UNIVERSITY OF CINCINNATI

Date: 1-Oct-2009

I, Deepti Nandiraju **,**

hereby submit this original work as part of the requirements for the degree of:

Doctor of Philosophy

in Computer Science & Engineering

It is entitled:

Efficient Traffic Diversion and Load-balancing in Multi-hop Wireless Mesh Networks

Student Signature: Deepti Nandiraju

This work and its defense approved by:

Committee Chair: Dharma Agrawal, DSc
Dharma Agrawal, DSc

Kenneth Berman, PhD
Kenneth Berman, PhD

Yiming Hu, PhD
Yiming Hu, PhD

Kelly Cohen, PhD
Kelly Cohen, PhD

Chia Han, PhD
Chia Han, PhD

Efficient Traffic Diversion and Load-balancing in Multi-hop Wireless Mesh Networks

A Dissertation submitted to the
Division of Research and Advanced Studies
of the University of Cincinnati

In partial fulfillment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in the Department of Computer Science
of the College of Engineering
September, 2009

By

Deepti V. S. Nandiraju

Master of Science (Computer Science)
Assam University,
Silchar, India, 2003

Thesis Adviser and Committee Chair: **Dr. Dharma P. Agrawal**

Abstract

Wireless Mesh Networks (WMNs) are one of the upcoming technologies which envision providing broadband internet access to users any where any time. WMNs comprise of Internet Gateways (IGWs) and Mesh Routers (MRs). They seamlessly extend the network connectivity to Mesh Clients (MCs) as end users by forming a wireless backbone that requires minimal infrastructure. For WMNs, frequent link quality fluctuations, excessive load on selective links, congestion, and limited capacity due to half-duplex nature of radios are some key limiting factors that hinder their deployment. Also, other problems such as unfair channel access, improper buffer management, and irrational routing choices are impeding the successful large scale deployment of mesh networks. Quality of Service (QoS) provisioning and scalability in terms of supporting large number of users with decent bandwidth are other important issues.

In this dissertation, we examine some of the aforementioned problems in WMNs and propose novel algorithms to solve them. We find that the proposed solutions enhance the network's performance significantly. In particular, we provide a traffic differentiation methodology, Dual Queue Service Differentiation (DQSD), which helps in fair throughput distribution of network traffic regardless of spatial location of its nodes. We next focus on managing the IGWs in WMNs since they are the potential bottleneck candidates due to huge volume of traffic that has to flow through them. To address this issue, we propose a load balancing protocol, LoaD BALancing (LDBAL), which efficiently distributes the traffic load among a given set of IGWs. We then delve into the aspects of load balancing and traffic distribution over multiple traffic paths in WMNs. To achieve this, we propose a novel Adaptive State-based Multipath Routing Protocol (ASMRP) that provides reliable and robust performance in WMNs. We also employ four-radio architecture for MRs, which allows them to communicate over multiple radios tuned to non-overlapping channels and better utilize the available spectrum. We show that our protocol achieves significant throughput improvement and helps in distributing the traffic load for efficient resource utilization. Through extensive simulations, we observe that ASMRP substantially improves the achieved throughput (~5 times gain in comparison to AODV), and significantly minimizes end-to-end latencies. We also show that ASMRP ensures fairness in the network under varying traffic load conditions.

We then focus on prudent user admission strategy for IGWs and other Wireless Service Providers (WSPs). WSPs typically serve diverse user base with heterogeneous requirements and charge users accordingly. In scenarios where a WSP is constrained in resources and have a pre-defined objective such as revenue maximization or prioritized fairness, a prudent user selection strategy is needed to optimize it. In this dissertation, we present an optimal user admission / allocation policy for WSPs based on yield management principles and discrete-time Markov Decision Process model to maximize its potential revenue. We finally conclude with a summary of our results and some pointers for future research directions.

Acknowledgement

I am very fortunate and thankful to have Prof. Dharma Agrawal as my advisor who has been extremely helpful and understanding. Dr. Agrawal has been an excellent advisor, advocate and inspiration and provided me fantastic support and conversation on both research and real life. Dr. Agrawal's guidance and direction towards this dissertation has been impeccable from all perspectives. Dr. Agrawal provided me with the necessary freedom to carry out my research, and encouraged, coached, and facilitated me in publishing various journal and conference papers.

I also express my sincere thanks to Dr. Kenneth Berman, Dr. Chia-Yung Han, Dr. Yiming Hu, and Dr. Kelly Cohen for taking the time to serve on my dissertation committee and offering valuable suggestions to enhance the quality of this dissertation.

I am grateful to my mother Mrs. N. Ananta Lakshmi who has been my key motivator to pursue Ph.D., and my father Prof. N.V.Satyanarayana Rao for his invaluable guidance and constant encouragement which elevated my performance bar. I am thankful to – Mrs. V. Mythili Shyam & Prof. V. Syama Sundar (my in-laws), Dr. Deepika, Dr. Madhavi, Mallika and Abhinay for their constant support and encouragement. My special thanks to Mrs. Purnima Agrawal, Dr. RangaSai and his family members for their inspiration, support and encouragement during my stay at Cincinnati.

Well, there is no boundary on how much I can write on how fortunate I am - to be a sister who was able to discuss, brainstorm, constructively argue and pursue parallel research and publish several co-authored papers with my brother, Dr. Nagesh Nandiraju. My body just trembles with thrill when I recollect those days and late nights of working together and struggling to generate solutions, and jumped together in our hearts when we found some for complex problems. I just want to say heartfelt thanks to him.

I am thankful to all my fellow CDMC lab mates who were very friendly, supportive and encouraging at all times. In particular, I have enjoyed the companionship of Lakshmi and Dave with whom I used to spend long hours of brainstorming discussions. I am thankful to Prof. K. Hemachandran for his sincere and constant support.

During the last and most crucial phase of my graduate career, I have been gifted with the love and companionship of my husband Vamsee Krishna Venuturumilli. He has put up endless discussions of my work with steady perseverance and I couldn’t have completed this work without his unstinting support and cooperation. I would like to express my heartfelt gratitude to him.

To my lovely new-born…

Ved Sameeraj

~*~*~*~*~*~*

Contents

LIST OF FIGURES iv

LIST OF TABLES vi

CHAPTER 1. INTRODUCTION 1

1.1 TRADITIONAL WIRELESS LOCAL AREA NETWORKS (WLANS) 2

1.2 WIRELESS MESH NETWORKS 6

1.3 MOTIVATION 8

1.3.1 Unfairness in Multi-hop Wireless Mesh Networks 8

1.3.2 Hot-zones at IGWs 10

1.3.3 Hot Paths and Route Flaps 10

1.3.4 Single Interface Scenario 13

1.3.5 Route Stability and Robustness 13

1.3.6 Source Routing Strategy 14

1.3.7 Optimization of Wireless Service Provider's (WSP) Utility 15

1.4 ORGANIZATION OF THE DISSERTATION 17

1.5 SUMMARY OF CONTRIBUTIONS 18

CHAPTER 2. SERVICE DIFFERENTIATION IN MESH NETWORKS: A DUAL QUEUE STRATEGY 20

2.1 INTRODUCTION 20

2.2 ILLUSTRATION OF UNFAIRNESS PROBLEM IN MULTI-HOP WMNS 21

2.3 DESIGN GOALS 25

2.4 DUAL QUEUE SERVICE DIFFERENTIATION (DQSD) 27

2.4.1 Data Structures......28

2.4.2 DQSD Algorithm......29

2.5 PERFORMANCE ANALYSIS......30

2.5.1 Aggregate Throughput......31

2.5.2 Delay Distribution......32

2.6 RELATED WORK......33

2.7 SUMMARY 34

CHAPTER 3. ACHIEVING LOAD BALANCING IN WIRELESS MESH NETWORKS THROUGH MULTIPLE GATEWAYS 36

3.1 INTRODUCTION......36

3.2 CONGESTION AWARE LOAD BALANCING......37

3.2.1 Gateway Discovery Protocol......37

3.2.2 Load Migration Procedure......38

3.3 PERFORMANCE ANALYSIS......41

3.4 RELATED WORK......43

3.5 SUMMARY 44

CHAPTER 4. MULTI-RADIO MULTI-PATH ROUTING IN WIRELESS MESH NETWORKS ..46

4.1 INTRODUCTION......46

4.2 MULTI-PATH ROUTING IN WIRELESS MESH NETWORKS......47

4.2.1 Network Model......47

4.2.2 Network Initiation......48

4.2.3 Congestion-aware Routing......53

4.3 NEIGHBOR STATE MAINTENANCE MODULE......54

4.4 MULTI-RADIO ARCHITECTURE......55

4.5 PERFORMANCE EVALUATION......58

4.5.1 Multi-rate Capability......61

4.5.2 Throughput Comparison......62

4.5.3 Fairness Comparison *64*

4.5.4 Delay Distribution *65*

4.5.5 Traffic Partitioning Strategies *68*

4.6 RELATED WORK 69

4.7 SUMMARY 72

CHAPTER 5. DYNAMIC ADMISSION POLICY FOR WIRELESS SERVICE PROVIDERS USING DISCRETE-TIME MARKOV DECISION PROCESS MODEL 74

5.1 INTRODUCTION 74

5.2 RELATED WORK 77

5.3 CHARACTERISTICS OF YIELD MANAGEMENT AND PARALLELISM TO PROPOSED MODEL 81

5.4 PROBLEM FORMULATION USING MARKOV DECISION PROCESS MODEL 82

5.4.1 Constant Service Charge for a Given Class over Allocating Time Horizon *86*

5.4.2 Varying Service Charge for a Given Class over Allocating Time Horizon *89*

5.5 ILLUSTRATION OF DECISION POLICY COMPUTATION THROUGH NUMERICAL EXAMPLES 91

5.5.1 Constant Service Charge over Allocating Time Horizon *92*

5.5.2 Varying Service Charge over Allocating Time Horizon *95*

5.6 PERFORMANCE ANALYSIS 98

5.6.1 Comparison with Greedy Allocation Strategy *98*

5.6.2 Expected Revenue using MDP with Varying Resources *102*

5.6.3 Cumulative Revenue using MDP over Varying Durations of Allocation Time Horizon *103*

5.7 SUMMARY 104

CHAPTER 6. CONCLUSIONS AND FUTURE RESEARCH 105

6.1 FUTURE WORK 107

BIBLIOGRAPHY 108

List of Figures

Figure No.	Name	Page No.
1.1	An example WLAN	3
1.2	An example ad hoc network	3
1.3	Hierarchical architecture of Wireless Mesh Networks	7
1.4	Spatial bias – unfair queue management	9
1.5	Illustration of congested high throughput link	11
2.1(a)	MRs connected in a linear scenario	23
2.1(b)	Aggregate throughput of flows from each MR	23
2.1(c)	CDFs of flows from each MR	23
2.2(a)	Aggregate throughput of flows	31
2.2(b)	CDF of the delay distribution	31
3.1	Illustrating load balancing in a WMN through gateway switching	39
3.2	Timing diagram depicting the sequence of actions while switching gateways	39
3.3	Instantaneous throughput obtained by the flows using the default scheme	42
3.4	Instantaneous throughput obtained by the flows using the proposed scheme	42
3.5	Packet delivery ratio for the flows f1, f2, and f3	43
3.6	Average delay for the flows f1, f2, and f3	43
4.1	Illustration of the proposed algorithm	49
4.2	Illustration of the route discovery, child and parent notification procedures	51
4.3	State machine of a neighbor	55
4.4(a)	Aggregate throughput multi-rate links vs. constant data rate links	61
4.4(b)	Delay distribution multi-rate links vs. constant data rate links	61

4.5(a) Aggregate throughput of flows from different MRs with varying traffic load 63

4.5(b) Fairness index for different MRs with varying traffic load 63

4.6(a) CDF of packet delays with varying traffic rate: Offered load of 400 Kbps 66

4.6(b) CDF of packet delays with varying traffic rate: Offered load of 500 Kbps 66

4.6(c) CDF of packet delays with varying traffic rate: Offered load of 1000 Kbps 66

4.7(a) CDF of packet delays with varying traffic load with the presence of some failed MRs: Offered load of 400 Kbps 67

4.7(b) CDF of packet delays with varying traffic load with the presence of some failed MRs: Offered load of 500 Kbps 67

4.7(c) CDF of packet delays with varying traffic load with the presence of some failed MRs: Offered load of 1000 Kbps 67

4.8 Aggregate throughput for a linear flow 68

4.9 Illustration of aggregate throughput: improvement with congestion-aware algorithm 69

5.1 Allocating and usable time horizons 83

5.2 Logic diagram of the proposed model 89

5.3 Varying service charge 90

5.4 Expected Revenue Comparison for MDP and Greedy policy with Constant Charge and Arrival Pattern 99

5.5 Expected revenue comparison for MDP and greedy policy 101

5.6 Revenue comparison in each simulation instance 102

5.7 Expected revenue comparison for MDP with varying resources 103

5.8 Cumulative revenue for varying durations of allocating time horizon 104

List of Tables

Table No.	Name	Page No.
4.1	Describing the purpose of different states in the proposed state machine	55
4.2	Conditions in State Machine	55
4.3	Simulation Parameters	59
5.1	Example Parameters for Constant Service Charge Scenario	92
5.2	Request Arrival Probabilities for the Service Classes	93
5.3	Computed Expected Revenue for Constant Service Charge Scenario	93
5.4	Decision Policy Logic	94
5.5	Decision Policy Computed at WSP for Constant Service Charge Scenario	95
5.6	Varying Service Charges for Different Service Classes	96
5.7	Computed Expected Revenue for Varying Service Charge Scenario	96
5.8	Decision Policy Computed at WSP for Varying Service Charge Scenario	97
5.9	Parameters Used in Simulation for the Constant Service Charges and Arrival Pattern Scenario	99
5.10	Service Charges and Arrival Probabilities for Varying Service Charge Scenario	100

Chapter 1. Introduction

Wireless networking technology has been growing tremendously in recent years [1][2] due to the growing demand for ubiquitous broadband Internet connectivity and a widespread use of applications such as multimedia streaming (VoIP services, video streaming etc.). Wireless Mesh Networks (WMNs) have drawn considerable attention due to their potential to supplement the wired backbone with a wireless support in a cost-effective manner. Some key advantages of WMNs include their self-organizing ability, self-healing capability, low-cost infrastructure, rapid deployment, scalability, and ease of installation. WMNs are capable of providing attractive services in a wide range of application scenarios such as broadband home/enterprise/community networking, disaster management, and public safety applications.

The mesh-networking technology attracted both academia and industry stirring efforts for their real-world deployment in a variety of applications. MIT deployed WMN in one of its laboratories for studying the industrial control and sensing aspects. Several companies like Nortel Networks, Strix Systems, Tropos Networks, MeshDynamics are offering mesh networking solutions for applications such as building automation, small and large scale internet connectivity, etc., using customary products. Strix systems has deployed a city-wide Wi-Fi mesh network in Belgium spanning an area of 17.41 KM^2 to provide wireless Internet access to its residents, tourists, businesses, and municipal and public-safety applications and advertising systems around the city. Strix also deployed a wireless tracking system called *project kidwatch* that traces the real-time location of a child in a beach area or around a city.

Further commercial interests in WMNs have prompted immediate and increasing attention for integrating WMNs with the Internet. IEEE has setup a task group 802.11s for specifying the PHY and MAC standards for WMNs. The current draft of the 802.11s standard targets defining an Extended Service Set (ESS) that provides reliable connectivity, seamless security, and assure interoperability of devices. It also proposes the use of layer-2 routing, frame forwarding and increased security in data transmission. Industry giants such as Motorola Inc., Intel, Nokia, Firetide, etc., are actively participating in these standardization efforts. Two main proposals, one each from consortiums SEEMesh and WiMesh Alliance, have been considered and successfully merged into a single draft version of the IEEE 802.11s standard in July 2007. The task group is refining the specifications and aiming to finalize the standards by the end of year 2009.

In this chapter, we first provide a brief overview of the conventional wireless networking paradigms in Section 1.1. In Section 1.2, we introduce one of the upcoming wireless technologies, Wireless Mesh Networks (WMNs) [2], which is an amalgamation of the existing network architectures. We then outline the motivating factors for our research in Section 1.3, highlighting some key issues that are impeding the wide scale deployment of WMNs. In Section 1.4, we explain how this dissertation is organized and finally, in Section 1.5, we summarize the main contributions of our work.

1.1 Traditional Wireless Local Area Networks (WLANs)

Traditional Wireless Local Area Networks (WLANs) are broadly characterized into two types [3][4]:

1. Infrastructure WLANs, and
2. Ad hoc WLANs, also called as Mobile Ad hoc Networks (MANETs)

This classification is based on whether or not there is a central controller providing Internet connectivity. Infrastructure WLANs, shown in Figure 1.1, are structured networks consisting of Access Points (APs) and the client-stations, or the subscriber units. APs are typically installed at fixed locations and are connected to a wired network, also known as Distribution System (DS), and relay data between wireless and wired devices. The clients that could be either stationary or mobile, communicate with each other through APs. These client nodes are connected to the APs through wireless links. In other words, all the information exchange among the clients in the network occurs via an AP and the AP is also responsible for providing Internet connectivity to the clients registered with it. Multiple APs can be interconnected to form a large network which allows the clients registered with them to switch between the APs.

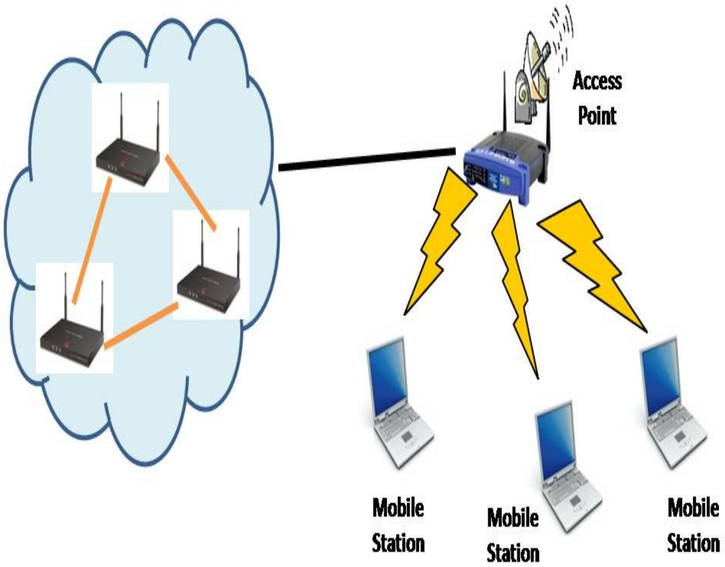


Figure 1.1

An Example Infrastructure WLAN

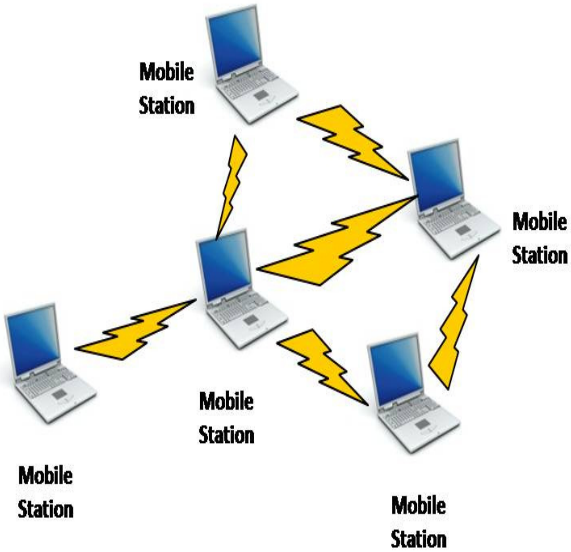


Figure 1.2

An Example Ad hoc Network

The other WLAN architecture, MANET, shown in Figure 1.2, is characterized by the absence of any infrastructure in terms of AP, and the client devices communicate directly with other close by devices and relay each other's traffic. MANETs are easier to install and to configure due to the absence of any needed infrastructure, but have limited connectivity options for other devices and weak security mechanism.

The IEEE 802.11 family of protocols standardizes WLAN technology and includes the three well known standards: 802.11a, 802.11b, and 802.11g. These standards operate in unlicensed Industrial Scientific Medical (ISM) bands. Specifically, IEEE 802.11a operates at a frequency of 5.8 GHz, while 802.11b and 802.11g operate at 2.4 GHz. The maximum data rate supported by 802.11a and 802.11g is 54 Mbps and the maximum data rate supported by 802.11b is 11 Mbps. However, in case of any losses or errors on the data links, 802.11b reduces the data rate to 5.5 Mbps or to 2 Mbps or to 1 Mbps depending on the loss rate of the links. This method, called automatic fallback, is used in order to operate over extended range of communication and in areas with high levels of interference. Also, Wi-Fi alliance has been created to enable compatibility and interoperability between products produced by different vendors in the industry.

These WLAN standards do not provide a significant improvement in achievable bandwidth for applications that span long distances such as mining industry. For instance, with 802.11b, the data rate of the wireless links drops off as the distance or the number of hops increases. The 802.11g standard intends to provide higher bandwidth in a confined space such as inside a building, so that it can be used as a replacement for wired networks. 802.11b and 802.11g both operating in the same frequency band and using identical signal propagation. 802.11g aims to achieve performance improvement by using an encoding scheme Orthogonal Frequency Division Multiplexing (OFDM) that incorporates detailed information into the signal. A receiver requires higher power to decode the signal encoded using OFDM. When the signal is transmitted over large distances, Signal to Noise Ratio (SNR) parameter measured at the receiver decreases. As a result, signals encoded using higher modulation techniques cannot be decoded at the receiver. Further, with increasing error rates in the medium, the radio employing 802.11g reverts back to 802.11b encoding scheme and its data rates. Also, with ever increasing wireless devices in the market operating in the same frequency band, interference from other sources cannot be avoided. Thus, the theoretical data rates specified in the standard are not achievable in a practical scenario.

A big leap in terms of achieved throughput of about 600Mbps and range greater than that provided by 802.11g is promised by the emerging standard called 802.11n [5][6]. This standard offers improvement in many aspects such as throughput, range, channel reliability, and transmission efficiency. It can operate in either 2.4GHz or 5GHz frequency bands and use Multiple Input Multiple Output (MIMO) antennas for data transfer. A single transmission stream can be split into multiple (4 in 802.11n) sub-streams and sent over the available antennae. Further, certain improvement at the physical layer, along with an increased channel band achieves an escalation of throughput for 802.11n.

Typically, increasing the number of nodes or the node density in WLANs can enhance the network coverage, connectivity options and consequently improve the reliability and robustness of the network. However, the disadvantage is that it may dramatically reduce the throughput and capacity of the network. As wireless communication is mostly broadcast in nature, a single channel is shared by all the nodes and transmission between a pair of nodes prevents several other potential transmissions within the communication range. It could potentially lead to increased number of collisions in the network and thus significantly limit the throughput and the capacity of the network. End users can experience unacceptable delays, and hence these networks are not yet suitable for large scale commercial deployment.

1.2 Wireless Mesh Networks

The architecture of Wireless Mesh Networks (WMNs) is derived largely as a combination of Infrastructure WLANs and MANETs described in the previous section. WMNs encompass Internet Gateways (IGWS), Mesh Routers (MRs) and Mesh Clients (MCs) and can be organized into a three-tier hierarchical architecture, as shown in Figure 1.3.

The first (or the top) tier includes a subset of MRs, called Internet Gateways (IGWs), which are connected to the wired network and these IGWs act as a bridge between the wireless mesh backbone and the wired network. IGWs also have an interface solely to communicate with the wired network. The second (or the middle) tier consists of relatively large number of wireless MRs which communicate with IGWs and with each other using a multi-hop communication paradigm, thus forming a multi-hop wireless mesh backbone network. The MRs organize autonomously and are self-healing, facilitating the addition and deletion of resources in the network on a dynamic basis. This backbone network of MRs is responsible for providing services to the MCs by transporting traffic either to/from IGWs by cooperatively relaying each others' traffic and facilitating interconnectivity. With their bridging property, MRs also enable integration of WMNs with other existing wireless networks such as cellular, Wi-Fi (Wireless Fidelity), and WiMAX (Worldwide Interoperability Microwave Access).

The third (or the bottom) tier includes the end users or the MCs, which use the network to access the Internet and other services such as Internet Protocol (IP) telephony, etc. In WMNs, MRs are mostly static and MCs are typically mobile and get registered with different MRs at different points of time. It should be noted that MRs and IGWs are similar in design, with the only one exception that an IGW is directly connected to a wired network, while MR is not. The links in a WMN can be either wired/wireless. In a WMN, only a subset of APs needs to be connected to the wired network in contrast to a traditional Wi-Fi network where each AP has to be connected to the wired network.

WMNs require minimal planning, marginal infrastructure support and are easily scalable. Specifically, WMNs can be deployed in places where either infrastructure is unavailable or where it is difficult to plant the APs. Also, WMNs can be deployed with few IGWs and numerous wireless MRs requiring low infrastructures for setting them up. WMNs provide a cost-effective alternative to other types of networks, requiring meticulous planning and indulge in huge expenses. Further, these networks are scalable, meaning they can be extended to thousands of MRs by just deploying new MRs which self-configure themselves in a dynamic manner. Large number of MRs in the mesh backbone of a WMN provides high connectivity, facilitating availability of multiple routes between any two users/end nodes. This feature can be used to increase reliability of the data transmission, allowing adequate fault tolerance.

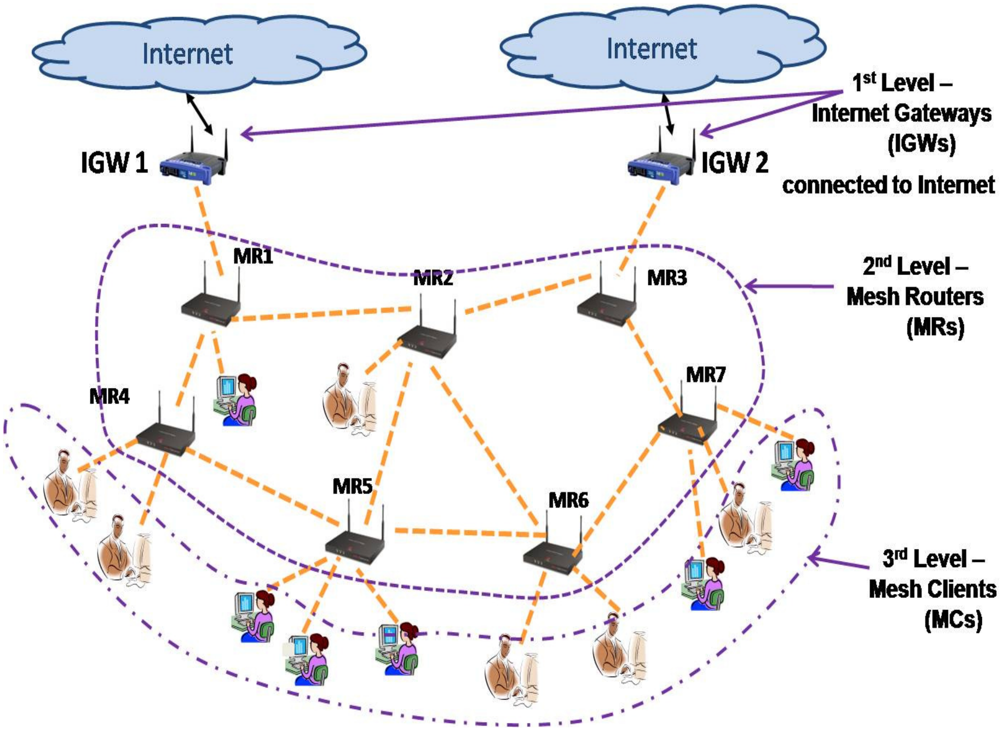

Internet
Internet
1st Level – Internet Gateways (IGWs) connected to Internet
IGW 1
IGW 2
MR1
MR2
MR3
2nd Level – Mesh Routers (MRs)
MR4
MR7
MR5
MR6
3rd Level – Mesh Clients (MCs)

Figure 1.3
Hierarchical Architecture of Wireless Mesh Networks

1.3 Motivation

WMNs are capable of providing attractive services in a wide range of application scenarios such as broadband home/enterprise/community networking and disaster management. However, unpredictable interference, excessive congestion, and half-duplex nature of radios may hinder their deployment.

WMNs are proven to provide ubiquitous broadband Internet access to support a large number of users at low costs. Though feasible, their performance is still considered to be far below the anticipated limits for practical applications. And so, unfortunately the companies involved in WMN deployments often face challenges in designing, deploying and ensuring their optimal performance due to underlying inherent problems of multi-hop networks. The multi-hop wireless communication is beset with several problems such as unpredictable/high interference, increased collisions due to hidden/exposed terminals [2][7], excessive congestion and its typical half-duplex nature of radios [8]. This results in poor performance of WMNs with low end-to-end throughput and high latencies, which are undesirable in the perceived applications of WMNs. Though envisioned applications of WMNs seem luring, considerable research is still needed in designing protocols used for WMNs before wide scale deployment of WMNs becomes practical.

In the following sections, we explain the issues that motivated us towards designing our proposed solutions.

1.3.1 Unfairness in Multi-hop Wireless Mesh Networks

In a multi-hop WMN, packets originated from MRs with larger number of hops experience poor performance compared to those from MRs with fewer hops (*spatial bias*). The link layer buffer/queue management scheme at the intermediate MRs plays a major role in causing spatial bias apart from other contributing factors such as hidden and exposed terminal problems [9][10].

Most of the existing queuing mechanisms do not consider the parameter - *number of hops a packet has traversed* - in their queuing logic and drop packets when there is no space in its Interface Queue (IFQ), independent of the number of hops they have already traversed. An IFQ is a queue maintained at a node to keep track of packets that are later transmitted over the medium one at a time. The packets in the queue comprise of those generated at the node as well as those arriving from other nodes in the network which need to be forwarded by this node.

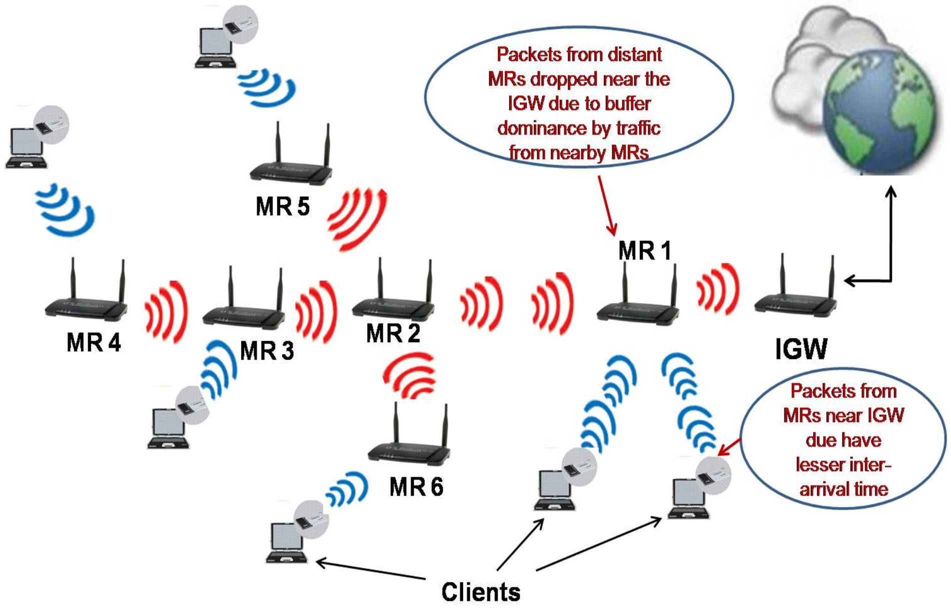

Packets from distant MRs dropped near the IGW due to buffer dominance by traffic from nearby MRs
MR 5
MR 1
MR 4
MR 3
MR 2
IGW
Packets from MRs near IGW due have lesser inter-arrival time
MR 6
Clients

Figure 1.4

Spatial Bias - Unfair Queue Management

The problem of spatial bias, shown in Figure 1.4, affects the network's performance in two ways. Firstly, it results in wastage of valuable network resources, and secondly, clients of a MR far away from IGW will get very low throughput and undergo starvation as compared to the clients connected to a MR that is near to an IGW. Thus, this motivates us to propose a service differentiation strategy for traffic that provides service guarantees to all users in the network irrespective of their spatial location.

1.3.2 Hot-zones at IGWs

In a WMN, the estimated traffic volume is anticipated to be very high which makes scalability and load balancing as important issues among others. WMNs are aimed to provide high bandwidth broadband connections to a large community and thus should be able to accommodate a large number of users with different application requirements for accessing the Internet. Usually, most of the traffic in WMNs is oriented towards the Internet, which may increase the traffic load on certain paths (leading towards the IGW). As the IGWs are responsible for forwarding all the network traffic, they are likely to become potential bottlenecks in WMNs resulting in *hot-zones* around IGWs. The high concentration of traffic at a gateway leads to saturation which in turn can result in packet drops due to potential buffer overflows. Dropping packets at the IGWs is highly undesirable and inefficient, especially after having consumed a lot of network resources en route from source to the IGW. Thus, to avert the danger of congestion, it is prudent to balance the traffic load over different IGWs and also possibly along the routes followed by the packets enroute to the IGW. This motivates us to devise a scheme which would enable sharing of the load among multiple gateways and improve the overall performance of the network.

1.3.3 Hot Paths and Route Flaps

Consider the IEEE 802.11a wireless network shown in Figure 1.5, and let the label on each link denotes the data rate supported by it. Let the individual optimal paths for MR6, MR7 and MR8 be {MR6-MR4-MR2-IGW}, {MR7-MR5-MR2-IGW}, and {MR8-MR5-MR2-IGW} respectively. It can be observed that all these individual optimal paths contain a common route segment {MR2-IGW}. Now, if MR6, MR7 and MR8 simultaneously send traffic through their optimal paths, then all this traffic will be directed through the segment {MR2-IGW}. If the required cumulative bandwidth exceeds the capacity of the path segment {MR2-IGW}, then needed demand over its supported capacity leads to congestion. Thus, {MR2-IGW} will eventually become the bottleneck segment, resulting in potential packet losses. Such segment is referred to as a *hot path*.

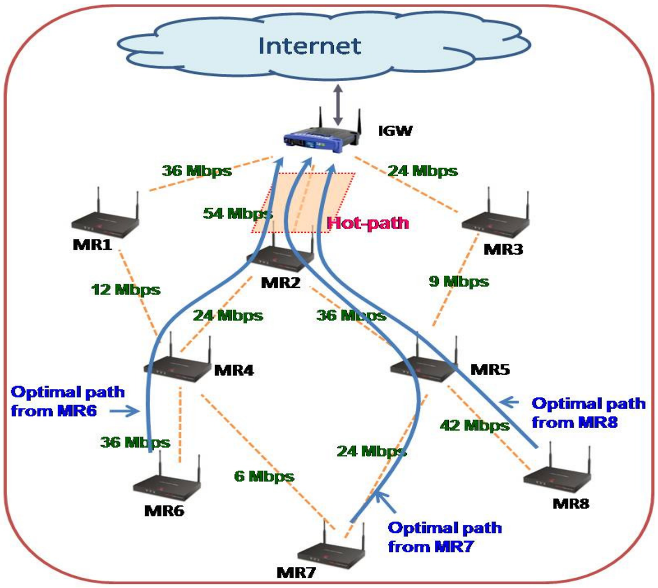


Figure 1.5

Illustration of Congested High Throughput Link

Whenever such a hot-path is formed, it could trigger MR6, MR7 and MR8 to look for an alternate route. If {MR2-IGW} is avoided, these MRs could simultaneously choose alternate paths, which could yet lead to another such common route segment, that will result in a hot-path scenario again, and such cycle results in oscillations if repeated. Thus, frequent route changes or flaps from one path to another leads to increased packet loss and delays due to route rediscovery. An efficient routing protocol should consider hot-path formation scenario, and limit their occurrence and resulting oscillations. One solution could be through the use of multiple near-optimal paths and distribute the traffic among them, instead of always using the best path, and thus balance the load over the network.

For several reasons, traditional routing solutions of MANETs are not directly useful for WMNs. Most of them are usually designed around single-path routing which can result in an unbalanced network load, with some links being highly utilized while others seldom used. Also, in single path routing, if a link in the chosen path fails, applications will be interrupted and rediscovering an alternate path results in delays. To increase the reliability, extensions to single-path routing protocols have been designed which typically use backup paths to route the traffic, in case primary path fails [11][12][13][14][15] . However, even these models mostly result in higher latencies due to path switching.

Further, traffic in WMNs is predominantly between IGWs and the MRs, in contrast to MANETs, where traffic is among peer nodes. This focused traffic flow of WMNs towards and from IGW places higher demand on certain paths, connecting IGWs and MRs, unlike that of MANETs where the traffic is more or less uniformly distributed. The advantage with WMNs is the high connectivity of the mesh backbone, which facilitates availability of multiple routes between any two end users.

Existing multi-path routing protocols advocate the use of disjoint paths and do not consider the delays (such as queuing delay) and congestion experienced over the links, once the paths are readily selected. Authors in [16] reveal that the multiple paths need not be disjoint and in fact, use of disjoint paths is counter-productive. Use of multiple paths offer a window of error resilience and traffic load distribution as the spatial diversity and data redundancy can be exploited. We extend MMESH [17] to increase reliability of data transmission, allowing adequate fault tolerance.

The distinguishing feature of our proposed protocol is to maintain multiple near optimal routes, not necessarily disjoint, with the unique property of opportunistically selecting them according to their congestion levels and quality of the links. Information is distributed among various routes to maximize the probability of information propagation.

1.3.4 Single Interface Scenario

MMESH presents a multipath routing protocol for WMNs where each of the MRs is equipped with a single radio. However, communication using a single radio could result in overall end-to-end transmission delays. For instance, in Figure 1.5, suppose that MR2 is equipped with a single radio, and that it has to receive data from MR4 and transmit the same to IGW. Then, the half-duplex nature of the radio does not permit MR2 to transmit data simultaneously to IGW while it receives data from MR4. Since the relaying load in a WMN is particularly higher on some MRs, such half duplex communication results in very high end-to-end latencies. If MR2 is equipped with multiple radios and each of these operate in non-interfering channels, then simultaneous transmission and reception can be accomplished with IGW and MR4, respectively. This improves overall end-to-end delays and minimizes collisions.

To overcome the half-duplex limitation, in our proposed multi-radio routing protocol, we extend MMESH and employ a multi-radio architecture in which all the MRs are equipped with more than one interface. Further, these radios are tuned onto non-overlapping channels to avoid interference caused at the MR.

1.3.5 Route Stability and Robustness

Though MRs in a WMN are relatively stationary, links between adjacent MRs could be unstable, typically due to variations[1.1] in the wireless link quality. Also, since the WMNs operate in an open ISM frequency band of 2.4/5 GHz range, there could be interference from external devices which is unpredictable. Link quality fluctuations, which are frequent, often result in

[1.1] Possible reasons are multi-path fading effects, weather conditions, and external interference route fluctuations in WMNs. Sometimes, these fluctuations may be temporary and the link quality could become better in few seconds. However, single path routing algorithms typically search for an alternate route as soon as they sense a bad link in the existing route. Temporary link quality fluctuations cause unnecessary overhead, trigger MRs to flap between routes, disrupt ongoing communication, and introduce instability to the network [18]. Maintaining multiple routes reduces the dependency on any single link or route and offers much needed flexibility for recovery.

Further, temporary link failures result in a subset of routes where a link could become stale, and choosing such routes for transmission leads to packet loss. Our proposed routing protocol improves the robustness and stability of a WMN by employing a Neighbor State Maintenance module that monitors the state of neighbors and the quality of the link connecting each neighbor and ensures validity of the route. This approach aids in preventing frequent oscillations, provides robustness to any link failure, and improves the network stability.

1.3.6 Source Routing Strategy

In source routing algorithms such as MR-LQSR [19], the entire route from source to its destination is appended to the packet payload. However, this procedure poses significant challenge for scalability of WMNs in terms of high message overhead. For instance, currently the IPv6 address size for a single MR is 16 bytes, and if a packet has to be transported using source routing technique and uses 10 hops to reach destination, then the overhead for this scenario would be 160 bytes. As more and more MRs are added to WMN, appending the route in every packet considerably increases the overhead of the network. To overcome this issue, one strategy could be to store routes and additional state information at intermediate MRs themselves. Owing to the recent advancements of digital technology, memory consumption at these intermediate MRs is not a concern these days, which decreases the cost of an on-chip memory. This also aids in maintaining the scalability of the protocol if the size of WMN increases.

In our proposed routing protocol, instead of sending the whole list of routes, the MRs maintain additional state information, by assigning labels to the routes and using these set of labels as periodic advertisements.

1.3.7 Optimization of Wireless Service Provider's (WSP) Utility

In WMNs, an IGW provides services to its registered users by forwarding their traffic to and from Internet. These services could be offered through service plans from which the users can choose a plan that suits their needs. When a user chooses a service plan and requests the respective services, an IGW can either accept or deny servicing those requests. Typically, IGW decides whether or not to accept arriving user requests depending upon its pre-defined utility optimization function. This function could be maximization of revenue, minimization of user migration or optimization of prioritized fairness. For instance, if the IGW charges the users for its offered services, then its optimization goal would be to maximize revenue accrued from its admitted users over a given period of time. Similar admission selection strategies are needed for any Wireless Service Provider (WSP) having parallel goals.

These days, users require wireless services for a variety of applications such as web-browsing, VoIP, webinars, streaming videos, IPTV, coordinated multi-player networked games, interactive voice and video, etc. Most often, the same type of resources (for example, bandwidth) are utilized by WSPs to serve these spectrum of applications. Typically, WSPs are constrained with limited availability of resources to support such a wide variety of applications. To serve these heterogeneous demands, WSPs offer portfolio of services targeting specific application requirements. The offered services of WSPs, which we call as service classes from here on, usually differ in terms of either its application type and/or Quality of Service (QoS) level. For instance, to explain QoS differentiated service plans, a broadband Internet based WSP may offer service plans to its residential and business users to choose from, which may differ in uplink / downlink data rates (like 28 Kbps, 54 Kbps, 100 Kbps connections for internet), resource usage limitations (like limited minutes vs. unlimited minutes phone service) or other such QoS aspects. Moreover, WSPs may also offer bundled packages of two or more applications together, (e.g. Internet and VoIP bundled offering). Similarly, in cellular networks, the service level or QoS differentiation could be in terms of call admission probability, i.e., calls belonging to a higher service class have higher call admission probability as compared to those of lower service classes.

Typically, each of the above mentioned service classes consume different amount of resources at a WSP. It is widely acceptable for WSPs to charge its users different prices, which we call service charges corresponding to these offered augmented service classes. WSPs set prices for these service classes based on their average resource consumption, service requirements, value-based pricing for a given application or corresponding market pricing.

The total revenue that will be earned by WSP depends on the mix of its subscribed user base as this mix dictates the obtained service charges gained from each of them. From WSP's standpoint, it is imperative to manage its limited resources and maximize its revenue through an optimal and prudent selection of its admitted user base. We use discrete-time Markov Decision Process model to formulate and optimize the admission / allocation policy.

Though we choose WSP's revenue as optimization parameter in this dissertation, other utility factors such as prioritized fairness, QoS can also be considered for optimization in a similar manner for respective applications.

1.4 Organization of the Dissertation

The remaining dissertation is organized as follows. In Chapter 2, we demonstrate the unfairness problem posed in multi-hop WMNs through simulations. We propose a dual queue strategy that provides service guarantees to all users in the network irrespective of their spatial location. The algorithm is designed to elegantly segregate and exclusively reserve queues for either of the traffic. We implement this module above the standard IEEE 802.11 MAC layer thus obviating any modifications to the legacy MAC. We perform simulations to study the effect of our proposed scheme on the performance of multi-hop flows.

In Chapter 3, we focus our research on routing layer and its performance with respect to load balancing. As the WMNs are envisioned to provide high bandwidth broadband service to a large community of users, the Internet Gateway (IGW) which acts as a central point of internet attachment for the MRs, it is likely to be a potential bottleneck because of its limited wireless link capacity and due to high traffic transfer demand from MRs. We propose a novel technique that elegantly balances the load among the different IGWs in a WMN. We then evaluate our proposed scheme to observe its efficiency in traffic load balancing.

As we have described in Section 1.3, most existing routing protocols are suboptimal and do not aptly exploit newer design choices and resources available in WMNs. Clearly, such protocols have not been designed with the focus on using the multi-rate and multi-channel capable multi-interface designs. In Chapter 4, we present a comprehensive multi-path routing discovery and maintenance protocol for multi-radio multi-channel WMNs. Our proposed protocol exploits multiple paths to synergistically improve the overall performance of the network. We analyze the performance of the protocol towards throughput, fairness and delay under various factors. We also investigate the effectiveness of various traffic splitting algorithms used for balancing the traffic load over multiple routes.

To maximize the obtainable revenue at WSPs with limited resources, a prudent user admission / selection policy is needed. In Chapter 5, we formulate a user request admission / allocation policy for WSPs such that their potential revenue is maximized. The proposed model is based on discrete-time Markov Decision Process model and computes the expected revenue and decision policy matrix for various combinations of available capacity and allocating time period. The WSP will accept / deny the arriving user requests in real-time dynamically based on its current network state and its pre-computed decision policy matrix.

Finally Chapter 6 concludes this dissertation offering significant inferences and suggestions for future research.

1.5 Summary of Contributions

The summary of contributions of our work is:

- We perform simulation based demonstrations of the spatial bias problem in multi-hop WMNs leading to unfairness and study its impact on the performance of these networks.
- We identify some key limiting factors hindering the large scale deployment of WMNs with regards to routing, and attempt to mitigate such factors in our proposed routing paradigm.
- We propose a novel service differentiation technique using dual queues for IEEE 802.11s based mesh networks [20].

- To address the hot-zone problem around IGWs in WMNs, we propose a load balancing routing scheme among different IGWs based on their current traffic serving capacity [21].
- We propose a novel Adaptive State-based Multi-path Routing Protocol (ASMRP), which constructs Directed Acyclic Graphs (DAGs) and effectively discovers multiple optimal path set between any given MR-IGW pair [22].
- We design a novel Neighbor State Maintenance (NSM) module that innovatively employs a state machine at each MR to monitor the quality of links connecting its neighbors in order to cope up with unreliable wireless links.
- We employ four-radio architecture for MRs, which allows them to communicate over multiple radios tuned to non-overlapping channels and better utilize the available spectrum.
- We propose a dynamic user request admission / allocation policy for WSPs to maximize their obtainable revenue through an optimal and prudent selection of its admitted user base [23].
- We apply yield management principles in building the framework for the proposed user admission model for WSPs.
- We use discrete-time Markov decision process model in the formulation and optimization of the admission / allocation policy.

All the above contributions are explained in detail in subsequent chapters.

Chapter 2. Service Differentiation in Mesh Networks: A Dual Queue Strategy

2.1 Introduction

Fairness in a network implies optimal allocation of available network resources such as channel access and bandwidth, to the flows originating from various nodes based on a pre-determined and balanced criterion. Users in conventional single hop networks such as a cellular network typically get fair access to resources and this process is managed by its Base-Station or a central controller. However, in a multi-hop network like a WMN, an IGW typically is neither assigned nor can perform the role of a centralized coordinator, as MRs are connected in a multi-hop fashion to the IGW. In such a scenario, MRs solely depend on cooperation of their peer MRs to relay their traffic. Thus, though multi-hop communication facilitates increased coverage, low deployment costs, and other such advantages, it suffers from drawbacks such as spatial bias, collisions, hidden/exposed terminal problems, which are further explained in detail.

Emerging applications such as video on-demand, VoIP, video conferencing have strict Quality of Service (QoS) requirements such as bounded delay, minimum bandwidth and minimal jitter. They are different from elastic applications such as file transfer which are tolerant to delays but demand high throughput gains. Providing enhanced QoS support to users with such application requirements is the major concern for researchers in the current era.

In a multi-hop WMN, the proximity of client's corresponding MR to the IGW plays a significant role in its obtained performance. Often the clients attached to MRs that are closer to the IGW receive greater throughput and experience lesser end-to-end delays when compared to the clients attached to MRs far away from the IGW. In other words, the longer hop length flows receive extremely lower throughput and experience higher end-to-end delays. The envisioned goal of WMNs to replace the wired backbone implies an implicit requirement of unbiased treatment to all flows regardless of their spatial origin.

We propose a dual-queue service differentiation algorithm to ensure fairness to the multi-hop traffic from the traffic originating from local neighborhood of a node. Broadly, this algorithm works by maintaining two queues at each node, which separately hosts locally generated traffic at the MR and the multi-hop traffic traversing through this node. The scheduling of the packets from either of these queues is based upon a service rate defined at each node, giving more priority to the forwarded traffic when compared to the locally generated traffic.

The remaining chapter is organized as follows: In Section 2.2, we present the motivation that guided our work and highlight the need for spatial fairness in WMNs. The major design goals and considerations are described in Section 2.3. Section 2.4 elaborates the architecture of our proposed dual-queue based scheme with the help of the algorithm. In Section 2.5, we provide a comprehensive performance evaluation of our scheme. Section 2.6 presents the various existing schemes to alleviate unfairness to longer hop length flows in WMNs. We finally conclude with the summary of our scheme in Section 2.7.

2.2 Illustration of Unfairness Problem in Multi-hop WMNs

In this section, we illustrate the aforementioned unfairness problem in multi-hop WMNs through simulations in *ns-2* [24]. In WMNs, most of the traffic is directed either towards the Internet or vice versa through the IGW. Thus, in order to enable Internet-driven communication, multi-hop forwarding is inevitable. Unfortunately, multi-hop forwarding is plagued with myriad of problems – one of the major concerns being the fairness in forwarding the traffic. In other words, packets coming from far away MRs need to contend with the packets originated from the MRs near the IGW. Often, due to MAC layer contention at the intermediate hops, packets from far away MRs have higher inter-arrival rate compared to others. In addition to this, the intermediate MRs usually employ a First In First Out (FIFO) drop-tail queuing mechanism. As each node has an additional responsibility to relay others' traffic, the MR's locally generated traffic[2.1] competes with the relayed traffic. The bounded buffer is shared between the local traffic and relayed traffic. Usually, the local traffic overwhelms this buffer because of the FIFO queuing policy and the higher inter-arrival times of relayed traffic. This sort of satiating the buffers at the intermediate MRs by the nearby MCs results in dropping of packets arriving from clients registered under far away MRs. This also results in wastage of network resources such as bandwidth and incurs lot of delay as the dropped packets need to be retransmitted.

This problem can be better explained using an example scenario. Consider a real-time video streaming session between a pair of nodes multiple hops away. During the session, if a set of the video packets are dropped due to buffer overflow/congestion at an intermediate MR that is closer to the IGW, then there is pronounced degradation in the video quality perceived by the end user. We consider a simple IEEE 802.11s based mesh network (with 25 MRs) in a grid scenario. All these MRs communicate with each other using the legacy IEEE 802.11 based interfaces, forming a wireless backbone. MR 0 is in the bottom left corner of the grid and acts as the attached gateway that provides Internet connectivity to the other MRs. As assumed in [10], we also consider the MRs communicate with their MCs using an alternative 802.11 interface that works

[2.1] By local traffic we mean the traffic generated by the clients under an MR and 'relayed' or 'multi-hop' traffic means the traffic generated by clients under a different MR.

in a non-interfering channel. Thus, the communications between a MR and its clients does not interfere with the communication among peer MRs.

Further, we assume that all clients employ IEEE 802.11 DCF operating at 11 Mbps with RTS-CTS handshake disabled. The radio propagation model used is the two-ray ground model with a transmission range of 250 m and carrier sensing range of 550m. As shown in Figure 2.1(a), we randomly choose four MRs in the grid topology, each of them having their clients generating traffic. This traffic is aggregated at the corresponding MR and forwarded towards the IGW. For ease of illustration, we consider that the clients generate only UDP flows and their rate is adjusted such that the aggregate offered load by each selected MR is up to 500 Kbps. Without loss of generality, we assume a constant packet size of 1024 bytes for all the UDP flows.

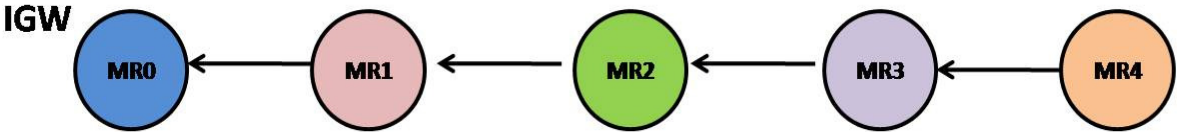
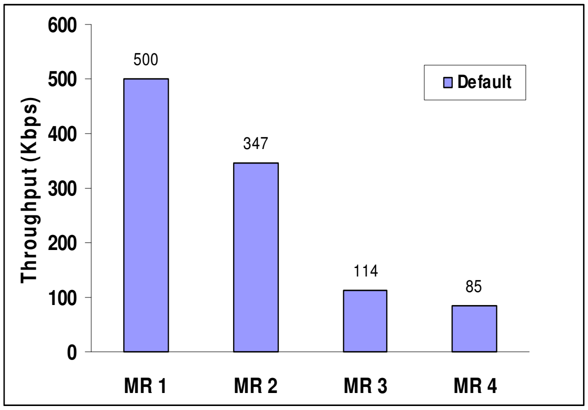


Figure 2.1 (a) MRs Connected in a Linear Scenario



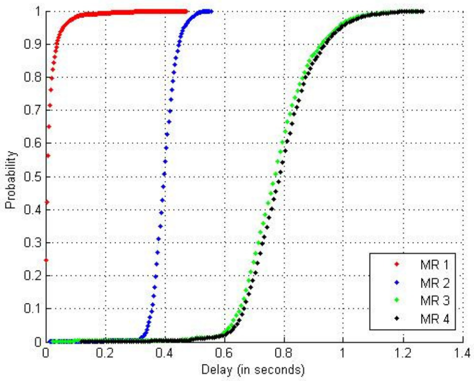


Figure 2.1 (b) Aggregate Throughput of Flows from each MR

Figure 2.1 (c) CDFs of Flows from each MR

We first measure the aggregate throughput of each MR. We define the aggregate throughput of a MR as the sum of individual throughput obtained by all the flows from the clients registered under that corresponding MR. Figure 2.1(b) shows the aggregate throughput obtained by each

MR. We notice that MR 1 which is 1-hop away from the IGW receives a throughput of 500 Kbps (100% of its offered traffic load) while MR 2 that is 2-hops away from IGW receives nearly 350 Kbps (70% of its offered traffic load). Flows with increasing hop count, i.e., MR 3 (3-hop) and MR 4 (4-hop) obtain 114 Kbps (22% of the offered traffic load) and 85 Kbps (17% of the offered traffic load) respectively. Clearly, we can notice pronounced spatial unfairness in terms of throughput obtained by each MR. There is severe degradation in the obtained throughput for the MRs that are located far away from the Internet attachment (IGW). This shows that the proximity of clients in a network to the IGW plays a significant role in the performance obtained. Clients attached to MRs that are located far away from the IGW receive low throughput which is highly undesirable and hence obtain poor quality of service.

We also investigate the per packet end-to-end delay experienced by the clients. In Figure 2.1 (c), we plot the distribution of delay for 1-, 2-, 3-, and 4-hop flows using the same scenario as described earlier in this section. As can be observed from the Cumulative Distribution Function (CDF), the delay incurred in transmitting packets of flows from 1-hop distance is much lower than other flows. We notice that 90% of the packets belonging to 1-hop flows experience a delay less than 100 ms, and 60% of the packets belonging to 2-hop flows experience delays less than 400ms. We further observe that the packets belonging to 3-hop and 4-hop flows encounter substantial latencies. More than 50% of the packets belonging to the 3-hop and 4-hop flows experience an average delay of more than 800ms. Such increased latencies are highly unacceptable for certain applications such as real-time sessions or applications involving critical and reliable information transfer.

It is also worthwhile to note here that the number of packets belonging to 3-hop and 4-hop flows that are transmitted through the network is substantially less which can be observed from the obtained lower throughput.

This kind of scenario is prevalent in any multi-hop network and WMNs are no exception. Additionally, in WMNs, the traffic volume in WMN can be large at an intermediate MR. Thus, it is very important to provide service differentiation among the traffic from local neighborhood and the traffic traversing more number of hops. In other words, traffic that has traveled larger number of hops has already consumed network resources and ought to receive a fair treatment. Considering the bounded buffer and the drop tail queuing mechanism at the nearby MR, it would be beneficial to isolate the local (own) traffic from the relayed traffic. This would in turn ensure guaranteed quality of service to users located far away from the internet attachment. Although IEEE 802.11e MAC protocol provides service differentiation, it focuses mainly on single hop networks and does not address multi-hop networks. Thus, we focus mainly on ensuring fair service to users in a multi-hop WMN.

2.3 Design Goals

In this section, we enlist the main design goals of our scheme. First, we plan to incorporate a flow level service differentiation for provisioning QoS. Applications that run over the internet today are varied, ranging from video-audio streaming, file sharing, peer-to-peer messaging, amongst others. These applications have contrasting resource requirements. For example, audio-video conferencing require minimal jitter and finite delay bounds while file sharing applications require large bandwidth. Thus, any proposed network must meet the requirements of a very general usage scenario in order to be successful in the end user market. As WMNs are expected to support such applications, QoS provisioning is an essential requirement and is a key challenge.

Thus, we provide packet level service differentiation to guarantee better QoS to the end user applications regardless of their spatial location.

Our second design goal is to consider the placement of our proposed Queue Management module in the network protocol stack. Installing new hardware or making hardware upgrades for enabling service provisioning for multi-hop traffic would be costly and may not be desirable. Considering the wide scale deployment of the legacy IEEE 802.11 devices, any changes at the MAC layer may not be ideal. Our Queue Management module is implemented above the standard IEEE 802.11 MAC layer, thus obviating any modifications to the legacy MAC. Our algorithm can be easily patched onto the device driver of the Network Interface Card (NIC).

Providing fair share of service to users with exogenous data rate requirements is one of the major concerns of future wireless networks. The objective of our scheme is two-fold: to fully utilize the resources available in a network such as bandwidth and to ensure proportional quality of service to end users. In our scheme, we maintain two queues, one each for local and multihop traffic. Even within a forwarded traffic queue, we may have packets belonging to flows from the same source, in which case if we give more priority to such flows, then the self-generated traffic may suffer from starvation. Thus, we need to embed a rate adapter or regulator to control such aggressive sources. However, the main focus in this chapter is to provide differentiated service to local and multi-hop traffic at an intermediate MR such that the local traffic does not monopolize the network resources. Thus, the primary responsibility of our proposed module is to shield traffic belonging to longer hop length flows from being throttled by the local traffic at a node; eventually enhancing the quality of service experienced by the end users.

2.4 Dual Queue Service Differentiation (DQSD)

Experiments in Section 2.2 indicate that sources in the close proximity of the IGW grab an unfair share of the buffer at the intermediate nodes and end up overwhelming the longer hop flows due to their spatial positioning. This leads to significant throughput degradation of longer hop length flows. In order to solve this problem, we put forward a mechanism to identify the aggressive flows and regulate the traffic from these flows. Our main goal is to provide proportional quality of service and fair performance to end users regardless of their spatial location and rate of their flows. The proposed algorithm guarantees a fair buffer share at each intermediate MR, for all flows traversing through the MR, irrespective of their hop length.

To cope with the abovementioned lack of guaranteed service and to alleviate unfairness, we propose a dual queue strategy which elegantly provides service guarantees to users located far away from the internet attachment. In this work, we propose a Queue Management (QM) module for the IEEE 802.11s based mesh networks to ensure proportional level of service to multi-hop traffic compared to the local traffic at each node. The algorithm works by elegantly segregating and exclusively reserving queues for either of the traffic. In other words, while one of the queues buffers self-originated packets at a node, called the local traffic; the other queue exclusively stores the multi-hop traffic; i.e., traffic traversing multiple hops.

Specifically, our scheme works by segregating the self-originated flows from the relayed traffic at each node. We use two queues to maintain the local traffic at a node and the multi-hop traffic traversing through this node. In our terminology, local traffic at a MR is the traffic generated from all the clients that are being served by the MR and can be maintained in the Local Queue (LQ). Traffic arriving from far away MRs which has to be relayed is called forwarded/relayed/multi-hop traffic and is stored in a separate queue, called the Multi-hop Queue (MQ), thus shielding from the local traffic.

We first propose the use of dual queue discipline to identify and segregate the local flows from relayed flows at each node. Upon identifying these set of flows, the QM module is responsible for efficiently scheduling the traffic from both these queues in a manner that would provide fair service to both. The main idea behind our scheduling algorithm is to assign proportional priority to relayed packets and the packets belonging to local traffic. For example, consider the case in which all the MRs have the same amount of traffic load. At intermediate MRs, packets originating at far away MRs have to traverse multiple hops thus having relatively higher inter-arrival times at the intermediate MRs. On the other hand, packets from local or nearby sources may arrive more frequently. If the packet generation rate of a near-by flow is higher, then more often the buffer at these intermediate MRs is dominated by the local traffic, thus overwhelming the forwarded traffic. Thus, continuously admitting such packets in the buffer would lead to a full buffer, resulting in no space for longer hop length flows when they arrive eventually. In order to prevent such unfair treatment, we devise a scheduling sequence for both these types of packets. In other words, at each node, the packets from the source are scheduled to gain the channel access in proportion to the number of relayed flows' packets currently buffered at the node, which is a reasonable way to enforce fair treatment of flows. In this way, our algorithm ensures that neither local flows nor relayed flows monopolize the buffer at any node.

Before we proceed to describe our scheme in detail, we first introduce the data structures and variables used at each MR for maintaining the per-active-flow state.

2.4.1 Data Structures

Each MR maintains a Flow Table that contains information about all the active nodes that are routing their packets through it. The fields are explained below:

Source Address: A source node which is having one or more flows routed through this MR.

Flow_ID: Unique flow id of each flow traversing through the node. A single source may have several active flows in session.

Own_Service_Rate: This specifies the rate at which the Local Queue (LQ) will be serviced, that is, the probability with which a packet from the local queue will be scheduled. This is computed according to the number of flows from the node denoted by *Num_Own_Flows* and the total number of flows traversing through this node which includes its own generated flows. The value of *Own_Service_Rate* is updated whenever a new flow arrives at the node.

Others_Service_Rate: This specifies the probability of scheduling a packet from Multi-hop Queue (MQ).

Total_Num_Flows: *Total_Num_Flows* keeps track of the total number of flows at the node including both the self generated (originated) flows as well as the relayed traffic flows.

2.4.2 DQSD Algorithm

In this section, we describe our algorithm which offers a fine-grain treatment to longer hop length flows and alleviates unfairness to them. The DQSD algorithm is summarized below.

```
When a packet p arrives:
If(p->source is not in Flow_table)
  Update the Flow_table
  Create an entry for source with the flow_id
  Buffer the packet in the corresponding queue, either LQ
  or MQ
End If
Whenever a packet p has to be dequeued:
If(p->source is in the Flow_table)
  Calculate the Own_Service_Rate and Others_Service_Rate at the node
  According to the service rates, deque a packet from the corresponding queue
End If
```

DQSD Algorithm

The main unit is the Queue Management (QM) module which governs the service schedule sequence of both the queues at each node. Whenever a packet arrives at the link layer, the MR checks whether there is an entry for the corresponding *source* in the Flow table. If there is no entry for that *source*, a new entry is created and the Flow table is updated with necessary parameters. Then the packet is buffered into one of the two queues depending on whether it belongs to a flow that is relayed through this node or generated by the clients registered with this node. Whenever a packet has to be dequeued, the service rate is computed for each of the queues. More clearly, at each node, the QM computes the service rate of the local queue (LQ), denoted by *Own_Service_Rate*, for this source based on the number of flows originating from this source divided by the total number of flows currently occupied in the buffer space in the Flow table. We then recalculate the service rate of flows in the other queue, the multi-hop queue (MQ), denoted by *Others_Service_Rate* for other sources in the table.

Our algorithm is capable of identifying and distinguishing the local flows from the multi-hop flows and correspondingly ensures proportional service fairness. The proposed algorithm guarantees a fair buffer share at each intermediate MR, for all flows traversing through the MR, irrespective of their hop length. The proportional service schedule of the LQ and MQ can be computed using the total number of flows currently being serviced at each node and the own flows at each node. The average service rate of own flows and forwarded flows are estimated as follows:

$$Own\_Service\_Rate = \left(\frac{1}{Total\_Num\_Flows} \right) * Num\_Own\_Flows$$

$$Others\_Service\_Rate = (1 - Own\_Service\_Rate)$$

2.5 Performance Analysis

In this section, we compare the DQSD algorithm with the default link layer drop-tail queue mechanism through simulations performed in *ns-2* (version 2.28). We have used the scenario and configuration as described in Section 2.2, in which each MR aggregates the traffic from 2-3 clients that are associated with it. This traffic is oriented towards the IGW (MR 0).

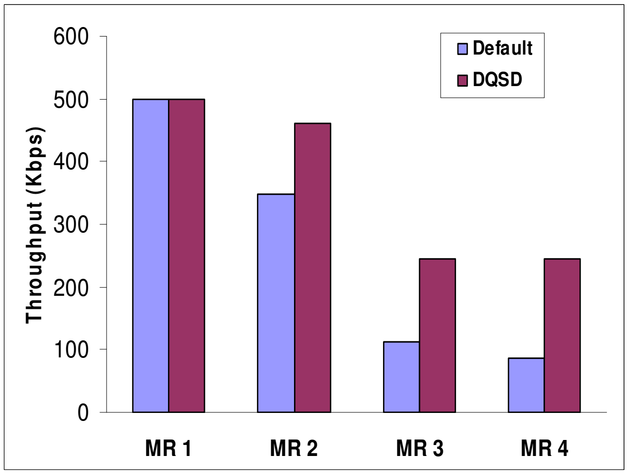


Figure 2.2 (a)
Aggregate Throughput of Flows

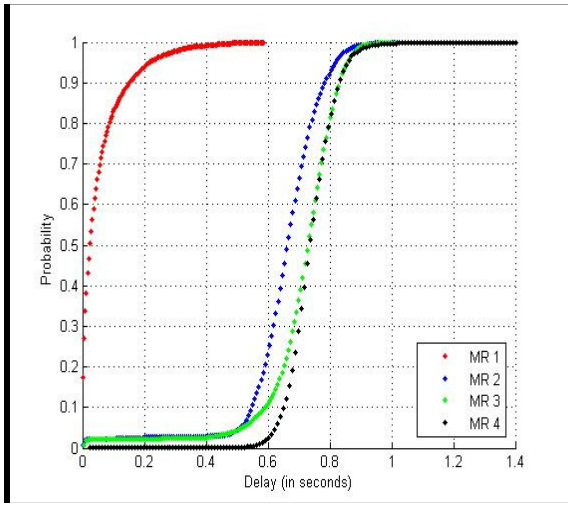


Figure 2.2 (b)
CDF of the Delay Distribution

2.5.1 Aggregate Throughput

We first compare the aggregate throughput obtained by each MR, which is defined as the combined throughput of all the clients registered with the MR. Figure 2.2(a) compares the aggregate throughput of the four MRs using the default queue management and using our proposed DQSD algorithm. As we can see from the figure, with the default queue management, MR 1 and MR 2 achieve relatively high throughput, while MR 3 and MR 4 starve. As explained earlier, in Section 2.2, these unacceptably low throughputs are mainly due to the spatial contention which is caused due to the proximity of high-traffic generating sources near the IGW. The meager throughput obtained is also partly due to the unfair buffer sharing between the local and multi-hop traffic at each of the intermediate MRs. Under high load, local flows from MR 1 quickly fill up the link layer buffer at this MR. When the packets from MR 2, MR 3 and MR 4 arrive at MR 1, they find a full buffer and are dropped. For similar reasons, MR 2 drops the packets from MR 3 and MR 4. Thus, flows from MR 4 experience a drastic decrease in their throughput. On the other hand, when we employ our dual queuing mechanism along with the queue service management module, we notice substantial improvement in the performance of the 3-hop and 4-hop flows. The 3-hop and 4-hop flows now receive up to 50% of their offered traffic load which is around 250% improvement when compared to the default case.

As we may recall from the previous section, with the presence of two different queues, the multi-hop/relay traffic is effectively shielded from the local traffic at each intermediate MR. As a result, relatively larger fraction of packets belonging to multi-hop traffic are scheduled for transmission to the next hops. However, it is important to note here that at any intermediate MR, all the multi-hop traffic is treated equally irrespective of the number of hops traversed.

2.5.2 Delay Distribution

Figure 2.2(b) depicts the delay distribution of the 1-, 2-, 3-, and 4- hop flows after employing the DQSD algorithm. We observe that while 90% of the packets belonging to 1-hop flows experience delay of less than 100ms, the latencies for 2-, 3-, and 4- hop flows are higher. More than 50% of the packets of 2-, 3-, and 4- hop flows have delays higher than 600ms. Compared to the delay CDF of the network when default queuing strategy is employed, the latencies are a bit higher. This can be attributed to the number of packets transmitted through the network belonging to these longer hop length flows.

We notice huge improvement in the number of packets belonging to 3-hop and 4-hop flows that are transmitted through the network when our dual queue strategy is employed (from the throughput graph). For the 3-hop and 4-hop flows, the number of packets transmitted is almost doubled compared to the default queuing strategy, as can be observed from the throughput graph. As there is substantial increase in the number of transmitted packets, the traffic load in the network is relatively higher when compared to the default case. The channel access delays and queuing dynamics at each intermediate MR contribute to the excessive delay for the longer hop length flows.

2.6 Related Work

The effect of spatial bias on the performance of flows originating from distant sources relative to the IGW has been widely studied. Specifically, the authors in [25] show through experiments that performance of sources far away from the IGW suffer drastically and obtain meager throughput as the offered load increases at each of the nodes that have traffic destined to the IGW. Much work in the literature has addressed this unfairness problem caused to farther hops. To resolve this unfairness problem, Jun and Sichitiu in [25] suggest maintaining a separate queue for each individual source at the intermediate relaying nodes. However, maintaining separate queues for individual sources may be infeasible considering the large scale of WMN deployments and the dynamic nature of traffic in WMNs. In [7], Gambiroza et al. propose an Inter-tap fairness algorithm in which the nodes exchange channel usage information and determine their maximal channel access times. Authors in [26] develop analytical models for hop by hop congestion control and propose mechanisms for controlling the traffic generated at the source nodes.

A different perspective, yet related problem that has been explored widely is the severe degradation of the network performance in the presence of aggressive flows. Aggressive flows that pump more number of packets into the network than the usual threshold; occupy high buffer space in the intermediate nodes and leave no room for packets belonging to other flows. Existing work has dealt with ensuring fair allocation of resources to multi-hop flows and the non-aggressive sources amidst the presence of aggressive sources [10][27][28]. In particular, [28] focuses on segregating the aggressive flows' packets from other flows' packets and applies a fair schedule to shorter queues compared to other queues. Though their work is closer to our effort, the main goal in this work is to segregate relayed traffic from local traffic and efficiently schedule the traffic. QMMN [10] elegantly manages the buffer at intermediate MRs by limiting the maximum buffer share occupied by any source node. However, the scheme provides fairness based on per-node allocation of the buffer space and ignores the degree of activity by each node while deciding the buffer allocation. CBTR [27] works by providing an impartial service to all flows irrespective of the number of hops they traverse. The authors employ most frequently seen cache discipline to identify the aggressive flows and drop packets from aggressive flows if they affect the performance of non-aggressive flows.

Although the above schemes address the deficiencies of drop tail queuing, they do not consider issues inherent to multi-hop wireless networks. The unfair behavior of the network and the eventual starvation of distant sources can be attributed to the sharing of the limited queue space, often a common queue, by the originated and the relayed traffic at intermediate nodes. Dropping traffic that has traversed multiple hops will result in wastage of valuable network resources. Thus, we suggest a more elegant mechanism of maintaining two separate queues for distinguishing the flows from nearby sources compared to flows originating from sources far away from the IGW.

2.7 Summary

In this chapter, we have illustrated the severe unfairness experienced by longer hop length flows in multi-hop WMNs through extensive simulations. We observe that the proximity of sources to an IGW has a profound impact on the performance of other flows from far away sources. Thus, we propose a dual queue based scheme for ensuring fairness to flows spanning several hops in the presence of flows from closer vicinity of a MR. We further propose a DQSD algorithm that employs a Queue Management module that efficiently manages two separate queues for the local and the multi-hop traffic. The algorithm provides fair treatment to the multi-hop flows using a service discipline where relayed/forwarded traffic is given higher priority relative to the traffic from local sources. The results obtained indicate substantial improvement in the throughput of 2-, 3-, and 4- hop flows without effecting 1-hop flows compared to when a default queuing strategy is applied. More specifically, the 3- and 4-hop flows experience a 250% improvement in their throughput compared to the default scenario.

Chapter 3. Achieving Load Balancing in Wireless Mesh Networks through Multiple Gateways

3.1 Introduction

WMNs are envisioned to serve a large community of users and thus average volume of traffic to be transported is significantly high. As discussed in Chapter 1, load balancing of the traffic at the gateway nodes is an important requirement to be addressed, considering fairly significant volume of traffic expected in WMNs. Existing mesh routing protocols do not focus on achieving fair load balancing at the IGW. Thus our main focus in this work is to build an approach for balancing load across the gateways in a WMN.

In a WMN, the estimated traffic volume is anticipated to be very high which makes scalability and load balancing as important issues among others. WMNs are aimed to provide high bandwidth broadband connections to a large community and thus should be able to accommodate a large number of users with different application requirements for accessing the Internet. Usually, most of the traffic in WMNs is oriented towards the Internet, which may increase the traffic load on certain MRs (close to the IGWs). As the IGWs are responsible for forwarding all the network traffic, they are likely to become potential bottlenecks in WMNs. High concentration of traffic at a gateway leads to saturation which in turn can result in packet drops due to potential buffer overflows. Dropping packets at the IGWs is highly undesirable and inefficient, especially after having consumed a lot of network resources en route from source till the IGW. Thus, to avert the danger of congestion, it is prudent to balance the traffic load over different IGWs and also possibly along the routes followed by the packets en-route to the IGW. This motivates us to devise a scheme which would enable sharing of the load among multiple gateways and improve the overall performance of the network.

Briefly, the logic behind our proposed solution is as follows. A potential congestion at an IGW is detected based on the average queue length estimated over a time period and an alert is raised, upon which selective active sources are sent notification messages to switch their internet attachment to a possible alternate less-congested gateway.

This chapter is organized as follows. We describe our proposed scheme for achieving load balancing in Section 3.2. In Section 3.3, we discuss the performance analysis of the scheme using *ns 2*. Section 3.4 gives an overview of the related work in this area. Finally, we conclude and summarize the chapter in Section 3.5.

3.2 Congestion Aware Load Balancing

The proposed scheme can be broadly classified into two phases - gateway discovery protocol and load migration procedure. The following subsections discuss these two phases in detail.

3.2.1 Gateway Discovery Protocol

In this phase, all the nodes discover their primary gateways. The gateways advertise their presence by sending beacons periodically. On receiving a beacon signal, a node registers itself to the gateway under two conditions:

- If it has not already selected a gateway node (i.e., its gateway ID is unknown), or
- If this new gateway is nearer than the already registered gateway. In such a case the node saves its originally registered longer hop gateway as its secondary gateway that could later be exploited for load balancing.

Initially, while detecting the gateways, we consider hop count as the basic metric for selecting a gateway as the other node-specific information (such as queue length, etc.) are unavailable at this stage.

3.2.2 Load Migration Procedure

Usually, each MR is serviced by a primary gateway through which it receives (sends) traffic from (to) the wired network. However, it also keeps track of other possible gateways in its Internet Gateway Table (IGT) by listening to the periodic beacons from the intermediate nodes. Each MR announces the list of IGW IDs it knows through periodic HELLO packets.

After the initial gateway discovery procedure, in the second phase, each IGW continuously monitors its queue length during each time window. If the average queue length rises above a certain threshold value in that time period, it is indicative of a possible impending congestion at the IGW. In such a case, the IGW identifies a set of active sources (sources with high traffic) serviced by it. In an attempt to reduce the load, it then sends a notification to these nodes intimating them to look for an alternative gateway that is relatively less congested.

The set of active (aggressive) sources can be easily identified by monitoring the packets handed over to the IFQ at the IGW. When the average queue length overshoots a certain threshold, the gateway selects these active sources and sends a notification message, NOTIFY. The NOTIFY message consists of useful information such as average queue length, the exceeded capacity, etc. We send these notifications essentially as unicast messages. We avoid broadcasting these messages because of two important reasons.

- Firstly, delivery of broadcast messages cannot be guaranteed as communication in wireless network is highly unreliable. With unicast messaging, the probability of loosing such packets is greatly reduced.

- Secondly, if the NOTIFY message is broadcasted; all the nodes routing through this IGW would try switching to a different IGW. As a result, the load on the current IGW will be drastically reduced and the new IGW would be highly overloaded. The greediness of load balancing would now force the MRs to frequently oscillate between these IGWs and severely degrade the network performance. Hence, any load balancing algorithm should be carefully designed to avoid such ping-pong effect in the network. On receiving a NOTIFY message from the primary IGW, the source node tries to join a different gateway listed in its gateway table. If its Internet Gateway Table indicates the presence of alternate IGWs, it sends a Gateway Request (IGW_REQ) message to the alternative gateway ID. The new gateway then decides whether to admit this node or not by observing its own load status (queue length, number of flows etc.). The gateway then sends a Internet Gateway Reply (IGW_REP) to the source if it accepts to serve the node. After receiving the IGW_REP, the source node switches to the new alternate gateway.

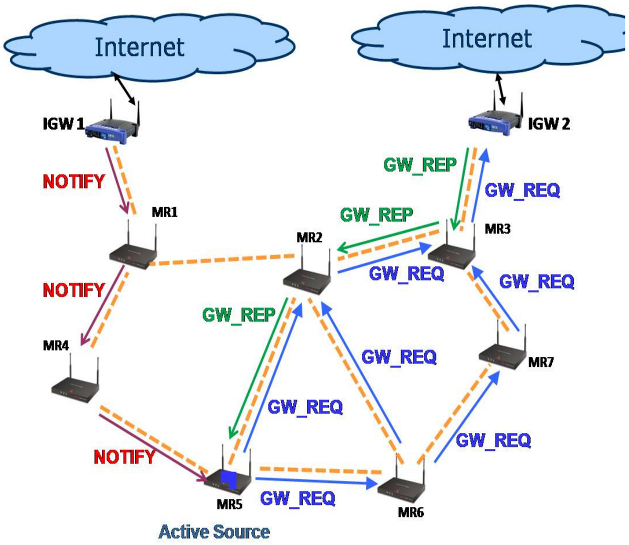


Figure 3.1
Illustration of Load Balancing in a WMN through Gateway Switching

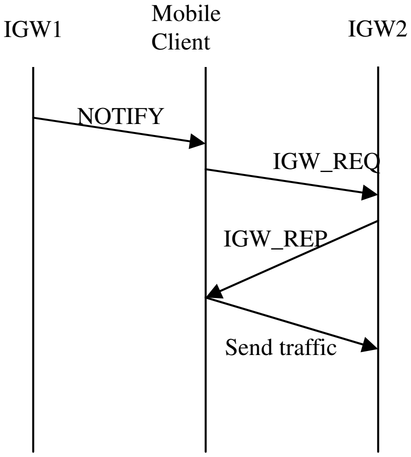


Figure 3.2
Timing Diagram Depicting the Sequence of Actions while Switching Gateways

We use Figure 3.1 to describe the load migration process. Initially, MR5 is connected to the Internet through IGW1 and starts some flows destined to an internet server. After some time, when IGW1 senses congestion (i.e., high packet drops or large queue lengths) for more than a threshold time period, it starts identifying the active sources. It identifies MR5 as one of the active sources that is sending a large amount of traffic and thus sends a NOTIFY message to it. MR6 upon receiving the NOTIFY message starts searching for an alternate IGW. If there are any IGWs listed in its IGT, MR5 sends a IGW_REQ message to its nearest alternate gateway (IGW2 in the figure). If no alternate IGW is found in the IGT, it sends an open request and waits for the reply. If IGW2 accepts to be its gateway, it sends a IGW_REP to the corresponding source node, here MR5. After MR5 receives the IGW_REP, it redirects all its further traffic flows towards this new gateway, IGW2. Thus, the load at the IGWs can be balanced by appropriately switching the primary IGW connectivity by the MRs. If however, there is no response to the open request, the node continues to send its flows through the current IGW. The timing diagram depicting the sequence of actions is shown in Figure 3.2 and the algorithm is summarized as follows.

```
At a IGW:
If the average IFQ length for a time period (Monitor_Cycle ) > Max_Permissible_Threshold
    Identify all the active sources
    For each active source
       Send a NOTIFY message to switch the gateway, if
       possible
    End for
 End if
 If a IGW_REQ message arrives from a node
    If the average IFQ length < Max_Permissible_Threshold
         Admit this node and send a IGW_REP to it
    End if
 End if
At a source node:
 Record the gateway information (IGW IDs) in the gateway table
 When a notification message from IGW arrives:
    For each gateway ID in the gateway table
       Send a IGW_REQ with the node’s estimated traffic
    End for
 When a IGW_REP message arrives from a gateway:
     Make the nearest gateway as the primary IGW
```

Queue-based Load Balancing Algorithm (LDBAL)

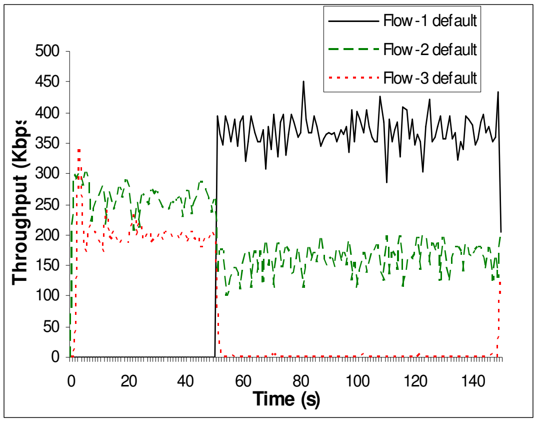
3.3 Performance Analysis

We use *ns-2* simulator to simulate our proposed scheme. We run the simulations for duration of 150s. We have used the same network scenario as in Figure 3.1. Clients registered with MRs: MR2, MR1 and MR5 generate traffic flows *f1*, *f2* and *f3* at the rate of 1200, 300 and 700 Kbps respectively. We start flows *f2* and *f3* just after the simulation starts. The packet size was set to 512 bytes. We use IEEE 802.11 as the underlying MAC protocol.

Initially, all the MRs are registered under the gateway IGW1 and the destination for all the flows is set to some Virtual server in the Internet (not shown in the figure). If the MRs later hear about any neighbor which is also a gateway, the nodes store that additional gateway ID into their Gateway Table. This additional IGW ID is later used for switching in case of congestion at any point of time.

The flows from MR1 and MR5 follow the path through IGW1. After the traffic corresponding to *f1* from MR2 starts, the throughput of the other flows suffers due to the congestion that gets accumulated at the gateway due to the high traffic generation rate of *f1*. Figure 3.3 shows the instantaneous throughput of the different flows when there is no load balancing applied. We can observe from Figure 3.3 that the throughput of flows *f2* and *f3* suffer once flow *f1* starts. The throughput of *f3* drops down to almost 0 while flow *f2* manages to obtain a meager throughput.

On the other hand, when we employ our load balancing scheme at the gateways, the throughput of all the flows are comparable and fair treatment is provided for all the MRs. Figure 3.4 shows the instantaneous throughput of the flows when we apply our scheme to balance the load across the gateways. As can be observed from Figure 3.4, even after flow *f1* with heavy traffic starts, the throughput of the other two flows does not suffer considerable loss. As mentioned earlier, when IGW1 senses increase in congestion resulting in packet losses, it informs MR2 (active source) and MR5 about the congestion.



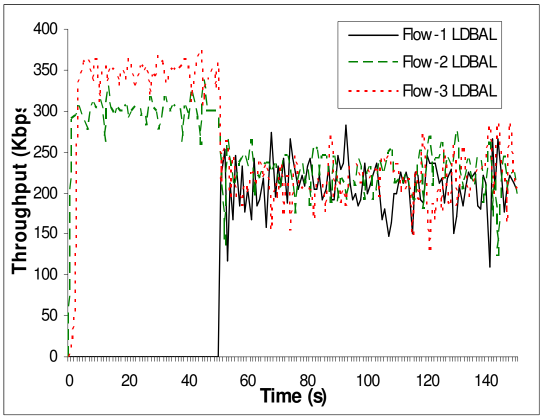


Figure 3.3
Instantaneous Throughput obtained by the Flows using the Default Scheme

Figure 3.4
Instantaneous Throughput obtained by the Flows using the Proposed Scheme

Upon receiving such a congestion notification message, MR2 would send a request message to the alternate gateway ID (IGW2 retrieved from its Gateway Table). However, MR5 does not have any information regarding any other alternate gateways; and thus it does not take any further action. Once IGW2 accepts MR2's request, MR2 diverts its traffic through IGW2. This decreases the load on IGW1 and thus helps in improving the throughput of all the flows in the network.

Figure 3.5 compares the Packet Delivery Ratio (PDR) of the flows in the default and the LDBAL schemes. If a heavy generating traffic flow starts, congestion builds up at the gateway, resulting in an increased packet loss. With the default scheme, as soon as *f1* starts its traffic, the congestion at IGW1 causes the packet loss for the other ongoing flows, *f2* and *f3*. However, in our scheme, as soon as the gateway detects congestion, it informs the active sources (MRs) about the congestion. The active sources, in turn, switch their gateways (if possible). As can be observed from the Figure 3.5, the PDR of all the flows improves once we employ our load balancing scheme in the network.

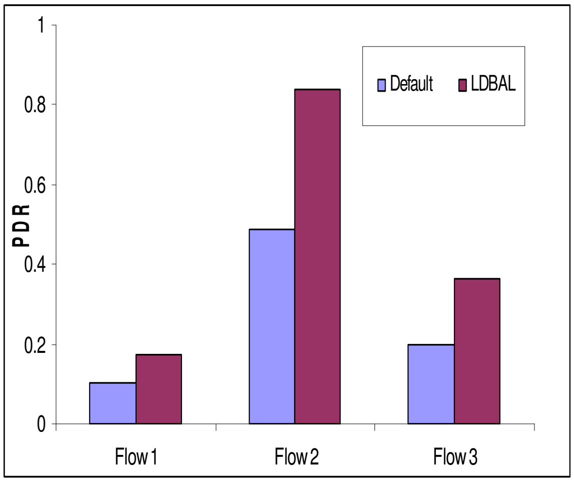


Figure 3.5
Packet Delivery Ratio for the Flows *f1*, *f2*, and *f3*

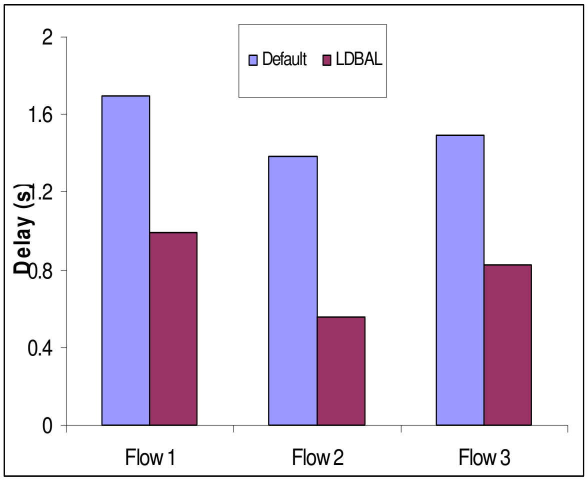


Figure 3.6
Average Delay for the Flows *f1*, *f2*, and *f3*

We also observe from Figure 3.6 that the average delay in the network for the flows using the LDBAL scheme is lesser than the default scheme. As the traffic load is divided among the different gateways, congestion at any IGW decreases. As a result, the end to end delay for all the flows decreases.

3.4 Related Work

A vast amount of research has been done in designing load balanced routing algorithms for wireless networks. These existing approaches differ in the metric considered for evaluating the load in a network. The authors in [29] propose a method for selecting the routes at a destination depending upon the extent of nodal activity (number of emanating paths from a node) coupled with the cost of transmission interference in a neighborhood. The proposed routing algorithm in [30] involves selection of a best route at the destination based on the number of queued packets at all the intermediate nodes. The route that has the least number of buffered packets is supposedly the less-congested route and thus, such a route can be considered the best path. The approach followed by the authors in [31] is concerned with the involvement of gateways in the process of load balancing in the network. In one of their proposed schemes, the gateway node coordinates the load balancing in the top part of the network (part of the network comprising of nodes nearest to the IGW) using the number of flows as the load metric. Essentially, they consider a load to be defined as a flow and proceed to balance the number of loads in the network. However, these existing techniques are not directly applicable to WMNs.

Due to a large number of nodes connected in a WMN, such algorithms are difficult to realize. Most of these schemes focus on load balancing over the links or intermediate nodes (MRs) connecting to the gateway and do not focus on alleviating the congestion at the gateway. Authors in [32] devise a scheme for load balancing over the links that uses path capacity and gateway link capacity as the cost metric. Recent work by Krishna et al. [33] suggests that balancing load across gateways is probably more effective than load balancing along the paths. According to their scheme, a better gateway is selected than the current servicing gateway if the Round Trip Time (RTT) for the new gateway is less than the registered one. However, we believe that delay is not an adequate metric for quantitatively weighing load in the network and propose to use a better measure or indication such as the queue length at the IGW as it is more relevant for such estimation. Thus, efficient load balancing techniques need to be designed that can be applicable to the WMNs for well-balanced network resource utilization.

3.5 Summary

In this chapter we have illustrated degradation in performance of the flows in a WMN network due to the congestion at an IGW. We propose an elegant load balancing mechanism so that traffic load is distributed among multiple gateways by switching the point of attachment by the underlying MRs that are actively generating high traffic. The basic objective is to utilize all the available gateways for balancing the traffic load and mitigate congestion at only some gateways. Through simulation it has been demonstrated that our proposed scheme is able to balance the load substantially and improve the performance of all the flows.

Chapter 4. Multi-radio Multi-path Routing in Wireless Mesh Networks

4.1 Introduction

As discussed in Chapter 1, traffic in WMNs is predominantly between IGWs and the MRs, in contrast to MANETs where traffic is among the peer nodes. This focused traffic flow of WMNs towards and from IGW places higher demand on certain paths, connecting IGWs and MRs, unlike that of MANETs where the traffic is more or less uniformly distributed. The advantage in WMNs is the high connectivity of the mesh backbone, which facilitates availability of multiple routes between any two end clients. We propose Adaptive State-based Multi-path Routing Protocol (ASMRP) to increase reliability of data transmission, allowing adequate fault tolerance. Our approach in this work is similar to that of MMESH [17]. However, the key difference between MMESH and our proposed ASMRP is that, we incorporate multi-radio architecture and Neighbor State Maintenance feature for MRs.

Several of the existing routing solutions for WMNs are derived from protocols which are designed for Mobile Ad hoc Networks (MANETs). Although routing protocols designed for MANETs can be directly applied to WMNs, they are not effective and could lead to sub-optimal performance. The focus of the routing algorithms designed for MANETs is primarily to minimize the power consumption, and to cope up with the mobility feature of the nodes. However, MRs in WMNs are either stationary or minimally mobile and are not power-constrained. Also, WMNs are envisioned to support applications such as Voice over IP (VoIP), real-time video surveillance, and broadband internet to communities which are required to have minimal jitter, finite delay bounds, guaranteed throughput and other such Quality of Service (QoS) constraints. Hence, proposed algorithms for MANETs are not fully appropriate for WMNs.

Our objective is to design a routing protocol for WMNs that:

- Provides resilience, robustness, and stability against fluctuating wireless links, transient/permanent channel outages, and occasional MR failures.
- Provides load balancing and cope up with congestion.

In a WMN, it is possible to reach an IGW or any MR through multiple paths. We propose ASMRP for WMNs that opportunistically exploits multiple paths to synergistically improve the overall performance of WMNs. The proposed algorithm is a novel multi-path hybrid routing protocol that effectively discovers multiple paths and employs an elegant traffic splitting algorithm for balancing traffic over these multiple paths. Through extensive simulations, we observe that our protocol works very well to cope up with variations in the network traffic. Our protocol also improves the performance of flows traversing multiple hops.

4.2 Multi-path Routing in Wireless Mesh Networks

4.2.1 Network Model

We model the WMN as a graph *G (N, E)* where *N* represents the set of MRs and IGWs and *E* denotes the set of links between them. Two MRs are said to be adjacent or neighbors if they are connected by a link. Let N_i represent the set of neighbors of MR_i; H_{ij}^{k} be the set of allowed next hops at MR_k for the traffic destined from $MR_i \rightarrow MR_j$ or $H_{ij}^{k} = \{h_{ij}^{k}(l), h_{ij}^{k}(m), h_{ij}^{k}(n), ...\}$, implying $\{MR_l, MR_m, MR_n, ...\}$ are the allowed next hops at MR_k for the traffic from

$MR_i \rightarrow MR_j$. When *k*=*i*, it can be intuitively noted that, H_{ij}^{i} represents the set of next hops at MR_i to reach MR_j

In Figure 4.1, let a Directed Acyclic Graph (DAG), G_i, of MR_i be a set of all possible routes from MR_i to IGW. Each route in this graph is a series of hops from the source MR_i to the destination IGW, which can be represented as $\{MR_i, h_{ij}^{i}(k), h_{ij}^{k}(l), ..., IGW\}$, where the term $h_{ij}^{i}(k)$ represents one of the next hops of MR_i (here, MR_k) for traffic from $MR_i \rightarrow MR_j$; $h_{ij}^{k}(l)$ represents one of the next hops of MR_k (here, MR_l) for traffic from $MR_i \rightarrow MR_j$, and so on.

In general, a DAG for a given MR, say MR_i, to IGW can be considered as a tree with a root at MR_i and having the successor hop set H_{ij}^{k} at each successor MR (here, MR_k), till IGW. Each successor hop set, H_{ij}^{k} consists of at least one MR. If *n* represents the cardinality of the successor hop set, H_{ij}^{k}, i.e., *n*= $|\ H_{ij}^{k}\ |$, then *n* is equal to 1 for single path routing for all successor hops from MR_i → IGW. And, for multipath routing *n* ≥ *1*, that is, there is an option of sending traffic through multiple routes at each MR. Such a DAG created at MRs possess the connectivity and loop-freedom properties.

4.2.2 Network Initiation

In WMNs, the mesh backbone can be hierarchically structured as illustrated in Figure 4.1, and intuitively divided into different levels, say, *level-1, level-2 … level-n*, where *level-i* is considered higher in the hierarchy compared to *level-(i+1)*. For instance, in Figure 4.1, let the MRs which are in the transmission range of IGW be *level-1* MRs, and among the remaining MRs, let the MRs which are in the transmission range of at least one *level-1* MR, be *level-2*

MRs. And with a similar logic, let the MRs in the transmission range of *level-2* MRs and which are not in it or its higher level, be *level-3* MRs, and so on.

When a WMN is deployed, an IGW first advertises its presence and internet connectivity through beacons, which are received by MRs in *level-1*. The beacons from an IGW could include information such as the link capacity, average load at the IGW, and other such needed data. In a WMN, there may be multiple IGWs providing internet connectivity.

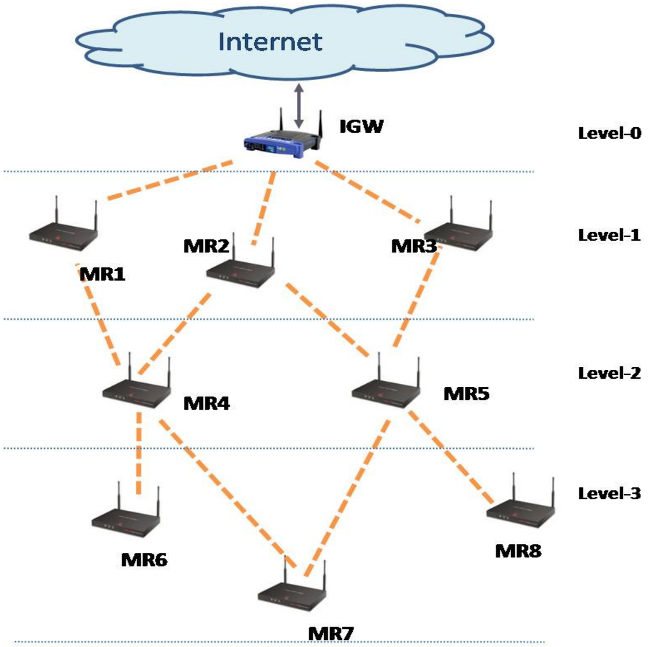


Figure 4.1

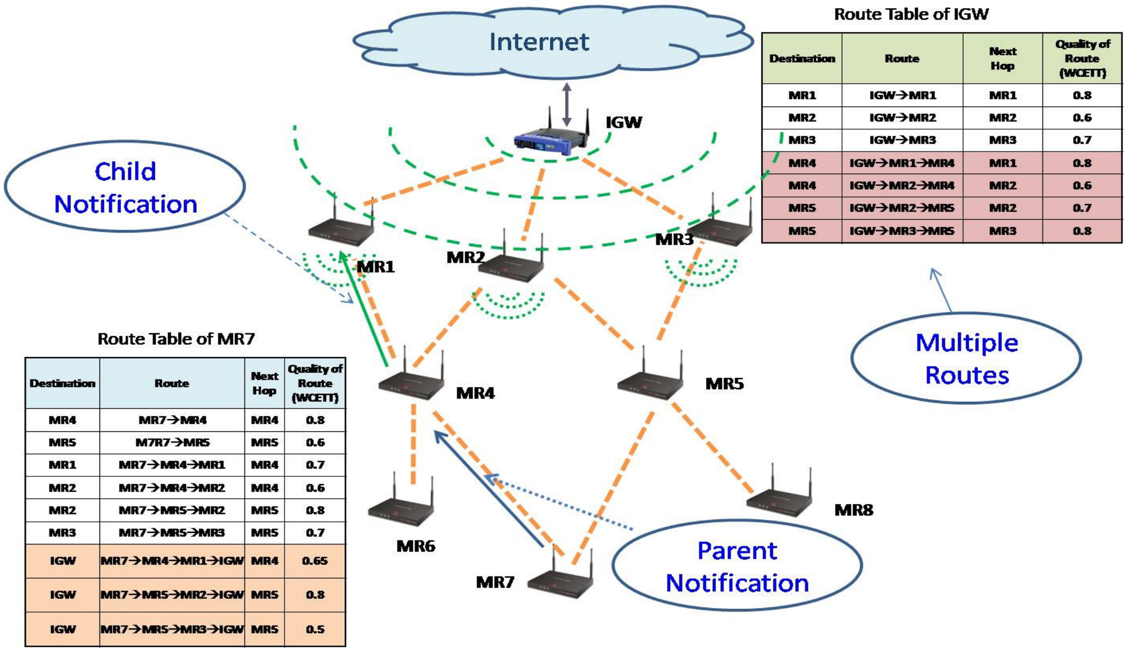
Illustration of the Proposed Algorithm

MRs in the WMN that have at least one path to the IGW will be able to broadcast the advertisements. Upon receiving the beacon, a *level-1* MR, will update its routing table with the route to reach the IGW. For instance, in Figure 4.1, MR1, MR2 and MR3 will update their routing tables with the routes {MR1 → IGW} and {MR2 → IGW}, and {MR3→IGW} respectively. After updating its routing table, a *level-1* MR informs IGW of it being the MR's parent, through *parent_notify* message. It then also broadcasts its connectivity with IGW to its neighboring MRs. The beacons from MRs include a set of routes to reach the IGW, along with their performance metrics. The quality of a route to the IGW depends on the parameters such as the link capacity, the channel diversity, and the number of hops needed. Depending upon specific application scenario, required route performance parameters are determined. For instance, an average end-to-end latency of the route is the performance metric for delay-sensitive applications such as VoIP, video streaming, etc.

These beacons of *level-1* MRs are then received by *level-2* MRs; say MR4, and MR5. A *level-2* MR will compute the performance metrics of these received routes and update its *optimal route set*, if needed. That is, a *level-2* MR chooses one or more relevant routes and will add to its routing table and/or update any changes in the route metrics of any previously known routes. For instance, in Figure 3, MR4 will add two routes, {MR4 → MR1 → IGW} and {MR4 → MR2 → IGW}, to its optimal route set. MR5 will also add two routes, {MR5 → MR2 → IGW} and {MR5 → MR3 → IGW}, respectively. MR7 will add routes {MR7 → MR4 → MR1 → IGW}, {MR7 → MR5 →MR2 → IGW}, and {MR7 → MR5 → MR3 → IGW} to its routing table, as shown in Figure 4.2.

These *level-2* MRs (child MRs) will then inform their respective *level-1* MRs (parent MRs) regarding their selection as parents, through *parent_notify* messages. Through this message, a child MR notifies its parent MR, the paths which can be used for forwarding its traffic. Parent MRs receiving this notification register the child MR in their routing tables and update the route(s) that should be used to forward the traffic from the child MR. The parent MRs also establish a reverse route to the child MR. The notification process continues by propagating further until it reaches the corresponding IGWs. This notification facilitates the use of multiple routes at each MR and the IGWs. Similar advertising and intimating logic is continued throughout all the MRs of the network, thus establishing the connectivity options at each MR.

After processing the *parent_notify* message, a parent MR will then unicast another notification message, called *child_notify*, to all the corresponding MRs that occur in the selected routes. This notification informs all the intermediate MRs along the route including the IGW about a child MR and the path that can be followed to reach this child MR. For instance, in Figure 4.2, MR7 (*level-3* MR) informs MR5 (*level-2* MR) of it being the parent and the selected routes, {MR7 → MR5 → MR2 → IGW} and {MR7 → MR5 → MR3 → IGW}, through *parent_notify* message. Then, MR5 informs *level-1* MRs, (MR2 and MR3) about the reachability of MR7 (its child) and the routes through individual *child_notify* messages. This message is also propagated all the way until it reaches the IGWs. On receiving the *child_notify* message, each parent MR registers this child MR and follows similar steps as described earlier, in registering the multiple route(s) to reach the child MR in their respective routing tables. Thus, each intermediate MR (including the IGW) that is in the path from a child MR to IGW now has one or more route(s) to the corresponding child MR.



Route Table of IGW

Destination	Route	Next Hop	Quality of Route (WCETT)
MR1	IGW→MR1	MR1	0.8
MR2	IGW→MR2	MR2	0.6
MR3	IGW→MR3	MR3	0.7
MR4	IGW→MR1→MR4	MR1	0.8
MR4	IGW→MR2→MR4	MR2	0.6
MR5	IGW→MR2→MR5	MR2	0.7
MR5	IGW→MR3→MR5	MR3	0.8

Route Table of MR7

Destination	Route	Next Hop	Quality of Route (WCETT)
MR4	MR7→MR4	MR4	0.8
MR5	M7R7→MR5	MR5	0.6
MR1	MR7→MR4→MR1	MR4	0.7
MR2	MR7→MR4→MR2	MR4	0.6
MR2	MR7→MR5→MR2	MR5	0.8
MR3	MR7→MR5→MR3	MR5	0.7
IGW	MR7→MR4→MR1→IGW	MR4	0.65
IGW	MR7→MR5→MR2→IGW	MR5	0.8
IGW	MR7→MR5→MR3→IGW	MR5	0.5

Figure 4.2
Illustration of the Route Discovery, Child and Parent Notification Procedures

At the end of network initiation, the network will now be in a near steady state. Each MR, say MR_i, will have multiple available routes to IGW (say MR_j) through its next hop set H_{ij}^{i}. And, a corresponding intermediate MR in its route, say MR_k will have its next hop set H_{ij}^{k} for routing the traffic from MR_i to IGW (or MR_j). When traffic needs to be sent from MR_i, it selects one of the next hops from H_{ij}^{i} for routing its traffic destined to IGW. Let MR_i select an arbitrary next hop MR, say MR_k. MR_k uses H_{ij}^{k} to select the next hop and the process continues until the packet reaches the destination IGW.

In ASMRP, we can limit the number of multiple paths by limiting the cardinality *n* of the next hop set, H_{ij}^{k}, to a restricted integer, say *q*; or, *{max(n)=q}*, indicating that the successor node set H_{ij}^{k} can consist of up to *q* MRs, where $q \geq 1$.

Among the multiple routes available, MR sorts these routes in the order of their route performance metrics which are computed in the following manner. For any adjacent MR pair, MR_i and MR_j, we assume that each link $\{l_{MR_i,MR_j} \mid MR_i, MR_j \in N\}$ is associated with a link weight vector $w(MR_i, MR_j) = \{w_1, w_2, w_3 ...., w_r\}$, in which w_i is an individual weight component, i.e., a single routing metric considered while selecting a route. Accordingly, any path from a source, MR_i, to a destination, MR_j, can be assigned a path weight vector $w^p = \{w_1^p, w_2^p, ..., w_r^p\}$, where $w_i^p = \sum_{l_{MR_i,MR_j} \in p} w_i(MR_i, MR_j)$ if w_i is an additive metric (e.g., delay); or $w_i^p = \min(w_i(MR_i, MR_j)), l_{MR_i,MR_j} \in p$, if w_i is a minimal metric (e.g., bandwidth). Typically, additive routing metric of a path is equal to the sum of the measured values of the metric over all the links along the path. Minimal routing metric of a path is obtained by taking the least value of the metric over all the links along the path. Multiplicative routing metric of a path is equal to the product of the measured values of the metric along all the links of the path. In ASMRP, we consider additive metric such as Expected Transmission Time (ETT) [19] and minimal metric such as congestion at next hop.

4.2.3 Congestion-aware Routing

Once multiple routes have been setup, to balance the network's load, a strategy for partitioning the traffic among these routes is required. With proper distribution of traffic, congestion can be minimized and reliable end-to-end packet delivery can be enhanced. Typically, when any MR needs to send traffic, it selects a next hop based on a routing parameter that is established by the network's routing protocol. One method of determining possible next hop at a MR is to evaluate the traffic load that the MR needs to transmit, and uses traffic distribution logic to balance the load of the potential next hops' links, as discussed in [34]. A strategy based on distance and link quality metrics [15] to determine the next hop could still result in decreased reliability if it encounters congestion at the next hop.

In our protocol, each MR say MR_k, proactively maintains its set of allowed next hops, H_{ij}^{k}, for sending its traffic destined from MR_i to MR_j. O*n*e important feature of the algorithm is its ability to effectively alleviate congestion by avoiding the traffic to route through the congested routes. Clearly, when choosing next hops, it is desirable to avoid neighbors with high congestion around them as well as those with low quality links. The routing layer periodically obtains information about a neighbor's congestion level, and uses this information in avoiding the congested routes. For instance, higher queue lengths at a neighbor indicate a congested MR, and can lead to possible packet loss. If a particular next hop is found to possess a large average queue length measured over a period of time, the protocol temporarily skips that MR and sends the traffic through other potential neighbors.

If we consider a MR, say MR_i, this has to adjudge the next hop for transmitting its traffic. Let the next hop be denoted by MR_k, which can be determined by applying the congestion-aware policy as given below:

$$MR_k = \min(\chi(MR_n))\forall(MR_n \in H_{ij}^{i}) ,$$

where, MR_n is any MR in the next hop successor set, H_{ij}^{i}, and $\chi(MR_n)$ denotes the load factor of MR_n.

4.3 Neighbor State Maintenance Module

In ASMRP, MRs use a state machine to determine promising neighbors and to maintain multiple reliable routes. Based on its link quality and reliability, we designate a *state* status to each neighbor to help combat with any intermittent links formed because of nodal mobility or unreliable wireless medium. We assess the history of a link connecting a given neighbor, based on metrics such as - the number of beacons (HELLOs) received; the number of data packets received; or its signal power over a given period of time. In our state maintenance module, each MR continuously monitors its links connecting to neighbors, and maintains their long-term and short-term histories.

Table 4.1 provides description of the possible states for MRs. Figure 4.3 depicts the transition mechanism of our neighbor state machine and Table 4.2 enlists the conditions that influence those state transitions.

Table 4.1: Describing the Purpose of Different States in the Proposed State Machine.

State	Description	Purpose
Initial (**I**)	By default, all neighbors will be in this state when the network boots up.	Bootstrapping the network.
Neighbor Candidate (**NC**)	State of a neighbor when a HELLO is *first* heard from it. A transition from state *I* to state *NC* happens when condition C1 is true.	A potential next hop, but has no history.
Neighbor (**N**)	State of a neighbor when the link is stable between the current MR and the neighbor. A transition from state *NC* to state *N* happens when condition C2 is true.	A MR forms routes using this neighbor.
Short-term History Bad (**SH_BAD**)	State of the neighbor when the MR observes that the short term history of the link connecting the neighbor is fluctuating and is bad. Transition occurs from state *N* to *SH_BAD* when condition C3 is true. If the short term history of the link improves again, then the neighbor transitions back to state *N*.	A neighbor in this state is temporarily disabled and all the routes through it are temporarily suspended. However, routes through this neighbor are used only when the MR doesn't have any other route. When the link connecting this neighbor gets better, the MR resumes transmission through the neighbor.
Long-term History Bad (**LH_BAD**)	State of the neighbor when the MR observes that the long term history of the link connecting this neighbor is bad. Transition occurs from state *SH_BAD* to *LH_BAD* when condition C4 is true.	All routes through this neighbor are deleted and the MR will attempt to form alternate routes. This neighbor is deleted from the MR's neighbor table.

Table 4.2: Conditions in State Machine

Condition	Description
C1	Any HELLO
C2	Repeated strong HELLOs
C3	History is bad for a defined short time period
C4	History is bad for a defined long time period

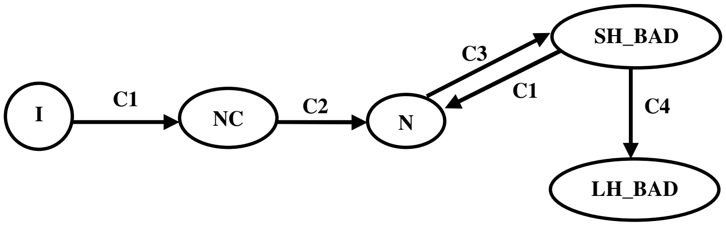


Figure 4.3
State Machine of a Neighbor

4.4 Multi-radio Architecture

In a single radio based multi-hop network that use only a single channel, effective bandwidth decreases drastically with increasing number of hops due to spatial contention [4]. Li et al. [35] have demonstrated that the achieved throughput of IEEE 802.11 in a multi-hop network is only 1/7 of the effective bandwidth of the channel. Typically, when a MR is equipped with only a single radio, this radio needs to switch its mode back and forth - for transmitting the backhaul traffic within the mesh backbone, and for communication with its registered wireless clients.

This back and forth switching results in significant latencies. Further, due to the half-duplex nature of the radio, a MR cannot send and receive traffic simultaneously.

Though incorporating a multi-channel transceiver that uses multiple non-interfering channels reduces the issue of spatial contention, it still requires complex MAC protocols and has the channel switching delays. Wu et al. [36] propose a MAC layer strategy, Dynamic Channel Assignment (DCA), which employs two transceivers - one for control packets transmission, and the other switches among different channels for data transmission with different receivers. However, such a methodology incurs considerable delays while initiating a communication session. Also, maintaining a dedicated control channel by a node can be expensive and results in wastage of bandwidth when the total number of available channels is limited. Usage of a time multiplexed control channel [37] addresses the limitations of dedicated control channel architecture, but it still suffers from synchronization problems.

One approach to overcome these limitations is to increase the number of radios at each MR, and balance the resource allocation for the needed backbone communication, and for relaying the traffic of its registered clients. In a dual-radio model, each MR is equipped with two radios - one is dedicated to the clients' access and the other is used for backbone communication. However, in this model, typically the radio used for backbone communication still results in considerable amount of channel switching due to several MRs in the mesh backbone, and simultaneous communication is still a problem. Thus, the performance improvement achieved compared to the single-radio architecture is marginal. This motivated the network service providers to increase the number of radios per MR so that majority of radios can be dedicated for the backbone communication and the remaining for the client access. Multi-radio extensions to the standard AODV routing protocol have been able to utilize the available spectrum efficiently under high traffic load conditions [38]. To overcome such limitations, Pathamasuntharam et al. [39] and Kyasanur et al. [40] propose a multi-interface architecture which employs three half-duplex interfaces each dedicated for transmitting, receiving and broadcasting. They present an interface switching strategy in which receiver interface is fixed on a specific channel for reception of data, and transmit interface gets tuned on demand according to the needed receiver's channel. We extend this model even further, as described below.

In our proposed ASMRP architecture, we equip each MR with four radios - one for communicating with the client nodes (MCs) to segregate its registered client communication from that of the mesh backbone; one to receive traffic from its peer MRs; one to send traffic to its peer MRs; and the remaining one for sending and receiving the broadcast messages. Since there are dedicated radios for a MR to transmit and receive traffic, it can now emulate a full-duplex behavior by simultaneously sending/receiving traffic to/from its neighbors in the network. Also in our model, we keep the receive interface fixed on a single channel, but we allow transmit interface to change its channel dynamically to synchronize with the receiver with which it needs to communicate. This tuning of the transmitter to a receiver's channel is determined using information in Neighbor Channel Table (NCT). Each MR maintains the database, NCT that contains the corresponding channel frequency to communicate with each of its neighbors. During communication, if the RTS/CTS feature is enabled, then the transmit interface at the transmitter MR switches to the respective channel frequency of the receiver and transmits the Request To Send (RTS), as well as data packets. The receive interface of the receiver uses it's receive channel for all the communication with the transmitter to avoid channel switching. The receive interface of the receiver sends Clear To Send (CTS) and ACK packets on the receive channel. If RTS/CTS feature is disabled, similar communication as described above occurs with the exception of RTS and CTS packets; only data and ACK packets are exchanged between transmitter and receiver MRs. In our simulations, we disable RTS/CTS feature. Also, a broadcast interface is added at each MR to prevent large switching delays while sending broadcast packets. In our work, we assumed homogeneity in both the number, and type of the wireless cards at the MRs and employed static channel assignment to the MRs in the WMN.

4.5 Performance Evaluation

We provide a simulation-based performance evaluation for ASMRP and compare it with the AODV, MMESH, MMR and CAM-ASMRP protocols. CAM-ASMRP is a modified ASMRP protocol implementation using the Channel Aware Multipath (CAM) [41] routing metric logic.

Simulation Parameters

For the simulation study, we use ns-2 simulator, version 2.31. For ASMRP, we change the ns-2's default singe radio implementation to include multi-radio and multi-channel capability. We enhance the default 802.11 MAC layer implementation of ns-2 to support the multi-rate transmissions similar to RBAR [42]. However, since ASMRP uses 4 radios at each MR, we assign the data rates to the channel per 802.11a standard. At the physical layer, for error model implementation, an error prone channel is simulated using the popular two-state Markov chain model [43][44] to reflect a bursty wireless channel state. This error module in our simulation study computes the Packet Error Rate (PER) for varying packet sizes, based on a given Bit Error Rate (BER).

We consider IEEE 802.11s based WMN with 20 MRs randomly deployed, and one among these is designated as the IGW. For the study, we assume that each MR serves up to 5-10 Mesh Clients (MCs) and establish connections to the external network through IGW. For the simulation, we choose 10 clients randomly that generate Constant Bit Rate (CBR) traffic, and measure the network's performance with varying traffic loads. These MRs communicate with each other using the legacy IEEE 802.11a based interface, forming a wireless backbone. In other words, we use 802.11a in the backhaul network for communication among peer-to-peer MRs that form the wireless WMN backbone, and we use 802.11b for the communication between the MRs and their registered MCs. As described in Section 4.4, each MR's receive channel is configured according to the channel assignment strategy (DCA [45]). DCA is a channel assignment strategy that uses a graph coloring mechanism to determine the channels for communication at each MR. It ensures that MRs that are in the interference range are assigned non-overlapping (different) channels. Thus, we assume a channel assignment strategy to MRs that limits the effect of neighboring MR interference on the performance of the network. In a network, as the node density increases, the number of orthogonal channels that have to be assigned ought to increase to avoid interference, thereby requiring more channels/colors for this operation. When the transmission to interference ratio increases in a network scenario; DCA will require fewer orthogonal channels for achieving an interference-free channel assignment. On the other hand, when the transmission to interference ratio decreases, the spatial interference region is relatively larger and more channels are required to avoid interference and collisions between their transmissions. The key parameters used in our simulation are summarized in Table 4.3.

Table 4.3 Simulation Parameters

Parameters	**Value**
Packet Size	1000 bytes
Simulation Time	150 seconds
Transmission Range	250m
Carrier sensing range	550m
IFQ Size	50
Radio Propagation Model	Two Ray Ground
Transport/Application Protocol	UDP(CBR)
RTS/CTS	Disabled

We develop the ASMRP routing module similar to AODV module. The HELLO heartbeat messages are periodically sent every 1000ms and include the route advertisements and connectivity information. Also, periodic updates to the routes (if any) are informed through the HELLO messages. MRs using ASMRP also communicate local congestion information (queue length) in the heartbeat messages. We reserve 1 byte for this parameter which is sufficient for communicating the queue lengths up to 256. We notice that this metric is reasonable as it avoids packet loss due to transient network congestion at the intermediate nodes. In ASMRP, each MR sorts the routes in optimal order based on the end-to-end path metric, Expected Transmission Time (ETT). In our simulations, we select top $k = 5$ routes which are used as multiple paths to route the traffic. This factor, k, is a system configurable parameter and can be assigned any non-negative integer. When $k = 1$, ASMRP behaves like a single-path routing algorithm.

To implement the neighbor state maintenance module at a MR described in Section 4.3, a neighbor transitions from state N to SH_BAD in our simulation when two consequent DATA packets to the neighbor are dropped. It is then suspended as a neighbor at the MR and routes through it are disabled. Initially, the neighbor continues in this state for a time period of 50ms, after which its state is enabled with it transitioning back to state N, which means that it is considered as a valid neighbor and routes through it can be used by the MR. However, if two DATA packets destined to it are dropped again, then the neighbor is transitioned to SH_BAD state and the neighbor is kept suspended for 100ms. Thus, with packets dropped in every attempt, the SH_BAD waiting time for the neighbor is incremented exponentially at the MR, which maintains the previous record of the waiting times. Further, in the simulation, the LH_BAD time is considered as 5s which means that when the neighbor has been suspended for about 5s (which the neighbor may reach after being in state SH_BAD for several times), it is permanently discarded as its potential neighbor and all the statistics corresponding to it are deleted.

4.5.1 Multi-rate Capability

Multi-rate capability of links in wireless networks is widely studied and depends on the underlying channel state and the Signal to Noise Ratio (SNR) with which a packet is received. Based on these two factors, the nodes in the network adapt their data transmission rate by dynamically switching data rates such that optimal throughput is achieved for the given channel conditions. Figure 4.4(a) charts the comparison of throughput obtained when the links in the network are not adapted to varying data rates, or in other words have constant data rate, and compare it to when multi-rate feature is enabled. Basically, MR's receive signal power should be high enough to accurately decode the received packet that has been transmitted at a high data rate. When the SNR decreases, the packet may not be properly decoded at the receiver due to complex modulation.

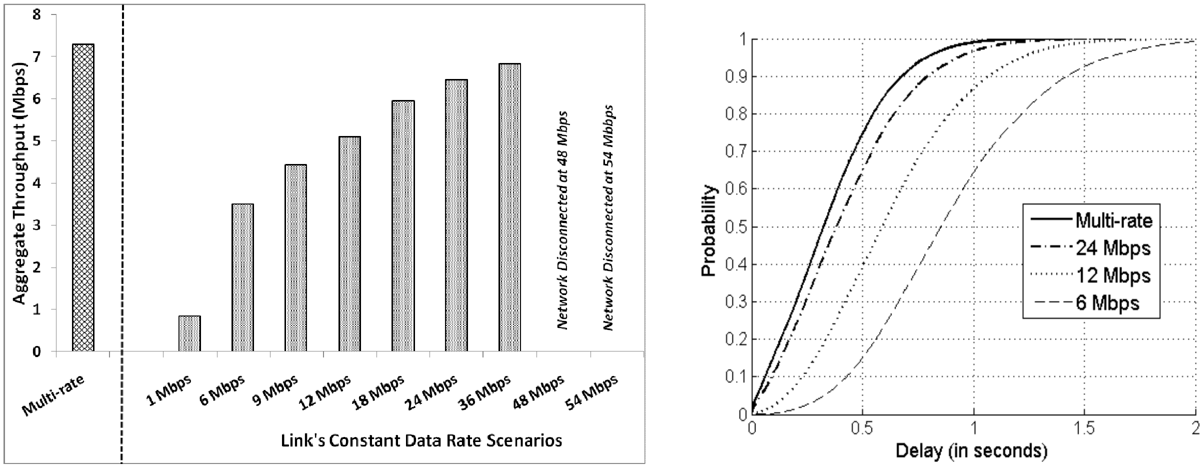


Figure 4.4(a)
Aggregate Throughput

Figure 4.4(b)
Delay Distribution

Multi-rate links vs. Constant data rate links
(Network got disconnected for links with constant 48 Mbps data rate)

In Figure 4.4(a), when the MRs transmit packets at a constant low data rate, the system achieves sub-optimal throughput, often lesser than its achievable capacity. We notice that higher throughput is achieved when the links are configured at 36 Mbps data rate compared to those scenarios with links supporting lesser data rates. However, when the links in the sample network operated at a constant high data rate of 48 or 54 Mbps, the network gets partitioned as some of the MRs are not able to decode the signal they receive. Thus, we exploit the multi-rate diversity of the links to improve the achieved throughput based on the prevalent channel conditions and SNR values. As observed in Figure 4.4(a), by enabling multi-rate feature of links, high throughput is achieved and still network connectivity is maintained, which is significant.

Figure 4.4(b) illustrates the delay distribution of packets when multi-rate capability of links is enabled versus scenarios with constant data rate links. The delay is high for the scenarios when links are configured with constant low data rate compared to those with constant high data rate links. The delay for the multi-rate link scenario is relatively less due to its optimal selection of links.

4.5.2 Throughput Comparison

Figure 4.5(a) compares the aggregate throughput of the network for AODV, MMESH, MMR, CAM-ASMRP and ASMRP models. CAM is a routing metric targeted for intelligent multi-path selection based on load distribution ratio over the routes, which we incorporated in the proposed ASMRP routing protocol and compared the results thus obtained with the congestion-aware routing metric which ASMRP originally uses.

From Figure 4.5(a), we notice that ASMRP outperforms MMR, AODV, and MMESH in the achieved aggregate throughput at all traffic loads. We notice that the average aggregate throughput of the network achieved for AODV protocol is 0.83 Mbps for an offered load of 500

Kbps at each of the 10 clients. For the same network and the traffic pattern scenario, when ASMRP with multi-radio architecture is employed, as shown in Figure 4.5(a), it results in a significant throughput improvement with about 460% over the AODV, 60% over MMESH and 280% over MMR.

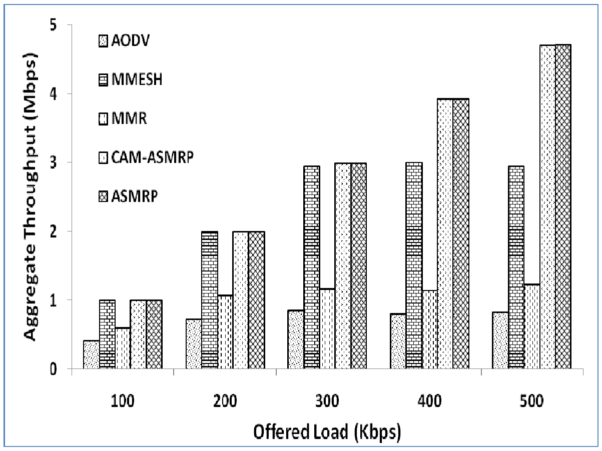


Figure 4.5(a)

Aggregate Throughput of Flows from Different MRs with Varying Traffic Load

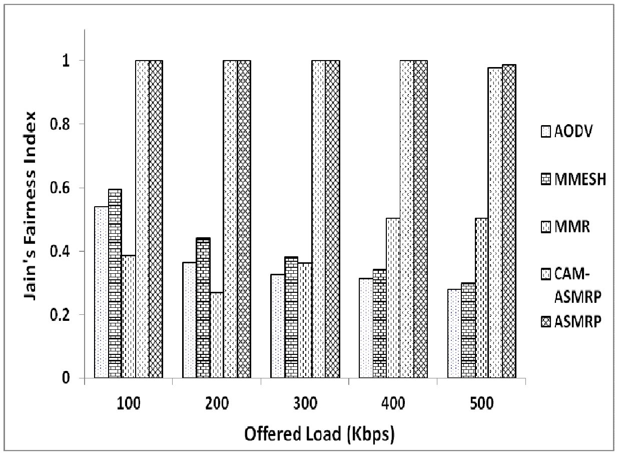


Figure 4.5(b)

Fairness Index for Different MRs with Varying Traffic Load

We also observe that MMR performs relatively poor in terms of throughput as compared to MMESH and ASMRP. The suboptimal throughput obtained for MMR could be due to two plausible reasons. First, the primary routing metric which determines the routes in MMR is the hop count, and as it is widely known, minimal hop count in networks with multi-rate links results in sub-optimal longer low-data rate hops, thus leading to performance degradation. Second, the reason lies in the protocol functionality itself. That is, in MMR, though each node forwards multiple route request packets for a single source-destination pair, the source selects only two of the multiple routes generated by the destination. This use of only two routes as considered by MMR could result in underutilization of network resources and limits the performance of the network. Further, in our simulation, we observed large number of collisions between the control packets and the data packets using MMR as routing protocol. Also, one of the channels is used both for control packets and routing data which becomes congested after sometime. Although, once routes are maintained, control packets may not be traveling over the route, but there can be collision between the cross traffic.

The throughput improvement with ASMRP can be attributed to the following factors. In ASMRP, the route selection process is based on optimizing two parameters – route selection criterion, and next-hop selection process. ASMRP chooses the routes with minimal estimated end-to-end transmission time for data transmit which is done based on global information. And, when the transmission is being done using these selected routes, ASMRP then dynamically chooses the minimal congested next hop at each MR and thus optimizes again based on local information. This way, any packet loss because of queue overflows is avoided. Frequent updates about the end-to-end statistics pertaining to routes may result in oscillations in routes chosen by the MRs while transmitting data, and hence to avoid these oscillations, the localized analysis of next-hop congestion is essential on a dynamic basis. In addition, simultaneous transmit and receive operations and non-interfering communication among MRs is possible which further enhances the throughput.

4.5.3 Fairness Comparison

Figure 4.5(b) compares the fairness index of the network for AODV, MMESH, MMR, CAM-ASMRP and ASMRP models. As seen in the Figures 4.5(a) and 4.5(b), simulation results suggest that ASMRP performs equally well with its congestion-aware metric as with CAM logic. Hence we will limit our simulation study to ASMRP using congestion aware routing metric for further simulation studies.

We use Jain’s fairness index [46] to measure the fairness among the flows. Jain’s fairness index, say f, is given by:

$$f = \left(\frac{\left(\sum_{i=1}^{n} x_i \right)^2}{\left(n * \sum_{i=1}^{n} x_i^2 \right)} \right) \qquad 1 \le i \le n ,$$

where, n is the number of flows in the network, and x_i is the throughput achieved by flow i. This fairness index is always positive and when it approaches 1, it implies that all the flows in the network get near equal share of the achievable network throughput.

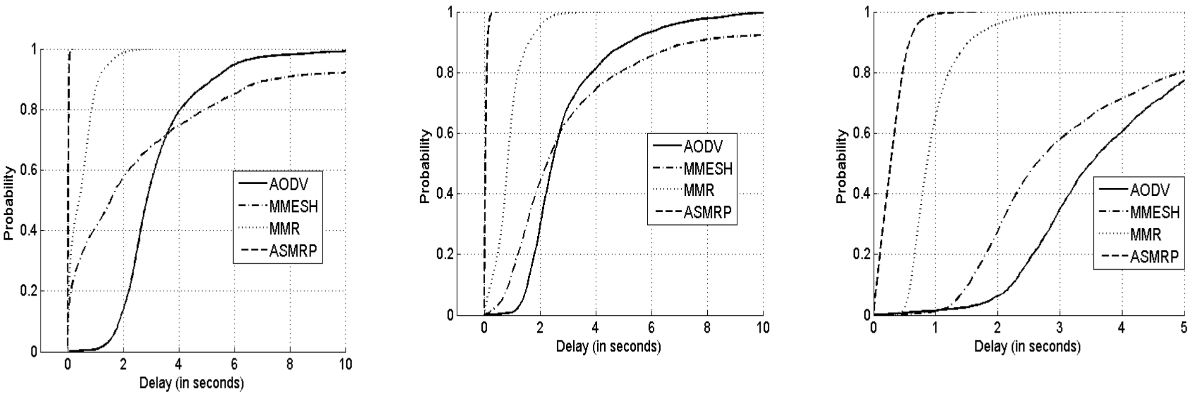
In a multi-hop WMN, proximity of client's corresponding MR to the IGW has a significant impact on the aggregate performance of the network [20]. Often, clients attached to MRs that are closer to a IGW receive greater throughput and experience lesser end-to-end delays as compared to the clients attached to MRs located far away from the IGW. In other words, the longer hop length flows receive extremely low throughput and experience high end-to-end delays.

In our simulation study, we observe that when AODV is employed, the network has relatively low fairness index; for instance, at a traffic load of 500 Kbps, the fairness index achieved is 0.3; meaning, only 3 out of 10 clients significantly contribute towards the aggregate throughput and the rest are starved with low throughput. The fairness index with MMESH is relatively close to AODV while the fairness index of MMR is marginally better compared to AODV. The fairness index with ASMRP is closer to 1, meaning, all the clients fairly contribute towards the aggregate throughput of the network. Use of multiple routes and concurrent transmissions along with the congestion-aware traffic partitioning algorithm result in uniform resource allocation and thus accomplish a fair distribution.

4.5.4 Delay Distribution

Figures 4.6(a-c) depict the per packet end-to-end delay distribution of AODV, MMESH, MMR and ASMRP for varied traffic rates. Among AODV, MMESH and MMR - MMR has less delay compared to others and this could be due to its multi-route and multi-channel transmission strategy.

It can be observed that ASMRP outperforms other protocols for delay distribution at all traffic loads where the end-to-end packet delays are much reduced, as illustrated in the Figures 4.6(a-c). For instance, the average delay of nearly 98% of the packets is less than 1 second at 1000 Kbps as in Figure 4.6(c). This shorter delay is because of the full-duplex nature of radios in ASMRP where all the MRs receive fair chance of transmission with very few collisions and have shorter queue lengths. Also, the pre-established routes for transmission at MRs coupled with congestion-aware traffic distribution contribute to this less delay.

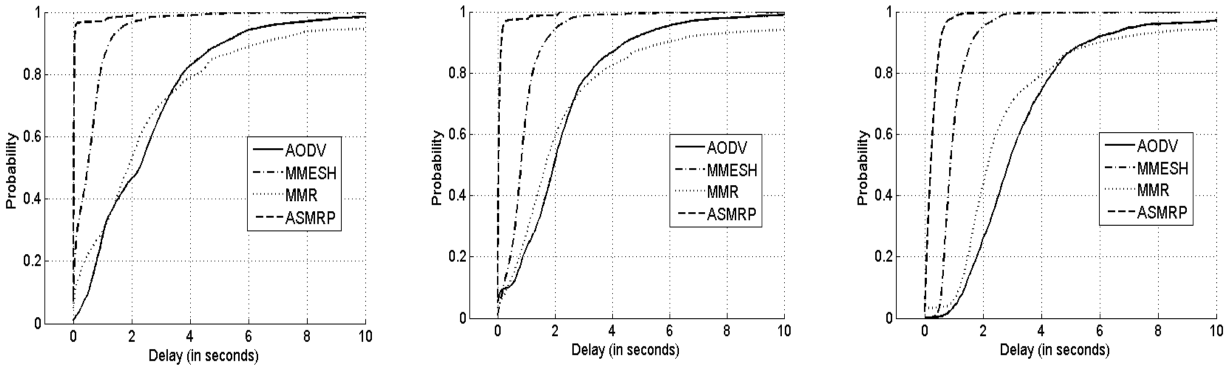


(a) Offered Load of 400 Kbps
(b) Offered Load of 500 Kbps
(c) Offered Load of 1000 Kbps

Figure 4.6 CDF of Packet Delays with Varying Traffic Rate

Delay Performance with Fluctuating MRs

Although WMNs are relatively static, topological changes, in the form of network upgrades by the service providers can occur, where new MRs may be added or existing MRs may be removed or shifted. These changes result in unstable links and the established routes may become stale and unsuitable for transmitting traffic. Figures 4.7(a-c) illustrate the delay distribution comparison of AODV, MMESH, MMR and ASMRP routing protocols in the presence of such occasional topological changes that result in temporary failure of MRs. From the figures, we notice that ASMRP protocol supersedes in its performance, even in the scenarios with failed MRs which can be attributed to its neighbor maintenance which learns about a failed neighbor and routes around it.



(a) Offered Load of 400 Kbps

(b) Offered Load of 500 Kbps

(c) Offered Load of 1000 Kbps

Figure 4.7
CDF of Packet Delays with Varying Traffic Load with the Presence of Some Failed MRs

Performance Benefit of Multi-channel vs. Multi-path

To illustrate the performance benefit from multi-channel vs. multi-path environment, we compare our standard ASMRP algorithm to ASMRP with single radio and single-channel, denoted by ASMRP-FC from hereon. This single channel is assigned with varied channel bandwidths of up to four times the bandwidth of ASMRP channel. Though we realize that it is not permitted by standards, for theoretical analysis, we simulate ASMRP-FC with a channel rate of 4 times that of 802.11a channels in ASMRP and compare its effective throughput to ASMRP with the default 802.11a channel rate and four radios. We also perform similar analysis with ASMRP-FC and intermediate channel rates of 2 and 3 times the channel rate of ASMRP.

For simulation, we used a linear scenario of 20 MRs with adjacent MRs separated by 100m. The MR to left end is designated as the IGW. The other simulation parameters are the same as specified in Table 4.3. Figure 4.8 illustrates the results we obtain for aggregate throughput of the following architectures – ASMRP, ASMRP-FC, ASMRP-FC with twice, thrice and four-times the channel rate of ASMRP. As represented in the figure, we compute results for a single 5-hop, 10-hop and 15-hop chain flow and with applied loads of 1 Mbps and 2 Mbps. We observe that the effective throughput of ASMRP is superior to the ASMRP-FC architecture even with a fat-channel of 4 times the channel rate of ASMRP, thus verifying the fact that a single fat channel of multi-path single radio network is still inferior to the ASMRP algorithm which uses multi-radio and multiple separate, orthogonal, narrow channels.

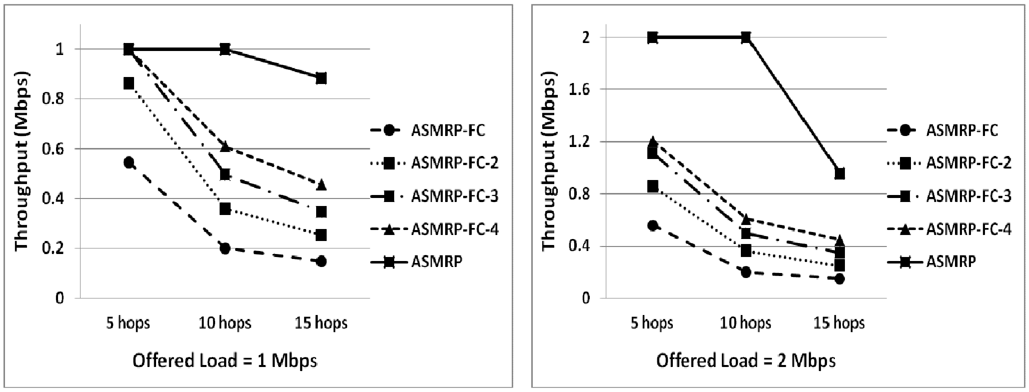


Figure 4.8
Aggregate Throughput for a Linear Flow
{Assume *x* to be the bandwidth of a channel in ASMRP}
ASMRP-FC : Single channel WMN with *x* bandwidth
ASMRP-FC – 2: Single channel WMN with 2*x* bandwidth
ASMRP-FC – 3: Single channel WMN with 3*x* bandwidth
ASMRP-FC – 4: Single channel WMN with 4*x* bandwidth

4.5.5 Traffic Partitioning Strategies

In the round-robin method of traffic partitioning, a MR selects its next hop out of a set of acceptable next hops in a successive manner. On the other hand, a congestion-aware splitting technique enables a MR to select its next hop based on the congestion level at those hops. From Figure 4.9, we observe nearly 15% improvement in the aggregate throughput of the network if congestion-aware algorithm is used rather than a round-robin strategy. Congestion aware strategy decreases the packet loss and ensures reliable data delivery to the destination.

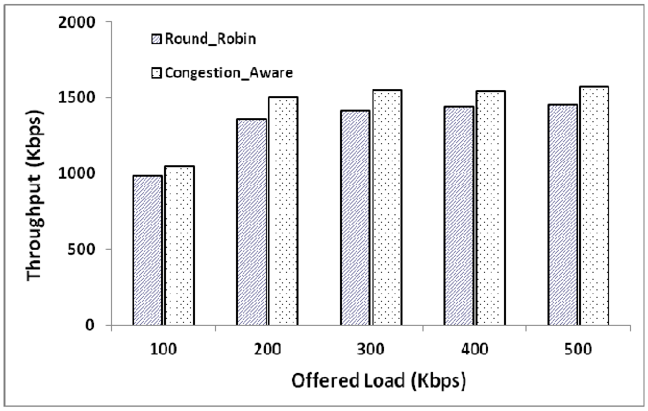


Figure 4.9
Illustration of Aggregate Throughput Improvement with Congestion-aware Algorithm

4.6 Related Work

Popular routing protocols for ad hoc networks such as AODV and DSR use the metric, number of hops, to decide their routing strategy. Shorter hop routes may not always prove optimal as revealed by Draves et al. in [47], particularly, when the wireless interface has multi-rate capability. In fact these shorter paths degrade the performance of the overall network. Draves et al. [19] also proposed Multi-Radio Link-Quality Source Routing (MR-LQSR) protocol for multi-radio multi-channel WMNs consisting of MRs equipped with equal number of interfaces and channels. MR-LQSR uses a novel routing metric, Weighted Cumulative Expected Transmission Time (WCETT), which enables the nodes to choose an optimal route that has a balance between channel variant hops and high bandwidth links. However, they do not consider the aspect of load balancing; involving traffic concentration and congestion on certain routes in the network. Also, MR-LQSR uses only a single path for transmitting traffic and will not utilize any of the additional available bandwidth on multiple channels in the neighborhood of a node.

Yang et al. [48] show that WCETT could result in routing loops in certain scenarios. They propose Load and Interference Balanced Routing Algorithm (LIBRA), which considers the intra and inter flow interference. However, even LIBRA does not consider congestion along its selected good paths and thus, may lead to suboptimal performance of the WMNs. The authors in [32] explore distributed channel assignment and a routing protocol, Hyacinth, for multi-channel WMNs consisting of MRs with two 802.11a Network Interface Cards (NICs). Hyacinth solves the problem of channel assignment by dividing the available NICs to communicate with the parent and with child MRs. The routing mechanism of Hyacinth creates a spanning tree with the gateway as root and considers three cost metrics - hop count, gateway link capacity and path capacity that influence the Quality of Service (QoS). Each node performs load balancing by periodically monitoring the channel usage in its neighborhood and uses a less loaded channel for transmission of traffic. Ramachandran et al. [33] propose a spanning tree based protocol, Ad-hoc On-demand Distance Vector Spanning-Tree (AODV-ST) which is a modification of the popular AODV protocol. AODV-ST uses Expected Transmission Time (ETT) [19] as the routing metric which is based on Expected Transmission Count (ETX) [49]. In their model, MRs construct a spanning tree corresponding to each gateway in the network and maintain a primary gateway through which they route their traffic. Load balancing in the network is achieved by periodically probing for a less loaded IGW and routing traffic through it. However, AODV-ST does not consider routing in multi-channel architecture, which leaves inter-flow and intra-flow interference as unexplored challenges.

MMR [50] is a multi-path source routing protocol that aims to eliminate co-channel interference between the routes by tuning onto different frequency bands. The route selection is based on the metrics - hop count, power budget, and number of disjoint nodes of different routes.

Specifically, MMR uses two routes for data transmission which are tuned onto different frequency bands. While the route discovery and maintenance takes place only on one channel, data transmission occurs using both channels. However, use of only two routes as considered by MMR could result in underutilization of network resources and limit the performance of a WMN, as the mesh connectivity could accommodate higher number of possible routes.

AOMDV [14] is another multi-path routing protocol that is proposed for ad hoc wireless networks. Although the protocol computes multiple loop-free and link-disjoint paths, only the primary single path is used for communications while the alternate paths are used in the case when primary path fails. In our work, we focus on load balancing and efficient resource utilization by spreading the traffic in the network. We believe AOMDV does not address any load balancing issues and it focuses on establishing routes for faster recovery from failures. We believe the performance gains of AOMDV vs. AODV to be inferior to those of ASMRP vs. AODV. For instance, in a static network, Marina et al. [14] illustrate that the throughput and delay performance of AODV and AOMDV are similar. As represented in the Figures 4.5(a) and 4.6, ASMRP significantly outperforms AODV in these performance measures for a static network. Even in a dynamic unstable network, it can be noticed from Figure 4.7 that delay performance of ASMRP over AODV is much higher than 2 fold improvement that AOMDV achieves; ASMRP is more towards 5 fold improvement as shown in Figure 4.7. We attribute this improvement to the fact that AOMDV still uses only one primary path at a time and uses one of the backup paths only when the primary path fails. Such a strategy often results in under-utilization of available network resources, as the available multiple routes are not being exploited simultaneously, which is inherent in ASMRP architecture.

Sheriff et al. [41] presents a routing metric, Channel Aware Multi-path (CAM), for multi-radio multi-channel WMNs which performs route selection based on inter-path and intra-path interference among the routes such that the end-to-end throughput is maximized. CAM assumes a static channel assignment scheme and computes the goodness of the paths based on the channel diversity and load distribution ratio on those paths. CAM is a weighted metric of independent path quality index and inter-path interference index. Based on WCETT, the independent path quality index component of CAM is the weighted average of the WCETT and the traffic distribution ratio on the paths under consideration. The inter-path interference index in CAM accounts for bottleneck channel transmission time which depends on the load (traffic) carried over the channel among all the channels on the set of paths considered.

4.7 Summary

In this chapter, we address two critical performance aspects of a WMN: (1) improving robustness and stability again rr0nxb23 st the weak wireless links, and transient or permanent channel outages; and (2) provisioning elegant load balancing technique to minimize congestion. To this end, we propose a novel protocol that opportunistically exploits multiple paths between MRs and IGWs and distributes traffic among them synergistically, to improve the overall performance of WMNs. We introduce a Neighbor State Maintenance module that periodically monitors the link quality between neighboring MRs and assists in ensuring route stability and better recovery from transient failures. We also introduce a smart traffic partitioning technique, congestion aware logic, which splits the flowing traffic at a MR to over multiple available routes in a manner that minimizes congestion in the network. We then incorporate the multi-radio architecture in the network where each MR is equipped with four radios. Through extensive simulations, we conclude that ASMRP considerably improves the performance of WMNs. We notice up to 460% improvement in the aggregate network throughput when compared to AODV.

Chapter 5. Dynamic Admission Policy for Wireless Service Providers Using Discrete-time Markov Decision Process Model

5.1 Introduction

Wireless Service Providers (WSPs) typically serve diverse user base with heterogeneous requirements and offer portfolio of services targeting their requirements. As explained in Section 1.3.8 of Chapter 1, service providers using WMNs can offer different service plans to their registered users and charge them service fee accordingly. The key network component in WMNs that would provide these services (by performing functions such as call admission control, traffic policing and shaping) is the IGW. When a user request arrives at the IGW, the IGW can either accept or deny such request based upon its pre-defined utility optimization function and its existent parameters like available resources, load level, etc. which we elaborate further later in this chapter. Such user admission policies can be broadened to any Wireless Service Providers (WSPs) with similar motive.

Consider the following scenario: A WSP plans to expand its market by introducing a new hub in a metropolitan / residential area. It desires to allocate its limited resources to various user requests within a time duration. The time duration set might be defined because of WSP’s preset goals by management or its stake holder preferences to maximize use of its resources, and be fully functional by end of such allocating time horizon. The objective of WSP is to maximize its revenue by the end of this finite time horizon, by optimally selecting admissible users and allocating corresponding resources. Let us assume that the expected demand is significantly larger than the available resources.

Depending on its offered service portfolio, requests for certain service classes would generate more revenue as compared to others. So for a WSP, if a user request for a service class arrives during the allocating horizon, a prudent selection strategy is required which decides whether or not to accept the request depending on several factors like - remaining available resources, service charge of the requested service class, remaining available time in allocating horizon, and the expected future demand pattern of other users' requests. This is because, if WSP exhausts all its available resources in advance by accepting more lower revenue generating requests as and when they arrive, any higher revenue service requests have to be denied later, thus resulting in net lower revenue. On the other hand, if WSP denies larger number of lower rewarding requests to reserve resources for probable future higher revenue generating requests which may subsequently not show up, then it will eventually result in under-utilized resources by the end of allocating time horizon. And so, WSP loses the corresponding marginal revenue from the excess denied lower revenue generating requests that it would otherwise have gained. Hence, WSP needs to have an optimal user admission selection strategy; correspondingly allocating its resources in such a way that maximizes the total expected revenue during the allocating time horizon.

In this work, our approach in building a framework for optimal user admission model for WSPs is based on yield management principles that are widely applied in the airline industry. Briefly, yield management or revenue management deals with maximization of revenue from relatively fixed perishable resources that allow price segmentation, by selling the individual resources to the right customers at the right time and for the right price.

We believe that there are several other similar fields where our proposed model could be applicable. Some examples include on-demand IT services such as Web content offloading, software application-server allowing users to run applications, etc. [51]. Additional application scenarios and services of yield management to telecommunication industry are provided in [52] and [53]. Another emerging field where our proposed model could be applicable is in the area of coordinated dynamic spectrum allocation [54]. In such a scenario, a centralized entity called spectrum broker monitors the allocation of unused and unallocated spectrum, called Coordinated Access Band (CAB) to WSPs or other users based on their need in a dynamic manner. From the spectrum broker's point of view, the CAB needs to be allocated within a specific time-frame, after which it may be considered not utilizable and thus may not yield any revenue. So, it is important for the spectrum broker to choose a right mix of WSPs based on the premiums they would pay so as to generate maximal incremental revenue. The admission policy proposed in this work is an approach which could be followed by the spectrum broker in this scenario to ensure maximal revenue while efficiently utilizing available excess spectrum.

The rest of the chapter is organized as follows. In Section 5.2, we present a brief overview of the related work in this field. In Section 5.3, we discuss about yield management framework and its application to WSPs. We then formulate the proposed Markov Decision Process revenue maximization model for WSPs in Section 5.4. In Section 5.5, we explain the proposed model and its working logic through some illustrative numerical examples. We present our simulation results and provide a detailed performance analysis of the proposed optimal user admission policy in Section 5.6. Finally, we conclude our chapter and discuss future research directions in Section 5.7.

5.2 Related Work

Hayel et al. [51] analyze a yield management model for IT on Demand services (e-commerce or data processing centers) with preset price structure and fixed job sojourn times. The authors consider a resource allocation framework with a pool of homogeneous nodes that are to be allocated to users willing to pay different fees based on their arrival pattern and their service/price preferences. For this, they employ a policy which maintains a maximal number of users for each fee class and once they are exhausted during the allocation time, the resources from the next higher fee class are available for further allocation in the system. Their key objective is to determine optimal quantities at each fee level and the available offerings for the arriving customers such that the potential revenue from customers is maximized. The authors illustrate that the revenue can be maximized by having higher number of fee offerings to users if the demand is higher. The same authors extend their work in [55] and propose a unified framework for real-time yield management of e-services. The framework considers dynamic change in an e-firm's offerings (number of service classes, corresponding service level and prices) based on volatile market conditions. They formulate an optimization problem which provides the optimal number of service classes that should be offered, their corresponding resource allocation and prices for the e-firm.

Yield management principles are employed in [56] proposing a method for Internet Service Providers (ISPs) that could increase their revenue from customers with stochastic arrival and departure patterns. The paper presents a model based on continuous time Markov Decision Process for allocating modem capacity to arriving user log-on requests for the internet access. Their model considers only two classes of arriving customers: Platinum and Gold customer segments.

In [57], online management of QoS and provider revenue is performed for CDMA cellular networks by adaptively controlling system parameters to changing traffic conditions. They propose a call admission controller based on a Markov model and a bandwidth degradation scheme for real-time traffic. Specifically, they consider two levels of priority for real-time calls arriving into the cellular system - high priority and low priority calls. To maximize the network's revenue, the bandwidth for existing low priority calls is degraded for a temporary time period if high priority calls arrive and the available bandwidth is not sufficient to support them.

The authors in [58] deal with dynamic pricing strategies for connection-oriented services in wireless systems where the network operators charge the users based on the network usage per time unit. They model the user demand and the call-duration as a function of service price. Further, they use Markovian techniques to represent such a system and devise an optimal linear pricing scheme. They show that the model provides better Quality of Service (QoS) and improves network operator's profit as compared to a flat-rate policy. An approach to maximize income for a telecommunication network provider is proposed in [59] by offering multiple service classes and controlling the demands using pricing and resource allocation technique.

The impact of user migration between service providers with resource management algorithms for service differentiated CDMA networks is studied in [60]. Basically, the proposed algorithm, called CBPAR, incorporates the user migratory behavior into admission control and power management algorithms such that the revenue loss due to user migration is minimized at a service provider. In their analysis of CBPAR, it has been noticed that when the air-interface congestion is severe, i.e., when a Base Station is serving a large number of users, then lower class users are dropped to maintain service quality for higher class users. They also show that the ideal customer composition in the system with which a service provider's utility is maximized is when the number of users requiring high quality service is fewer as compared to users requiring lower quality service.

Auction based price discovering models are proposed in [61] for dynamic pricing of differentiated wireless services in a cellular network. The uniform pricing auction and discriminatory pricing auction models are compared against flat-rate pricing model and the authors establish through simulation that these models generate higher revenue than a flat-rate pricing model.

In [62], the authors describe a model for resource allocation and pricing of downlink resources at Base Station for either time-slotted or CDMA systems. They show that in order to maximize revenue, the Base Station should consider discriminatory pricing based on varying channel quality of users. Further, the users compete for resources through bidding process in the considered model. Base Station selects users based on their bids and determines an optimal resource allocation strategy for maximizing its total revenue.

A communication network with heterogeneous customers such as data and voice users which are delay-tolerant and delay-sensitive respectively is analyzed in [63]. Users in their model join the network as long as their utility normally being a function of queuing delay is greater than the price of offered service. The model determines the price of the services for these two types of users such that the provider's profit is maximized.

The authors in [64] study pricing of differentiated services and its impact on the choice of service priority at equilibrium. Performance of two types of connections, TCP and CBR, is considered over a bottleneck link. The choice of a particular service class by an application depends on the utility (obtained by various performance measures such as throughput, average queue size etc.) and the cost for a given priority class. The authors model the problem as a noncooperative game and establish conditions for equilibrium to exist. They numerically study the pricing problem of how the network should set prices for offered service classes so as to maximize the network's benefit.

Shamik et al. [65] propose an economic framework for dynamic spectrum allocation to service providers. The framework also includes pricing mechanism for service providers. They use a knapsack based auction model for the allocation of dynamic spectrum to the service providers such that the spectrum usage and received revenue are maximized.

Lee et al. [66] develop a discrete-time dynamic programming model for finding an optimal booking policy for airline seat inventory control with various fare classes. At each instance of booking request arrival, a decision is made whether to accept or deny a request for a seat based on the type of fare class to which the request came, remaining available seats in the flight, and the time at which the request arrived. They also extend their model to determine an optimal booking policy for the case when a request for multiple seat bookings arrives.

Our objective in this chapter is to propose an optimal user admission / allocation policy for a WSP to maximize its total accrued revenue. The WSP is assumed to have finite capacity and there is a defined time limit for the allocation of the capacity. Our approach is similar to one used in [66], however following are the key differences. First, we apply the yield management principles to the field of service provisioning by WSPs and model it relevantly incorporating its applicable parameters. Second, in our model, each service class would consume varied amount of resources as compared to models proposed for the airline industry. Third, in our model the service charge for a given class could vary over the allocating time period. The characteristics of our proposed model are further explained in the next section.

5.3 Characteristics of Yield Management and Parallelism to Proposed Model

In this section, we explain how WSPs fit into yield management model by explaining parallelism between the characteristics for WSPs and to those of an airline industry where the yield management principles are widely accepted.

1) **Finite Amount of Resources** - One of the key characteristics for yield management is to have limited pre-defined number of resource units that can be allocated to various users. For instance, in airline industry, the number of seats available in a given flight is bounded. Similarly, WSPs possess limited resources such as bandwidth, number of available channels in a cellular network, limitation on total number of customers they can serve, downlink power, percentage of time a user can have exclusive access of a channel, or codes in a CDMA cellular system. So, WSPs need to manage available scarce resources in a judicious manner so as to maximize their obtainable profit.

2) **Perishable Resources** - A resource is considered perishable if it becomes unusable or if its value deteriorates significantly after a preset time. In airline industry, an empty seat in a flight after its departure is considered perished as it can no longer be used for that travel segment. Similarly, for WSPs which need to allocate their resources within a preset time frame, the bandwidth at the WSPs can be considered perishable as any un-allocated bandwidth is considered unutilized after the end of the allocation time period.

3) **Limited Resource Allocation Time Horizon** - In airline industry, the seat reservations for a given flight are done over a fixed booking / reservation horizon.

Similarly, we assume that the resource allocation to users at the WSPs in our proposed model occurs over a finite time horizon.

4) **Ability to Segment Market Space** - WSPs can use their resources to provide heterogeneous services for applications such as web-browsing, VoIP, webinars, streaming videos, etc. This possibility of user requests with regards to different service requirements enables us to perform segmentation.

5) **Ability to Price Identical Resources Differently** - WSPs can charge users differently based on offered services and/or competitive aspects. WSPs may use the same resource such as bandwidth to serve variety of user applications. Thus, WSPs have the ability to gain different revenue from identical resources, based on the application type and the services for which the resources are used.

6) **Fluctuating Demand** - Variability is present in the arrival of user requests at the WSPs and the services they need. Thus, the demand for resources from users fluctuates over time. For example, in cellular networks which are resource-constrained, the user demand for the radio resource fluctuates due to the presence of some peak and off-peak periods [67].

5.4 Problem Formulation Using Markov Decision Process Model

In this section, we formulate the observed problem and explain its features.

Consider a WSP with fixed pool of resources that needs to maximize its total expected revenue within a finite time horizon. In our model, we formulate the problem as a finite-horizon, discrete-time Markov Decision Process (MDP) in which the state variable is the available resources at WSP at a given point of time. Following are the key characteristics of our model:

1) We assume two time horizons at the WSP - the first being resource allocating time horizon $H_{allocating}$, during which the resources of WSP are allocated to the users' requests, and the second being resource usable time horizon, H_{usable}, the time up to which the allocated users can access the resources of WSP from the time they are allocated. Therefore, $H_{allocating} \leq H_{usable}$. This is illustrated in Figure 5.1.

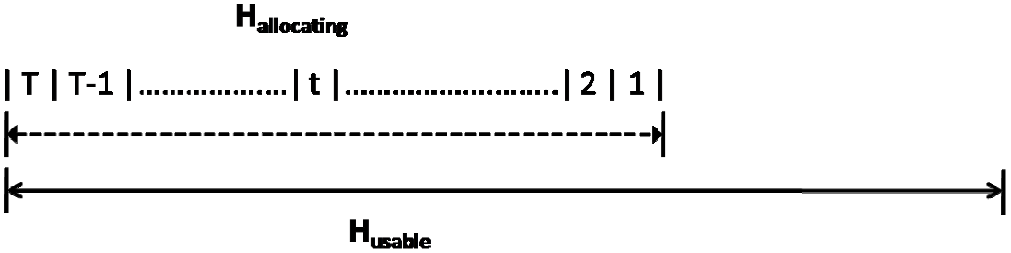

Hallocating
| T | T-1 |...................| t |..........................| 2 | 1 |
Husable

Figure. 5.1. Allocating and Usable Time Horizons

2) We assume that the resource allocating time horizon, $H_{allocating}$ at the WSP is finite and is divided into a number of tiny discrete time periods. These time periods are small enough such that in each time period, either there is a chance for WSP to receive at most one user request, or no such request at all. We index these time periods at the WSP in reverse chronological order from T to 1, $\{T, T-1, ...t, ..., 2, 1\}$, where T corresponds to the start of allocating period and 1 corresponds to the end of allocating period, after which no further resource allocations can be done. Any un-allocated resources after end of the allocation period (meaning, after time period 1) are considered as perished, as they cannot be further allocated and thus unutilized.

3) Let the total resources that are available for WSP at the start of allocating time horizon, i.e., time period T be R units, and available resource units at a given time period t be R_t units. We assume that the WSP offers K service classes, say QoS levels, $\{1, 2, ..., K\}$ to its users. In our model, let $\{r_1, r_2, ..., r_K\}$ be the resource

requirement at the WSP corresponding to service classes $\{1,2,...,K\}$. Any arriving user request at the WSP is for one of these offered service classes. For instance, an arriving request belonging to class *i* would require r_i units of resources from the WSP. We assume in our model that these resource units are integers. We also assume that users of the network individually choose among the service classes offered, i.e., choose which level of service he/she desires from the set of service classes offered by the WSP.

4) Each arriving user request will need to pay a corresponding service charge, if it is accepted by the WSP. Let the service charges at WSP to its users for the *K* offered service classes be $\{c_1, c_2, ..., c_K\}$ respectively, and the service classes are organized such that $\{c_1 \geq c_2 \geq ... \geq c_K\}$. We initially assume that service charge c_k for a given class *k* will not change over the allocating time horizon; however, we will later relax this assumption.

5) We consider the arrival of user requests at WSP as stochastic and independent in nature. In each time period $t \in H_{allocating}$, let the probability with which a user request belonging to service class *k* arrives is denoted by p_k^t .In other words, the user request arrival probabilities belonging to service classes $\{1,2,...,K\}$ at time period *t* is given by $\{p_1^t, p_2^t, p_3^t, ..., p_k^t\}$. Let the probability of there being no arrival of any request at WSP during time period *t* be denoted by p_0^t, where

$$p_0^t = 1 - \sum_{i=1}^{K} p_i^t$$

To summarize, at any given time period $t \in H_{allocating}$, there is a probability p_i^t with which a user request belonging to service class $i \in \{1,2,...,K\}$ arrives. If such a request is honored, then it depletes r_i resource units from the available resource pool and would generate a service charge of c_i for the WSP at that time period.

Some additional assumptions that we make while formulating our model are:

- The set of arrival probabilities corresponding to the users' requests belonging to various service classes at each time period are known *a priori* at the WSP. For instance, a WSP will have knowledge about the past demand patterns of users' requests from other similar established hubs.
- The WSP has the ability to either accept or deny an arriving user request at any given period of time.
- Once a user request is accepted by the WSP, it will be allocated resource units dictated by its respective service class and are dedicated through remaining allocating time horizon $H_{allocating}$, and hence assumed unavailable. We also assume that once the user is admitted, the user has access to the resources from the point of admission.
- The set of service classes and their corresponding service charges are advertised by the WSP and the users are aware about available service classes to choose from.

A WSP obtains revenue from all accepted user requests during the allocating time horizon, based on the requests' service charges. The goal of our work is to develop an optimal accept / deny decision policy for users' requests to the WSP during the allocating time horizon, so as to maximize its total overall expected revenue. The optimal accept / deny decision policy of WSP is computed based on the *(i)* request's service class, *(ii)* current time period, *(iii)* remaining available resources, *(iv)* arriving request's service charge, and *(v)* expected future revenue if or not the user request is accepted.

We formulate our problem as a Markov Decision Process (MDP) which follows a discrete-time dynamic programming approach for making the accept / deny decisions of the requests in order to obtain an optimal user admission / resource allocation policy to maximize the expected revenue over the allocating horizon. Let $\Phi_T(R)$ represent the total expected revenue the WSP earns from admission of user requests during the entire allocating time horizon, $H_{allocating}$, with *R* available resource units. Our goal is to maximize $\Phi_T(R)$.

Now, let $\phi_t(r)$ denote the expected revenue that WSP can get with *r* available resource units and from time period *t* through the end of allocating horizon, i.e., $\{t, t-1, ..., 2, 1\}$. We first develop our model for constant service charge scenario for a given class over allocating time horizon and later develop the model for varying service charge scenario.

5.4.1 Constant Service Charge for a Given Class over Allocating Time Horizon

During a decision period *t*, WSP will have *r* available resource units and if a request belonging to a service class *i* arrives during this time period *t*, WSP needs to decide whether or not to accept the arriving request.

- If the request is accepted, r_i resource units are allocated to it and the remaining resource units at the WSP will be $(r - r_i)$ for the next $(t-1)$ decision periods of the allocating time horizon. The revenue that the WSP earns from accepting this request is c_i. So, the total expected revenue that WSP earns by accepting the request at time *t* can be given by:

$$\phi_t(r) = \phi_{t-1}(r - r_i) + c_i.$$

- On the other hand, if the request is denied, then the remaining resource units at the WSP will still be *r* for the next $(t-1)$ decision periods and there will be no added revenue that the WSP earns during this time period. In this case, the total expected revenue that WSP earns by denying the request at time *t* can be given by:

  $$\phi_t(r) = \phi_{t-1}(r).$$

To maximize WSP's total expected revenue, such a user request belonging to service class *i* during time period *t* that consumes r_i resource units with a service charge value of c_i is *accepted* if and only if:

$$\phi_{t-1}(r - r_i) + c_i \geq \phi_{t-1}(r). \qquad (5.1)$$

On the other hand, if

$$\phi_{t-1}(r - r_i) + c_i < \phi_{t-1}(r), \qquad (5.2)$$

the arriving request is *denied*.

To summarize, the optimal $\phi_t(r)$ for the above scenario can be written as:

$$\phi_t(r) = \max(\phi_{t-1}(r - r_i) + c_i, \phi_{t-1}(r)). \qquad (5.3)$$

Using the Markov Decision Process principles, the maximal expected revenue from time period *t* through the end of allocating time horizon $H_{allocating}$, when there are *r* resource units available at the WSP can be computed recursively with the following equations:

$$\phi_t(r) = \sum_{i=1}^{K} p_i^t \max(\phi_{t-1}(r - r_i) + c_i, \phi_{t-1}(r)) + p_0^t \phi_{t-1}(r) \qquad \text{for} \qquad R \geq r > 0; t > 0, \qquad (5.4)$$

$$\phi_0(r) = 0 \qquad \text{for} \qquad R \geq r > 0, \qquad (5.5)$$

$$\phi_t(0) = 0 \qquad \text{for} \qquad t > 0, \qquad (5.6)$$

where,

- r = number of resource units remaining at the WSP at time of consideration, i.e., time period *t*.
- R = maximum amount of resource units at the WSP.
- K = number of possible service classes supported by the WSP and into which arriving user requests can be categorized.
- i = service class to which the arriving request belongs to, $i \in \{1,2,...,K\}$.
- r_i = number of resource units required from the WSP for a user request belonging to service class *i*.
- c_i = service charge that would be collected from the arriving user request belonging to service class *i*, $i \in \{1,2,...,K\}$.
- p_i^t = probability that the arriving user request at the decision period *t* belongs to *i*th service class.
- p_0^t = probability of having no user request arrival during decision period *t*.
- $\phi_t(r)$ = maximal expected net revenue that can be earned by the resource allocation to the service requests over the periods *t* through 0 with *r* resource units still available at time period *t*.

Objective Function for Maximizing the Expected Revenue:

For the WSP with available resource units of *R*, the objective is to maximize the total expected revenue over the allocating time horizon from period *T* through 1, starting with the state variable *R*. So, the expected total optimal revenue would be $\Phi_T(R)$ which can be obtained by substituting these relevant parameters in Equation 5.4 and computing recursively. Figure 5.2 shows the logical flow used in this model.

5.4.2 Varying Service Charge for a Given Class over Allocating Time Horizon

In the previous section, we assume that each arriving user request will pay a corresponding service charge c_k for a given class *k*. This service charge for class *k* is assumed to remain the same over the allocating horizon $H_{allocating}$. However, in practical WSP applications, since the users start utilizing the resource units from the time of allocation through the usable time horizon H_{usable}, the maximal amounts of time for which a user can access the WSP's resources depend on when the user arrives. To explain further, as illustrated in Figure 5.3, a user request accepted at a time period closer to start of the allocating time horizon $H_{allocating}$, here time period *T*−1, will have larger time window to access the WSP's resources (i.e., from period *T*−1 through end of H_{usable}),

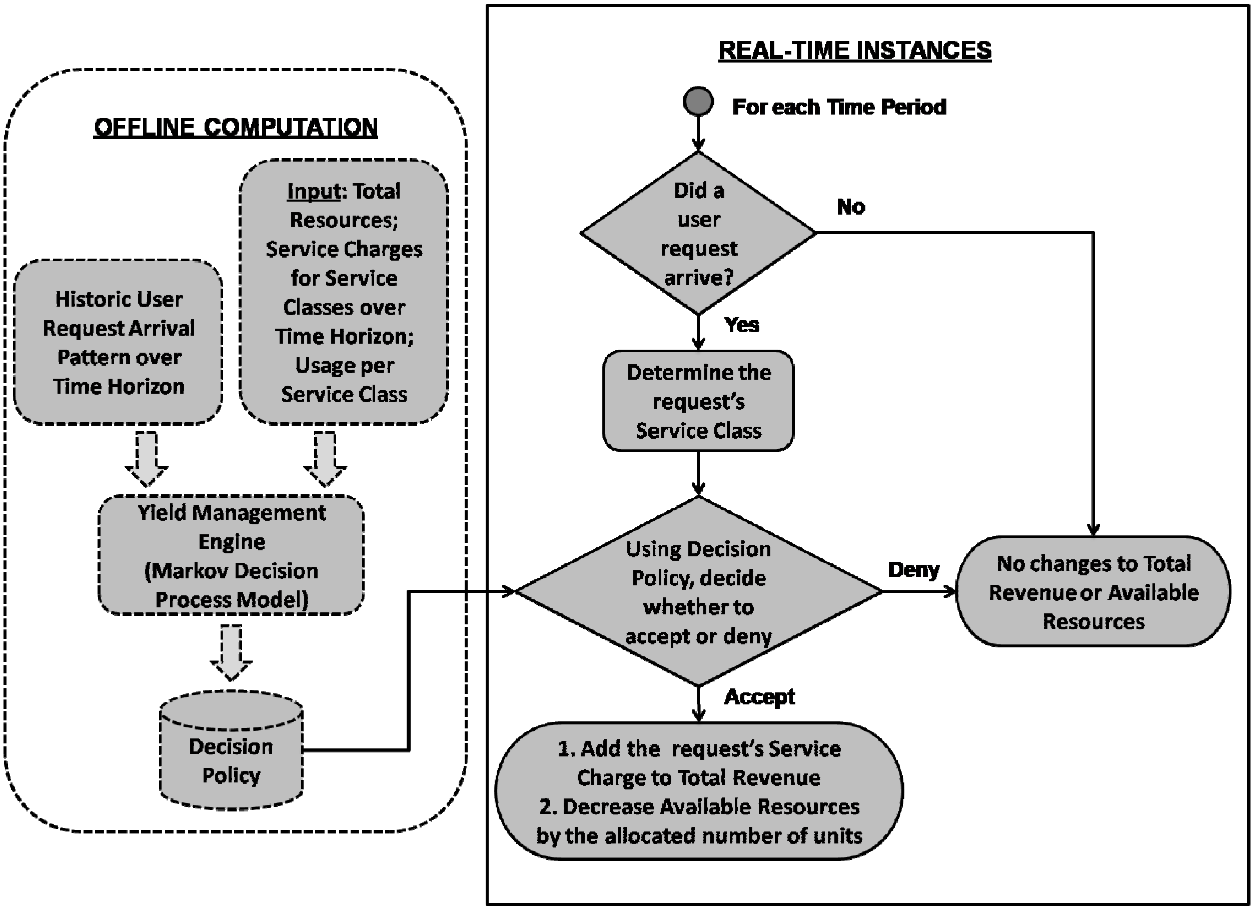

OFFLINE COMPUTATION
Historic User Request Arrival Pattern over Time Horizon
Input: Total Resources; Service Charges for Service Classes over Time Horizon; Usage per Service Class
Yield Management Engine (Markov Decision Process Model)
Decision Policy
REAL-TIME INSTANCES
For each Time Period
Did a user request arrive?
No
Yes
Determine the request's Service Class
Using Decision Policy, decide whether to accept or deny
Deny
No changes to Total Revenue or Available Resources
Accept
1. Add the request's Service Charge to Total Revenue
2. Decrease Available Resources by the allocated number of units

Figure. 5.2. Logic Diagram of the Proposed Model

where as a user request accepted at time period towards the end of $H_{allocating}$, here time period 2, will have relatively shorter accessible time window (i.e., from period 2 through end of H_{usable}). Or, there could be a scenario where WSP may offer advertised specials over certain periods of time and may collect lower service charges for that time period.

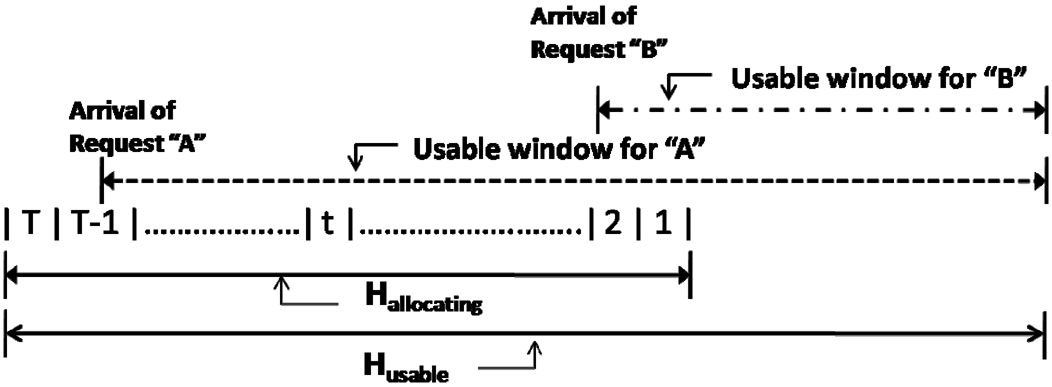


Figure 5.3. Varying Service Charge

Due to possibility of such scenarios, it is acceptable for the WSPs to charge different prices at various time periods of allocating horizon. So, for the same service class, we relax our earlier assumption of having fixed service charge for class *k* over the allocating horizon $H_{allocating}$, and vary the pricing based on the considered time period. Let c_k^t be the service charge that would be collected from the arriving user request belonging to service class *k* at time period *t*. Then, using the Markov Decision Process principles, the maximal expected revenue from time period *t* through the end of allocating time horizon, when there are *r* resource units available at the WSP, can be computed recursively with the following modified equations:

$$\phi_t(r) = \sum_{i=1}^{K} p_i^t \max(\phi_{t-1}(r-r_i) + c_i^t, \phi_{t-1}(r)) + p_0^t \phi_{t-1}(r) \qquad \text{for} \qquad R \geq r > 0; t > 0, \qquad (5.7)$$

$$\phi_0(r) = 0 \qquad \text{for} \qquad R \geq r > 0, \qquad (5.8)$$

$$\phi_t(0) = 0 \qquad \text{for} \qquad t > 0, \qquad (5.9)$$

where,

- c_i^t = service charge that would be collected from the arriving user request belonging to service class *i*, $i \in \{1,2,...,K\}$ at time period *t*.
- and, the other parameters of this equation have the same meaning as those in Equation 5.4.

In this case, to maximize WSP's total expected revenue, a user request arriving during time period *t* in service class *i* and consuming r_i resource units and pays a service charge value of c_i^t - is *accepted* if and only if

$$\phi_{t-1}(r - r_i) + c_i^t \geq \phi_{t-1}(r). \tag{5.10}$$

On the other hand, if

$$\phi_{t-1}(r - r_i) + c_i^t < \phi_{t-1}(r), \tag{5.11}$$

the arriving request is *denied*.

The optimal $\phi_t(r)$ for this above scenario is given by:

$$\phi_t(r) = \max(\phi_{t-1}(r - r_i) + c_i^t, \phi_{t-1}(r)). \tag{5.12}$$

The expected total optimal revenue over the allocating time horizon from period *T* through 1, starting with the state variable *R* would be $\Phi_T(R)$ which can be obtained by substituting relevant parameters in the Equation 5.7.

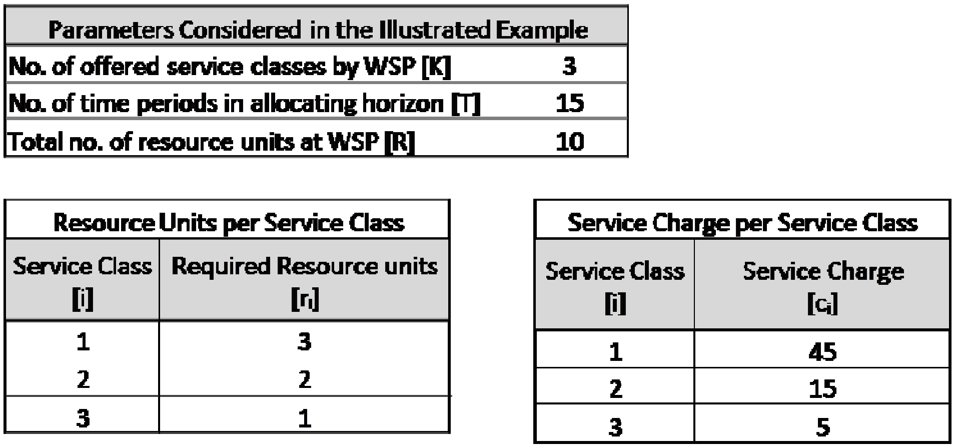
5.5 Illustration of Decision Policy Computation through Numerical Examples

We explain the proposed discrete time Markov Decision Process algorithm methodology with the following examples. First, we discuss the computation of the optimal decision policy at WSP for the scenario where the service charge for a given class remains constant over the allocating time horizon as explained in Section 5.4.1. Later, we illustrate an example scenario where the service charge for a given class varies over the allocating time horizon as explained in Section 5.4.2, which is more relevant to the applications supported by WSPs in real world. We present a simpler example in this section to help illustrate this decision policy, and in a later section, we use a larger numerical model to analyze the performance results.

5.5.1 Constant Service Charge over Allocating Time Horizon

In this example, we consider a WSP with R = 10 resource units, offering K = 3 different service classes which needs to allocate these resources over the allocating time horizon with T = 15 discrete time periods. The service charges and the required resource units corresponding to the offered service classes are assumed as given in Table 5.1.

Table 5.1. Example Parameters for Constant Service Charge Scenario



Parameters Considered in the Illustrated Example	
No. of offered service classes by WSP [K]	3
No. of time periods in allocating horizon [T]	15
Total no. of resource units at WSP [R]	10

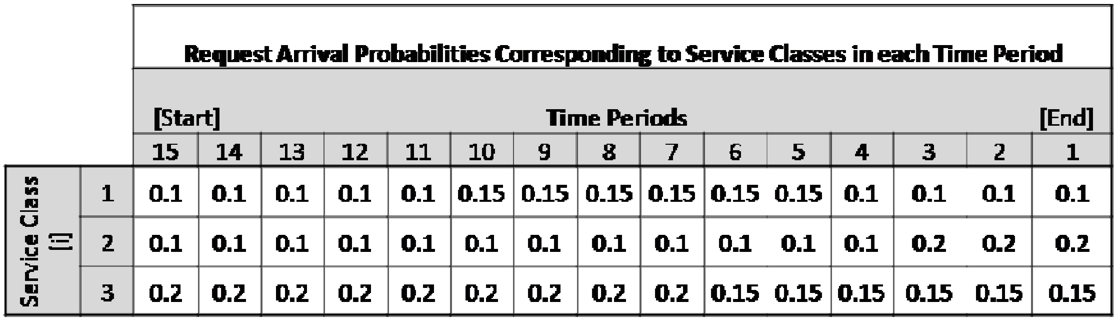
Resource Units per Service Class	
Service Class [i]	Required Resource units [r_i]
1	3
2	2
3	1

Service Charge per Service Class	
Service Class [i]	Service Charge [c_i]
1	45
2	15
3	5

As explained in Section 5.4, we assume that in each decision time period, at most one user request arrives at the WSP and that the request arrival probabilities of various service classes over the allocating time horizon are known in advance. The request arrival probabilities considered in this example are given in Table 5.2.

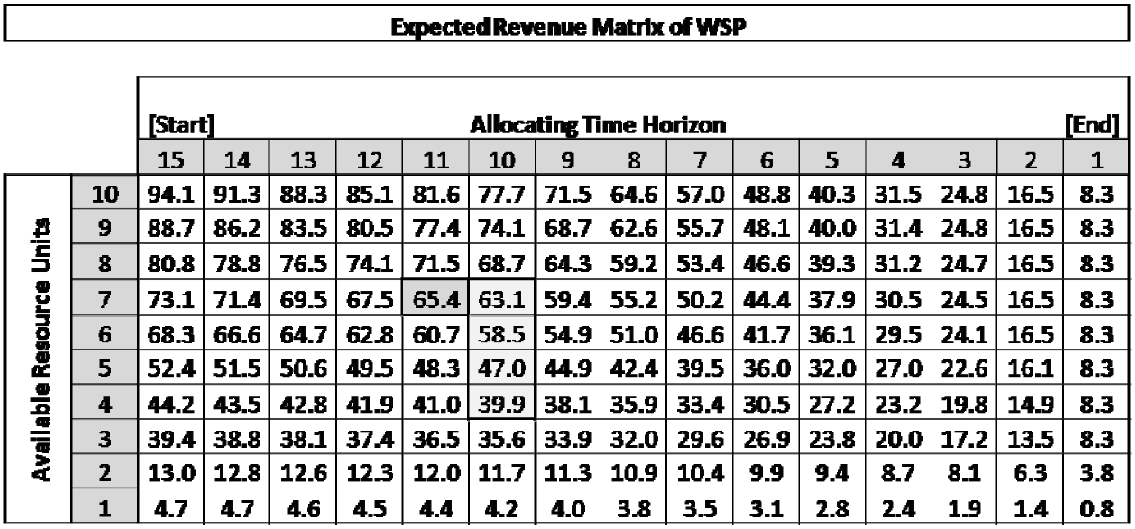
Using the recursive optimizing Equations 5.4, 5.5, and 5.6, we now compute the maximal expected revenue for the combination of each time period and available resource units, which facilitates determining a dynamic allocation policy for the WSP. The computed expected revenue matrix for the example is shown in Table 5.3.

Table 5.2. Request Arrival Probabilities for the Service Classes



		Request Arrival Probabilities Corresponding to Service Classes in each Time Period														
		[Start]						Time Periods								[End]
		15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
Service Class [i]	1	0.1	0.1	0.1	0.1	0.1	0.15	0.15	0.15	0.15	0.15	0.15	0.1	0.1	0.1	0.1
	2	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2
	3	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.15	0.15	0.15	0.15	0.15	0.15

Table 5.3. Computed Expected Revenue for Constant Service Charge Scenario

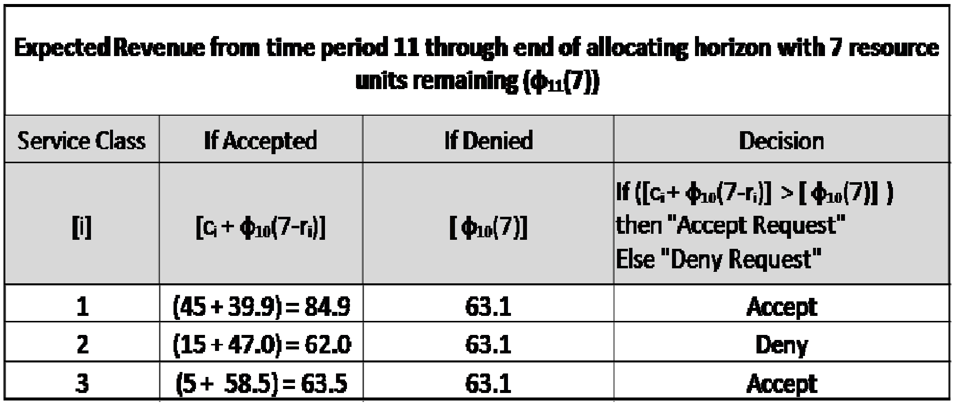


Expected Revenue Matrix of WSP

		[Start]						Allocating Time Horizon								[End]
		15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
Available Resource Units	10	94.1	91.3	88.3	85.1	81.6	77.7	71.5	64.6	57.0	48.8	40.3	31.5	24.8	16.5	8.3
	9	88.7	86.2	83.5	80.5	77.4	74.1	68.7	62.6	55.7	48.1	40.0	31.4	24.8	16.5	8.3
	8	80.8	78.8	76.5	74.1	71.5	68.7	64.3	59.2	53.4	46.6	39.3	31.2	24.7	16.5	8.3
	7	73.1	71.4	69.5	67.5	65.4	63.1	59.4	55.2	50.2	44.4	37.9	30.5	24.5	16.5	8.3
	6	68.3	66.6	64.7	62.8	60.7	58.5	54.9	51.0	46.6	41.7	36.1	29.5	24.1	16.5	8.3
	5	52.4	51.5	50.6	49.5	48.3	47.0	44.9	42.4	39.5	36.0	32.0	27.0	22.6	16.1	8.3
	4	44.2	43.5	42.8	41.9	41.0	39.9	38.1	35.9	33.4	30.5	27.2	23.2	19.8	14.9	8.3
	3	39.4	38.8	38.1	37.4	36.5	35.6	33.9	32.0	29.6	26.9	23.8	20.0	17.2	13.5	8.3
	2	13.0	12.8	12.6	12.3	12.0	11.7	11.3	10.9	10.4	9.9	9.4	8.7	8.1	6.3	3.8
	1	4.7	4.7	4.6	4.5	4.4	4.2	4.0	3.8	3.5	3.1	2.8	2.4	1.9	1.4	0.8

Once the expected revenue values are computed, the *accept / deny* decisions for arriving user requests belonging to different service classes are determined using logical comparative Equations 5.1 and 5.2. Example decision logic is shown in Table 5.4 for the [*time period, available resource units*] combination of [11, 7]. At this instant, if a request from service class 1 arrives and is accepted, the service charge that WSP earns in that time period will be $c_1 = 45$ corresponding to the service class, and the expended resource units for that service class would be r_1 = 3, thereby leaving 7 − 3 = 4 available resource units for the remaining time periods. In this case, the total expected revenue earned by the WSP from time period 11 is the sum of the service charge c_1 = 45, and the expected revenue from the next time period t = 10 with 4 available resource units or $\phi_{10}(4)$ = 39.9. Thus, the total expected revenue if accepted is 45 + 39.9 = 84.9.

Table 5.4. Decision Policy Logic

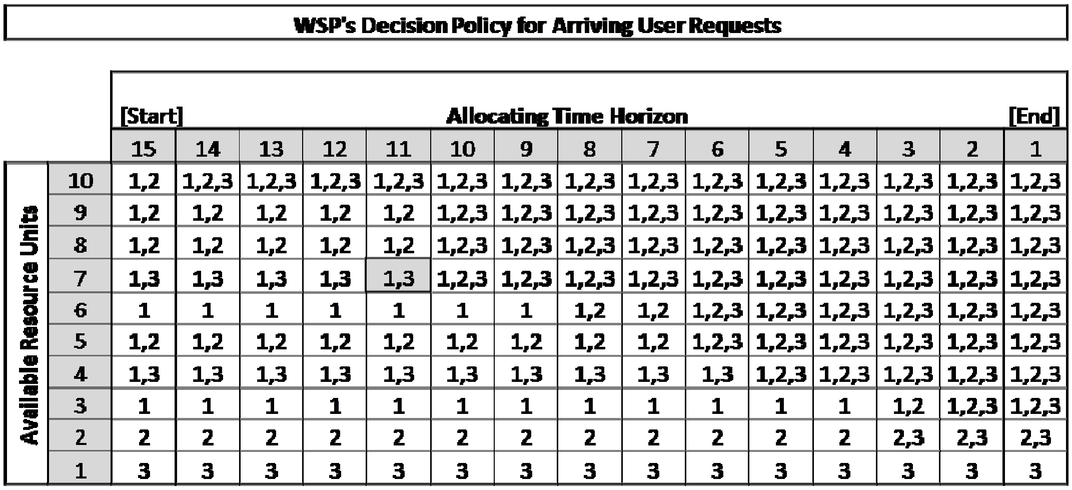


Expected Revenue from time period 11 through end of allocating horizon with 7 resource units remaining ($\phi_{11}(7)$)

Service Class	If Accepted	If Denied	Decision
[i]	$[c_i + \phi_{10}(7-r_i)]$	$[\phi_{10}(7)]$	If ($[c_i + \phi_{10}(7-r_i)] > [\phi_{10}(7)]$) then "Accept Request" Else "Deny Request"
1	(45 + 39.9) = 84.9	63.1	Accept
2	(15 + 47.0) = 62.0	63.1	Deny
3	(5 + 58.5) = 63.5	63.1	Accept

On the other hand, if the request from service class 1 arrives and is denied, then the expected revenue earned from time period 11 would be the expected revenue from time period 10 with all 7 resource units still available, which is $\phi_{10}(7) = 63.1$. From the Equations 5.1 and 5.2, since the total expected revenue by accepting this request (84.9) is greater than the expected revenue by denying the request (63.1), this arriving request from service class 1 is *accepted*. The decisions for the requests corresponding to other service classes are computed in a similar manner. The optimal decision policy matrix computed for each combination of time period and available resource units for the above example is shown in Table 5.5, which is used by the WSP for its decision making in real-time. In the table, each entry represents a decision parameter which indicates the service class requests that can be accepted at that time period and available resource unit's combination. For instance, the decision parameter [1,3] for [*time period, available resource units*] combination of [11,7] in the table implies that the WSP can accept the arriving user request if it belongs to either service class 1 or 3, and should deny the request if it belongs to service class 2.

Table 5.5. Decision Policy Computed at WSP for Constant Service Charge Scenario



WSP's Decision Policy for Arriving User Requests

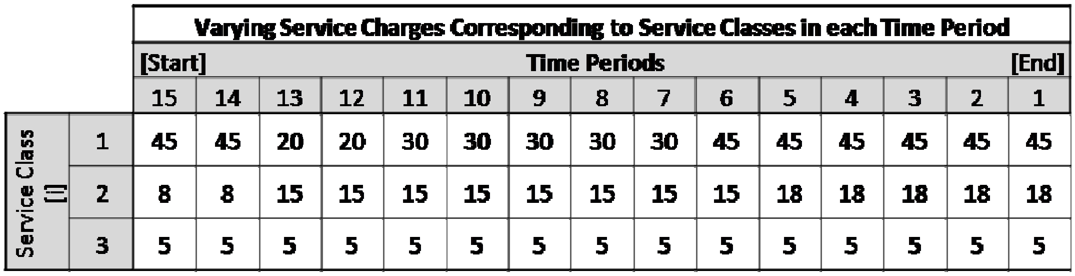
		[Start]							Allocating Time Horizon							[End]
		15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
Available Resource Units	10	1,2	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3
	9	1,2	1,2	1,2	1,2	1,2	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3
	8	1,2	1,2	1,2	1,2	1,2	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3
	7	1,3	1,3	1,3	1,3	1,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3
	6	1	1	1	1	1	1	1	1,2	1,2	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3
	5	1,2	1,2	1,2	1,2	1,2	1,2	1,2	1,2	1,2	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3
	4	1,3	1,3	1,3	1,3	1,3	1,3	1,3	1,3	1,3	1,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3
	3	1	1	1	1	1	1	1	1	1	1	1	1	1,2	1,2,3	1,2,3
	2	2	2	2	2	2	2	2	2	2	2	2	2	2,3	2,3	2,3
	1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3

In the decision policy matrix of such fixed service charge scenarios, it may be noted that, for a given capacity and service class combination, there exists a critical allocating time period, after which the service class's request is always acceptable [16]. It may be observed from Table 5.5 that for a resource level of 6 units - a service class 1 request is acceptable after time period 15; a service class 2 requests is acceptable after time period 8; and a service class 3 request is acceptable after time period 6 through the end of the allocating horizon.

5.5.2 Varying Service Charge over Allocating Time Horizon

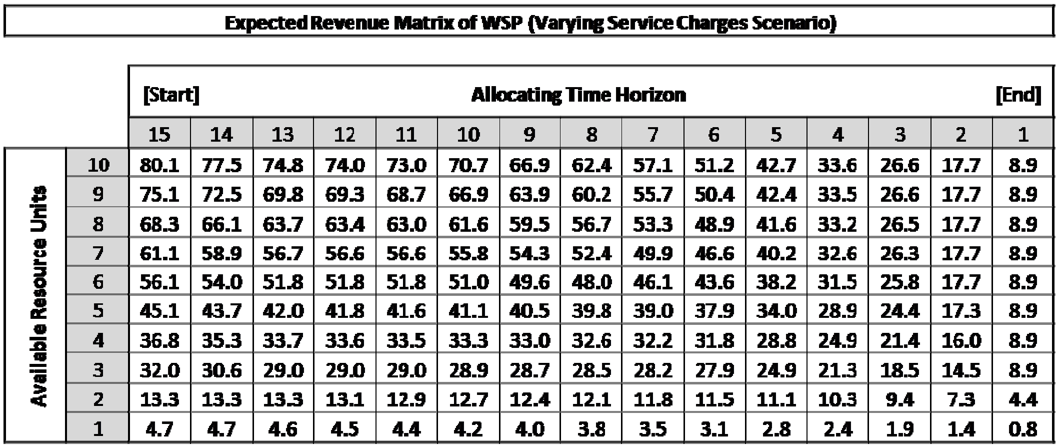
In this example, we relax our earlier assumption of having fixed service charge for a given service class over the allocating horizon. As discussed in Section 5.4.2, we vary the pricing of service classes over the time periods as assumed in Table 5.6 for this example scenario. Except for the service charges, we assume all other parameters for this example to be same as in Tables 5.1 and 5.2 of Section 5.5.1.

Table 5.6. Varying Service Charges for Different Service Classes



		Varying Service Charges Corresponding to Service Classes in each Time Period														
		[Start]						Time Periods								[End]
		15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
Service Class [l]	1	45	45	20	20	30	30	30	30	30	45	45	45	45	45	45
	2	8	8	15	15	15	15	15	15	15	15	18	18	18	18	18
	3	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5

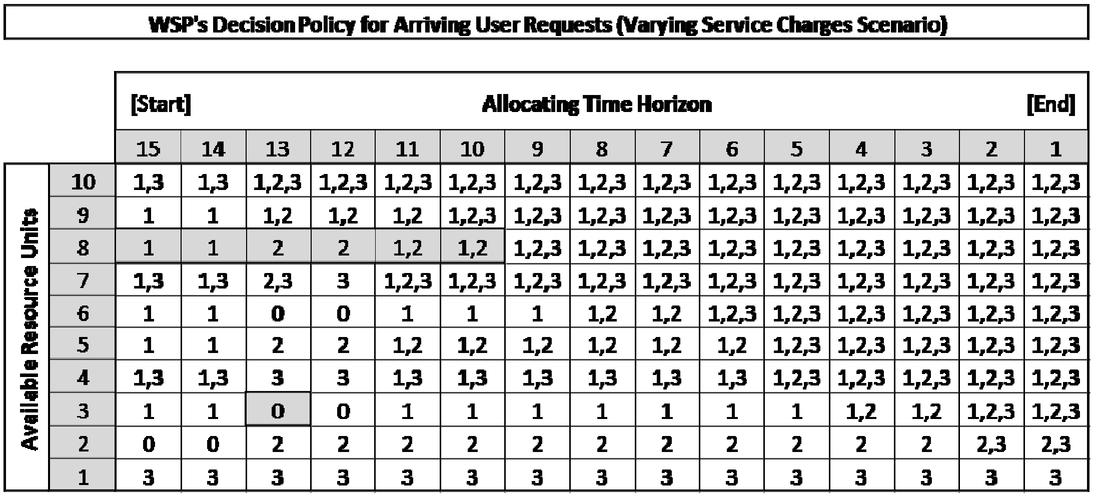
Table 5.7. Computed Expected Revenue for Varying Service Charge Scenario



		Expected Revenue Matrix of WSP (Varying Service Charges Scenario)														
		[Start]						Allocating Time Horizon								[End]
		15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
Available Resource Units	10	80.1	77.5	74.8	74.0	73.0	70.7	66.9	62.4	57.1	51.2	42.7	33.6	26.6	17.7	8.9
	9	75.1	72.5	69.8	69.3	68.7	66.9	63.9	60.2	55.7	50.4	42.4	33.5	26.6	17.7	8.9
	8	68.3	66.1	63.7	63.4	63.0	61.6	59.5	56.7	53.3	48.9	41.6	33.2	26.5	17.7	8.9
	7	61.1	58.9	56.7	56.6	56.6	55.8	54.3	52.4	49.9	46.6	40.2	32.6	26.3	17.7	8.9
	6	56.1	54.0	51.8	51.8	51.8	51.0	49.6	48.0	46.1	43.6	38.2	31.5	25.8	17.7	8.9
	5	45.1	43.7	42.0	41.8	41.6	41.1	40.5	39.8	39.0	37.9	34.0	28.9	24.4	17.3	8.9
	4	36.8	35.3	33.7	33.6	33.5	33.3	33.0	32.6	32.2	31.8	28.8	24.9	21.4	16.0	8.9
	3	32.0	30.6	29.0	29.0	29.0	28.9	28.7	28.5	28.2	27.9	24.9	21.3	18.5	14.5	8.9
	2	13.3	13.3	13.3	13.1	12.9	12.7	12.4	12.1	11.8	11.5	11.1	10.3	9.4	7.3	4.4
	1	4.7	4.7	4.6	4.5	4.4	4.2	4.0	3.8	3.5	3.1	2.8	2.4	1.9	1.4	0.8

With the above parameters and using the recursive optimizing Equations 5.7, 5.8, and 5.9, we compute the maximal expected revenue for the combination of each time period and available resource units. As done before, this expected revenue is used later to determine a dynamic admission / allocation policy for the WSP. The computed expected revenue matrix for this example is shown in Table 5.7. Once the expected revenue values are computed, the *accept / deny* decisions for arriving user requests belonging to different service classes are determined using logical comparative Equations 5.10 and 5.11.

Table 5.8. Decision Policy Computed at WSP for Varying Service Charge Scenario



WSP's Decision Policy for Arriving User Requests (Varying Service Charges Scenario)

		[Start]							Allocating Time Horizon							[End]
		15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
Available Resource Units	10	1,3	1,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3
	9	1	1	1,2	1,2	1,2	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3
	8	1	1	2	2	1,2	1,2	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3
	7	1,3	1,3	2,3	3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3
	6	1	1	0	0	1	1	1	1,2	1,2	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3
	5	1	1	2	2	1,2	1,2	1,2	1,2	1,2	1,2	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3
	4	1,3	1,3	3	3	1,3	1,3	1,3	1,3	1,3	1,3	1,2,3	1,2,3	1,2,3	1,2,3	1,2,3
	3	1	1	0	0	1	1	1	1	1	1	1	1,2	1,2	1,2,3	1,2,3
	2	0	0	2	2	2	2	2	2	2	2	2	2	2	2,3	2,3
	1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3

The optimal decision policy matrix computed for each combination of time period and available resource units for this example is shown in Table 5.8. Unlike fixed service charge scenario, it can be observed that critical allocating time period for a given capacity and service class combination after which the service class's request is always acceptable *does not* exist in the decision policy matrix. For instance, if we refer to Table 5.8, we observe that at a available resource level of 8 units - the service class 1 request is acceptable during time periods {15,14}, is not acceptable during time periods {13,12}, and is again acceptable during time periods {11,··· ,1}. This is because, due to varied service charges for a given service class over allocating horizon, there could be intermediate periods during which an arriving request is denied as the expected revenue that can be obtained in future for the resource units under consideration is greater than the revenue that would be obtained by accepting the user requests in the considered time periods.

For varying service charge scenario, it may be noted that there could be possible instances of time period and available resource units combination in which *none* of the user requests are acceptable, even with available resources for the same reasoning given above. For instance, as shown in Table 5.8, for time period 13 and available resource level of 3 units, none of the arriving user requests are accepted.

5.6 Performance Analysis

In this section, we evaluate the performance of our proposed algorithm through simulations. We analyze the impact on WSP's revenue by varying the duration of allocating time horizon, total available resource units, service charges and request arrival probabilities of offered service classes over the allocating horizon. For most of our following analysis, we compute our performance metrics over 100 simulation runs and plot corresponding charts with obtained average metrics.

5.6.1 Comparison with Greedy Allocation Strategy

We compare our proposed algorithm with a simple greedy resource allocation strategy. A greedy allocation strategy is one that accepts any incoming user request as long as the needed resources are available at the WSP and denies if sufficient resources are not available.

1) *Constant Service Charges and Arrival Probabilities Scenario*:

We consider a WSP with R = 30 resource units offering K = 3 different service classes, and that the WSP needs to allocate these resources over the allocating time horizon. The service charges and the request arrival probabilities corresponding to the offered service classes are assumed to be constant over the allocation time horizon, and are as shown in Table 5.9.

Table 5.9. Parameters Used in Simulation for the Constant Service Charges and Arrival Pattern Scenario

Parameters for Simulations			
Service Class [i]	Required Resouce Units [r_i]	Service Charge [c_i]	Request Arrival Probability [p_i^t]
1	3	45	0.1
2	2	15	0.1
3	1	8	0.3

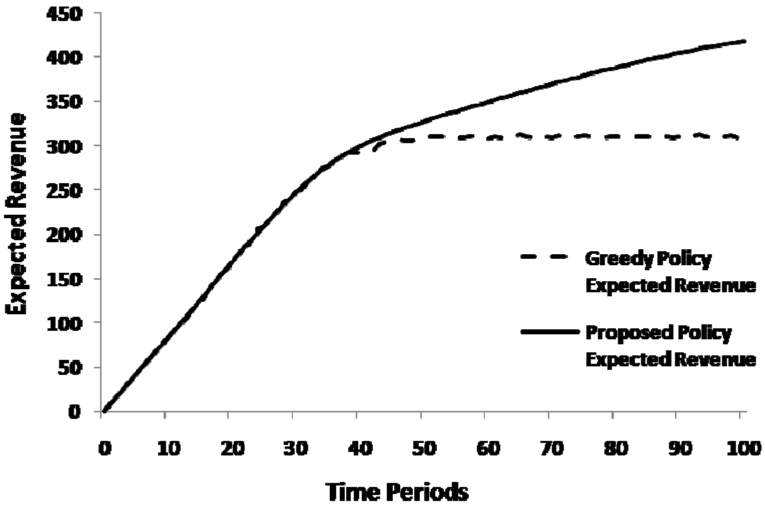


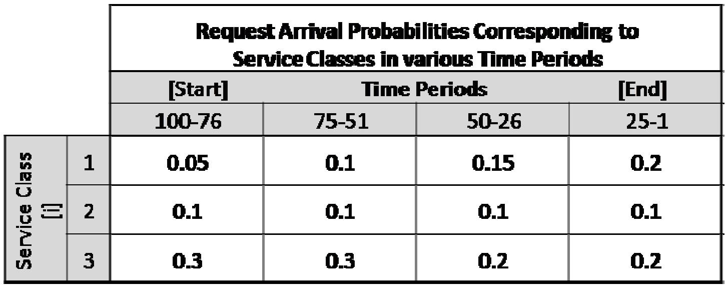
Figure 5.4. Expected Revenue Comparison for MDP and Greedy policy with Constant Charge and Arrival Pattern

Figure 5.4 represents the expected revenue that will be obtained by WSP using the above parameters, for various durations of allocating horizon. We observe that, both the proposed allocation policy and the greedy policy perform similarly if the duration of allocating horizon is small enough such that the demand is relatively less compared to the available resources - as both policies will accept all incoming user requests without denying any. We observe such a trend till about time period 40 in the figure. As the duration of the allocation time horizon increases, or in other words when the expected demand over the allocating time horizon is more than the available resources at WSP, the proposed allocation policy performs better than that of greedy policy as it optimally accepts / denies the incoming user requests, as discussed in Section 5.4.1.

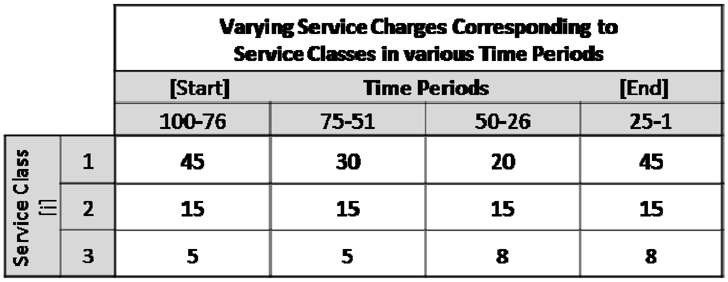
2) *Varying Service Charges and Arrival Probabilities Scenario*:

In this scenario, we vary the pricing of service classes and the request arrival probabilities over the time periods as given in Table 5.10. Similar to earlier scenario, the total available resource units at the WSP are R = 30 units and the individual required resource units corresponding to offered service classes are the same as assumed in Table 5.9.

Table 5.10. Service Charges and Arrival Probabilities for Varying Service Charge Scenario



		Request Arrival Probabilities Corresponding to Service Classes in various Time Periods			
		[Start]	Time Periods		[End]
		100-76	75-51	50-26	25-1
Service Class [i]	1	0.05	0.1	0.15	0.2
	2	0.1	0.1	0.1	0.1
	3	0.3	0.3	0.2	0.2



		Varying Service Charges Corresponding to Service Classes in various Time Periods			
		[Start]	Time Periods		[End]
		100-76	75-51	50-26	25-1
Service Class [i]	1	45	30	20	45
	2	15	15	15	15
	3	5	5	8	8

Figure 5.5 compares the expected revenue that will be obtained by WSP with the above parameters using the proposed and greedy allocation policies, for various durations of allocating horizon. We observe that both the proposed allocation policy and the greedy policy perform similarly if the duration of allocating horizon is small enough such that the demand is relatively less as compared to the available resources. However, it may be noted that the expected revenue for the greedy algorithm could fluctuate for increased durations of allocating time horizon.

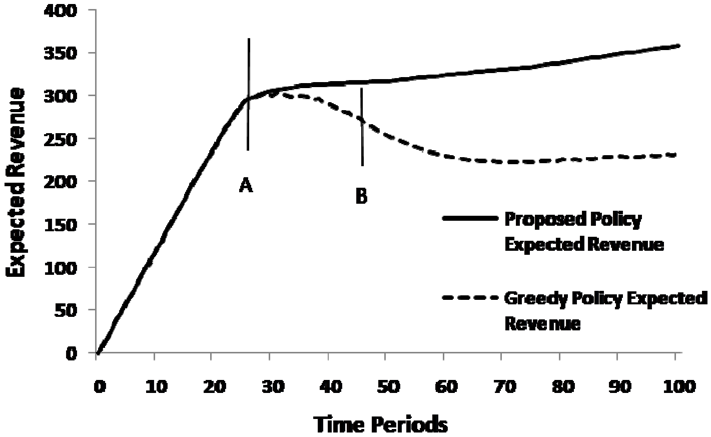


Figure 5.5. Expected Revenue Comparison for MDP and Greedy Policy

For instance, we observe that the expected revenue for greedy algorithm for allocating horizon of about 25 periods, depicted in figure as point ‘A’, is higher than the expected revenue obtained for allocating horizon of about 45 periods, depicted in figure as point ‘B’. This is because; the greedy policy accepts the series of first arriving requests as long as there are available resources at WSP. As a result, in this case, the revenue generated by the combination of accepted user requests and their corresponding service charges for allocating horizon of {45···1} is lower than that of revenue generated for the horizon of {25···1} periods. As mentioned earlier, for greedy approach, not all the requests in allocating horizon are accepted and the WSP starts denying incoming user requests after exhausting available resources. On the other hand, our proposed admission / allocation policy performs optimally throughout the allocation time horizon by accepting / denying the incoming user requests judiciously, as discussed in 5.4.2.

Figure 5.6 represents total revenue obtained for the proposed and greedy allocation policies for each of the 100 different simulation runs. The model for this analysis assumes the parameters of Table 5.10, available resource units of $R = 30$ and allocating time horizon of $T = 100$ time periods.

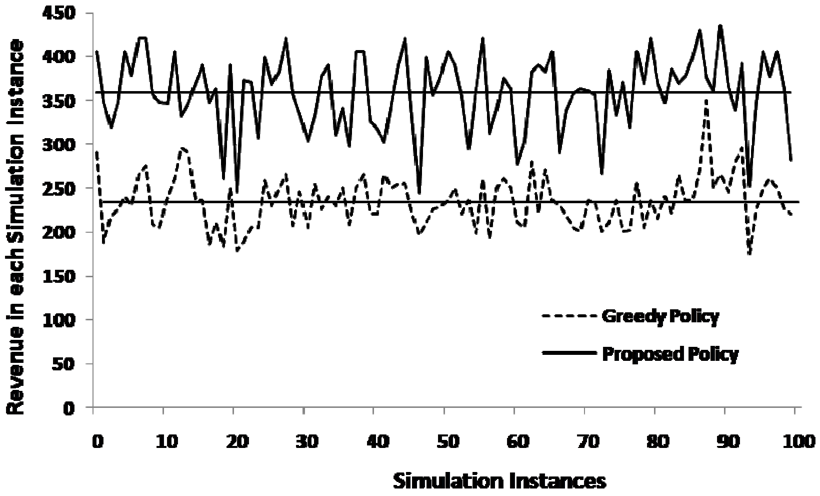


Figure 5.6. Revenue Comparison in Each Simulation Instance

5.6.2 Expected Revenue using MDP with Varying Resources

With all other parameters remaining the same as in the previous section, we study the performance of our proposed allocation algorithm by varying the total available resource units at the WSP for $R = \{15, 20, 25, 30\}$. Figure 5.7 represents the corresponding expected revenue obtained over various durations of allocation time horizon for the proposed admission strategy.

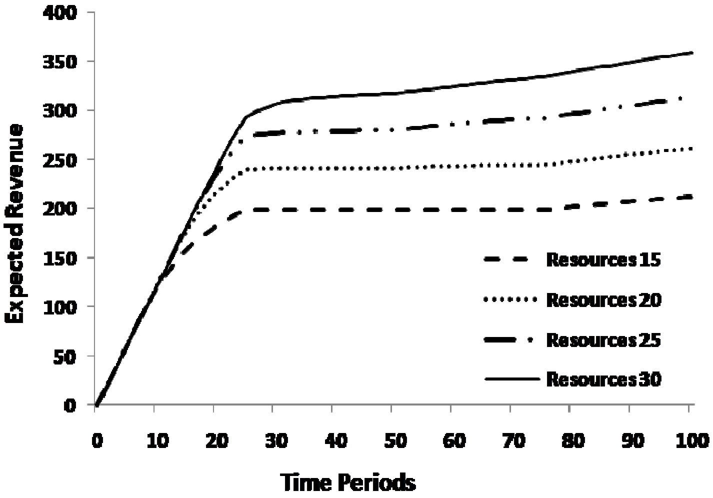


Figure 5.7. Expected Revenue Comparison for MDP with Varying Resources

5.6.3 Cumulative Revenue using MDP over Varying Durations of Allocation Time Horizon

Using the parameters of Table 5.10 and R = 30 resource units, we compute the cumulative revenue from simulations for different durations of allocation time horizons, T = {40, 60, 80, 100}. We plot these curves in Figure 5.8 along with the expected revenue curve of the proposed policy for reference. The cumulative revenue curves are plotted in reverse chronological order. For instance, the cumulative revenue obtained for T = 40 time periods is plotted starting from time period 40 and ending at time period 1. From the graphs, we observe that the obtained cumulative revenue from the simulations is close to the expected revenue generated by the proposed optimal policy, as expected.

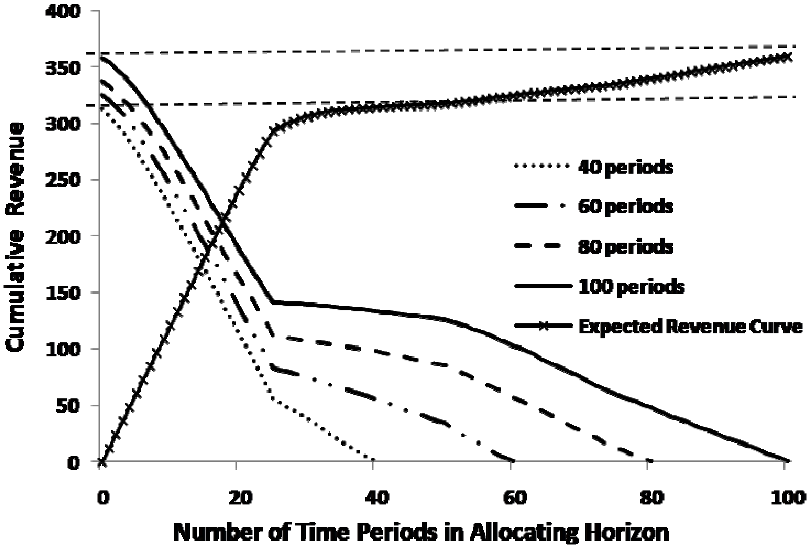


Figure 5.8. Cumulative Revenue for Varying Durations of Allocating Time Horizon

5.7 Summary

In this chapter, we describe a model for WSP to maximize its overall revenue from its limited resources and within finite time duration. The WSP offers various service classes and charge users correspondingly. The service charge for a given class can vary over the allocating time horizon. We use discrete-time Markov Decision Process model to formulate and optimize the allocating policy.

We analyze the formulated model through simulations and compare it to a basic greedy allocation policy. For scenarios where the expected demand is much lower than the total capacity at the WSP, both models perform similarly. However, when the expected demand is higher than the capacity, the proposed model prudently admits only the appropriate user requests and thus performs significantly better compared to the basic model. We also study the performance of the proposed model with regards to net generated revenue for WSP by varying the available resources at the WSP and also for varying the duration of the allocating horizon.

Chapter 6. Conclusions and Future Research

Recently, the increased demand for ubiquitous internet connectivity and broadband internet service has spurred the need for new innovative wireless technologies [4]. WMNs are one such upcoming technology that offer wireless broadband internet connectivity and would provide varied functionalities. They offer cost-effective and flexible solution for extending broadband services to the residential areas without any necessity for line-of-sight communication. WMNs are formed by a set of mesh routers (MRs), among which a small subset is directly connected to the wired network called the Internet Gateway (IGW). Communication between MRs is based on the ad hoc networking paradigm and thus adopts a self-configurable and self-healing approach.

WMNs are certainly one of the key topics for research in both academia and industry owing to its alluring features and innumerable advantages. Many industry's bigwigs such as Motorola, Intel, and Nokia are developing their own proprietary mesh devices with customary protocols for the WMNs [2]. The increased commercial interest in WMNs has driven the IEEE to establish a new task group, IEEE 802.11s, for standardizing the PHY and MAC layer protocols.

However, their real-world deployment and performance is often hindered by certain problems such as spatial bias, hot zones, and excessive congestion as described in Chapter 1. These problems are typically due to issues with wireless nature of communication (e.g. interference) and multi-hop communication paradigm employed by the constituent routers in WMNs.

In this dissertation, we have addressed such problems that affect WMNs' performance, and proposed following solutions. We illustrated the severe unfairness experienced by longer hop length flows in multi-hop WMNs. To address this issue, we proposed a novel service differentiation technique using dual queues that provides service guarantees to all users in the network irrespective of their spatial location. To resolve the hot-zone problem around IGWs in WMNs that result in excessive packet drops, we devised a load balancing routing scheme among different IGWs based on their current traffic serving capacity. We proposed a novel Adaptive State-based Multi-path Routing Protocol which constructs Directed Acyclic Graphs and effectively discovers multiple optimal path set between any given MR-IGW pair. We also proposed a congestion aware traffic splitting algorithm to balance traffic over multiple paths which synergistically improves the overall performance of the WMNs. We designed a novel Neighbor State Maintenance module that innovatively employs a state machine at each MR to monitor the quality of links connecting its neighbors in order to cope up with unreliable wireless links. We employed four-radio architecture for MRs, which allows them to communicate over multiple radios tuned to non-overlapping channels and better utilize the available spectrum.

To address the scenarios where an IGW/WSP is constrained in resources and have a pre-defined objective such as revenue maximization or prioritized fairness, a prudent user selection strategy is needed. In this dissertation, we proposed an optimal user admission / allocation policy model based on yield management and discrete-time Markov Decision Process principles. The proposed model computes expected revenue and decision policy matrix for a WSP for various combinations of available capacity and allocating time period. The WSP will accept / deny the arriving user requests in real-time in a dynamic manner based on its current network state and its pre-computed decision policy matrix.

6.1 Future Work

We believe an interesting extension to the proposed multipath routing protocol (ASMRP) would be to devise an adaptive transmission scheduling mechanism that splits the network's traffic among two or more possibly different paths to reduce latency, improve throughput, and balance traffic load. Further modification to the state machine employed in our proposed routing protocol (ASMRP) would be to investigate certain aspects of its functionality such as the mode of neighbor suspension, the neighbor suspension duration, etc.

In the proposed admission policy model for WSPs, once a user request is accepted by the WSP, the user stays in the system for the entire allocating time horizon and hence the corresponding allocated resources are assumed unavailable for any later requests. A possible extension to this work would be to relax this assumption and plan to consider the scenario where the users can leave the system at any time and pay a service charge for the used time and resources.

Bibliography

[1] http://www.earthlink.net/

[2] N. Nandiraju, D. Nandiraju, L. Santhanam, B. He, J. Wang and D. Agrawal, "Wireless Mesh Networks: Current Challenges and Future Directions of Web-in-the-sky," *in IEEE Wireless Communication Magazine 2007.*

[3] D. P. Agrawal and Q. A. Zeng, *Introduction to Wireless and Mobile Systems*. Brooks/Cole (Thomson Publishing), 2003.

[4] C. Cordeiro and D. P. Agrawal, *Ad hoc and Sensor Networks - Theory and Applications*, World Scientific Publishing, 2006

[5] http://www.wirevolution.com/2007/09/07/how-does-80211n-get-to-600mbps/

[6] http://i.i.com.com/cnwk.1d/html/itp/burton_80211nbeyon.pdf

[7] N. Nandiraju, Deepti Nandiraju, and D. P. Agrawal, "*Medium access control in Wireless Mesh Networks*" Nova Science Publishers, Inc.

[8] D. Nandiraju, N. Nandiraju, and D. P. Agrawal, "*Network Architecture and Flow Control in Multi-Hop Wireless Mesh Networks*," Encyclopedia of Ad hoc and Ubiquitous Computing, World Scientific Publishers Co., Inc.

[9] V.Gambiroza, B.Sadeghi, and E. W. Knightly, "End to End Performance and Fairness in Multihop Wireless Backhaul Networks," *in the Proc. of ACM Mobicom 2004.*

[10] N. Nandiraju, D. Nandiraju, D. Cavalcanti, and D. P. Agrawal, "A Novel Queue Management Mechanism for Improving Performance of Multihop Flows in IEEE 802.11s based Mesh Networks," *in Proc. of IEEE IPCCC 2006.*

[11] D. B. Johnson and D. A. Maltz, “Dynamic Source Routing in Ad hoc Wireless Networks,” *In Mobile Computing, volume 353, 1996.*

[12] S.-J. Lee and M. Gerla, “AODV-BR: Backup Routing in Ad hoc Networks,” *In Proc. of WCNC, 2000.*

[13] S.-J. Lee and M. Gerla, “Split Multipath Routing with Maximally Disjoint Paths in Ad hoc Networks,” *In Proc. of ICC, 2001.*

[14] M. Marina and S. Das, “On-demand Multipath Distance Vector Routing in Ad hoc Networks,” *In Proc. of IEEE International Conference on Network Protocols (ICNP), 2001.*

[15] P. Sambasivam, A. Murthy, and E. Belding-Royer, “Dynamically Adaptive Multipath Routing based on AODV,” *In Proc. of MedHocNet, June 2004.*

[16] M. Mosko and J. J. Garcia-Luna-Aceves, “Multipath Routing in Wireless Mesh Networks,” *WIMESH 2005.*

[17] N. Nandiraju, D. Nandiraju, and D.P. Agrawal, “Multipath Routing in Wireless Mesh Networks,” *in the Proc. of IEEE International Workshop on Heterogeneous Multi-Hop Wireless and Mobile Networks (MHWMN), 2006.*

[18] K. Ramachandran, I. Sheriff, E. Belding-Royer, and K. Almeroth, “Routing Stability in Static Wireless Mesh Networks,” *Passive and Active Measurement Conference, 2007.*

[19] R. Draves, J. Padhye, and B. Zill, “Routing in Multi-Radio, Multi-Hop Wireless Mesh Networks,” *in the Proc. of MOBICOM 2004.*

[20] D. Nandiraju, N. S. Nandiraju, and D. P. Agrawal, “Service Differentiation in IEEE 802.11s Mesh Networks: A Dual Queue Strategy,” *in Proc of IEEE MILCOM 2007.*

[21] D. Nandiraju, L. Santhanam, N. Nandiraju, and D. P. Agrawal, "Achieving Load Balancing in Wireless Mesh Networks Through Multiple Gateways," *in the Proc. of IEEE International Workshop on Wireless Mesh-Networks and Applications (WiMa06), 2006.*

[22] D. Nandiraju, N. Nandiraju, and D. P. Agrawal, "Adaptive State-based Multi-radio Multi-channel Multi-path Routing in Wireless Mesh Networks," *accepted in Journal of Pervasive and Mobile Computing, 2009.*

[23] D. Nandiraju and D. P. Agrawal, "Dynamic Admission Policy for Wireless Service Providers Using Discrete-time Markov Decision Process Model," *IEEE Transactions on Mobile Computing*, Under Review

[24] UCB/LBNL/VINT Network Simulator (ns-2), Available at http://www.isi.edu/nsnam/ns/index.html

[25] J. Jun and M. L. Sichitiu, "Fairness and QoS in Multihop Wireless Networks," *in Proc. of the IEEE Vehicular Technology Conference (VTC 2003), 2003.*

[26] Yi Y. and S. Shakkottai, "Hop-by-hop congestion control over a wireless multi-hop network," *In Proc. of IEEE INFOCOM 2004.*

[27] N. Nandiraju, D. Nandiraju, L. Santhanam, and D. P. Agrawal, "A Cache Based Traffic Regulator for Improving Performance in IEEE 802.11s based Mesh Networks, " *in the Proc. of IEEE Radio and Wireless Symposium (RWS), 2007.*

[28] V. Liberatore, "Local Flow Separation," *International Workshop on Quality of Service (IWQoS), 2004*

[29] H. Hassanein and A. Zhou, "Routing with load balancing in wireless Ad hoc networks," *in Proc. of ACM MSWiM '01.*

[30] S. J. Lee and M. Gerla, "Dynamic Load-Aware Routing in Ad hoc Networks," *in Proc. of*

ICC '01.

[31] P. Hsiao, A. Hwang, H. Kung, and D. Vlah, "Load-Balancing Routing for Wireless Access Networks," *in Proc. of IEEE INFOCOM '01.*

[32] A. Raniwala and T. Chiueh, "Architecture and Algorithms for an IEEE 802.11-based Multi-channel Wireless Mesh Network," *in Proc of IEEE INFOCOMM, 2005.*

[33] K. Ramachandran, M. Buddhikot, G. Chandranmenon, S. Miller, E. Belding-Royer, and K. Almeroth, "On the Design and Implementation of Infrastructure Mesh Networks," *in the Proc. of IEEE Workshop on Wireless Mesh Networks (WiMesh) 2005 (invited paper).*

[34] S. Vutukury and J. J. Garcia-Luna-Aceves, "A Traffic Engineering Approach based on Minimum-Delay Routing," *in Proc. IEEE IC3N 2000.*

[35] J. Li, C. Blake, D. S. De Couto, H. I. Lee, and R. Morris, "Capacity of Ad hoc Wireless Networks," *in Proc of ACM MOBICOM, 2001.*

[36] S.-L. Wu, C.-Y. Lin, Y.-C. Tseng, and J.-P. Sheu, "A New Multi-channel MAC Protocol with On-demand Channel Assignment for Multi-hop Mobile Ad hoc Networks," *ISPAN 2000.*

[37] J. So and N. H. Vaidya, "Multi-channel MAC for Ad hoc Networks: Handling Multi-channel Hidden Terminals using a Single Transceiver," *in Proc of ACM MobiHoc, 2004.*

[38] A. A. Pirzada, M. Portmann, and J. Indulska, "Evaluation of Multi-Radio Extensions to AODV for Wireless Mesh Networks," *International Workshop on Mobility Management and Wireless Access (MobiWAC), 2006.*

[39] J. S. Pathmasuntharam, A. Das, and A. K. Gupta, "Primary channel assignment based MAC (PCAM) – A Multi-channel MAC Protocol for Multi-hop Wireless Networks," *in Proc. of IEEE WCNC 2004.*

[40] P. Kyasanur and N. H. Vaidya, “Routing in Multi-channel Multi-interface Ad hoc Wireless Networks,” *in Proc. of IEEE WCNC 2005.*

[41] I. Sheriff and E. Belding-Royer, “Multipath Selection in Multi-radio Mesh Networks,” *Broadnets 2006.*

[42] G. Holland, N. H. Vaidya, and P. Bahl, “A Rate-Adaptive MAC Protocol for Multi-Hop Wireless Networks,” *in MOBICOM, 2001.*

[43] H. S. Wang and N. Moayeri, “Finite-state markov channel - a useful model for radio communications channels,” *IEEE Transactions on Vehicular Technology, vol. 44, no. 1, pp. 163–171, 1995.*

[44] ——, “On verifying the first-order markovian assumption for a rayleigh fading channel model,” *IEEE Transactions on Vehicular Technology, vol. 45, no. 2, pp. 253–357, 1996.*

[45] K. R. Chowdhury, P. Chanda, D. P. Agrawal, and Q. A. Zeng, “DCA- A novel Distributed Channel Allocation scheme for wireless sensor networks,” *in Proc. of the 16th IEEE PIMRC, 2005.*

[46] R. Jain, *The Art of Computer Systems Performance Analysis*, John Wiley and Sons, 1991.

[47] R. Draves, J. Padhye, and B. Zill, “Comparison of Routing Metrics for Static Multi-Hop Wireless Networks,” *SIGCOMM 2004.*

[48] Y. Yang, J. Wang, and R. Kravets, “Designing Routing Metrics for Mesh Networks,” *in the Proc. of IEEE Workshop on Wireless Mesh Networks (WiMesh), 2005.*

[49] D. D. Couto, D. Aguayo, J. Bicket, and R. Morris, “A High-Throughput Path Metric for Multi-Hop Wireless Routing,” *in the Proc. of ACM MobiCom 2003.*

[50] B. Yan and H. Gharavi, “Multi-Path Multi-Channel Routing Protocol,” *Proceedings of the Fifth IEEE International Symposium on Network Computing and Applications, 2006.*

[51] Y. Hayel, L. Wynter, and P. Dube, “Analysis of a Yield Management Model for On Demand Computing Centers,” *in Proc. of IFIP World Computer Congress (WCC), 2004.*

[52] S. Humair, “Yield Management for Telecommunication Networks : Defining a New Landscape,” *Thesis (Ph.D.), Operations Research Center, Massachusetts Institute of Technology, 2001.*

[53] F. Jallat and F. Ancarani, “Yield Management, Dynamic Pricing and CRM in Telecommunications,” *Journal of Services Marketing, 2008.*

[54] M. M. Buddhikot and K. Ryan, “Spectrum Management in Coordinated Dynamic Spectrum Access Based Cellular Networks,” *in Proc IEEE DySpan, 2005.*

[55] P. Dube and Y. Hayel, “A Real-Time Yield Management Framework for E-Services,” *in Proc. of the The 8th IEEE International Conference on E-Commerce Technology and The 3rd IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services, 2006.*

[56] S. K. Nair and R. Bapna, “An Application of Yield Management for Internet Service Providers,” *Naval Research Logistics, 2001.*

[57] C. Lindemann, M. Lohmann, and A. Thummler, “Adaptive Call Admission Control for QoS/Revenue Optimization in CDMA Cellular Networks,” *Wireless Networks, 2004.*

[58] E. Viterbo and C.F. Chiasserini, “Dynamic Pricing for Connection-Oriented Services in Wireless networks,” *12th International Symposium on Personal, Indoor and Mobile Radio Communications, 2001.*

[59] G. Zachariadisa and A. J. Barria, “Demand management for Telecommunications Services,” *Computer Networks, 2007.*

[60] H. Lin, M. Chatterjee, and S. K. Das, "Utility based Service Differentiation in CDMA Data Networks," *Wireless Networks, 2006.*

[61] S. Mandal, D. Saha, and M. Chatterjee, "Dynamic Price Discovering Models for Differentiated Wireless Services," *Journal of Communications, 2006.*

[62] P. Marbach and R. Berry, "Downlink Resource Allocation and Pricing for Wireless Networks," *Proc. of IEEE INFOCOM, vol. 3, 2002.*

[63] M. Mandjes, "Pricing Strategies Under Heterogeneous Service Requirements," *Proc. of IEEE INFOCOM, vol. 2, 2003.*

[64] E. Altman, D. Barman, R. Azouzi, D. Ros, and B. Tuffin, "Pricing Differentiated Services: A Game-theoretic Approach," *Computer Networks, 2006.*

[65] S. Sengupta and M. Chatterjee, "An Economic Framework for Dynamic Spectrum Access and Service Pricing," *To appear in ACM/IEEE Transactions on Networking, 2009.*

[66] T. Lee and M. Hersh, "A Model for Dynamic Airline Seat Inventory Control with Multiple Seat Bookings," *Transportation Science, 1993.*

[67] E. D. Fitkov-Norris and A. Khanifar, "Dynamic Pricing in Cellular Networks, A Mobility Model with a Provider-Oriented Approach," *Proc of 3G Mobile Communication Technologies, 2001.*