the hop that has the AP with the highest traffic load as transmitter along the route of connection i

5: Find $\mathcal{D}_{i,h^*}$ and then find τ_{i,h^*} using Pseudocode 5.2

Let $h = h^*$

while ENDALL=0 and $h > 1$ **do**

  Let $h = h - 1$

  Find $\mathcal{D}_{i,h}$ and then find $\tau_{i,h}$ using Pseudocode 5.2

end while

Let $h = h^*$

while ENDALL=0 and $h < H_i$ **do**

  Let $h = h + 1$

  Find $\mathcal{D}_{i,h}$ and then find $\tau_{i,h}$ using Pseudocode 5.2

end while

$\mathcal{N} = \mathcal{N} - \{i\}$

if $\tau_{i,H_i} - \tau_{i,1} + 1 > d_i$ **then**

  Let ENDALL = 1

end if

end while

We then describe how to find $\tau_{i,h}$ for a typical hop h of packet i with APs m and n as the transmitter and receiver, respectively. Similar to ABS, before finding $\tau_{i,h}$ we need to find the candidate set, $\mathcal{D}_{i,h}$. The process to find $\mathcal{D}_{i,h}$ in CBS is exactly the same as described in Section 4.1.

There are three different cases to be considered, depending on whether the immediate upstream or downstream hop has been scheduled for the packet at the time of determining $\tau_{i,h}$. Case 1: no hop has been scheduled for the packet; Case 2: the immediate upstream hop, i.e., the $(h-1)$-st hop, has been scheduled for the packet; and Case 3: the immediate downstream hop, i.e., the $(h+1)$-st hop, has been scheduled for the packet. Based on the outline in Pseudocode 5.1, it is impossible to have both $\tau_{i,h-1}$ and $\tau_{i,h+1}$ determined before determining $\tau_{i,h}$.

Pseudocode 5.2 gives details of the scheduling, where Lines 4-12 are for Case 1, Lines 13-23 are for Case 2, and Lines 24-34 are for Case 3. In Case 1 (Lines 4-12), any timeslot in $\mathcal{D}_{i,h}$ can be $\tau_{i,h}$ as long as both $T_{rt,m}$ and $T_{rt,n}$ can still be kept within one SI after $\tau_{i,h}$ has been determined. In Case 2 (Lines 13-23), besides the same considerations as in Case 1, the Pseudocode selects the timeslot for $\tau_{i,h}$ from all available timeslots so that i) $\tau_{i,h} > \tau_{i,h-1}$ to ensure the processing time in the h-th hop is later than that in the $(h-1)$-st hop, and ii) $\tau_{i,h} - \tau_{i,h-1}$ is minimized in order to minimize the delay between successive hop transmissions. The scheduling in Case 3 (Lines 24-34) is similar to that in Case 2 except that $\tau_{i,h} < \tau_{i,h+1}$ and $\tau_{i,h+1} - \tau_{i,h}$ is minimized. If $\tau_{i,h} = \infty$ after this process, there is no feasible solution to schedule this packet, and ENDALL=1 (Line 39). Otherwise, the real-time portion of AP m and AP n are updated in Line 37.

Pseudocode 5.2: Find $\tau_{i,h}$

1: Let $\tau_{i,h} = \infty$

Let $T_{\min} = t_{m,\max} - T_{SI}$ and $T_{\max} = t_{m,\min} + T_{SI}$
Let $t = T_{\min}$ and ENDM=0
if Case 1 **then**
  while $t \leq T_{\max}$ and ENDM=0 **do**
    Find $\tilde{T}_{rt,n}$
    if $\tilde{T}_{rt,n} \leq T_{SI}$ and $t \in \mathcal{D}_{i,h}$ **then**
      Let $\tau_{i,h} = t$ and ENDM=1
    else
      $t = t + 1$
    end if
  end while
else if Case 2 **then**
  for $t = T_{\min}$: 1: $T_{\max}$ **do**
    Find $\tilde{T}_{rt,n}$
    if $\tilde{T}_{rt,n} \leq T_{SI}$ and $t \in \mathcal{D}_{i,h}$ **then**
      if $t > \tau_{i,h-1}$ **then**
        if $t - \tau_{i,h-1} < \tau_{i,h} - \tau_{i,h-1}$ **then**
          Let $\tau_{i,h} = t$
        end if
      end if

end if

end for

else

for $t = T_{\min}: 1: T_{\max}$ **do**

Find $\tilde{T}_{rt,n}$

if $\tilde{T}_{rt,n} \leq T_{SI}$ and $t \in \mathcal{D}_{i,h}$ **then**

if $t < \tau_{i,h+1}$ **then**

if $\tau_{i,h+1} - t < \tau_{i,h+1} - \tau_{i,h}$ **then**

Let $\tau_{i,h} = t$

end if

end if

end if

end for

end if

if $\tau_{i,h} \neq \infty$ **then**

Update $T_{rt,m}$ and $T_{rt,n}$

else

Let ENDALL=0

end if

Now we analyze the complexity of the CBS scheme. For each hop of connection i, CBS checks at most all timeslots in one scheduling interval to find $\tau_{i,h}$. Thus the total running time for determining $\tau_{i,h}$ is at most $a_{CBS} \cdot T_{SI}$, where a_{CBS} is the running time for checking the feasibility of one timeslot in CBS and assumed to be a constant. For a connection with H_i hops, the total running time of scheduling all hops is at most $a_{CBS} \cdot H_i \cdot T_{SI}$, and the total running time of scheduling all N connections is

$$r_{CBS} \leq a_{CBS} \sum_{i=1}^{N} H_i \cdot T_{SI}. \tag{5.1}$$

Due to $H_i \leq M$, the complexity of CBS is

$$O(N \cdot M \cdot T_{SI}). \tag{5.2}$$

5.2 Connection-based optimum scheduling

In order to examine the performance of the CBS scheme, we formulate a connection-based optimum scheduling (OPT-CB). The basic idea is similar to CBS. That is, connections with more hops are given a higher priority in making scheduling decisions. The connection-based optimum scheduling for connection i is performed after the scheduling decisions for all higher priority connections are made. We use the same notations, $\tau_{i,h}$ and $\mathcal{D}_{i,h}$, respectively, to denote processing time at the hth hop of connection i and the candidate set of timeslots for $\tau_{i,h}$. An outline of the connection-based optimum scheduling is shown in Pseudocode 5.3.

Pseudocode 5.3: Connection-based optimum scheduling

Let $\mathcal{N} = \{1, 2, \ldots, N\}$ and ENDALL $= 0$
while $\mathcal{N}$ is not empty and ENDALL $= 0$ **do**
  Find connection i, which is the connection with the largest number of hops in set $\mathcal{N}$
  Find $\mathcal{D}_{i,h}$'s for $h = 1, 2, \ldots, H_i$
  Find $\tau_{i,h}$'s for all $h = 1, 2, \ldots, H_i$ by solving Problem 5.1
  if no feasible solution **then**
    Let ENDALL $= 1$
  else
    Let $\mathcal{N} = \mathcal{N} - \{i\}$
  end if
end while

Since values of $\tau_{i,h}$'s for all $h = 1, 2, \ldots, H_i$ of connection i are jointly determined in the connection-based optimum solution, we use $m_{i,h}$ to denote the transmitting AP of the h-th hop of connection i in order to distinguish the APs at different hops of the connection. The objective of the optimum scheduling is to minimize the maximal real-time portion of the APs along the route of connection i, since the network capacity is limited by the AP with the highest traffic load. That is, minimize $T_{rt} = \max_{h=1}^{H_i} T_{rt,m_{i,h}}$. The connection-based optimization problem is formulated as

Problem 5.1:

$$\min_{[\tau_{i,h}],\, T_{rt}} T_{rt} \tag{5.3}$$

$$\text{s. t.} \quad \tau_{i,h} < \tau_{i,h+1}, \quad h = 1, ..., H_i \tag{5.4}$$

$$\tau_{i,H_i} - \tau_{i,1} + 1 \leq d_i \tag{5.5}$$

$$\tau_{i,h} - t_{m_{i,h},\min} + 1 \leq T_{rt}, \quad h = 1, ..., H_i \tag{5.6}$$

$$t_{m_{i,h},\max} - \tau_{i,h} + 1 \leq T_{rt}, \quad h = 1, ..., H_i \tag{5.7}$$

$$T_{rt} \leq T_{SI} \tag{5.8}$$

$$\tau_{i,h} \in \mathcal{D}_{i,h}, \quad h = 1, ..., H_i, \tag{5.9}$$

where $[\tau_{i,h}]$ is a vector with $\tau_{i,h}$ $(h = 1, ..., H_i)$ as its entry.

Note that such an optimization problem is solved for each of the N connections. In the above formulation, the constraint in (5.4) is to ensure that an upstream AP transmits the packet before a downstream one does; the constraint in (5.5) is to guarantee the delay performance of the packet; the constraint in (5.9) defines the domain of $\tau_{i,h}$, and the constraint in (5.8) indicates the upper bound of the real-time portion. The left-hand side of constraints in (5.6) and (5.7) defines the real-time portion of each AP along the route of connection i after $\tau_{i,h}$'s are determined, where $t_{m_{i,h},\min}$ and $t_{m_{i,h},\max}$ are defined in the same way as in Section 4.1. When $\tau_{i,h} < t_{m_{i,h},\min}$, (5.7) gives the updated real-time portion of AP $m_{i,h}$ and $\tau_{i,h} - t_{m_{i,h},\min} + 1 \leq 0$ (i.e., (5.6) is true); when $\tau_{i,h} > t_{m_{i,h},\max}$, (5.6) gives the updated real-time portion of the AP and $t_{m_{i,h},\max} - \tau_{i,h} + 1 \leq 0$ (i.e., (5.7) is true); when $t_{m_{i,h},\min} < \tau_{i,h} < t_{m_{i,h},\max}$, the real-time portion of the AP is unchanged and both (5.6) and (5.7) are true.

The objective function and all constraints except (5.9) are linear. For (5.9) all values in $\mathcal{D}_{i,h}$'s are integers, but $\mathcal{D}_{i,h}$ may not include consecutive integers. We divide $\mathcal{D}_{i,h}$ into multiple subsets, each of which includes a sequence of consecutive integers. That is $\mathcal{D}_{i,h} = \bigcup_{v=1}^{V_{i,h}} \mathcal{D}_{i,h,v}$, where $\mathcal{D}_{i,h,v}$ is the v-th consecutive subset of $\mathcal{D}_{i,h}$, and $V_{i,h}$ is the total number of such consecutive integer sets in $\mathcal{D}_{i,h}$. This is done for all $h = 1, 2, \ldots, H_i$. The original optimization problem is then divided into $V_{i,1} \times V_{i,2} \times \ldots \times V_{i,H_i}$ subproblems, each of which is an integer linear programming problem. We can then use the branch-and-bound method to solve all the subproblems. We use $[\tau_{i,1}, \tau_{i,2}, \ldots, \tau_{i,H_i}] = \mathbf{X}^*_{v_1,v_2,\ldots,v_{H_i}}$ to represent the solution to the subproblem when $\mathcal{D}_{i,1}$ is replaced with $\mathcal{D}_{i,1,v_1}$, $\mathcal{D}_{i,2}$ is replaced with $\mathcal{D}_{i,2,v_2}$, ..., $\mathcal{D}_{i,H_i}$ is replaced with $\mathcal{D}_{i,H_i,v_{H_i}}$, and the corresponding T_{rt} is denoted as $T^*_{v_1,v_2,\ldots,v_{H_i}}$. The final solution to Problem 5.1 is the solution to the subproblem that gives the minimum $T^*_{v_1,v_2,\ldots,v_{H_i}}$ among all the subproblems.

5.3 Numerical results

In this section, we compare the CBS scheme with the connection-based optimum scheduling (OPT-CB), ABS, and OPT. We also compare these scheduling solutions with the earliest deadline first (EDF) scheduling scheme [46]. EDF is known to

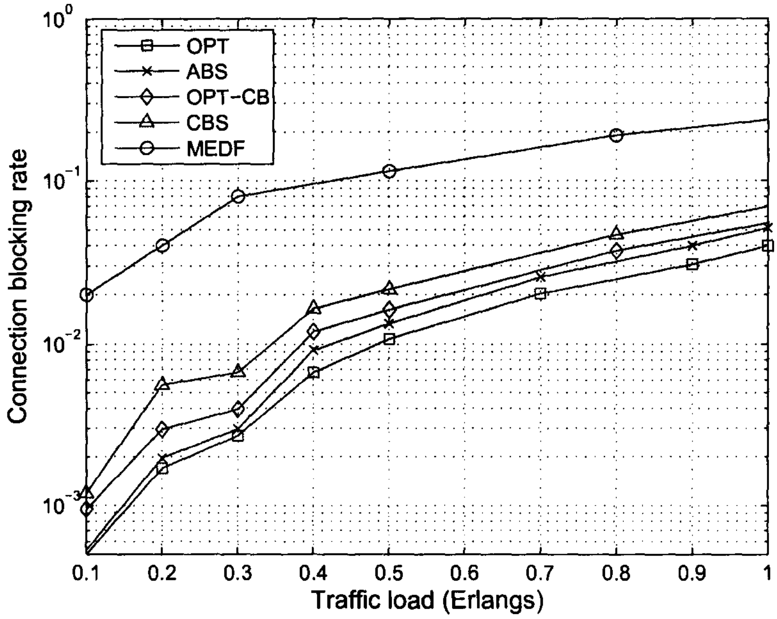

Figure 5.1: Connection blocking rate: different schemes (C =1Mbps, $c = 9$)

provide the optimal delay performance in the deterministic environment. When it is applied in a wireless mesh network, packets with less delay budgets available are given a higher priority in the timeline coordinations. This may happen when multiple packets from different links compete for the same AP, or transmissions of two links interfere with each other. We refer to this version of EDF as a modified earliest deadline first (MEDF) scheme. Real-time CBR connections arrive to the system according to a Poisson process (except for Fig. 5.19 which is based on snapshot experiments). We use the same default system parameters as in Section 3.5.1.

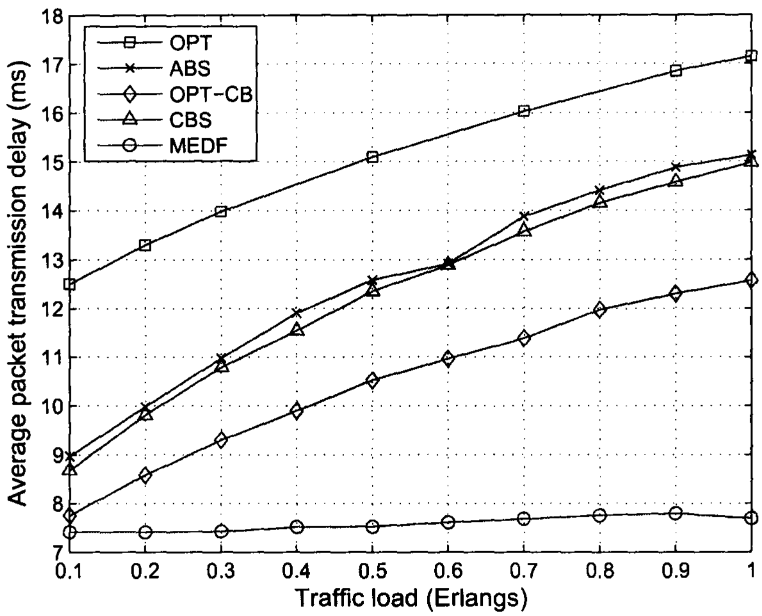

Figure 5.2: Average packet transmission delay: different schemes(C =1Mbps, c = 9)

5.3.1 Performance vs. traffic load

We first consider the same 3×3 topology as shown in Fig. 3.1. Fig. 5.1 compares the connection blocking rates of different schemes. From this figure we observe that OPT achieves the best blocking performance, the blocking performance of ABS is slightly higher, the connection-based schemes (CBS and OPT-CB) have higher blocking rate than ABS, and MEDF has the worst blocking performance among all the schemes. As explained in Chapter 4, ABS first considers how to efficiently utilize the radio resources of individual APs, especially the ones with heavier traffic load, and leaves the performance of individual packet transmissions as the second consideration. Although

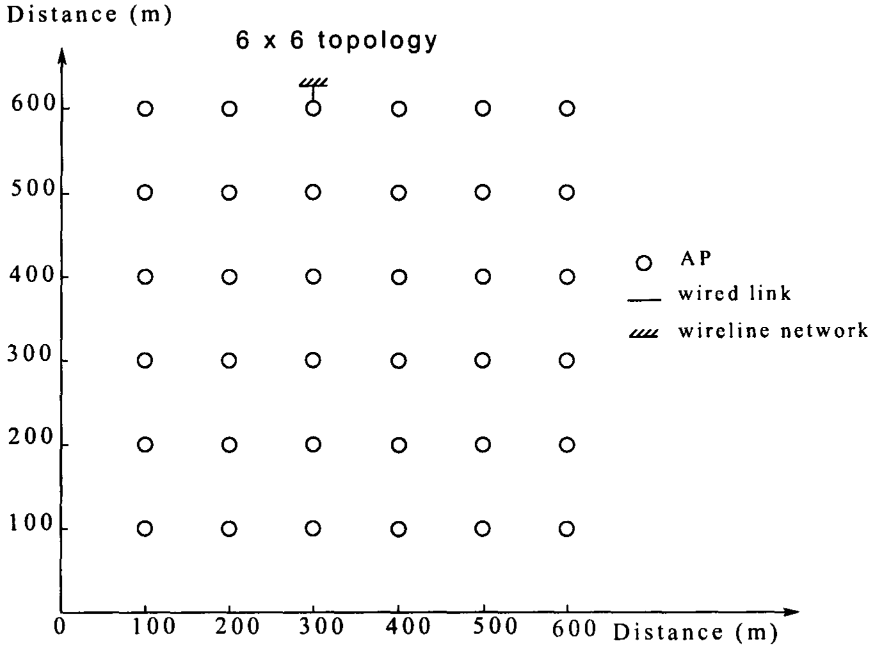

Figure 5.3: A WMN with 6 × 6 topology

this makes ABS more complicated, it does improve the network capacity. In contrast, CBS must complete scheduling all hops of a connection before scheduling the next connection. During this process, it may have to sacrifice the resource utilization of the APs. Therefore, ABS can achieve better blocking rate performance (for a given offered traffic load) or capacity performance (for a given blocking rate requirement) than CBS.

Meanwhile, we can also see from Fig. 5.1 that the connection blocking rate performance using CBS is very close to that using ABS. For a given mesh topology, there are two main aspects that can affect the packet transmission performance: the number of hops of the connections and traffic load distribution among the APs.

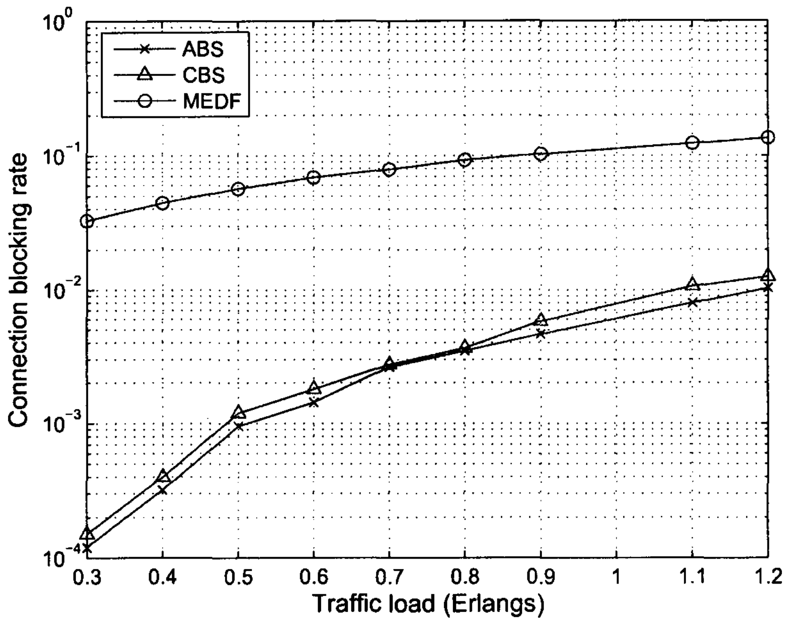

Figure 5.4: Connection blocking rate: 6×6 topology (C =2Mbps, $c = 36$)

First, by giving a higher priority to the connections traversing more hops, the proposed CBS scheme reduces the probability that a connection is rejected because its delay requirement cannot be satisfied. Second, by giving the highest priority to the AP with the highest traffic load along the route of each connection, the proposed CBS scheme reduces the probability of connection rejection because there are not any available timeslots at the APs.

It is also seen from Fig. 5.1 that the proposed CBS scheme achieves much lower connection blocking rate than MEDF, as MEDF does not take into consideration multihop transmissions of the packets and therefore results in poor connection blocking performance. For a given connection blocking rate requirement, using CBS

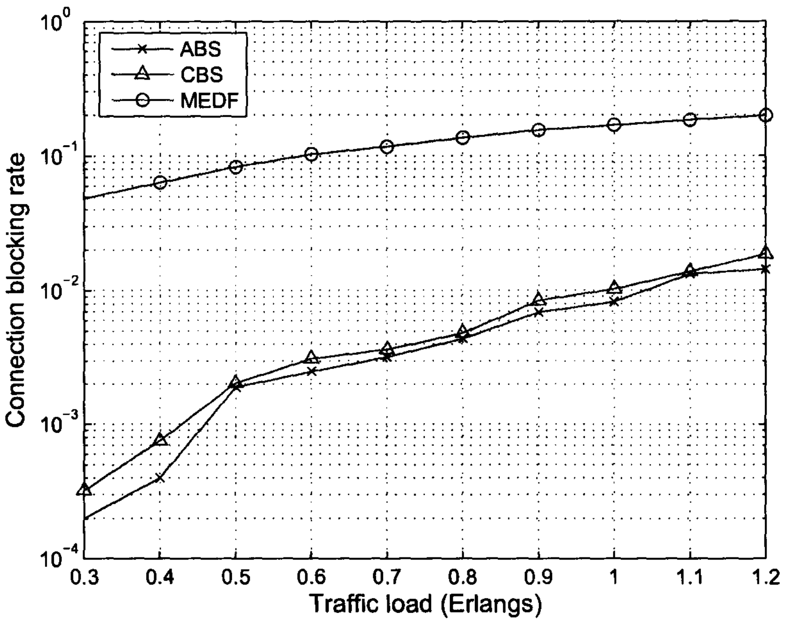

Figure 5.5: Connection blocking rate: 6×6 topology (C =2Mbps, $c = 3$)

can accept more real-time CBR connections than using MEDF.

We then look at the packet level performance. The maximum delay of all packets is guaranteed by admission control. Therefore in Fig. 5.2 we only show the average delay performance of admitted connections. CBS has better delay performance than ABS because CBS minimizes the delay at successive hops. We also observe that OPT has larger delay than all other schemes. This is because OPT sacrifices the delay performance in order to obtain the best capacity performance. Although MEDF provides the best mean delay performance, it is at a price of significantly higher connection blocking rate as shown in Fig. 5.1.

Next we consider a WMN with more APs. Fig. 5.3 shows a WMN with 6×6

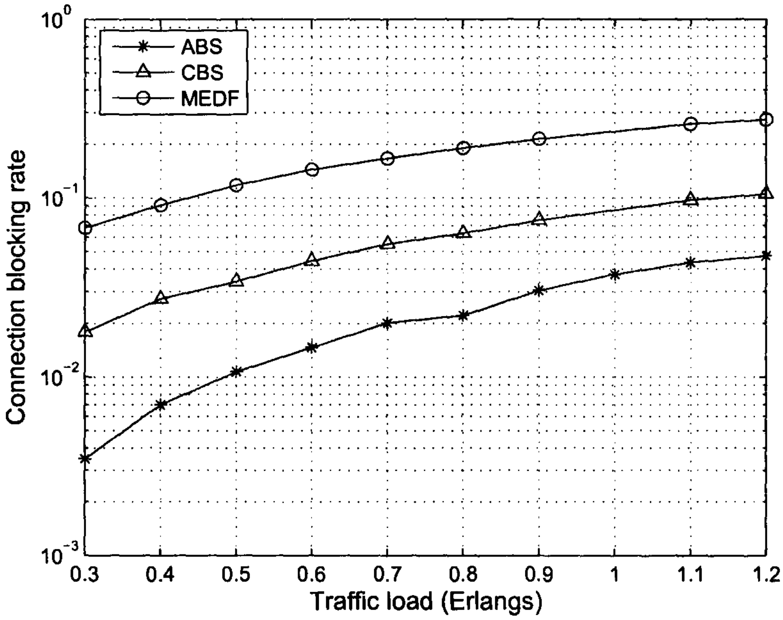

Figure 5.6: Connection blocking rate: 6×6 topology (C =2Mbps, $c = 1$)

grid topology. Figs. 5.4, 5.5 and 5.6 show the connection blocking rate when c=36, 3 and 1, respectively. With more frequency channels available, the connection blocking rate decreases since more timeslots are available for different APs to transmit simultaneously. All three figures show that the connection blocking rate using ABS and CBS is much lower than that using MEDF. This observation is consistent with that based on the 3×3 topology. When co-channel interference is moderate ($c = 3$) or there is no co-channel interference ($c = 36$), ABS achieves a slightly lower connection blocking rate than CBS, or the network capacity using CBS is very close to that using ABS. When there is strong co-channel interference, e.g., all APs share the same frequency channel, the performance gap between CBS

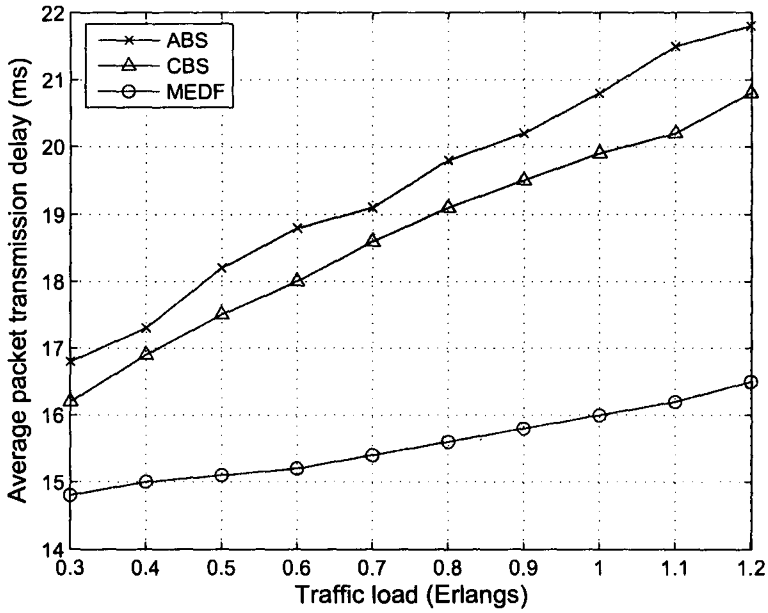

Figure 5.7: Average packet transmission delay: 6×6 topology (C =2Mbps, $c = 3$)

and ABS becomes obvious. The system capacity is mainly limited by the AP with the highest traffic load. With ABS the AP with the highest traffic load is scheduled first, its scheduling is not subject to co-channel interference, and therefore its radio time is utilized with the highest efficiency. When co-channel interference becomes stronger, efficiently utilizing the AP resources becomes more important. In this case, ABS achieves much better capacity performance than CBS.

Fig. 5.7 shows that CBS achieves lower average packet transmission delay than ABS, and MEDF has the lowest average packet transmission delay among the three schemes. This is consistent with the delay performance in the 3×3 topology.

We now look at the performance of WMNs with randomly generated topologies.

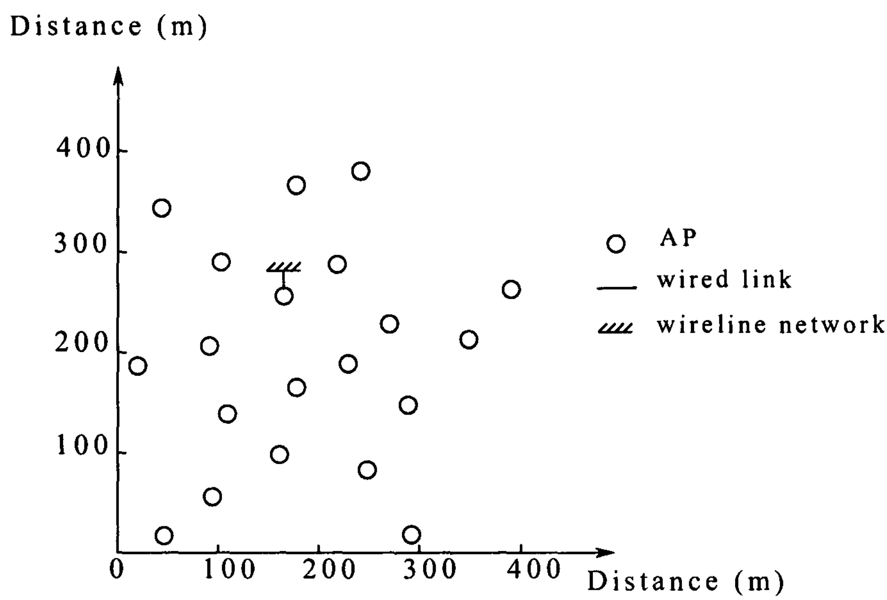

Figure 5.8: A WMN with a random topology

We consider that co-channel interference is not present. This is either due to careful frequency planning, or one frequency channel is exclusively assigned to each AP. We randomly place 20 APs in a 400m $\times$ 400m area. These APs form an irregular topology. An example of such topology is shown in Fig. 5.8. Twenty different such random topologies are generated. In each topology, one AP is randomly selected as the root AP. MSs are uniformly distributed in the network area, and each MS is associated to the AP closest to it. The system performance shown in Figs. 5.9 and 5.10 is averaged over the twenty random topologies. Fig. 5.9 shows that the connection blocking rates of ABS and CBS are very close to each other and a lot lower than that of MEDF. Fig. 5.10 shows that CBS achieves lower average packet

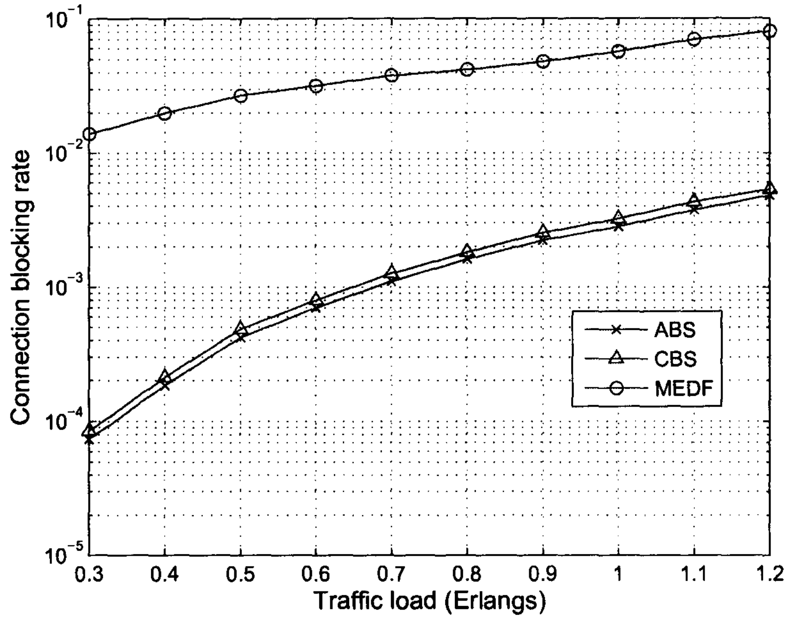

Figure 5.9: Connection blocking rate: random topology (C =2Mbps)

transmission delay, and MEDF achieves lower delay than both ABS and CBS.

In the following, we investigate the capacity of WMNs and packet transmission performance based on both ABS and CBS, and find the effects of physical channel transmission rate, network topology, real-time traffic delay requirement on the system performance.

5.3.2 Performance vs. physical transmission rate

Fig. 5.11 shows that the connection blocking rate in the 6 $\times$ 6 topology is decreased significantly with the increase of the channel transmission rate. For example, when C is doubled from 5.5Mbps to 11Mbps, the connection blocking rate of ABS is

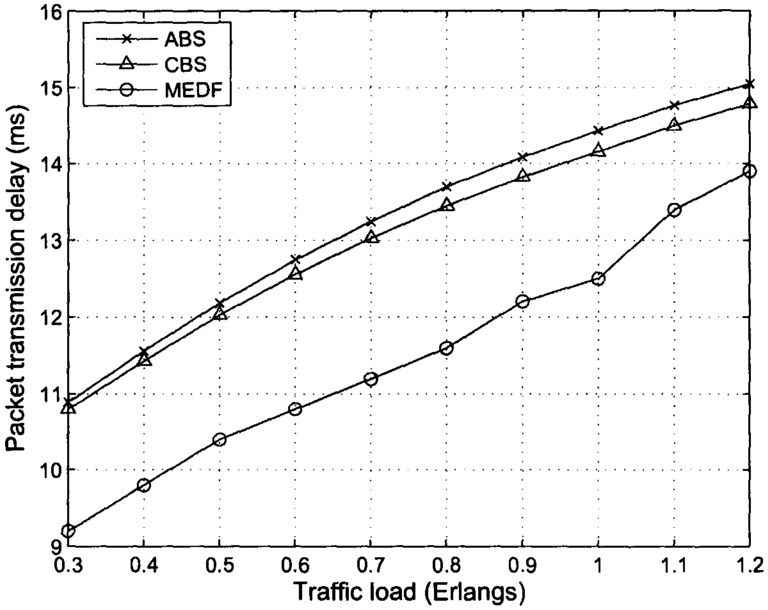

Figure 5.10: Average packet transmission delay: random topology (C =2Mbps)

decreased from 0.2% to 0.01%. Another observation is that with the increase of channel transmission rate, the performance gap between the three heuristic schemes increases. This is especially obvious for the difference between MEDF and the two proposed schemes.

5.3.3 Performance vs. latency requirements

Now we look at the system performance for real-time CBR traffic with different delay requirements. Fig. 5.12 shows the connection blocking rate of real-time CBR traffic with delay budget from 30ms to 80ms in the 6 $\times$ 6 topology. We observe that when the delay budget is relatively tight, such as between 30ms and 50ms, the connection

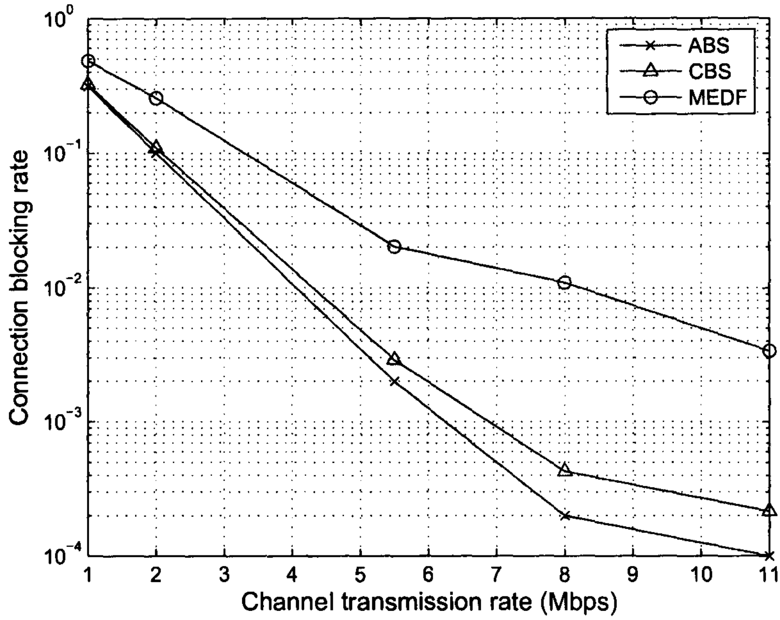

Figure 5.11: Connection blocking rate: different channel transmission rates (A = 4 Erlangs, $c = 3$)

blocking rate using both ABS and CBS is relatively high, and a more relaxed delay budget results in lower connection blocking rate. This implies that when the delay budget is tight, a new connection request can be rejected because its delay requirement cannot be satisfied. When the delay budget is 60ms or higher, further increasing the delay budget does not affect the connection blocking performance, meaning that the non-zero connection blocking rate is caused by insufficient bandwidth (number of timeslots).

It is also shown that the blocking rate of MEDF is not much affected by the delay budget requirement. This is because MEDF always schedules the packet with

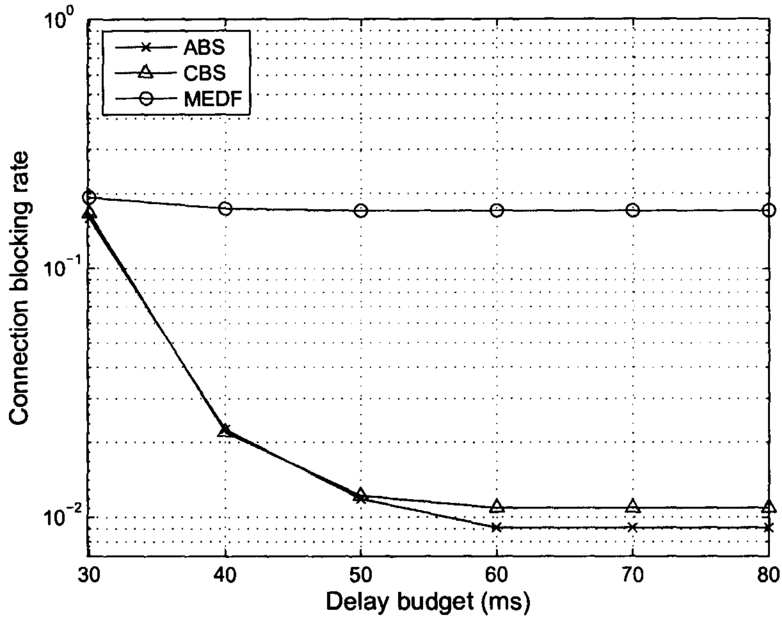

Figure 5.12: Connection blocking rate: different delay budgets (A =1 Erlang, $c = 3$)

the smallest delay budget first. As a result, packet losses are seldom caused by unsatisfactory latency performance. This is true even when the delay requirement is as low as 30ms.

5.3.4 Two types of connection blocking

A connection may be rejected due to either of the two reasons: i) there is no sufficient bandwidth (or number of timeslots) available to serve the connection; and ii) the delay requirement of the connection cannot be satisfied. In the experiment below, we count the number of connections rejected due to each of the two reasons separately, then divide the number by the total number of rejected connections and generate the

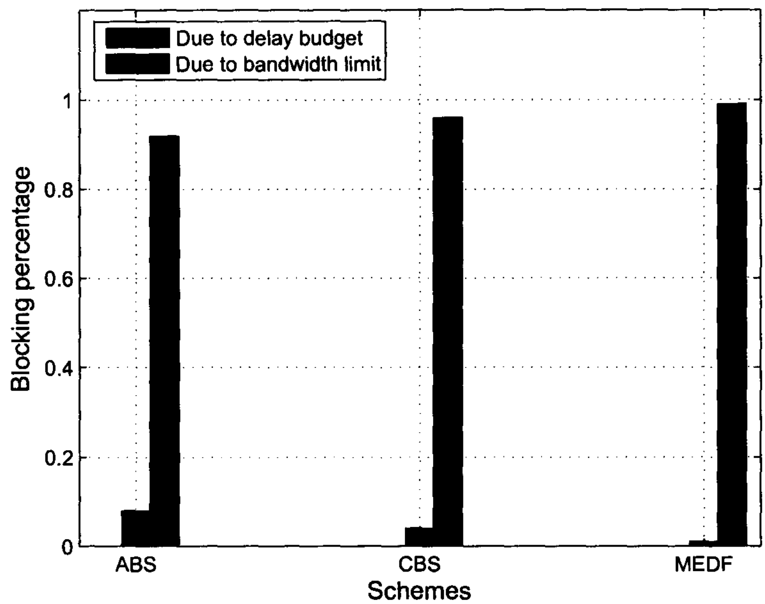

Figure 5.13: Two types of connection blockings ($A = 1$ Erlang, $c = 3$)

blocking percentage due to bandwidth limit and blocking percentage due to delay budget. Fig. 5.13 shows that for all the three schemes, i.e., ABS, CBS, and MEDF, the majority of connection rejections is due to the bandwidth limitation in the 6 $\times$ 6 topology. This demonstrates that bandwidth is the dominant factor restricting the system capacity in the considered system scenario. This is consistent with the conclusion drawn in Chapter 3 based on the optimal numerical results in the 3 $\times$ 3 topology.

In addition, Fig. 5.13 explains the capacity performance gap among ABS, CBS and MEDF. ABS has the lowest blocking percentage due to bandwidth limit. This implies that ABS is able to utilize the radio resources more efficiently. The blocking percentage due to delay budget is the lowest for MEDF, since MEDF always gives the packet with the minimum delay budget the highest priority. On the other hand, MEDF has the highest blocking percentages due to bandwidth limit, and thus the highest connection blocking rate. The blocking percentages due to bandwidth and delay budget of CBS are between the corresponding ones of ABS and MEDF. By giving a higher priority to the AP with the highest traffic load along the route of each connection, CBS has better capacity performance than MEDF. By minimizing the latency at each hop, CBS has better delay performance than ABS.

5.3.5 Performance vs. network topologies

We then consider WMNs with three different topologies, the 3×3 topology as shown in Fig. 3.1, and the $1+2\times4$ and $1+4\times2$ topologies as shown in Fig. 4.10. In Fig. 5.14, we compare the connection blocking rate of CBS in the three mesh topologies. It is seen that the connection blocking rate is the lowest in the 1+2×4 topology, and highest in the 1+4×2 topology. For access traffic, the bottleneck of a WMN is the root AP or the APs one hop away from it [84]. The system bottleneck can be any AP from AP 1 to AP 5 in the 1+2×4 topology, APs 1, 2, 4, and 5 in the 3×3 topology, and from AP 1 to AP 3 in the 1+4×2 topology. We define the system bottleneck region size (BRS) as the number of possible system bottleneck APs in a WMN. The BRS is 5 in the 1+2×4 topology, 4 in the 3×3 topology, and 3 in the 1+4×2 topology.

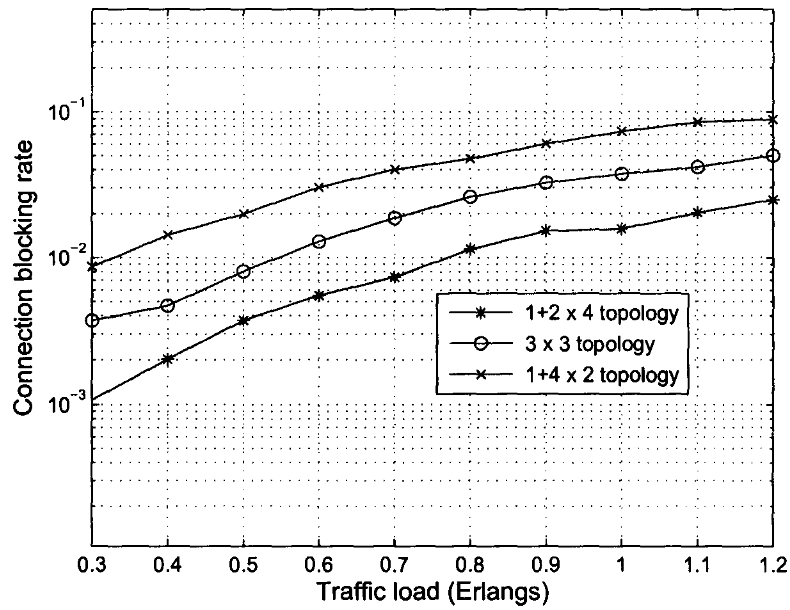

Figure 5.14: Connection blocking rate: different topologies ($c = 1$)

A larger BRS allows traffic to be more evenly distributed in the network, resulting in higher network capacity and lower connection blocking rate.

Fig. 5.15 shows the average packet transmission delay in different topologies. The 1+2×4 topology has the lowest average delay and the 1+4×2 one has the highest average delay. This is mainly due to the number of hops that connections have to traverse—a larger number of hops results in a longer delay.

5.3.6 Performance vs. network sizes

We compare the connection blocking rate in three topologies, 4×4, 5×5, and 6×6 grids with 16, 25 and 36 APs, respectively. The 4×4 and 5×5 topologies are shown

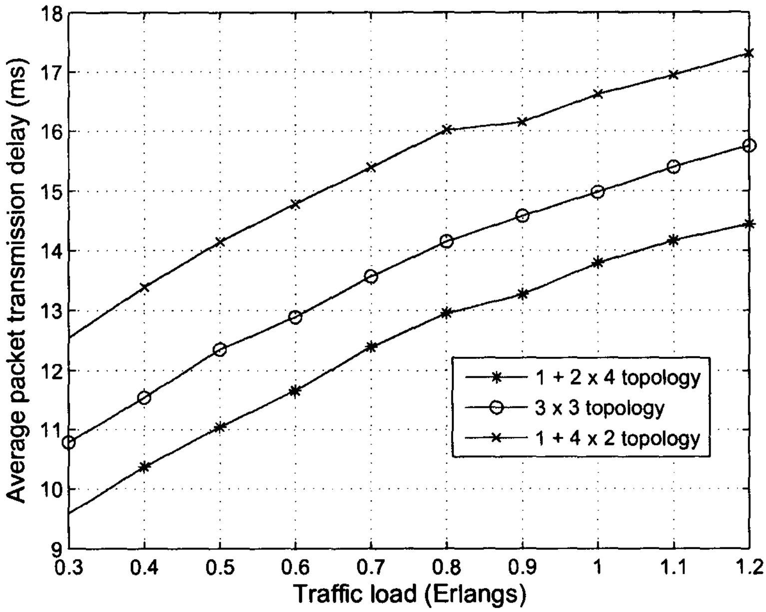

Figure 5.15: Average packet transmission delay: different topologies ($c = 1$)

in Fig. 5.16, and the 6×6 topology is shown in Fig. 5.3. Fig. 5.17 shows that the connection blocking rate of CBS increases slightly with the network size for a given traffic load. In other words, the system capacity decreases slightly with the increase of the network size. This is because with the increase of network size, the delay of each connection can be increased, which may increase the connection blocking rate.

5.3.7 Performance vs. number of frequency channels

Co-channel interference prevents the APs that share the same frequency channel from transmitting simultaneously. Fig. 5.18 shows that with the increase of the number of frequency channels, the connection blocking rate is decreased accordingly, indicating

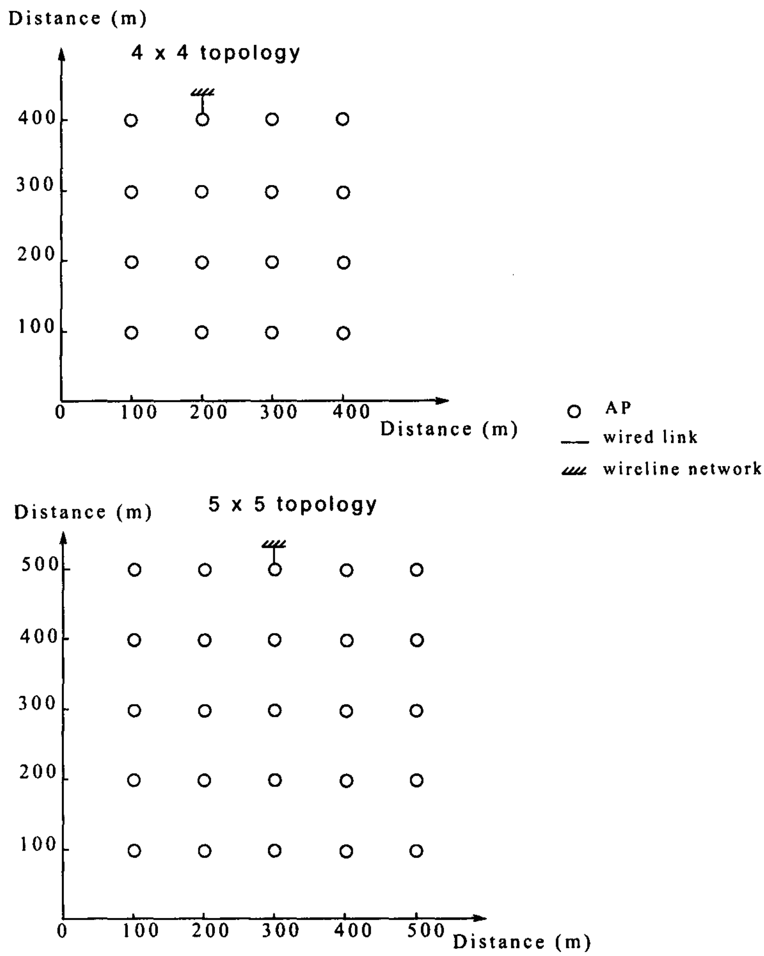

Figure 5.16: WMNs with 4 × 4 and 5 × 5 topologies

an increase in the system capacity. With more channels available, more APs can transmit at the same time using different frequency channels. As a result, more connections can be supported in the WMN.

We also observe that the connection blocking rate does not decrease much when the number of frequency channels is increased from 3 to 36. This indicates that 3 frequency channels are sufficient to prevent most of the co-channel interference in

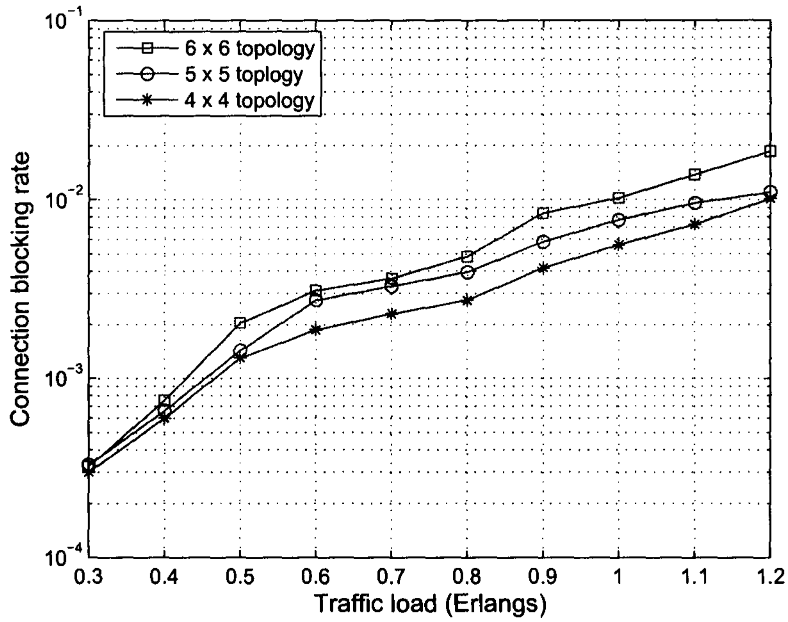

Figure 5.17: Connection blocking rate: different network size ($c = 3$)

the considered network using the proposed scheduling scheme. Further increasing the number of frequency channels does not improve the system capacity, which is now limited by the number of available timeslots in each scheduling interval, or the capacity of individual APs. Adding more radios to the APs may further improve the system capacity as will be shown in Section 6.2.

5.3.8 Running time of CBS and ABS

Now we compare the running time of ABS and CBS. We consider the 6×6 topology without co-channel interference. The system setting and parameters are the same as described in Section 4.4.3. Fig. 5.19 shows the running time of CBS and

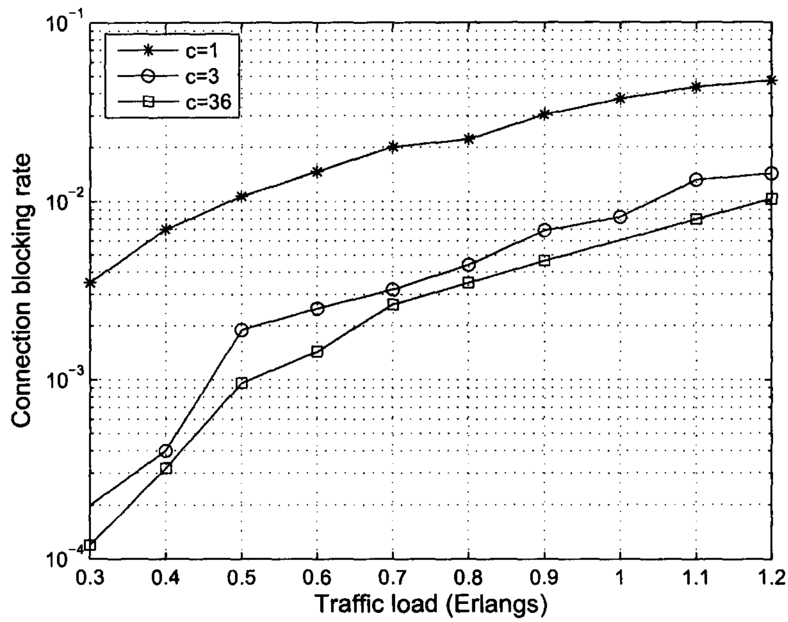

Figure 5.18: Connection blocking rate: different number of frequency channels (ABS)

ABS. The running time of CBS is lower than that of ABS, and the difference increases significantly with the traffic load. The figure shows that the running time of ABS may be larger than the tolerable delay (e.g. 1 or 2 seconds) for establishing a new connection when the number of connections is relatively large. In contrast, the running time of CBS increases linearly with the number of connections, as derived in (5.2). Therefore, CBS is more suitable for WMNs with a relatively heavy traffic load and is able to rapidly generate schedules for real-time connections.

In summary, ABS can be used as an approximation of the optimum scheduling, and CBS archives slightly lower capacity performance than ABS but with much lower complexity.

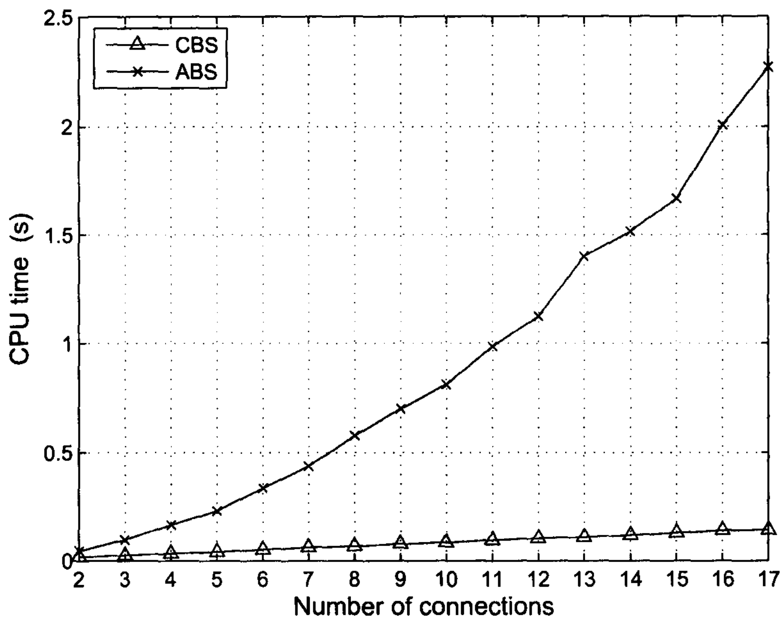

Figure 5.19: Running time of ABS and CBS (C=11Mbps, $c = 3$)

5.4 Summary

In this chapter, we have proposed a connection-based scheduling (CBS) scheme for supporting real-time CBR traffic in a WMN. The scheme achieves connection blocking performance close to the ABS scheduling while providing satisfactory delay performance at lower complexity. Both ABS and CBS schemes can be extended to a WMN with multi-radio APs and to support VBR real-time traffic. These extensions are presented in the next chapter.

Chapter 6

Extensions of Proposed Scheduling Schemes

The proposed ABS and CBS schemes in the previous two chapters are limited to schedule CBR traffic in a WMN with single-radio APs. However, these conditions can be relaxed. In this chapter, both the ABS and CBS schemes are extended to i) schedule real-time VBR traffic, and ii) schedule real-time traffic in a WMN with multi-radio APs.

6.1 Scheduling of real-time VBR traffic

6.1.1 Introduction

There are two basic approaches to serving real-time VBR traffic. One is to dynamically allocate network resources based on the instantaneous packet arrival rate and buffer status; the other is to allocate a constant service rate to the VBR traffic. The objective for the resource allocation methods is to satisfy the delay and packet loss requirements of the VBR traffic, while efficiently utilizing the network resources.

Dynamic resource allocation for serving VBR traffic can utilize the network resources more efficiently and has been studied extensively for wireline networks, for example in [85] [86], and single-hop wireless networks, for example in [87] [88]. However, it is very difficult to implement these methods in WMNs. First of all, a packet transmission at any given time involves two APs, one transmitting and one receiving. In a WMN, the timeline arrangement of an AP affects that of the APs one hop away from it, i.e., the APs directly communicating with it, and may further affect the timeline arrangement of other APs in the same network. Moreover, APs sharing the same frequency channel may cause co-channel interference to each other. Thus adjusting the packet transmission time at one AP may change the interference condition of other APs and probably further the interference condition in the entire network. In addition, signaling exchange for coordinating the timelines of the APs and distributing the scheduling decisions can be time consuming in the multihop mesh environment. Therefore, dynamically adjusting wireless resources allocations on a packet-by-packet basis is not practical for supporting real-time VBR traffic in WMNs.

A more practical approach is to use relatively static resource allocation methods. In this section we propose to use a constant rate to serve each VBR connection so that a number of timeslots are allocated to each VBR connection periodically. In this way, updating the scheduling decisions is only required when the new connection arrives, but not needed before every packet transmission. This is reasonable since the initial access delay (e.g., 1 or 2 seconds) can be much longer than the tolerable packet transmission delay (e.g., a few tens of milliseconds).

6.1.2 Scheme

We consider homogeneous real-time VBR traffic. Each connection is modelled by a superposition of N_p ON-OFF mini-sources, each of which has two states—ON state and OFF state [89] [90]. During the ON state the mini-source generates packets at a constant rate, R_p packets per second, and during the OFF state there is no packet generated. For each mini-source, let α and β be the transition probabilities from the ON state to the OFF state and from the OFF state to the ON state, respectively.

Both the ON and OFF intervals are exponentially distributed, and thus the average ON and OFF intervals are $T_{on} = 1/\alpha$ and $T_{off} = 1/\beta$, respectively. Each connection has strict delay and packet loss rate requirements. Each connection has a delay budget, denoted as d, and packet loss rate, denoted as p. A packet is dropped if its experienced delay exceeds d. Packet losses can also be caused by background noise and co-channel interference. Background noise is not considered, as it is usually much less than co-channel interference. Let p_d be the packet loss rate due to long delay, and p_c the packet loss rate due to co-channel interference. We assume that the buffer space at each AP is sufficiently large so that there is no packet loss due to buffer overflow. The required buffer size at the APs in order to achieve this objective will be given below. The relationship between p_d, p_c and p is given by $p = p_d + (1 - p_d)p_c$. We consider that a packet transmission is successful only if there is no other AP transmitting within the interference range of the considered receiver. Both ABS and CBS can ensure this condition. Therefore, $p_c = 0$ and $p = p_d$.

Our approach to serving a VBR connection is to first convert it into a CBR connection. This is achieved by having the first hop AP to serve the VBR connection at a constant rate, C_e. In order to simplify the presentation we define the home hop as the first hop of the VBR connection. This is the hop from the MS to its home AP for the uplink, and from the root AP to the next downstream AP in the downlink. Other hops along the route of the connection are referred to as the mesh hops of the connection. Once the connection has passed the home hop, it is converted to a CBR connection, and can be served based on scheduling decisions using either ABS or CBS. The QoS performance of the VBR connection can be found by combining the service quality along the home hop and along the mesh hops. Both the home hop and the mesh hops involve transmission delay. Let d_h and $d_m = d - d_h$, respectively, represent the delay budgets allowed for the home hop and the mesh hops. For an admitted connection, packet losses only occur along the home hop, because beyond that the connection is equivalent to a CBR connection and the service rate is equal to the packet generation rate of the CBR connection. Therefore, only the home hop transmissions contribute to the value of p_d. Based on this, the rate C_e can be found by using the concept of effective bandwidth (EB) [91]:

$$\frac{C_e}{N_p} = \frac{R_p}{2} - \frac{(\alpha+\beta)b}{2\ln(1/p_d)} + \sqrt{\left[\frac{R_p}{2} - \frac{(\alpha+\beta)b}{2\ln(1/p_d)}\right]^2 + \frac{\alpha R_p b}{\ln(1/p_d)}}, \tag{6.1}$$

where

$$b = C_e d_h \tag{6.2}$$

is the required buffer size at the AP of the home-hop. C_e is the constant rate to serve the VBR connection in order to guarantee the packet delay (for the home hop) and loss rate requirement and its value can be found numerically from (6.1) and (6.2).

Given the EB of each VBR connection, the scheduler virtually performs the ABS or CBS scheduling scheme along the mesh hops of the connections assuming each VBR connection is a CBR connection with rate C_e. The scheduling is done for all existing connections and the new connection. The transmission delay along all the mesh hops for connection i, $w_{m,i}$, then can be found after the virtual scheduling is done. If $w_{m,i} \leq d_m$ for all existing connections and the new connection, the new connection can be accepted. Otherwise, the connection is rejected. Note that upon the time when the new connection is admitted in the system, the scheduling decisions for the new connection have been done and those for existing connections have been updated. The overhead for making admission control and scheduling decisions occurs before a new connection starts.

6.1.3 Numerical results

We consider voice traffic with alternate active and silent periods. Each voice connection generates packets at a constant rate during the active periods and no packet is generated during the silent periods. We consider the WMN with the 6×6 grid topology as shown in Fig. 5.3. Co-channel interference is not considered. Default parameters are shown in Table 6.1, where parameters related to the ON-OFF mini-sources are from reference [92]. We present numerical results of both connection and packet level performance.

Fig. 6.1 shows the connection blocking rate of ABS, CBS and MEDF, where $d_h/d = 25\%$. It is shown that both the extended ABS and CBS result in much lower connection blocking rate than MEDF. That is, for a given connection blocking rate

Table 6.1: Default parameters

Number of mini-sources in each VBR connection	1
Average ON time of a voice connection	240 ms
Average OFF time of a voice connection	400 ms
Maximum packet loss rate	1%
Packet generation rate	50
Physical channel transmission rate	11Mbps
Maximum packet transmission delay budget	60ms

requirement, using the extended ABS and CBS can accommodate much more VBR traffic than using MEDF.

We then take the extended ABS as the scheduling scheme and show the packet level performance. When a packet experiences a longer delay than the delay budget, it is dropped, and the packet loss rate is shown in Fig. 6.2. It is seen that the packet loss rate increases with the traffic load, but is always below the maximum tolerance of 1%. This demonstrates that the CAC and packet transmission scheduling using extended ABS can guarantee the packet loss rate for VBR traffic. The gap between the actual packet loss rate ($< 0.8\%$) and the required value (1%) is due to the conservative estimation of EB in (6.1), which tends to over-allocate resources [91].

Fig. 6.3 shows the relationship between the home hop delay budget d_h and connection blocking rate. There is an optimum range of d_h, denoted as d_h^*. The connection blocking rate is minimized when d_h is in this range. We observe that d_h^* is from 21% to 25% of d, i.e., from 12.6 to 15ms in the figure. Varying d_h in the

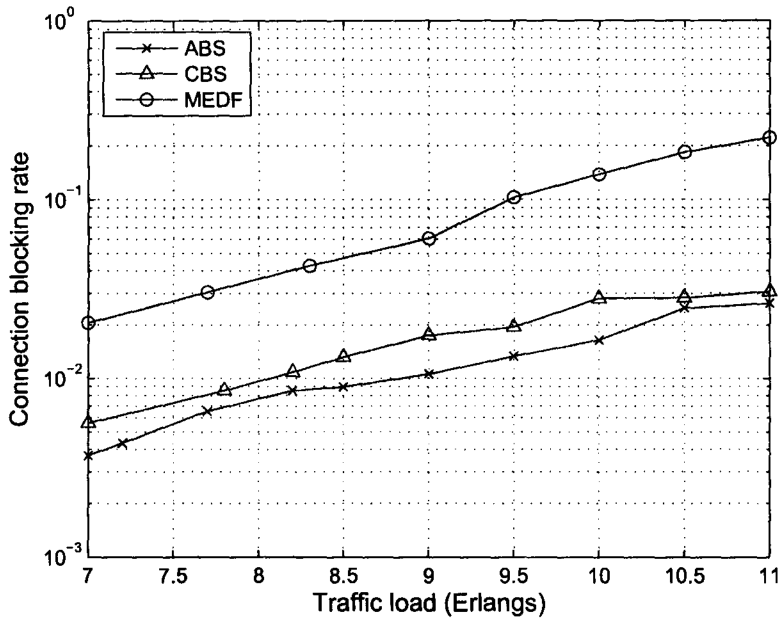

Figure 6.1: Connection blocking rate of extended ABS, CBS and MEDF: VBR traffic

range does not affect the connection blocking rate very much. This observation can be explained by looking at the two contradictory effects of d_h on the system performance. First, when d_h is increased, the effective bandwidth C_e is decreased due to the relaxed delay requirement at the home hop. Thus more connections can possibly be accommodated in the network. That is, the system capacity can potentially be improved, as long as the overall (home and mesh hops together) delay performance is not a problem. Because of this, increasing d_h can potentially decrease the connection blocking rate. Second, when d_h is increased, d_m is decreased, leaving less delay budget for the mesh hops. This may increase the connection blocking rate due to the delay constraint. When d_h is small, d_m is

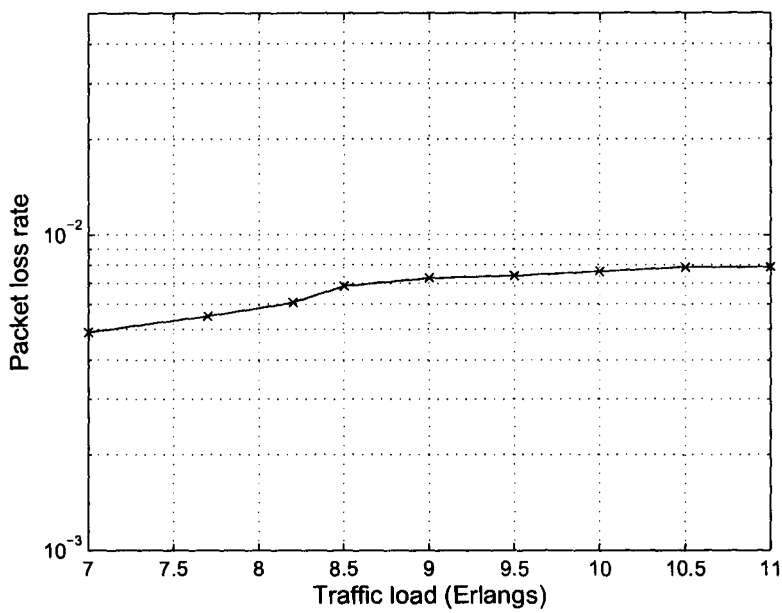

Figure 6.2: Packet loss rate vs. traffic load

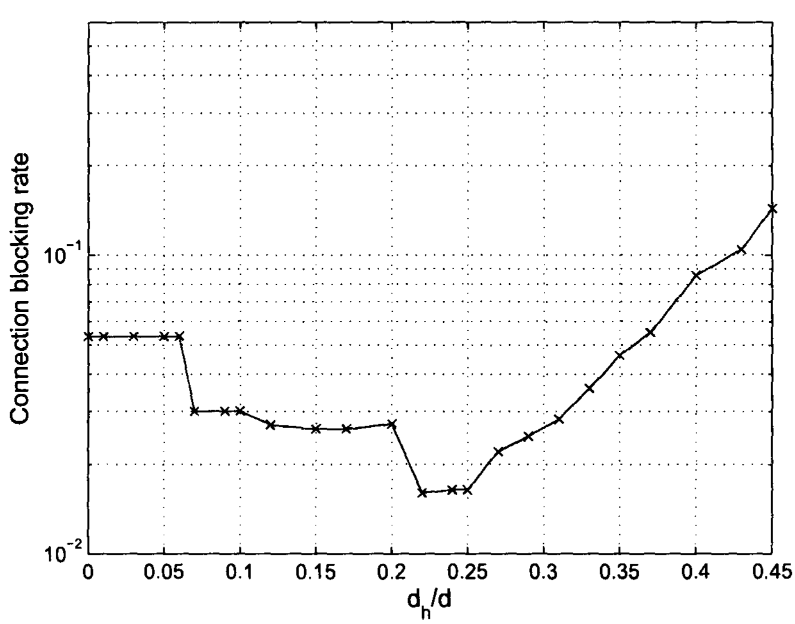

Figure 6.3: Connection blocking rate vs. d_h/d ($A = 10$ Erlangs)

sufficiently large so that the connection blocking rate due to stringent delay budget along the mesh hops is very small, and the first effect of d_h on the blocking rate dominates. As a result, the overall blocking rate decreases with the increase of d_h, which is shown in Fig. 6.3 when $d_h < d_h^*$. When d_h is sufficiently large, further increasing d_h results in too small d_m, which then results in very high blocking rate due to the fact that the delay requirement cannot be satisfied, even though the system may have enough bandwidth. In this case, the connection blocking rate increases with d_h, which is shown in Fig. 6.3 when $d_h > d_h^*$.

We also observe from Fig. 6.3 that the connection blocking rate does not decrease monotonically with d_h when d_h is relatively small. This is mainly due to the TDMA effect in packet transmissions. Although the EB of each VBR connection decreases continuously with d_h, the number of timeslots in each SI can only take discrete values. Therefore, until the number of timeslots can be increased by 1, increasing d_h does not increase the network capacity and therefore does not affect the connection blocking rate. We also find the blocking rate curve is smoother when d_h is relatively large, because connection blocking is mainly caused by tight d_m in the mesh hops, rather than limited by the available timeslots or bandwidth.

6.2 Scheduling in a WMN with multi-radio APs

Having multiple frequency channels can increase the network capacity. However, the results in Section 5.3.7 have shown that the system capacity for access traffic is limited by the capacity of bottleneck APs, which is independent of the number of frequency channels. Therefore, in order to further increase the mesh network capacity, the capacity at the bottleneck APs should be increased. Having more radios working at different frequency channels in the APs can potentially achieve this objective. This method becomes possible and practical with the low price of wireless adapters [93]. In this section, we extend the ABS and CBS scheduling schemes to WMNs with multi-radio APs.

6.2.1 Schemes

Both ABS and CBS make scheduling decisions based on the traffic load of the APs. For scheduling in a WMN with multi-radio APs, we define the relative traffic load of AP m as

$$L_m = \frac{|C_{Rm}| + |C_{Tm}|}{Z_m}, \qquad (6.3)$$

where Z_m is the number of radios at AP m. This is to reflect the fact that in a WMN with multi-radio APs, an AP with more radios should be able to process more packets.

Since all Z_m radios at the same AP work at different frequency channels, at most

Z_m packets can be processed at AP m at any given time. We define a set of binary variables $A_{m,z,t}$ with $A_{m,z,t} = 1$ representing that the z-th radio at AP m is available at timeslot t and $A_{m,z,t} = 0$ otherwise. Similar to the single-radio AP, the real-time portion in one typical SI at the z-th radio of AP m is defined as

$$T_{rt,z,m} = t_{m,z,\max} - t_{m,z,\min} + 1, \tag{6.4}$$

where $t_{m,z,\max} = \text{argmax}_t\{A_{m,z,t} = 0\}$, $t_{m,z,\min} = \text{argmin}_t\{A_{m,z,t} = 0\}$, and t is in the considered typical SI of the z-th radio of AP m. Then the real-time portion of the AP is defined as

$$T_{rt,m} = \max_{1 \leq z \leq Z_m} T_{rt,z,m}. \tag{6.5}$$

The scheduling process should ensure that $T_{rt,m} \leq T_{SI}$, or $T_{rt,z,m} \leq T_{SI}$ for all $z = 1, 2, \ldots, Z_m$. The real-time portion of AP m can also be defined using the same expression as in (4.1) with $t_{m,\max}$ and $t_{m,\min}$, where $t_{m,\max} = \max_{z=1}^{Z_m} t_{m,z,\max}$ and $t_{m,\min} = \min_{z=1}^{Z_m} t_{m,z,\min}$.

Consider the hop h of connection i with AP m and AP n as the transmitter and receiver, respectively. Before determining $\tau_{i,h}$ using ABS or CBS, $\mathcal{D}_{i,h}$ should be found first. In a WMN with multi-radio APs, if $t \in \mathcal{D}_{i,h}$, then at least one radio in AP m and one radio in AP n are available at t. Thus each timeslot t in $\mathcal{D}_{i,h}$ may correspond to multiple pairs of radios of AP m and AP n. Let the z-th radio of AP m and the y-th radio of AP n be one of such radio pairs, then $A_{m,z,t} = A_{n,y,t} = 1$. Let f represent the frequency channel assigned to the z-th radio of AP m to transmit. If $t \in \mathcal{D}_{i,h}$, then except for the z-th radio of AP m, no other radio of any AP that is within the interference range of AP n and assigned frequency channel f has been scheduled to transmit at time t. Meanwhile, except for the y-th radio of AP n, no other radio that is within the interference range of AP m has been scheduled to receive at frequency channel f at time t.

When extending ABS to a WMN with multi-radio APs, the basic steps are exactly the same as in Pseudocode 4.1, but only slight changes are required to Pseudocode 4.2. Each timeslot t in $\mathcal{D}_{i,h}$ may correspond to multiple pairs of radios of AP m and AP n. Without loss of generality we assume that AP m has a higher relative traffic load than AP n, then ABS chooses the radio pair that results in the minimum real-time portion of AP m among all the available radio pairs. This is shown in Pseudocode 6.1, where $\mathcal{P}_t = \{(z, y)|A_{m,z,t} = 1, A_{n,y,t} = 1\}$ denotes a set of all radio pairs of AP m and AP n corresponding to the timeslot $t \in \mathcal{D}_{i,h}$. This code is to replace Lines 5-10, 16-21, 28-33 and 40-44 in Pseudocode 4.2 so that the selected pair of radios in the transmitting and receiving APs results in the minimum real-time portion of AP m. In Line 6, the z^*-th radio of AP m is selected as the transmitting radio and the y^*-th radio of AP n is selected as the receiving radio.

Pseudocode 6.1: Selecting transmitting and receiving radios

1: $\mathcal{P}_{t,0} = \mathcal{P}_t$

2: **if** $t \in \mathcal{D}_{i,h}$ **then**

while $\mathcal{P}_{t,0}$ is not empty **do**
    if $\tilde{T}_{rt,m} \leq T_{SI}$ and $\tilde{T}_{rt,n} \leq T_{SI}$ **then**
        if $\tilde{T}_{rt,m} < T_{temp}$ **then**
            Let $T_{temp} = \tilde{T}_{rt,m}$, $\tau_{i,h} = t$, $z^* = z$, and $y^* = y$
        end if
        Let $\mathcal{P}_{t,0} = \mathcal{P}_{t,0} - (z, y)$
    end if
end while
end if

When extending CBS to the multi-radio case, the basic steps are still the same as outlined in Pseudocode 5.1. When determining $\tau_{i,h}$ for a given hop of connection i, the scheme chooses the timeslot which minimizes the delay between successive hop transmissions. After such a timeslot is found, if there are multiple transmitting-receiving radio pairs available, then the extended scheme chooses the pair of radios that minimizes the real-time portion of AP m (assuming AP m has a heavier load than AP n). This can be implemented by replacing Lines 8, 19, and 30 in Pseudocode 5.2 with Pseudocode 6.1.

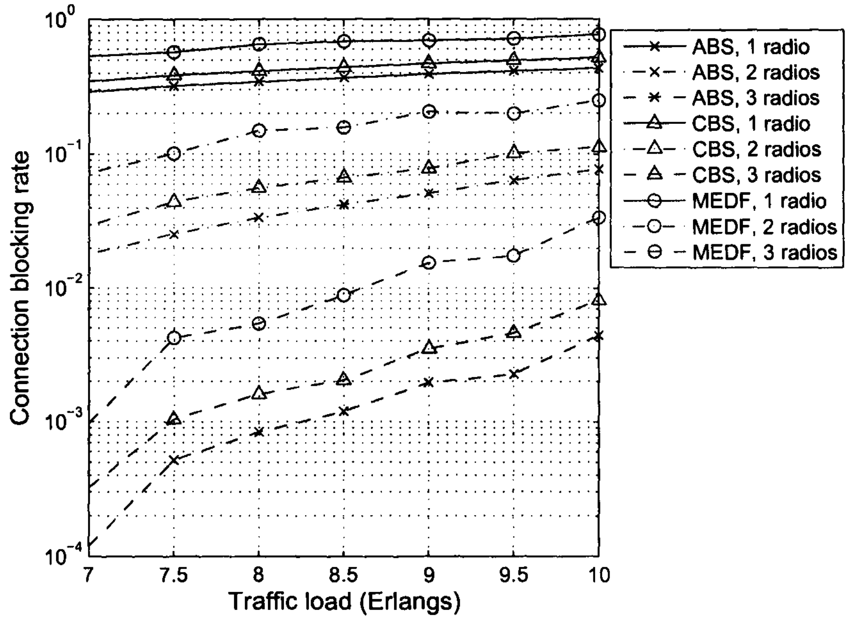

Figure 6.4: Connection blocking rate vs. number of radios at each AP

6.2.2 Numerical results

We consider the WMN with a 3 $\times$ 3 topology as shown in Fig. 3.1 and 3 frequency channels available. We adopt parameters as described in Section 3.5.1 but vary the number of radios at each AP. Fig. 6.4 compares the connection blocking rate of MEDF, CBS and ABS when each AP has 1 to 3 radios. Adding more radios to the APs can reduce the blocking rate significantly. It is shown that the connection blocking rates of the extended CBS and ABS schemes are close to each other and lower than that of the MEDF. This observation is consistent with that in the WMN with single-radio APs. The figure also shows that the difference between the blocking rates of ABS/CBS and that of MEDF is larger as more radios are in each AP. This shows that the proposed schemes can take better advantage of the increased radio resources than MEDF.

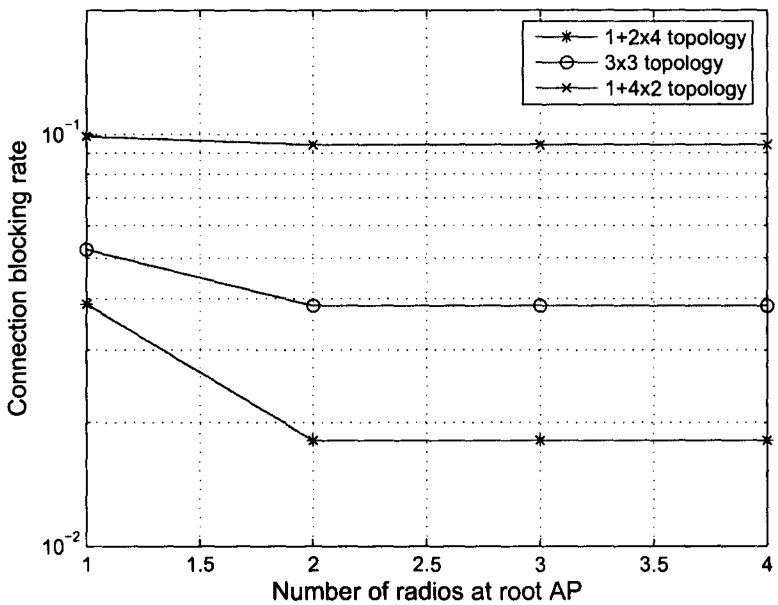

Figure 6.5: Connection blocking rate vs. number of radios at root AP (A= 3 Erlangs)

Next we consider 3 different mesh topologies: the 3×3 topology shown in Fig. 3.1, the $1 + 2 \times 4$ and $1 + 4 \times 2$ topologies shown in Fig. 4.10. We perform a different experiment by varying the number of radios at the root AP (AP 1) and fixing the number of radios at each of the other APs in the network to one. Fig. 6.5 shows the connection blocking rate vs. the number of radios (Z_1) at the root AP. Increasing Z_1 from 1 to 2 in the 3×3 and 1+2×4 topologies reduces the connection blocking rate. However, further increasing Z_1 to 3 and 4 does not lower the blocking rate. In the

$1 + 4 \times 2$ topology, increasing Z_1 does not affect the connection blocking rate.

This phenomenon can be explained as follows. First, we define a level-h AP as an AP that is h hops away from the root AP. Then the root AP with Z_1 radios can process at most $K_0 = Z_1 \lfloor T_{SI}/2 \rfloor$ two-way connections, and each of the other (single-radio) APs in the network can process at most $\lfloor T_{SI}/4 \rfloor$ two-way connections, where T_{SI} is normalized to the duration of a timeslot (or one packet transmission time). Note that each two-way connection requires 2 timeslots in the root AP and 4 timeslots in a non-root AP in every SI. The maximum number of two-way connections that all level-h APs can process is $K_h = M_h \lfloor T_{SI}/4 \rfloor$, where M_h is the total number of level-h APs. When $h_1 \leq h_2$, $M_{h_1} \leq M_{h_2}$, thus $K_{h_1} \leq K_{h_2}$. Therefore, the system capacity bottleneck is either at the root AP or the level-1 APs, depending on the values of K_0 and K_1. If $K_0 = K_1$, the root AP and the level-1 APs together form the capacity bottleneck of the network. Table 6.2 shows the capacity bottleneck in the three mesh topologies. It can be seen from the table that in the 3×3 topology, the capacity bottleneck is in the root AP when $Z_1 = 1$. Therefore, increasing Z_1 from 1 to 2 increases the system capacity. Meanwhile, the capacity bottleneck is also moved from the root AP to the level-1 APs, and further increasing Z_1 does not increase the system capacity. The blocking rate curves for the $1+2\times4$ and $1+4\times2$ topologies can also be explained similarly. These results show that having more radios at the capacity bottleneck can increase the network capacity. However, care must be taken

Table 6.2: System capacity bottleneck

Mesh topology	$Z_1 = 1$	$Z_1 = 2$	$Z_1 = 3$	$Z_1 = 4$
1+2×4 topology	Root AP	Level-1 APs	Level-1 APs	Level-1 APs
3×3 topology	Root AP	Level-1 APs	Level-1 APs	Level-1 APs
1+4×2 topology	Root AP & Level-1 APs	Level-1 APs	Level-1 APs	level-1 APs

as the bottleneck can be changed when more radios are added.

6.3 Summary

Similar to scheduling for real-time CBR traffic, the scheduling for VBR traffic can also be done at the time of new connection arrivals and the results can be used for making admission control decisions. By combining the concept of effective bandwidth with the scheduling schemes for CBR traffic, both delay and packet loss requirements of VBR traffic can be satisfied. As the scheduling is done before a new connection is admitted into the system, the results can be easily used for admission control purpose.

It is interesting to find that in a WMN with multi-radio APs, additional radios should be added in the capacity bottleneck in order to maximize the network capacity, and the bottleneck can be changed as radios are added to the network.

Chapter 7

Conclusions and Future Work

7.1 Summary and Conclusions

In this thesis we have studied packet transmission scheduling in wireless mesh networks for supporting real-time traffic. The main contributions of the thesis are summarized as follows.

The packet transmission scheduling problem is formulated as a standard integer linear programming problem by taking into consideration timeline availability of mesh APs, co-channel interference, and packet transmission delay. The solution to the optimization problem provides a theoretical benchmark and direction for designing heuristic scheduling schemes.

An AP-based scheduling scheme (ABS) is proposed. By giving a higher priority to the APs with higher traffic loads, ABS utilizes the wireless resources efficiently. It achieves capacity performance close to the optimum scheduler and satisfies packet transmission delay requirement with lower complexity than solving the optimization problem.

A connection-based scheduling scheme (CBS) is proposed. By giving a higher priority to the connections with more hops, CBS achieves better delay performance than ABS. Although the capacity performance of CBS is slightly poorer than ABS, the connection-based scheme simplifies the timeline coordinations among mesh APs and requires much lower complexity than ABS. For each individual connection, the scheme can efficiently utilize the remaining resources from serving higher priority connections. It achieves capacity performance close to the connection-based optimum scheduling.

The proposed ABS and CBS schemes are extended to handle VBR traffic. Both ABS and CBS guarantee the packet loss rate and delay performance of VBR connections once the connections are accepted in the WMN.

The proposed ABS and CBS schemes are extended to WMNs with multi-radio APs. Both of the extended schemes achieve good system performance as in the WMN with single-radio APs.

Packet transmission scheduling in a WMN can require much higher complexity than in single hop wireless networks. By giving a priority to APs with higher traffic loads and to connections with more hops, it is possible to reduce the complexity significantly with little compromise to the system capacity. The proposed heuristic schemes, both ABS and CBS, update transmission schedules only at the time when new connections arrive. This not only avoids updating schedules on a timeslot basis, which can be time consuming in a WMN, but also makes it easy to perform admission control, which otherwise can be a very complicated problem. Therefore, the contribution of this thesis is not only for packet transmission scheduling but also for admission control in WMNs.

A real-time connection can be rejected, or a packet can be lost, due to either insufficient bandwidth or unsatisfactory QoS. As most current WMNs have a relatively small number of mesh APs, especially when traffic with strict QoS is to be supported, latency performance is not very critical when using the proposed scheduling schemes. Therefore, arranging the timeline of bottleneck APs is important for accommodating more real-time traffic in a WMN. On the other hand, if there is a need to support traffic with very stringent latency requirement in a relatively large WMN, connections are more likely to be rejected due to unsatisfactory latency performance, even though the radio resources are not fully utilized.

Although co-channel interference reduces the network capacity, its effect can be dramatically mitigated by carefully coordinating the timelines of mesh APs. By using

ABS and CBS, it is possible for a WMN with a moderate and low level of co-channel interference to achieve the capacity close to a WMN without co-channel interference. The capacity of a WMN can be further increased by adding more radios (working at different frequency channels) in the bottleneck APs.

7.2 Future work

The complexity of solving the global optimum scheduling problem in Chapter 3 is NP-hard. This makes it only possible to find optimum scheduling solutions for small size WMNs. Reducing the complexity of the optimum scheduling for large size WMNs will be helpful in the future as it will provide an important performance benchmark for heuristic scheduling schemes.

Throughout this thesis we assume there is a central station available for collecting network and traffic information and making scheduling decisions. This type of centralized scheduling for WMNs is necessary, given the very high complexity of the scheduling problem. On the other hand, it is also desirable to design distributed scheduling schemes so that QoS support is possible in WMNs without a central station. The design of such scheduling schemes that can work with reasonable complexity is going to be a big challenge.

The scheduling in this thesis assumes that the route for each connection is given and unchanged. As a result, a connection can be rejected due to insufficient resources available at APs along the specified route, even though a different route can accept it. Joint routing and admission control/scheduling can potentially improve the resource utilization and reduce the connection blocking rate.

Scheduling for multicast and broadcast traffic in WMNs is another important and interesting topic. Multicast traffic is expected to expand in the near future, especially with the increasing video applications over WMNs; and broadcast in the WMN also plays an important role for exchanging signaling messages and other purposes. Different from the scheduling for unicast traffic, scheduling multicast and broadcast traffic should consider that one transmitter may communicate with multiple receivers at the same time.

Bibliography

[1] I. Akyildiz and X. Wang, "A survey on wireless mesh networks," *IEEE Communications Magazine*, vol. 43, no. 9, pp. S23–S30, September 2005.

[2] J.-H. Huang, L.-C. Wang, and C.-J. Chang, *Architectures and Deployment Strategies for Wireless Mesh Networks.* Springer US, 2007.

[3] Y. Zhu, W. Zeng, H. Liu, Y. Guo, and S. Mathur, "Supporting video streaming services in infrastructure wireless mesh networks: architecture and protocols," *Proceedings of IEEE International Conference on Communications*, pp. 1850–1855, May 2008.

[4] R. Bruno, M. Conti, and E. Gregori, "Mesh networks: commodity multihop ad hoc networks," *IEEE Communications Magazine*, vol. 43, no. 3, pp. 123–131, March 2005.

[5] P. Giacomazzi, L. Musumeci, G. Caizzone, G. Verticale, G. Liggieri, A. Proietti, and S. Sabatini, "Quality of service for packet telephony over mobile ad hoc

networks," *IEEE Network,*, vol. 20, no. 1, pp. 12–20, Janurary-Feburary 2006.

[6] I. F. Akyildiz, X. Wang, and W. Wang, "Wireless mesh networks: a survey," *Computer Networks*, vol. 47, pp. 445–487, 2005.

[7] X. Wang, A. Patil, and W. Wang, "VoIP over wireless mesh networks: challenges and approaches," *Proceedings of the 2nd annual international workshop on Wireless Internet*, 2006.

[8] H. Zhu, M. Li, I. Chlamtac, and B. Prabhakaran, "A survey of quality of service in IEEE 802.11 networks," *IEEE Wireless Communications*, vol. 11, no. 4, pp. 6–14, August 2004.

[9] H. Fattah and C. Leung, "An overview of scheduling algorithms in wireless multimedia networks," *IEEE Wireless Communications*, vol. 9, no. 5, pp. 76–83, October 2002.

[10] J. Blazewicz, W. Domschke, and E. Pesch, "The job shop scheduling problem: conventional and new solution techniques," *European Journal of Operational Research*, vol. 93, pp. 1–33, 1996.

[11] E. Lawler, A. R. Kan, and D. Shmoys, *Sequencing and scheduling: algorithms and complexity, Vol.4: Logistics of Production and Inventory.* Elsevier, 1993.

[12] J. Lenstra and A. R. Kan, "Complexity of machine scheduling problems," *Annals of Discrete Mathematics*, vol. 4, pp. 121–140, 1979.

[13] Y. Sotskov, "The complexity of scheduling problems with two and three jobs," *European Journal of Operational Research*, vol. 53, pp. 326–336, 1991.

[14] M. Lee, J. Zheng, Y.-B. Ko, and D. Shrestha, "Emerging standards for wireless mesh technology," *IEEE Wireless Communications*, vol. 13, no. 2, pp. 56–63, April 2006.

[15] IEEE Std 802.11i-2003 Edition, "Local and metropolitan area networks - specific requirements part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications," pp. 1–721, 2005.

[16] G. Hiertz, Y. Zang, S. Max, T. Junge, E. Weiss, and B. Wolz, "IEEE 802.11s: WLAN mesh standardization and high performance extensions," *IEEE Network*, vol. 22, no. 3, pp. 12–19, May-June 2008.

[17] S. Faccin, C. Wijting, J. Kenckt, and A. Damle, "Mesh WLAN networks: concept and system design," *IEEE Wireless Communications*, vol. 13, no. 2, pp. 10–17, April 2006.

[18] L. Chen and W. Heinzelman, "A survey of routing protocols that support QoS in mobile ad hoc networks," *IEEE Network*, vol. 21, no. 6, pp. 30–38, November-December 2007.

[19] L. Hanzo-II and R. Tafazolli, "A survey of QoS routing solutions for mobile ad hoc networks," *IEEE Communications Surveys and Tutorials*, vol. 9, no. 2, pp. 50–70, Second Quarter 2007.

[20] H. Aoki, N. Chari, L. Chu, W. Conner, S. Faccin, D. Guretich, V. Hasty, J. Jetcheva, S. Kangude, S. Saito, and R. Taori, "802.11 TGs simple efficient extensible mesh (SEE-Mesh) proposal," *IEEE. 802 11-05/0562r01*, July 2005.

[21] C. Perkins and E. Belding-Royer, "Ad hoc on-demand distance vector (AODV) routing," *IETF RFC 3561*, July 2003.

[22] T. Clausen and P. Jacquet, "Optimized link state routing protocol (OLSR)," *IETF RFC 3626*, October 2003.

[23] V. Loscrı, "MAC protocols over wireless mesh networks: problems and perspective," *Journal of Parallel and Distributed Computing*, vol. 68, no. 3, pp. 387–397, 2008.

[24] T.-J. Tsai and J.-W. Chen, "IEEE 802.11 MAC protocol over wireless mesh networks: problems and perspectives," *Proceedings of 19th International Conference on Advanced Information Networking and Applications*, vol. 2, pp. 60–63, March 2005.

[25] S. Xu, "Advances in WLAN QoS for 802.11: an overview," *Proceedings of 14th IEEE Proceedings on Personal, Indoor and Mobile Radio Communications*, vol. 3, pp. 2297–2301, September 2003.

[26] G.-H. Hwang and D.-H. Cho, "Performance analysis on coexistence of EDCA and legacy DCF stations in IEEE 802.11 wireless LANs," *IEEE Transactions on Wireless Communications*, vol. 5, no. 12, pp. 3355–3359, December 2006.

[27] D. Rawat, D. Popescu, and M. Song, "Performance enhancement of EDCA access mechanism of IEEE 802.11e wireless LAN," *Proceedings of IEEE Radio and Wireless Symposium,*, pp. 507–510, January 2008.

[28] B. Sikdar, "An analytic model for the delay in IEEE 802.11 PCF MAC-based wireless networks," *IEEE Transactions on Wireless Communications*, vol. 6, no. 4, pp. 1542–1550, April 2007.

[29] N. Ramos, D. Panigrahi, and S. Dey, "Quality of service provisioning in 802.11e networks: challenges, approaches, and future directions," *IEEE Network*, vol. 19, no. 4, pp. 14–20, July-August 2005.

[30] H. Trsek, J. Jasperneite, and S. Karanam, "A simulation case study of the new IEEE 802.11e HCCA mechanism in industrial wireless networks," *Proceedings of IEEE Conference on Emerging Technologies and Factory Automation*, pp. 921–928, September 2006.

[31] S. Xu and T. Saadawi, "Does the IEEE 802.11 MAC protocol work well in multihop wireless ad hoc networks?" *IEEE Communications Magazine*, vol. 39, no. 6, pp. 130–137, June 2001.

[32] H.-Y. Hsieh and R. Sivakumar, "IEEE 802.11 over multi-hop wireless networks: problems and new perspectives," *Proceedings of IEEE 56th Vehicular Technology Conference*, vol. 2, pp. 748–752, 2002.

[33] M. A. Munawar, "Multi-interface multi-channel wireless mesh networks," *Master Thesis, University of Waterloo, Canada*, 2004.

[34] IEEE Std 802.16-2001, "IEEE standard for local and metropolitan area networks, part 16: air interface for fixed broadband wireless access systems," April 2002.

[35] IEEE Std 802.16a-2003, "IEEE standard for local and metropolitan area networks part 16: air interface for fixed broadband wireless access systems amendment:2: Medium access control modifications and additional physical layer specifications for 2-11 GHz," January 2003.

[36] C. Eklund, R. Marks, K. Stanwood, and S. Wang, "IEEE standard 802.16: a technical overview of the WirelessMANTM air interface for broadband wireless access," *IEEE Communications Magazine*, vol. 40, no. 6, pp. 98–107, June 2002.

[37] C. Hoymann, "Analysis and performance evaluation of the OFDM-based metropolitan area network IEEE 802.16," *Computer Networks*, vol. 49, no. 3, pp. 341–363, October 2005.

[38] J. He, K. Yang, K. Guild, and H.-H. Chen, "Application of IEEE 802.16 mesh networks as the backhaul of multihop cellular networks," *IEEE Communications Magazine*, vol. 45, no. 9, pp. 82–90, September 2007.

[39] C. Cicconetti, L. Lenzini, E. Mingozzi, and C. Eklund, "Quality of service support in IEEE 802.16 networks," *IEEE Network*, vol. 20, no. 2, pp. 50–55, March-April 2006.

[40] C. Cicconetti, A. Erta, L. Lenzini, and E. Mingozzi, "Performance evaluation of the IEEE 802.16 MAC for QoS support," *IEEE Transactions on Mobile Computing*, vol. 6, no. 1, pp. 26–38, Janurary 2007.

[41] IEEE Std 802.16-2004, "IEEE standard for local and metropolitan area networks, part 16: air interface for fixed broadband wireless access systems," October 2004.

[42] N. Abu Ali, A.-E. Taha, H. Hassanein, and H. Mouftah, "IEEE 802.16 mesh schedulers: issues and design challenges," *IEEE Network*, vol. 22, no. 1, pp. 58–65, January-February 2008.

[43] H. Zhang, "Service disciplines for guaranteed performance service in packet-switching networks," *Proceedings of the IEEE*, vol. 83, no. 10, pp. 1374–1396, October 1995.

[44] L. Zhang, "Virtual clock: a new traffic control algorithm for packet switching networks," *ACM SIGCOMM Computer Communication Review*, vol. 20, no. 4, pp. 19–29, 1990.

[45] C. L. Liu and J. W. Layland, "Scheduling algorithms for multiprogramming in a hard-real-time environment," *Journal of the ACM*, vol. 20, no. 1, pp. 46–61, 1973.

[46] D. Kandlur, K. Shin, and D. Ferrari, "Real-time communication in multi-hop networks," *Proceedings of 11th International Conference on Distributed Computing Systems*, May 1991.

[47] A. Parekh and R. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: the single-node case," *IEEE/ACM Transactions on Networking*, vol. 1, no. 3, pp. 344–357, June 1993.

[48] D. Verma, H. Zhang, and D. Ferrari, "Guaranteeing delay jitter bounds in packet switching networks," *Proceedings of IEEE Tricomm Conference*, pp. 35–46, April 1991.

[49] S. J. Golestani, "A stop-and-go queueing framework for congestion management," *Proceedings of the ACM symposium on Communications architectures and protocols*, pp. 8–18, 1990.

[50] A. K. F. Khattab and K. M. F. Elsayed, "Channel-quality dependent earliest deadline due fair scheduling schemes for wireless multimedia networks," *Proceedings of the 7th ACM international symposium on Modeling, analysis and simulation of wireless and mobile systems*, pp. 31–38, 2004.

[51] P. Bhagwat, P. Bhattacharya, A. Krishma, and S. K. Tripathi, "Using channel state dependent packet scheduling to improve TCP throughput over wireless LANs," *Wireless Network*, vol. 3, no. 1, pp. 91–102, 1997.

[52] S. Lu, V. Bharghavan, and R. Srikant, "Fair scheduling in wireless packet networks," *IEEE/ACM Transactions on Networking*, vol. 7, no. 4, pp. 473–489, August 1999.

[53] N. Vaidya, A. Dugar, A. S. Gupta, and P. Bahl, "Distributed fair scheduling in a wireless LAN," *IEEE Transactions. on Mobile Computing*, vol. 4, no. 6, pp. 616–629, November 2005.

[54] H. Fern, C. Lee, J. Huang, and G. Chiu, "Designing a fair scheduling mechanism for IEEE 802.11 wireless LAN," *IEEE Communications Letters*, vol. 9, no. 4, pp. 301–303, April 2005.

[55] D. J. K. Vinay, N. Sreenivasulu and D. Das, "Performance evaluation of end-to-end delay by hybrid scheduling algorithm for QoS in IEEE 802.16 network," *Proceedings of IFIP International Conference on Wireless and Optical Communications Networks*, pp. 1–5, April 2006.

[56] T. K. H. Lee and D.-H. Cho, "An enhanced uplink scheduling algorithm based on voice activity for VOIP services in IEEE 802.16d/e system," *IEEE Communications Letters*, vol. 9, pp. 691–693, August 2005.

[57] V. Singh and V. Sharma, "Efficient and fair scheduling of uplink and downlink in IEEE 802.16 OFDMA networks," *Proceedings of IEEE Wireless Communications and Networking Conference*, pp. 984–990, April 2006.

[58] T. Issariyakul and E. Hossain, "Optimal radio channel allocation for fair queuing in wireless data networks," *Proceedings of IEEE International Conference on Communications*, pp. 11–15, May 2003.

[59] S. Borst and P. Whiting, "Dynamic channel-sensitive scheduling algorithms for wireless data throughput optimization," *IEEE Transactions on Vehicular Technology*, vol. 52, no. 3, pp. 569–586, May 2003.

[60] E. Amaldi, A. Capone, F. Malucelli, and G. Villa, "Optimization of packet scheduling in wireless systems with smart antennas: geometric models and

algorithms," *Proceedings of IEEE International Conference on Communications*, pp. 4238–4242, June 2004.

[61] K. K. R. Iyengar and B. Sikdar, "Scheduling algorithms for point-to-multipoint operation in IEEE 802.16 networks," *Proceedings of 4th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks*, pp. 1–7, April 2006.

[62] J. Cheng and S. Lu, "Achieving delay and throughput decoupling in distributed fair queueing over ad hoc networks," *Proceeding of the 12th International Conference on Computer Communications and Networks*, pp. 217–222, May 2003.

[63] A. K. Somani and J. Zhou, "Achieving fairness in distributed scheduling in wireless ad-hoc networks," *Proceedings of the IEEE International Performance, Computing, and Communications Conference*, pp. 95–102, April 2003.

[64] B. Tavli and W. B. Heinzelman, "MH-TRACE: multihop time reservation using adaptive control for energy efficiency," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 5, pp. 942–953, June 2004.

[65] R. P. Pope, R. A. Meyer, and S. J. Iuorno, "Exploring link activation scheduling with varying network organizations," *Proceedings of IEEE Military Communications Conference*, pp. 927–932, October 1993.

[66] K. Liu, J. Li, and W. Zhang, "User-dependent perfect-scheduling multiple access (UPMA) protocol in wireless multihop mobile ad hoc networks," *Proceedings of IEEE 55th Vehicular Technology Conference*, pp. 65–69, May 2002.

[67] J.-H. Ju and V. Li, "An optimal topology-transparent scheduling method in multihop packet radio networks," *IEEE/ACM Transactions on Networking*, vol. 6, no. 3, pp. 298–306, June 1998.

[68] K. Kang, W. Guo, and R. Liu, "A topology-independent TDMA transmission scheduling algorithm in wireless multihop networks," *Proceedings of International Conference on Communications, Circuits and Systems*, pp. 457–461, June 2004.

[69] C.-C. Tseng and K.-C. Chen, "QoS-guaranteed polling-based 2-layer integrated multihop scheduling algorithm for wireless ad hoc networks," *Proceedings of 59th IEEE Vehicular Technology Conference*, pp. 2144–2148, May 2004.

[70] J. Zhang, Y. P. Chen, and I. Marsic, "MAC scheduling using channel state diversity for high-throughput IEEE 802.11 mesh networks," *IEEE Communications Magazine*, vol. 45, no. 11, pp. 94–99, November 2007.

[71] S. Jain, S. R. Das, and H. Gupta, "Distributed protocols for scheduling and rate control to achieve max-min fairness in wireless mesh networks," *Proceedings of*

IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks, pp. 1–8, 18-21 June 2007.

[72] E. H. S. Motahari and V. Valaee, "Spatio-temporal schedulers in IEEE 802.16," *Proceedings of IEEE Global Telecommunications Conference*, pp. 984–990, Novermber 2005.

[73] H. Shetiya and V. Sharma, "Algorithms for routing and centralized scheduling to provide QoS in IEEE 802.16 mesh networks," *Proceedings of the 1st ACM workshop on Wireless multimedia networking and performance modeling*, pp. 140–149, 2005.

[74] G. Narlikar, G. Wilfong, and L. Zhang, "Designing multihop wireless backhaul networks with delay guarantees," *Proceedings of 25th IEEE International Conference on Computer Communications*, pp. 1–12, April 2006.

[75] B. Han, W. Jia, and L. Lin, "Performance evaluation of scheduling in IEEE 802.16 based wireless mesh networks," *Computer Communications*, vol. 30, pp. 780–788, 2007.

[76] A. Herms, S. Ivanov, and G. Lukas, "Precise admission control for bandwidth reservation in wireless mesh networks," *Proceedings of IEEE Internatonal Conference on Mobile Adhoc and Sensor Systems*, pp. 1–3, October 2007.

[77] Q. Shen, X. Fang, P. Li, and Y. Fang, "Admission control for providing QoS in wireless mesh networks," *Proceedings of IEEE International Conference on Communications*, pp. 2910–2914, May 2008.

[78] X. Wang, "Traffic scheduling with efficient channel assignment in WLAN mesh networks," *Master's thesis, McMaster University*, December 2007.

[79] A. Raniwala, K. Gopalan, and T. cker Chiueh, "Centralized channel assignment and routing algorithms for multi-channel wireless mesh networks," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 8, no. 2, pp. 50–65, 2004.

[80] H. Skalli, S. Ghosh, S. Das, L. Lenzini, and M. Conti, "Channel assignment strategies for multiradio wireless mesh networks: issues and solutions," *IEEE Communications Magazine*, vol. 45, no. 11, pp. 86–95, November 2007.

[81] L. A. Wolsey, *Integer Programming*. John Wiley Sons, 1988.

[82] A. H. Land and A. G. Doig, "An automatic method of solving discrete programming problems," *Econometrica*, vol. 28, no. 3, pp. 497–520, July 1960.

[83] M. Gast, *802.11 wireless networks the definitive guide*. O' Reilly & Associates, 2002.

[84] D. Zhao, J. Zou, and T. Todd, "Admission control with load balancing in IEEE 802.11-based ESS mesh networks," *ACM Wireless Networks*, vol. 13, no. 3, pp. 351–359, June 2007.

[85] T. Kocak, "Approximate analysis of a dynamic scheduler for self-similar video traffic in ATM networks," *IEE Proceedings-Communications*, vol. 153, no. 2, pp. 189–194, April 2006.

[86] P. Goyal, H. Vin, and H. Cheng, "Start-time fair queueing: a scheduling algorithm for integrated services packet switching networks," *IEEE/ACM Transactions on Networking,*, vol. 5, no. 5, pp. 690–704, October 1997.

[87] J.-F. Frigon, V. Leung, and H. Chan Bun Chan, "Dynamic reservation TDMA protocol for wireless ATM networks," *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 2, pp. 370–383, February 2001.

[88] M. Yuang and P. Tien, "Multiple access control with intelligent bandwidth allocation for wireless ATM networks," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 9, pp. 1658–1669, September 2000.

[89] P. Brady, "A model for generating on-off. speech patterns in two-way conversation," *System Technical Journal*, vol. 48, pp. 2445– 2472, April 1969.

[90] T. E. Stern, "A queueing analysis of packet voice," *Proceedings of IEEE Global Telecommunications Conference*, pp. 2.5.1–2.5.6, December 1983.

[91] R. Guerin, H. Ahmadi, and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 7, pp. 968–981, September 1991.

[92] M. Schwartz, *Broadband Integrated Networks.* Prentice Hall, 1996.

[93] D. Soldani and S. Dixit, "Wireless relays for broadband access," *IEEE Communications Magazine*, vol. 46, no. 3, pp. 58–66, March 2008.