Providing quality of service in wireless ad hoc networks

Wang, Huiqing

2008

Wang, H. Q. (2008). Providing quality of service in wireless ad hoc networks. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/13579>

<https://doi.org/10.32657/10356/13579>

NANYANG TECHNOLOGICAL UNIVERSITY

Providing Quality of Service in Wireless Ad Hoc Networks

A thesis submitted to the Nanyang Technological University
in fulfilment of the requirement of the degree of
Doctor of Philosophy

by

Wang Huiqing

School of Computer Engineering

May, 2008

Abstract

Unlike conventional wireless networks where the mobile users can only connect to the access points (APs) or routers that are within their transmission range, wireless ad hoc networks do not rely on any existing infrastructure and can be formed dynamically. With its unique characteristics such as the self-forming, self-healing and self-organization, wireless ad hoc networks have recently received a lot of interest for commercial deployment. In this thesis, we study the provision of Quality of Service (QoS) in wireless ad hoc networks.

We first consider a high mobility scenario which is common in Mobile Ad Hoc Networks (MANETs). A QoS framework is proposed to enhance QoS for real-time traffic in MANETs which consists of two components, a QoS routing protocol and a local scheduling algorithm. To reduce QoS violation like large delay and low delivery ratio, which can be caused by link breakage as well as excessive contention, a QoS routing algorithm is first proposed to discover multiple node-disjoint paths that also satisfy the QoS requirement of real-time traffic. Re-route discovery latency due to frequent link failures is minimized with the alternate path maintained at the sources. With passive acknowledgement, QoS violation due to link failure or excessive contention can be detected if a sent packet is not acknowledged for a certain time. To further enhance the QoS of admitted real-time traffic, a packet scheduling algorithm is implemented locally at each node which gives higher priority to real-time traffic over best-effort traffic. The simulation results show that our proposal can enhance the QoS of real-time traffic in terms of both low end-to-end transmission delay and high delivery ratio, even under high mobility scenario.

We then consider a wireless network with infrastructure support (MWNI) that can be constructed from an existing IEEE 802.11 based WLAN with multihop extension. Assuming that only a single network interface card operating on a single channel MAC

is equipped in each mobile host, the channel assignment is incorporated in the routing algorithm to maximally utilize all the available channels. A new load metric is first proposed to estimate the network load by considering possible spatial reuse. Based on that metric, a QoS-aware routing protocol is then proposed which can enhance the QoS of end users by exploring the channel diversity in the face of dynamic traffic and unreliable wireless channels. A distributed channel switch algorithm is proposed that avoids switching oscillation and with little overhead incurred. Through simulation evaluation, our routing algorithm is shown to be able to achieve more stable performance comparing with existing routing algorithms in such networks, while achieving a better resource utilization.

Finally, to support integrated services in MWNIs, a QoS framework is proposed to achieve the following objectives: (1) allow more real-time flows to be admitted into the network, (2) optimize the performance of best-effort traffic without losing fairness and (3) enhance QoS performance of admitted real-time traffic. There are three main components in our framework to provide corresponding solutions. First, a new QoS routing protocol is proposed that discovers global optimized topology based on the network interference modeled by flow contention graph. Second, admission control is performed at the AP so that a flow will be admitted only if the network has enough bandwidth to support its minimum bandwidth requirement. Third, a two-level hierarchical scheduling algorithm based on weighted fair queueing is implemented at the APs. On the first level, the available link bandwidth is divided between two virtual servers and each of them will schedule packets from the same type. On the second level, each virtual server performs weighted fair queueing algorithm so that the fairness among all backlogged flows of the same type is optimized. The efficiency of individual components, as well as the whole QoS framework is evaluated through simulation studies, and the results show our proposed QoS framework is able to achieve all the objectives.

Acknowledgments

I would like to express my special appreciation to my supervisor, Dr. Yow Kin Choong, for his patient guidance and care in the past years. Fortunately to have the chance to be his student, his smile and encouragement help me overcome the obstacles in my PhD studies at NTU.

I also want to show special thanks to Dr. Sukumar for his selfless help for my research work during the first year as a visiting professor. The discussions with him always benefit me a lot.

I must thank the colleagues in CeMNeT, they are Li Zhifei, Yang Peng, Ye Minghua, Xie Jiawei and Zhou Zhi. Their valuable comments and advice always inspire me.

For the other colleagues and friends in NTU, I also get a lot of help from them, and the resource they shared with me greatly facilitate my research work and the life, they are: Yu Renjun, Qiu Jie, Chen Quan, Dr. Gao Deyun, Song Peng, Feng Zhou, Wu Jianhua, Zhang Yuquan, Zhou Chen, Hu Yiqun, Liu Song, Zhou Zongzhao, Zhang Yu and Zhou Zixuan.

Finally, I want to dedicate this thesis to my parents, my brother and my friends. Without their love and support, I cannot make through with this.

Contents

Abstract	i
Acknowledgments	iii
List of Figures	viii
List of Tables	xi
 1 Introduction	 1
1.1 Wireless Ad Hoc Networks	1
1.2 QoS Provisioning in Wireless Ad Hoc Networks	4
1.3 Research Objectives and Contributions	6
1.3.1 Research Objectives	6
1.3.2 Assumptions	8
1.3.3 Contributions	8
1.4 List of Publications	11
1.5 Thesis Organization	11
 2 Literature Review	 13
2.1 Background	13
2.1.1 Concepts and Terminology	13
2.1.2 QoS in the Internet	14
2.2 Routing in MANETs	19
2.2.1 Single Path Routing	19
2.2.2 Multipath Routing	22
2.3 QoS in Wireless Ad Hoc Networks	23
2.3.1 QoS Model	24
2.3.2 QoS Signaling	26

2.3.3	QoS Routing	28
2.3.4	QoS MAC	31
2.4	Summary	34
3	Enhancing QoS for Real-time Traffic in Mobile Ad Hoc Networks	36
3.1	Introduction	36
3.2	Framework Overview	39
3.3	QDMR: QoS routing with Multipath Discovery	39
3.3.1	Route Discovery	40
3.3.1.1	Disjoint Paths	40
3.3.1.2	Route Exploration	42
3.3.1.3	QoS Metrics	43
3.3.1.4	Route Establishment	44
3.3.1.5	Bandwidth Reservation	45
3.3.1.6	Traffic Admission	46
3.3.2	Route Maintenance	47
3.3.2.1	QoS Violation Detection	47
3.3.2.2	Alternate Path Maintenance	48
3.3.2.3	Route Recovery	48
3.4	Scheduling Policies	50
3.5	Performance Evaluation	52
3.5.1	Simulation Environment	53
3.5.2	Routing Performance	53
3.5.3	Performance under Heterogeneous Traffic	59
3.6	Summary	61
4	QoS-aware Routing in Multihop Wireless Networks with Infrastructure Support	64
4.1	Introduction	64
4.2	Related Work	67
4.3	Probabilistic Distribution based Load Estimation	68
4.3.1	Assumptions	69

4.3.2	Preliminaries	70
4.3.3	Model the Interference in MWNIs	72
4.3.4	Load Estimation	74
4.4	PDQR: A Probabilistic Distribution based QoS-aware Routing Protocol for MWNIs	75
4.4.1	Route Establishment	75
4.4.2	Routing Metric	77
4.4.3	Route and Channel Switch	79
4.4.4	Enhancing Routing Efficiency	81
4.5	Performance Evaluation	82
4.5.1	Simulation Setup	82
4.5.2	Simulation Results	83
4.6	Summary	87
5	A QoS Framework to Support Integrated Services (QFIS) in Multihop Wireless Networks with Infrastructure Support	89
5.1	Introduction	89
5.2	Related Work	92
5.2.1	QoS Framework	92
5.2.2	Packet Scheduling	93
5.3	Preliminaries	95
5.3.1	Characterization of Traffic	95
5.3.2	Network Model	96
5.4	Estimate the Bottleneck Bandwidth in MWNIs	98
5.4.1	Procedures to Construct the Subflow Constraint Clique for a Flow	99
5.4.2	Optimized Algorithm to Identify the Subflow Constraint Clique for the Flows	103
5.5	FCGQR: Flow Contention Graph based QoS Routing to Support Inte- grated Services in MWNIs	105
5.5.1	Local State Update	105
5.5.2	Routing Metric	106

5.5.3	Association	108
5.5.4	Reassociation	108
5.6	Admission Control	109
5.7	Hierarchical Scheduling	111
5.7.1	Overview	111
5.7.2	Design Issues	112
5.7.3	Scheduling for Virtual Servers	115
5.7.4	Scheduling for Packets	118
5.7.5	Enhanced AIMD Control (EAIMD)	119
5.7.6	Parameter Analysis	121
5.8	Performance Evaluation	123
5.8.1	Simulation Setup	123
5.8.2	Efficiency of Admission Control	124
5.8.3	Efficiency of Routing Algorithm	126
5.8.4	Efficiency of Routing Metrics	127
5.8.5	Efficiency of Hierarchical Scheduling	132
5.8.6	Efficiency of QoS Framework	138
5.9	Summary	139
6	Conclusion and Future work	142
6.1	Main Research Contributions	142
6.2	Future Work	146
6.2.1	Predictive Routing	146
6.2.2	Support Peer-to-peer Communications in MWNIs	146
6.2.3	Support Multiple Transceivers in MWNIs	147
	References	149

List of Figures

1.1	Wireless network	2
2.1	RSVP signaling	16
2.2	DiffServ architecture	18
2.3	SWAN model	25
2.4	INSIGNIA QoS framework	27
2.5	A is hidden from C and exposed to B	31
2.6	RTS/CTS access scheme	33
2.7	Enhanced Distributed Coordination Function in 802.11e	34
3.1	System model	40
3.2	Disjoint paths	41
3.3	RREQ propagation in DSR	42
3.4	Passive acknowledgment	48
3.5	Priority queueing	52
3.6	End-to-end delivery ratio of real-time traffic	55
3.7	Admission ratio of real-time traffic	55
3.8	Average end-to-end delay	57
3.9	Average hop count of routes	57
3.10	Number of real-time traffic rejection	58
3.11	Normalized control overhead	59
3.12	End-to-end delivery ratio of real-time traffic	60
3.13	End-to-end delay of real-time traffic	61
3.14	Average end-to-end delay of best-effort traffic	62
3.15	Number of best-effort traffic packet	62

4.1	Multihop wireless network with infrastructure support (MWNI)	66
4.2	A multihop flow	70
4.3	Interference in multihop wireless networks	71
4.4	Contending flows in MWNIs	71
4.5	Spatial reuse around AP	72
4.6	Route establishment in MWNIs	76
4.7	Dynamic traffic in MWNIs	78
4.8	Route and channel switch in MWNIs	80
4.9	Simulation scenario	82
4.10	Aggregate throughput with various network load	84
4.11	Average end-to-end delay with various network load	85
4.12	Average route length with various network load	86
4.13	Average number of channel switch with various network load	86
4.14	Aggregate throughput over time	87
5.1	QoS framework for MWNIs	91
5.2	Subflow contention graph	97
5.3	Flow contention graph	98
5.4	Coexistence of real-time traffic and best-effort traffic in the network	107
5.5	Routing procedure in MWNIs	110
5.6	Hierarchical scheduling for real-time flows and best-effort flows	112
5.7	Multiple links contend for the shared medium in wireless networks	114
5.8	End-to-end delay as function of traffic load	115
5.9	Placement of the APs	124
5.10	Efficiency of admission control: average delivery ratio of real-time traffic	126
5.11	Efficiency of admission control: average end-to-end delay of real-time traffic	126
5.12	Efficiency of admission control: aggregate throughput of real-time traffic	127
5.13	Aggregate traffic throughput with different routing metrics	130
5.14	Average route length of best-effort flows with different routing metrics	132
5.15	Average end-to-end delay of real-time traffic with various AIMD parameters	133
5.16	Aggregated throughput of best-effort traffic with various AIMD parameters	134

5.17	Weight assigned to best-effort traffic over time	135
5.18	Aggregated throughput of best-effort traffic over time	135
5.19	Average end-to-end delay of real-time traffic with various shaping factors	137
5.20	Aggregate throughput of best-effort traffic with various shaping factors .	137
5.21	Fairness index of best-effort traffic with various shaping factors	138
5.22	Efficiency of QFIS: average end-to-end delay of real-time traffic	140
5.23	Efficiency of QFIS: average delivery ratio of real-time traffic	140
5.24	Efficiency of QFIS: average network throughput	141

List of Tables

2.1	Traffic behavior and QoS requirements	14
2.2	Terminologies for Internet QoS	15
3.1	Node-disjoint degree of the paths	58
4.1	The probability $P(N_i, N_j \leq 2R)$ for nodes in different tiers	73
4.2	Notations	74
4.3	per-AP aggregate throughput (Kbps)	84
5.1	Constraint clique construction for topology in Figure 5.2	103
5.2	Notations used in Section. 5.4	104
5.3	Notations used in Section. 5.7	123
5.4	per-AP performance: use the metric LC without admission control . . .	128
5.5	per-AP performance: use the metric LC with admission control	128
5.6	per-AP performance: use the metric $AEAS$ with admission control . . .	129
5.7	per-AP admitted traffic: use the metric LC with admission control . . .	131
5.8	per-AP admitted traffic: use the metric $AEAS$ with admission control . .	131

Chapter 1

Introduction

1.1 Wireless Ad Hoc Networks

Since the early 1970s, wireless technology has become increasingly popular in the network industry. Wireless networks can provide mobile users with ubiquitous communication and information access at any location. However, conventional wireless networks only provide the last hop connectivity, i.e. mobile users must stay within the transmission range of the access point (AP) or the wireless router which is usually fixed and connected with the backbone. However, such pre-installed infrastructure support may not be available in many application environments, such as battlefield communication and emergency rescue etc. Instead, wireless ad hoc network does not rely on any existing infrastructure and can be formed in a dynamic way. The nodes intercommunicate in a peer-to-peer fashion through other intermediate nodes which act as the routers to forward the packet to the next hop. The packets reach the destination through a multihop path. Figure 1.1 shows a wireless network with and without infrastructure.

Comparing with the wired networks, wireless ad hoc networks have the following distinctive characteristics:

- **Ad hoc networking:** Self-forming, self-healing and self-organization can be achieved with the cooperation among all peers which is further enhanced with multihop transmission.

Figure 1.1: Wireless network

- **Shared medium:** In wireless network, all the traffic transmitted in the same channel will contend for the shared medium. Simultaneous transmission within each other's interference range can cause collision which thus needs appropriate MAC layer design.
- **Low bandwidth capacity:** Due to the power constraint, wireless link has a rather low bandwidth comparing with the gigabit router in wired network. Also, the quality of wireless links is often affected by channel dynamics like fading, shadowing, noise and path losses etc.
- **Node mobility:** Without the mobility restriction caused by the infrastructure, nodes in wireless ad hoc networks can move freely which causes the topology to change frequently. The traffic has to be re-routed once such route breakage happens. Partition is likely to happen which may disrupt the communication totally.
- **Power constraint:** The mobile nodes rely on battery power or other portable power supply for operation. The power constraint limits the transmission range

and signal strength of the nodes.

- **Security:** The broadcast nature in wireless network potentially results in more exposure to security risks.

Designed for different application environments, types of wireless ad-hoc networks include Mobile Ad Hoc Networks, Wireless Sensor Networks, and Wireless Mesh Networks.

- **Mobile Ad Hoc Networks (MANETs):** MANETs have received intense research interest since the beginning of wireless communications, considering its unique characteristics. Such networks usually operate in a standalone fashion within a bounded area with homogeneous nodes working in an autonomous manner, which can serve as a complementary network where infrastructure or centralized administration is lacking. The network topology may change rapidly and unpredictably since all the nodes can move freely with various mobility. This determines that the routing efficiency is of primary importance to provide end-to-end connectivity for the entire duration of the communication.
- **Wireless Sensor Networks (WSNs):** The main applications of WSNs focus on providing surveillance or monitoring with spatially distributed sensor devices, and it is especially useful for tough or unattended environments. Besides military applications such as battlefield surveillance which first motivates the development of WSNs, commercial applications are now also benefiting from the deployment of WSNs, such as environment and habitat monitoring, healthcare applications, home automation, and traffic control. Depending on the application and the scale of deployment, the size and cost of sensor nodes are usually constrained and these characters result in corresponding constraints on available resources such as energy, memory, computational speed and bandwidth. Such constraints must be taken into consideration in designing communication mechanisms or algorithms for WSNs.

- **Wireless Mesh Networks (WMNs):** WMNs consist of mesh routers and mesh clients, where mesh routers with minimal mobility form the backbone of WMNs. The most attractive feature of WMN is its capability of providing the “last mile” access to existing networks such as the Internet, cellular, and WLANs, etc., which can be accomplished through the gateway and bridging functions in the mesh routers. Other advantages can also be provided by WMNs. With mesh routers forming a wireless backbone, the coverage, connectivity and reliability can be greatly enhanced. With the routing and management functionalities implemented at the mesh router, the burden on the end-user devices is significantly alleviated. Since the traffic in WMNs is usually to/from the backbone network, flexibility in such network is further enhanced that the mobile users can be associated with any router which is reachable through multiple wireless links. Load-balance can also be achieved in the network and all the available channels can be efficiently utilized as proposed recently [SV04, SV05]. Recent research work [RGC04, KV05] also suggests that the capacity improvement in such networks can be achieved by deploying multiple radios at mesh routers. Considering the characteristics of WMN, such network has recently received more attention for realistic deployment and it also inspires our studies in Chapter 4 and 5. The multihop wireless networks with infrastructure support (MWNIs) we will discuss in those two chapters can be categorized as one type of WMNs.

1.2 QoS Provisioning in Wireless Ad Hoc Networks

Recently, the rising popularity of multimedia applications among end users in various networks and the potential large scale commercial deployment of wireless networks have led to research interest in providing QoS support in wireless ad hoc networks. RFC 2386 [CNRS98] characterizes QoS as a set of service requirements to be met by the

CHAPTER 1. INTRODUCTION

network while transporting a packet stream from source to destination. Intrinsic to the notion of QoS is an agreement or a guarantee by the network to provide a set of measurable pre-specified service attributes to the user in terms of delay, delay variance (jitter), available bandwidth, probability of packet loss, and so on. The Internet of today operates in a connectionless and stateless mode, and all flows are treated in the same way as best effort traffic. “Best effort” means “just send as fast as possible” and there is no guarantee for the quality of the service. The importance of QoS can be recognized that however large the network capacity may be, it is still prone to congestion if the traffic is not well scheduled, e.g. all traffic goes through the same path during a short time. The limited bandwidth in wireless networks and the increasing demand for high bandwidth by the emerging applications also make it one of the most critical issues in wireless networks.

Considering the unique characteristics of wireless ad hoc networks, the following challenges must be considered in designing algorithms to provide QoS in such networks.

- Dynamics of the network and lack of a central management entity render it difficult to track up-to-date state information in the network. Thus even QoS of a established session (e.g. delay, bandwidth etc.) is likely to be violated frequently during a traffic session. Such QoS violation may be caused by route failure which is in turn caused by node movement. Meanwhile, multihop transmission over shared medium will also aggregate QoS violation if the contention in some bottleneck areas goes beyond the level that is supported by channel capacity. To minimize the impact of network dynamics is the key for QoS provisioning in wireless ad hoc networks.
- Due to the lack of infrastructure and a central management entity, there is no core and edge distinction that all nodes are equal in the roles of QoS provisioning. The conventional client/server architecture does not exist in ad hoc networks that

requires that all network management be carried out in a distributed manner. It is thus very important to reduce any global information updating and maintenance which may incur huge overheads and further burdens.

- The limited bandwidth in wireless ad hoc networks makes it more meaningful to study QoS in terms of the resource efficiency. Such efficiency should be considered from three aspects: First, with multihop transmission, more bandwidth will be used as the number of transmission hops increases. However, most QoS metrics are end-to-end measured, such as delay, delivery ratio, and throughput, etc.. So the achievable capacity in multihop wireless networks is determined by the topology which to a large extent also depends on the routing efficiency. Second, unlike in wired networks where congestion only affects an overloaded link, congestion in wireless ad hoc networks occurs at an area where multiple wireless links are contending for the shared medium. Congestion avoidance is thus desired which also requires accurate estimation of the load in the network. Third, considering applications in WMNs that most traffic is to/from the backbone network, channel assignment must be taken into consideration in designing routing algorithms. This is because most commodity wireless transceivers cannot receive and send at the same time, i.e. half-duplex, thus channel synchronization between the sender and the receiver is required to form a valid link on the route. Routing algorithm should be designed to improve the global resource efficiency.

1.3 Research Objectives and Contributions

1.3.1 Research Objectives

In this work, we are concerned with three major objectives:

- Facing the unique challenges in wireless ad hoc networks, enhance the QoS for real-time traffic is our first objective. QoS is violated if the QoS requirements,

CHAPTER 1. INTRODUCTION

such as delay bound and minimal bandwidth requirement, which are set up at the connection establishment phase cannot be held during the period of connection. Unlike wired network, the multihop connection in wireless ad hoc network is more vulnerable to node movement and flow contention. For scenarios where all nodes are homogeneous and can move freely, such as MANETs that will be first studied in this thesis, frequent link failures are expected which will cause traffic interruption. To support real-time traffic which usually comes with a strict delay and bandwidth requirements, the current routing protocols must be enhanced to minimize such QoS violation, and it requires our algorithms to quickly detect QoS violation and minimize the re-routing latency.

- Our algorithms should be able to achieve global optimization, i.e. by enhancing QoS experience of certain flows, it should not sacrifice global network performance. In wireless ad hoc networks, multihop flows will not only contend in time domain for the access to network interface, but also in spatial domain for access to the shared medium. The congestion thus occur will affect an area and multiple flows that pass through the area will be affected. So our proposal should be able to achieve global optimization by exploring the existing traffic and topology characteristics.
- Our algorithms should be compatible with existing 802.11 based WLANs and can be implemented in a distributed manner. In this thesis, we study QoS provisioning problems in wireless ad hoc network through a pragmatic approach and hope to minimize the gap between theory and application. By focusing on layer 3 and above, our algorithms should be easily implemented in and work with existing 802.11 based devices. Since nodes in wireless ad hoc networks are homogeneous, each node can be sender, receiver and router, distributed algorithms are favored in our proposal.

1.3.2 Assumptions

Before fully presenting our proposals, the key assumptions used in our work are first introduced:

- All nodes are assumed to bear identical characteristics.
- All nodes are equipped with an omni-directional antenna which cannot transmit and receive concurrently.
- All nodes are cooperative in forwarding traffic for other nodes.
- The transmission power is fixed, so the transmission range is also fixed.
- There is no power failure.
- The network is assumed to be *combinatorially stable* [CM01], that is, the topology cannot change too fast that no transmission can be finished.
- Bit error rate (BER) is not considered in our work, a packet is corrupted only upon collision.

1.3.3 Contributions

We summarize our main contributions in this thesis as follows:

- **A QoS framework to enhance QoS for real-time traffic in mobile ad hoc networks** In this part, a high mobility scenario is investigated with all nodes operating autonomously. The QoS framework consists of two components, a QoS routing protocol and a local scheduling algorithm. To reduce QoS violation which can be caused by link breakage as well as excessive contention, a QoS routing algorithm is first proposed which is able to discover routes with enough resources

to support the traffic. Re-route discovery latency due to frequent link failure is alleviated with multiple node-disjoint paths maintained at the sources so that any individual path failure is an independent event that does not affect the effectiveness of another path. With passive acknowledgement, QoS violation due to link failure or excessive contention can be detected if a sent packet is not acknowledged for certain time. With our QoS routing algorithm, the persistence of QoS provisioning is greatly enhanced even with high mobility. To further enhance the QoS of admitted real-time traffic, a packet scheduling algorithm is implemented locally at each node to provide differentiated services.

- A QoS-aware routing protocol for multihop wireless network with infrastructure support** In this part, a multihop wireless network with infrastructure support (MWNI) is investigated. It can be constructed from an existing IEEE 802.11 based WLANs with multihop extension. Assuming that only a single network interface card operating on a single channel MAC is equipped in each mobile host, the channel assignment must be considered with the routing algorithm to maximally utilize all the available resources in the different channels. A new load metric is first proposed that exploits the traffic and topology patterns in MWNI, which will be more accurate than other existing methods by considering possible spatial reuse. A QoS-aware routing protocol (QDMR) is then proposed to enhance the QoS for the MHs. Based on the load estimation at the APs, the MHs will associate with APs that provide more residual bandwidth in the face of network dynamics. A distributed channel switch algorithm is proposed to avoid oscillation and with little overhead incurred. Our routing algorithm can achieve more stable performance comparing with existing routing algorithms, while efficiently utilizing the global network resources.

- A QoS framework to support integrated services in MWNIs** To further support integrated services in MWNIs, various QoS requirements of the traffic such as delay bound and minimum bandwidth should be satisfied. A QoS framework is therefore proposed to achieve the following objectives: (1) allow more real-time flows to be admitted into the network, (2) optimize the performance of best-effort traffic without losing fairness and (3) enhance QoS performance of admitted real-time traffic. There are three main components in our framework to provide corresponding solutions. First, a new QoS routing protocol that discovers global optimized topology based on the network interference modeled by flow contention graph. Based on the estimation of the load in the bottleneck areas in the networks, a novel routing metric is proposed that chooses routes with higher additional end-to-end available resource. Second, admission control is performed at the AP so that a flow will be admitted only if the network has enough bandwidth to support its minimum bandwidth requirement. Third, to further enhance the QoS of admitted real-time traffic and to optimize the fairness among all best-effort flows, a two-level hierarchical scheduling algorithm based on weighted fair queueing is implemented at the APs. On the first level, the available link bandwidth is divided between two virtual servers and each of them will schedule packets from the same type. Since the available link bandwidth is determined by the topology considering contention in both time domain and spatial domain, only the AP will perform such scheduling which requires a lot of resources. The actual share to be allocated to best-effort traffic is controlled by an additive increase multiplicative decrease (AIMD) algorithm based on the performance of admitted real-time traffic, and its stability is further enhanced with a shaping control mechanism. The performance of admitted real-time traffic will not be affected by best-effort traffic and its delay performance can be improved by controlling best-effort traffic. On the second level, each vir-

tual server performs weighted fair queueing algorithm so the fairness among all backlogged flows of the same type is optimized.

1.4 List of Publications

Papers published in conference

- H. Wang, K. C. Yow, S. Nandi, “Providing Quality of Service to Real-time Traffic through Multipath Discovery”, ICMU 2004, Japan, January 2004
- H. Wang, K. C. Yow, “QoS Routing in Multi-channel Multihop Wireless Networks with Infrastructure Support”, INTERSENSE 2006, France, May 2006
- H. Wang, K. C. Yow, “A QoS Framework to Support Integrated Services in Multihop Wireless Networks with Infrastructure Support”, IPCCC 2007, USA, April 2007
- K. C. Yow, H. Wang, “Multihop Wireless Networks with Infrastructure Support (MWNIs): Extending the Coverage of WLANs with Multihop Networks”, submitted to ICAIT 2008 (Invited Paper), China, July 2008

Papers submitted to journal

- H. Wang, K. C. Yow, “A QoS Framework to Support Integrated Services in Multihop Wireless Networks with Infrastructure Support”, accepted to publish in Computer Communications, May 2008

1.5 Thesis Organization

The rest of the thesis is organized as follows. In Chapter 2, basic concepts and terminologies about QoS, as well as the routing in MANETs are introduced to provide background

CHAPTER 1. INTRODUCTION

knowledge, followed by a brief literature review including some seminal work on QoS provisioning in wireless ad hoc networks. In Chapter 3, a QoS framework is proposed to enhance QoS for real-time traffic in mobile ad hoc networks. In Chapter 4, a QoS-aware routing protocol is proposed to enhance QoS support in multihop wireless networks with infrastructure support. Based on the lessons learned from Chapter 4, a QoS framework is then proposed in Chapter 5 to provide QoS support for integrated services in such networks. Finally, conclusions and future work are given in Chapter 6.

Chapter 2

Literature Review

In this chapter, basic concepts and terminology of QoS are first introduced, followed by an overview of QoS provisioning in the Internet, as background. Considering the importance of routing algorithms in wireless ad hoc networks where links are vulnerable to breakdown due to node mobility or excessive contention, a brief review of the on-demand routing protocols and multipath routing protocols in MANETs is provided separately. We then review the works of literature on QoS provisioning in wireless ad hoc networks.

2.1 Background

2.1.1 Concepts and Terminology

QoS is a set of attributes and their values that, taken together, characterize the performance experienced by a user of the service. The network needs are governed by the service requirements specified by the end user applications. The network is expected to guarantee a set of measurable pre-specified service attributes to the user in terms of end-to-end performance. Common metrics include:

Bandwidth - the rate at which an application's traffic must be carried by the network

Delay - the latency that a packet experiences in the end-to-end transmission

Table 2.1: Traffic behavior and QoS requirements

Applications	Traffic Behavior	QoS Requirements
Electronic Mail (SMTP) File Transfer (FTP) Remote Terminal (Telnet)	Small, batch file transfers	Very tolerant of delay B/w requirement: low Best effort
HTML Web Browsing	Series of small, bursty file xfer	Tolerant of moderate delay B/w requirement: varies Best effort
IP-based Voice (VoIP) Real Audio	Constant or variable bit rate	Very sensitive to delay/jitter B/w requirement: low Requires predictable delay/loss
Streaming Video	Variable bit rate	Very sensitive to delay/jitter B/w requirement: High, variable Requires predictable delay/loss

Jitter - the variation in delay

Loss - the percentage of lost data

Various requirements are needed for different applications. Table 2.1 illustrates some common applications and their requirements for QoS.

2.1.2 QoS in the Internet

The Internet Engineering Task Force (IETF) has proposed many service models and mechanisms to meet the demands for QoS. Before introducing some of the well-known models, frequently used terminologies in the domain of Internet QoS are given in Table 2.2.

Integrated Service

The basic idea of the Integrated Service (IntServ) [BCS94] is that resources are reserved beforehand for a flow. A flow is an application session between a pair of end users. A session must first declare its QoS requirements including bandwidth, delay bound, and cost of the flow. RSVP [BZB⁺98] is adopted as the signaling protocol to reserve the

Table 2.2: Terminologies for Internet QoS

Flow	A stream of packets with the same source IP address, source port number, destination IP address, destination port number, and protocol ID.
Service level agreement	A service contract between a customer and a service provider that specifies the forwarding service a customer should receive. A customer may be a user organization or another provider domain (upstream domain).
Traffic profile	A description of the properties of a traffic stream, such as rate and burst size.
Differentiated services (DS) field	The field in which the differentiated services class is encoded. It is the Type of Service (TOS) octet in the IPv4 header or the traffic class octet in the IPv6 header.
Per-hop behavior (PHB)	The externally observable behavior of a packet at a DS-compliant router.
Mechanism	A specified algorithm or operation (e.g. queuing discipline) that is implemented in a router to realize a set of one or more per-hop behaviors.
Admission control	The decision process of whether to accept a request for resources (link bandwidth plus buffer space).
Classification	The process of sorting packets based on the content of packet headers according to defined rules.
Behavior aggregate (BA) classification	The process of sorting packets based only on the contents of the DS field.
Multifield (MF) classification	The process of classifying packets based on the content of multiple fields such as source address, destination address, TOS byte, protocol ID, source port number, and destination port number.
Marking	The process of setting the DS field in a packet.
Policing	The process of handling out of profile traffic (e.g., discarding excess packets).
Shaping	The process of delaying packets within a traffic stream to cause it to conform to source defined traffic profile.
Scheduling	The process of deciding which packet to send first in a system of multiple queues.
Queue management	Controlling the length of packet queues by dropping packets when necessary or appropriate.
Traffic trunk	An aggregation of flows with the same service class that can be put into an MPLS label-switched path.

Figure 2.1: RSVP signaling

resources in IntServ. The signaling process is illustrated in Figure 2.1. The sender sends a PATH message to the receiver specifying the characteristics of the traffic. Every intermediate router along the path forwards the PATH message to the next hop which is determined by the routing protocol. Upon receiving a PATH message, the receiver responds with a RESV message to request resources for the flow. Every intermediate router along the path can reject or accept the request of the RESV message. If the request is rejected, the router will send an error message to the receiver, and the signaling process will terminate. If the request is accepted, link bandwidth and buffer space are allocated for the flow, and related flow state information will be stored in the router.

In addition to Best Effort Service that provides no QoS guarantee, IntServ proposes two service classes, i.e. Guaranteed Service and Controlled Load Service. The Guaranteed Service provides firm quantitative bound on end-to-end packet delay by controlling the queuing delay along the path. The Controlled Load Service is for applications requiring reliable and enhanced best effort service. Several traffic control mechanisms must be implemented in IntServ-enable routers: packet classifier, packet scheduler, admission control. Applications with Guaranteed Service or Controlled Load Service requirements use RSVP to reserve resources before transmission. Admission control will determine whether the router has enough resources to accommodate a new flow. Every router along the path will make a local accept/reject decision whether the QoS requirements

can be granted or not. The source starts to transmit the data packet after receiving the accept notification. The packet classifier is used to identify the flows that are able to receive a certain level of service. The packet scheduler provides scheduling service for different packet flows to meet different QoS requirements.

The biggest problem of IntServ is that it cannot scale well in the Internet core since the amount of state information increases proportionally with the number of flows. This places huge overhead on the routers. The requirement for guaranteed service will also require all routers to be IntServ-compatible which is hard to implement in current network.

Differentiated Service

Differentiated Service (DiffServ) [BBC⁺98] provides a limited number of aggregated classes of services by defining the layout of the Type Of Service (TOS) bits in the IP header, called the DS field, and a base set of packet forwarding rules, called Per-Hop-Behavior (PHB). In order for a client to receive differentiated services from its Internet service provider (ISP), it must have a service level agreement (SLA) with its ISP. An SLA basically specifies the service classes supported and the amount of traffic allowed in each class. Figure 2.2 shows the architecture of DiffServ.

A packet's DS field will be marked when it first comes into the DiffServ-enabled domain to indicate the desired service. At the ingress of the ISP networks, packets are classified, policed, and possibly shaped according to the rules derived from the SLAs. The amount of buffering space needed for these operations is also derived from the SLAs. When a packet enters one domain from another domain, its DS field may be remarked which is determined by the SLA between the two domains.

DiffServ provides a number of services. Premium Service is provided through expedited forwarding which provides the highest level of aggregate quality of service with

CHAPTER 2. LITERATURE REVIEW

Figure 2.2: DiffServ architecture

minimized delay and jitter. Assured Service is for applications requiring better reliability than Best Effort Service. Olympic service provides three tiers of services: *gold*, *silver*, and *bronze*, with decreasing quality.

Comparing with IntServ, DiffServ is more scalable since service is allocated in the granularity of a class rather than as a flow in IntServ. Sophisticated operations, such as classification, marking, policing, and shaping are only needed at the boundaries of the networks. Therefore, it is easier to implement and deploy DiffServ. The limitation of DiffServ is that only PHB routing is provided with no end-to-end QoS guarantee. Aggregated traffic may exceed service rate at core routers. Traffic engineering or constraint based routing can be used to avoid flow congestion caused by uneven traffic distribution.

2.2 Routing in MANETs

Unlike wired network, the links in wireless network are subject to change due to node movement or channel dynamics. Without reliable routing support in MANETs, no QoS can be guaranteed. Therefore, finding and maintaining routes in MANETs is actually a big issue. Many protocols have been specially proposed for MANETs, with the aim of achieving efficient routing. We first review single path routing protocols, then multipath routing protocols are introduced which are usually built upon the single path routing protocols.

2.2.1 Single Path Routing

One way to classify routing protocols in MANETs is according to when the routes are determined. A protocol is called proactive if the routes are determined before a node makes a request for transmission and the protocol can provide a route immediately by keeping the global information on a continuous basis. The proactive routing protocols are also called table-driven protocols in the literature [PB94, MGLA96]. The routing information is stored in one or more tables and are refreshed by propagating updates throughout the network in order to maintain a consistent view of the network. This implies high overhead incurred since every node has to update the table when a node moves, and such frequent exchange of routing table could take up a large part of the capacity of the network. When the mobility rate of nodes is high, proactive protocols are less efficient due to the outdated information being kept. An attempt to overcome the limitations of proactive protocols is to discover a route in an on-demand fashion. This type of protocols are called reactive routing protocols or on-demand routing protocols. In this type of routing protocols [PR99, JMHJ02, PC98], nodes are not required to update the routing information periodically. A control message is sent to find the route to a given destination only upon request. Once a route has been established, it is maintained

CHAPTER 2. LITERATURE REVIEW

by a route maintenance process until the route is no longer desired or the destination is unreachable. The disadvantage of such reactive protocols is that a sender may experience a long delay in waiting for the route to be discovered.

The proactive and the reactive routing protocols can be seen as flat protocols since all nodes act an equal role in routing. A hierarchical routing protocol combines both the proactive and reactive approach by organizing the nodes in groups and then assign nodes with different functionalities inside and outside a group. Inside the group, proactive mechanism is adopted to make fast decisions while on-demand request is used for inter-group communication. Therefore, the routing table size, update packet size, as well as the control overhead, are also reduced. Such hierarchical routing protocols include CGSR [CG97], ZRP [HP97] and HSR [PGHC99].

The protocols introduced above only use the information of links and network topology to make the routing decisions. With the development of GPS devices nowadays, several location-aware routing protocols have also been proposed and the routing decision at each node is based on the location of its destination and neighboring nodes. LAR [KV98] and DREAM[BCSW98] use such types of mechanisms.

We then introduce two on-demanding routing protocols in details since they represent two different routing strategies, i.e. source routing and hop-by-hop routing. The basic mechanisms discussed will serve as the base to build our proposed routing algorithms in this work.

DSR

Dynamic Source Routing protocol (DSR)[JMHJ02] is a source routing protocol, where the complete, ordered list of nodes in the discovered route is carried in the header of every data packet. Nodes in DSR are not required to exchange routing information periodically. Instead, they maintain route caches to store the source routes that they are aware of,

CHAPTER 2. LITERATURE REVIEW

and update entries in the route caches when they find new routes. The protocol consists of two processes, namely route discovery and route maintenance.

When initiating a route request, a node will first check its cache. If the node does not have a route to the destination, it will initiate route discovery by broadcasting a route request packet (RREQ). The route request packet contains the address of the source and destination nodes, and an unique sequence number. The intermediate nodes will add their addresses into the source header of the packet and rebroadcast the packet if they don't have route information in their cache and if they haven't received the route request before. A route reply (RREP) is generated when either the destination receives the route request packet or an intermediate node with current information about the destination receives the route request packet.

In the route maintenance process, a node can detect link failure by using feedback from the MAC layer or by using other mechanisms such as passive acknowledgment. A route error packet (RERR) will be sent to the source once such a link break is detected. All nodes that receive the route error packet will purge the bad link in their caches. The source will have to re-initiate the route discovery if it cannot find any other route information to that destination in its route cache.

AODV Ad hoc On-demand Distance Vector Routing protocol (AODV)[PR99] is a distance vector protocol based on a hop-by-hop routing approach. The routing procedure is similar to DSR, mainly consisting of route discovery and route maintenance. Comparing with DSR, several distinctions can be noted. First, the packet is routed based on the information kept in a routing table which is maintained by each node. In the routing table, main fields include the id of the destination, the id of the next hop through which the node can connect with the destination, the hop count to the destination and a sequence number which is necessary to avoid looping. Second, after a route request is received,

the routing table is updated by recording the reverse route information which will later be used to send the route reply message back to the source. Third, for the route maintenance part, link failures can be detected by periodic hello messages exchange, which will then cause a route error packet to be sent back to all sources to erase the route entries containing the failed link.

2.2.2 Multipath Routing

For single path routing protocols such as DSR and AODV, when a link failure happens, nodes of the broken route simply drop data packets and ask for a new route establishment. Such disruption is not suitable for the ad hoc environment, which is vulnerable to link failure caused by node mobility, interference and packet collision. Multipath routing protocols are thus proposed to enhance the route reliability by providing alternate routes. In case of primary route failure, the traffic can be immediately shifted to the alternate path which reduces the route discovery latency. We then introduce two multipath routing protocols that use routing strategies like DSR and AODV.

SMR Split Multipath Routing (SMR) [LG01] is similar to multipath DSR using source routing. Instead of dropping duplicated RREQs which mostly generate overlapped paths, intermediate nodes forward the duplicated RREQs that come from different incoming link to destination and two routes that are maximally disjoint can be chosen. To find disjoint paths, intermediate nodes are not allowed to send REPLY back. When a link failure occurs, two policies are defined to rediscover routes: *SMR-1* which performs the route recovery when any route to the destination is invalidated and *SMR-2* which performs the route recovery only when both routes to the destination are invalidated.

AOMDV On-demand Multipath Distance Vector Routing (AOMDV) [MD01], on the other hand, is a multipath extension to AODV which can provide multiple loop-free and

link-disjoint paths. Several changes are thus needed in the route discovery phase. First, duplicate RREQs are examined to see if they provide new node-disjoint paths to the source. Second, to guarantee loop freedom, multiple routes are accepted and maintained at the intermediate nodes, but only alternate routes with lower hop counts are accepted. Third, at the destination, it replies up to k copies of RREQ. For the route maintenance phase, AOMDV will try to find a new path only when all paths to the destination fail.

2.3 QoS in Wireless Ad Hoc Networks

Recently, the rising popularity of multimedia applications among end users in various networks and the potential usage of wireless ad hoc network commercially have led to research interest in providing QoS support in such networks. This, however, is a big challenge considering all the constraints:

- There is no core and edge distinction and all nodes are homogeneous in QoS provision roles. The conventional client/server architecture does not exist in MANETs, and this requires all computation to be conducted in a distributed manner.
- The broadcast nature of the shared medium makes the available capacity of wireless link vary with changing topology.
- Imprecise information is another big challenge for supporting QoS in MANETs. The link state information such as delay and bandwidth is changing unpredictably over time, so the information obtained at session setup will become outdated as soon as the node moves out of transmission range.

In the literature, researchers report work on various aspects of QoS provisioning in MANETs including QoS models, QoS resource reservation signaling, QoS routing, and QoS Medium Access Control (MAC). All of these research work provides possible

solutions to certain extents. For a better comprehension of the relationship of these works, a comprehensive introduction to a selection of seminal work on providing QoS support in MANETs is presented in this section.

2.3.1 QoS Model

The QoS model specifies the architecture in which certain service could be provided in the network. A QoS model for MANETs should first consider the challenges of MANETs, e.g. dynamic topology and time-varying link capacity. In addition, the potential commercial applications of MANETs require seamless connection to the Internet. Thus the QoS model for MANETs should also consider the existing QoS architecture in the Internet. Two QoS models for MANETs are proposed by Xiao [XSLC00] and Ahn [ACVS02].

FQMM

A Flexible QoS Model for MANETs (FQMM) is proposed [XSLC00]. The model considers the characteristics of MANETs and combines both the per-flow service granularity in IntServ and the service differentiation in DiffServ.

FQMM supports dynamic roles of nodes. As in DiffServ, three kinds of nodes are defined. An ingress node is the source node that sends data. Interior nodes are the nodes that forward data to other nodes. An egress node is a destination node. The roles of the nodes change with the topology and network traffic.

The provision in FQMM refers to the determination and allocation of resources needed at various points in the network. Unlike per-flow granularity in IntServ and per-class granularity in DiffServ, a hybrid per-flow and per-class provisioning scheme is proposed for FQMM. A traffic of highest priority is given per-flow provisioning while other priority classes are given per-class provisioning. Not all portions of traffic in MANETs should be treated in per-flow granularity due to bandwidth limitations and other constraints.

Figure 2.3: SWAN model

A traffic conditioner is put at the ingress node which consists of a traffic profile, meter, marker and dropper. It is responsible for re-marking the traffic streams, discarding or shaping packets according to the traffic profile, which describes the temporal properties of a traffic stream such as rate and burst size.

FQMM is the first attempt at proposing a QoS model for MANETs. The biggest problem in FQMM is that making a dynamically negotiated traffic profile is a very difficult task in MANETs. Also, the details of the implementation is not discussed in depth in their work.

SWAN

SWAN [ACVS02] is a stateless network model which uses distributed control algorithms to deliver service differentiation in MANETs in a simple, scalable and robust manner. Comparing with FQMM, SWAN works on a more specific scenario where only two classes of traffic exist, best effort traffic and real time traffic.

CHAPTER 2. LITERATURE REVIEW

The SWAN model is illustrated in Figure 2.3. Before a real time session is set up, the source sends a special probing packet to probe the bottleneck bandwidth of the path, and the real time session can be admitted only if the end-to-end available bandwidth is enough to accommodate the new session. The end-to-end bandwidth violation can be detected by each node continuously and independently measuring the utilization of its real time traffic. When a node detects such a violation, it starts marking the ECN bits in the IP header of the real time packets. The destination node monitors the ECN bits and notifies the source using a regulate message. When a real-time flow can no longer be supported, it re-establishes the route and probes the path again. For the class of best-effort traffic, traditional AIMD-based rate control algorithm is used. Intermediate node independently control the sending rate of best-effort traffic based on the delay performance monitored at the MAC layer. In this way, differentiated service is provided for two classes of traffic in MANETs.

In SWAN, if a session is broken due to node movement or congestion, the source must wait for the sent probe to return before resuming the real time traffic. Since SWAN relies on the underlying routing protocol to transmit the packet, the pause time of the real time traffic will depend on the speed of the new route discovery. The re-routing can make the AIMD algorithm spend multiple round-trip-time (RTTs) to align itself to the new share of the bandwidth. In the case of frequent re-routing, the algorithm may never be able to catch up with the dynamics of the network, resulting in sub-optimal performance and poor fairness. The local rate control of best effort traffic requires the MAC layer to be able to regulate the sending rate which adds more burden to lower layers.

2.3.2 QoS Signaling

QoS signaling is used to reserve and release resources, set up, tear down, and renegotiate flows in the network. The QoS signaling system can be divided into in-band signaling and

CHAPTER 2. LITERATURE REVIEW

Figure 2.4: INSIGNIA QoS framework

out-of-band signaling. In-band signaling refers to the approach that control information is carried along with data packets, and out-of-band signaling refers to the approach that uses explicit control packets. RSVP [BZB⁺98] is an out-of-band signaling system in conventional wired networks. An in-band signaling system in MANETs is introduced below.

INSIGNIA

The primary design goal of the INSIGNIA QoS framework [LAZC00] is to support adaptive services which can provide base QoS (i.e. minimum bandwidth) assurances to real-time voice or video flows, and allow for enhanced levels of service (i.e. maximum bandwidth) to be delivered when resources become available. INSIGNIA is designed to adapt user sessions to the available level of service without explicit signaling between source-destination pairs.

Figure 2.4 shows the position and the role of INSIGNIA in wireless flow management

CHAPTER 2. LITERATURE REVIEW

at a mobile host. The in-band signaling module controls the establishment, restoration, adaptation, and destruction of adaptive QoS-aware paths between source-destination pairs. Admission control is responsible for allocating bandwidth to flows based on the minimum/maximum bandwidth (i.e. base and enhanced QoS) request. The admitted traffic is transmitted in a conformed way through the packet scheduler which can use a wide variety of scheduling algorithms such as Weighted Round-Robin (WRR) [LBS97, SV95]. The routing protocol and MAC layer can be any existing protocols, but the performance of the framework is strongly coupled with specific methods.

The flow state in INSIGNIA is managed in a soft-state manner. Unlike in wireline networks that support QoS and state management, quality of services is guaranteed for the duration of the session once it is admitted. However, QoS in MANETs is fragile and route discovery and resource reservation need to adapt to the topology changes dynamically. Soft state is a flexible approach, and it is suitable for the management of reservations at intermediate routing nodes in MANETs. The reservation state at an intermediate node is associated with a timer which is refreshed by the on-going traffic. If a node fails to receive the expected data for a defined period of time, the state will become obsolete and the reserved resources will be released automatically. The soft state is appropriate for distributed system where no central management exists, such as MANETs.

As a whole, the INSIGNIA framework can provide assured adaptive QoS levels to real-time applications, based on the QoS requested by the applications and the resource availability in the MANETs.

2.3.3 QoS Routing

QoS routing protocols search for routes with sufficient resources for QoS requirements. The QoS routing protocols should work together with resource management to establish

paths through the network that meet end-to-end QoS requirements, such as delay or delay jitter bounds, bandwidth demand, or multi-metric constraints. The literature in QoS routing includes Ticket-based Probing [CN99] and CEDAR [SSB99].

Ticket-based Probing

A ticket-based probing (TBP) algorithm with imprecise state model was proposed by Chen and Nahrstedt [CN99]. The basic idea is to discover a QoS-aware routing path with limited amount of flooding (routing) messages by issuing a certain amount of logical tickets. One ticket corresponds to the search of one path; and one probe message should carry at least one ticket. So the maximum number of the searched paths is bounded by the tickets issued from the source. When an intermediate node receives a probe message with n tickets, it decides whether to and how to split the n tickets and where to forward the probes based on its local state information. More tickets are issued for connections that have tighter or higher requirements. When the destination receives a probe message, a possible path from the source to the destination is found.

To deal with imprecise information in MANETs, history and current (estimated) delay variation are used by a smoothing formula to calculate the current delay. To adapt to the dynamic topology of ad hoc networking, three levels of route redundancy are provided. The highest level of redundancy is applied by sending every data packet along each path independently. The second level of redundancy only uses the primary path for transmission and resources are reserved at the secondary path while the third level does not even reserve the resources in the secondary path. Re-routing and path-repairing techniques are adopted for route maintenance. When a link breakage is detected, it will notify the source to reroute the connection to a new feasible path, and notify the intermediate nodes along the old path to release the corresponding resources. Unlike the re-routing technique, the path-repairing technique does not find a completely new path. Instead, it tries to repair the path using local reconstructions.

CHAPTER 2. LITERATURE REVIEW

In TBP, every node relies on the local state and the end-to-end state to search for the route. As we know, it is very difficult to maintain such states in a high mobility network. The level of imprecision can have a direct impact on the routing overhead and the success ratio of route discovery.

CEDAR

Sinha et al. proposed a Core-Extraction Distributed Ad-hoc Routing (CEDAR) algorithm [SSB99], which is proposed as a QoS routing scheme for small to medium size ad-hoc networks consisting of tens to hundreds of nodes. The basic idea is to dynamically establish the cores of the network, and then incrementally propagate the link states of stable high-bandwidth links to the core nodes. The route computation in CEDAR is on-demand which is performed by the core nodes using only the local state. The core broadcast approach is used to limit the flooding and to efficiently update the link information with topology change. CEDAR has three key components:

Core Extraction: The dominating set is a subset of the network in which every node not in the set is adjacent to at least one node in the set. A minimum dominating set is one such set with minimum cardinality. The core of the network in CEDAR is an approximation of a minimum Dominating Set (DS) of the network. A set of nodes is elected to form the core that maintains the local topology of the nodes in its domain, and also performs route computations.

Link State Propagation: QoS routing in CEDAR is achieved by propagating the bandwidth availability information of stable links to all core nodes. The basic idea is that information about stable high-bandwidth links can be made known to nodes far away in the network, while information about the dynamic or low bandwidth links remains within the local area.

Route Computation: Route computation first establishes a core path from the domain of the source to the domain of the destination. Using the directional information

Figure 2.5: A is hidden from C and exposed to B

provided by the core path, CEDAR iteratively tries to find a partial route from the source to the domain of the furthest possible node in the core path satisfying the requested bandwidth. This node then becomes the source of the next iteration.

In CEDAR, the core provides an efficient and low-overhead infrastructure to perform routing, while the state propagation mechanism ensures the availability of link-state information at the core nodes without incurring high overhead. However, the core may be broken at transient time period during which the routing cannot be effectively done and it's common in the dynamic nature of ad hoc networks. Furthermore, searching for a QoS-constrained path is directed by the core. The tree structure of the core may not lead to find the shortest path.

2.3.4 QoS MAC

QoS supporting components at upper layers, such as QoS signaling and QoS routing, assume the existence of a MAC protocol, which solves the problems of medium contention and supports reliable unicast communication. In multihop wireless networks, a fully distributed scheme is needed that should first solve the hidden terminal problem.

In Figure 2.5, the dot circle indicates the transmission range of a mobile host. Host A and C are out of range from each other but both within the range of B . When A

is transmitting a packet to B , C cannot sense the transmission from A . Thus C may transmit a packet to B and cause a collision at B . This is the *hidden terminal* problem since A is hidden from C . Similarly, when B is transmitting a packet to C , A cannot initiate a transmission to other nodes at the same time even if that node is out of range of B since A can sense the transmission within its own range and thus defers its transmission. This is called the *exposed terminal* problem as B is exposed to A . In the following we will show how the IEEE 802.11 [IEE99] solves the hidden/exposed terminal problems and we will also introduce an extension to 802.11 that can provide differentiate service.

IEEE 802.11 Distributed Coordination Function (DCF)

The IEEE 802.11 is a carrier sense multiple access with collision avoidance (CSMA/CA) protocol. In the DCF mode, a node with packet to transmit must first sense whether the medium is idle for a time period longer than the DCF interframe space (DIFS). If this is not satisfied, the node defers the transmission and starts the backoff procedure. The duration of the backoff time is decided by the value of contention window (CW) which is randomly chosen from a bound between CW_{min} and CW_{max} . The backoff time duration will decrease if the medium is sensed to be idle for a period longer than DIFS. As soon as the timer expires, the node starts to transmit. To reduce packet collisions, an optional RTS/CTS scheme is used as shown in Figure 2.6. The sender will first send a short packet called request-to-send (RTS) if the data packet is longer than a certain threshold value. If the intended receiver grants the request, it will return another short packet called clear-to-send (CTS). Upon receiving the CTS, the sender will start sending the data packet, while other nodes will try to avoid collision with the upcoming data packet by updating their Network Allocation Vector (NAV) with the duration value carried in the overhead frame.

Figure 2.6: RTS/CTS access scheme

The IEEE 802.11 is not specially designed for multihop wireless network and past work [XS01] revealed that it does not function well in MANETs. There is still a lot of work to be done to improve the MAC in MANETs so that algorithms in the upper layers can be properly implemented.

IEEE 802.11e Enhanced Distributed Coordination Function (EDCF)

The IEEE 802.11 DCF is a good example of a best-effort type control algorithm. It has no notion of service differentiation and provide no support for real-time traffic. But as we can see from above, differentiated service can be provided by setting various waiting time for nodes when they contend for the same resource. In IEEE 802.11e [IEE05], enhanced version of DCF (EDCF) is introduced to support QoS which is realized with the introduction of Traffic Categories (TCs) that are associated with different channel access parameters. Each TC will start a backoff after detecting the channel being idle for an Arbitration Interframe Space (AIFS) which is at least DIFS and can be enlarged individually for each TC. In EDCF, CW_{min} and CW_{max} are parameters set based on the priority of each TC. The process of several TCs contending for the medium in IEEE 802.11e EDCF is shown in Figure 2.7.

Figure 2.7: Enhanced Distributed Coordination Function in 802.11e

2.4 Summary

In this chapter, an overview of QoS related literature is presented. As we can see, QoS can be provided at the application layer, transport layer, network layer, MAC layer, and the physical layer of the network model. Although QoS provisioning has been widely studied in wired networks, providing QoS support in wireless ad hoc networks is still a challenge. Many existing techniques developed in the traditional wired network can be used in wireless ad hoc networks directly or with little modification, such as resource reservation, classifier, packet scheduler, etc. However, the dynamic nature of the network topology and the changing behavior of the communication medium in wireless ad hoc networks make the precise maintenance of network state information very difficult. With the routing protocol operating with imprecise information, frequent QoS violations are likely to happen during a connection. And if we consider node movement as well, the reduction of the re-routing latency is also very important to provide consistent service. We argue that QoS routing is indispensable in the QoS framework for multihop wireless ad hoc networks since the achievable capacity in such ad hoc networks is determined by the topology being formed. Another important issue is how to characterize the QoS

CHAPTER 2. LITERATURE REVIEW

requirement for heterogeneous traffic and also new scheduling policies should be designed to support integrated services.

Chapter 3

Enhancing QoS for Real-time Traffic in Mobile Ad Hoc Networks

3.1 Introduction

Mobile Ad Hoc Networks (MANETs) have been widely studied by researchers due to the unique characteristics. In MANETs, without infrastructure support, all nodes act in an autonomous manner and rely on the intermediate nodes to help forward the traffic. The easy deployment and self-adaptation make such networks suitable for tough situations such as emergency rescue, military fields, etc.. MANETs can also provide extension to existing infrastructure networks. Many applications in traditional wired network can be implemented in MANETs which can give more flexibility and convenience to the end users. Quality of Service (QoS) support in MANETs has earned a lot of attention in the research community with the rising popularity of multimedia applications and potential commercial usage of MANETs. A lot of work have been done in supporting QoS in the Internet, but when trying to apply those mechanisms to MANETs, many challenges arise. All nodes in MANETs are homogeneous and have the same roles in QoS provisioning, and it renders any central control mechanism not suitable for such networks. It is also not feasible to maintain global information at any node since it will incur high overhead in the message exchange. Considering the dynamic topology, it is difficult to maintain and

CHAPTER 3. ENHANCING QoS FOR REAL-TIME TRAFFIC IN MOBILE AD HOC NETWORKS

manage accurate link state information such as delay and bandwidth. Even if a resource reservation request is accepted at the start time with sufficient resources available, it cannot be guaranteed during the session when nodes could move in an arbitrary way and the frequent link breakages may interrupt the connection totally.

In this chapter, we consider a real-time application which requires a small delay and reliable transmission even in high mobility environment. Such application is more popular now, such as VOIP, video conference, multimedia-on-demand, etc., and there is a strict minimum bandwidth requirement and a delay/jitter bound to be met. The user usually will be willing to pay more for something better than best-effort service. On the other hand, for best-effort traffic such as data services like ftp, web browsing, etc., no hard QoS is required and they are just delivered with best effort, i.e., with no guarantee. Obviously, real-time traffic should get higher priority over best effort traffic. In our proposal, the QoS requirement of a real-time traffic is characterized as a delay bound and a minimum bandwidth requirement which are included in its traffic profile and the failure to achieve them will lead to QoS violation.

Some early works on QoS framework in MANETs, such as SWAN [ACVS02] and INSIGNIA [LAZC00], try to build a QoS framework to manage the resources and schedule traffic through in-band signaling (Wu et al. [WH01]). Their schemes have the advantage that their architecture is kept separated from the routing protocol which reduces the implementation complexity and allows step-by-step deployment. However, unlike traditional wired networks where the links are fixed once the session is set up, frequent node movement in MANETs may cause the topology to change and the communication interrupted. In the other hand, the additional signaling in their QoS framework also incurs more overhead. Other researchers [LG01, MD01, NCD01] work on providing more than one path in route discovery. Their works show that providing more than one path in MANETs can better support the high mobility network by reducing connection

CHAPTER 3. ENHANCING QoS FOR REAL-TIME TRAFFIC IN MOBILE AD HOC NETWORKS

interruption. However, these protocols are only designed for best-effort traffic. In MP-DSR [LLP⁺01], a new QoS metric called end-to-end reliability is included in multipath discovery, but it is not specifically designed to support real-time traffic. Chen [CN99] also proposed three levels of route redundancy to reduce the impact of mobility. However, some key issues such as how to find disjoint paths and how to maintain multiple paths are not discussed in his paper.

In MANETs, the QoS requirements established at the beginning for real-time traffic, such as delay bound and minimal bandwidth, are likely to be violated from time to time due to the dynamics of such networks. In this chapter, the objective is to enhance QoS for real-time traffic and minimize such QoS violations. We argue that a reliable and effective routing protocol should be our first consideration since end-to-end performance relies on a reliable route. Any interruption in the route can incur additional delay for recovery and overhead as well. The admitted real-time traffic in the network should be served with higher priority. A QoS framework is then proposed to achieve the objective, which consists of a QoS routing protocol with multipath support and a scheduling algorithm to provide differentiated services. The following features in our framework are highlighted: (1) effective disjoint paths discovery and maintenance, (2) instant QoS violation detection, (3) fast route recovery, (4) efficient resource reservation and release with route changing and (5) enhanced QoS for real-time traffic with differentiated services provision.

The rest of the chapter is organized as follows. Section 3.2 gives an overview of our framework. Section 3.3 and Section 3.4 propose the QoS routing protocol and differentiated services provision. The performance of our framework is then evaluated through simulations in Section 3.5. Section 3.6 gives a summary of the chapter.

3.2 Framework Overview

Our proposed QoS framework has two major parts: a QoS routing protocol and a scheduling algorithm. The routing protocol is responsible for providing multiple reliable and QoS qualified routes and the scheduling algorithm then provides differentiated services. In our routing protocol, two node-disjoint paths are discovered with QoS requirements satisfied. The traffic will be admitted into one of the path (primary path), meanwhile the other path (alternate path) that is not used is also maintained up-to-date. When QoS violation is detected in the primary path, i.e. a large delay incurred by link failure or excessive contention, the source can quickly switch real-time traffic to the alternate path without invoking another route discovery. In this way, the re-routing latency is minimized. After the admission into the network, there is no central control for the traffic. The packet scheduler then provides effective management for both real-time traffic packet and best-effort packet. The QoS for real-time traffic is enhanced through local packet scheduling in a totally distributed manner.

The system model of the proposed system is shown in Figure 3.1. When a request for real-time traffic comes from the upper application layer, the source performs admission control by invoking the routing module to search for the paths satisfying the requirements of the requester. Resource reservation is then performed by updating the resource table of every node along the discovered route. After the source gets a successful reply (which means the routes satisfying the required QoS are found), real-time traffic is admitted into the network.

3.3 QDMR: QoS routing with Multipath Discovery

In this section, we propose a new QoS disjoint multipath routing protocol (QDMR) using source routing type mechanism. To start traffic transmission, a node needs to invoke route

Figure 3.1: System model

discovery to find multiple routes with enough resources. After traffic is admitted, route maintenance is needed to minimize the impact of QoS violation due to either link failure or excessive contention.

3.3.1 Route Discovery

3.3.1.1 Disjoint Paths

To maximize the independence of two paths and thus improve the reliability, disjoint paths are preferred in QDMR. To describe how two paths are overlapped, disjoint degree is defined as the number of overlapping nodes of two paths, excluding the source node and the destination node. For two paths with least overlapping nodes, i.e. disjoint degree equals to 0, they are completely node-disjoint. It can be easier seen that paths with a larger disjoint degree are more likely to fail simultaneously if the overlapping nodes move out of range, assuming that node movement is an independent behavior. Figure 3.2 shows

Figure 3.2: Disjoint paths

two types of disjoint paths in the same scenario, a link-disjoint path with disjoint degree of 1 and a node-disjoint path with disjoint degree of 0. As we can see, the node-disjoint path can provide better reliability that there is still a path available after a link on the other path fails. For those scenarios with group node movement, the probability of a link breakage within the group is low, so the disjoint criteria can be applied only to the part of the route that stays outside the group.

According to the definition, the number of completely node-disjoint paths is bounded by the number of one-hop neighbors of the source or of the destination, whichever is smaller. In DSR, the source can obtain multiple routes if the destination replies to more than one request. However, the nodes in these paths are likely to be overlapping as being pointed out in [LG01]. It can be explained that the faster propagating RREQs will block those later arrived RREQs at the intermediate nodes, so the probability of those later arrived RREQs carrying disjoint paths is greatly reduced. It's illustrated in Figure 3.3 in which the propagation of RREQs is shown. The data in the figure is obtained from a simulation using DSR as the routing protocol. The path travelled by the first arrived

Figure 3.3: RREQ propagation in DSR

RREQ (indicated by solid arrow with arrow head pointing to the propagation direction) to the destination intersects the other possible shortest paths (e.g. connected by dashed arrows) that go through a different first hop neighbor of the source (node *B* in the figure), so these later discovered paths cannot be completely node-disjoint with the first one unless taking extra hops. In QDMR, the first hop information in RREQ is used to discover more node-disjoint paths.

3.3.1.2 Route Exploration

When the source node gets a request from a real-time application, it starts the route discovery by broadcasting a route request (RREQ). The following information will be piggybacked in RREQ: source id, destination id, a unique sequence number, route information and QoS requirements for real-time traffic. A RREQ from a certain source can be uniquely identified by the pair of source id and sequence number. The route information will be updated by the intermediate nodes by inserting their own node id in RREQ (same as source routing protocols like DSR).

Each node will maintain a routing table which stores the routes that have been discovered, and a resource table which records the bandwidth that is reserved for the real-time traffic. When the intermediate nodes receive a new RREQ, it will record the first hop node id and the sequence number. A duplicate RREQ will be rebroadcast by the intermediate node only if: that RREQ contains a different first hop node comparing with that in previous received RREQs; and the hop count of the path contained in RREQ is not larger than those of previous received RREQs. The latter principle guarantees that RREQ will not travel in the reverse direction and thus route-looping be avoided.

3.3.1.3 QoS Metrics

Taking QoS requirements of real-time traffic into consideration, we choose end-to-end delay and minimum bandwidth as the metrics for QoS routing. In highly dynamic MANETs, it is not feasible to have accurate measurement of the end-to-end delay which can be influenced by many factors. The end-to-end delay tends to increase with the length of the route since traffic will experience queueing, contention and so on at each hop. Most routing protocols are optimized to find the shortest path to minimize the end-to-end delay. In QDMR, we set a *one-hop-delay* limit which is a conservative estimation of the average one hop traversal time for packets which considers the queueing delay, MAC backoff time and the propagation delay. Its value is set based on simulation results with a protective margin added. For example, the value of it is set to 30 ms in AODV (included in NS-2 [NS2]). We then use the hop count of the path to estimate the end-to-end delay that may be experienced, which is the result of dividing the delay bound of real-time traffic by *one-hop-delay*. The Time To Live (TTL) field in RREQ is then set to this value, and RREQ will be dropped at intermediate nodes if it has travelled more than TTL hops.

The bandwidth estimation is also an open issue in MANETs especially when we take the *hidden terminal* problem (Section 2.3.4) into consideration. The accurate mea-

surement of the bandwidth is not the research issue in our work. Similar to the delay estimation, a conservative *admission rate* is used as the threshold for every node in our model. We should not admit real-time traffic up to this threshold rate for a number of reasons: First, best-effort traffic would be starved of resources should real-time traffic consumed bandwidth up to the threshold rate. Second, there would be no flexibility to tolerate network dynamics like those due to changes in traffic patterns and host mobility. For example, due to host mobility, there is more overhead for route maintenance which will consume more resources. The available bandwidth for real-time traffic in a node can be calculated as the difference between the *admission rate* and the minimum bandwidth required by real-time traffic. RREQ will be dropped if the available bandwidth at the intermediate node cannot satisfy the minimum bandwidth requirement of real-time traffic.

3.3.1.4 Route Establishment

As studied by Nasipuri and Castaneda in [NCD01], keeping more than 2 paths makes little improvement to the total performance considering the trade-off with the overhead. In our scheme, the destination will choose 2 paths with the largest disjoint degree from the paths contained in its received RREQs, and send replies (RREPs) separately to the source. The path contained in RREQ will be piggybacked in RREP, as well as the sequence number of RREQ. To reduce the propagation delay, the pair of paths with the least hop count is preferred.

To get more paths from RREQs, the destination will set a timer which starts after the first RREQ from a source is received. To reduce the latency of the route discovery, replies can be sent back immediately once two completely node-disjoint paths are found before the timer stops. However, the computational load will not increase exponentially with the number of received RREQs, because the first hop information in RREQ can be

Algorithm 3.1 Route Discovery in QDMR

```

/*when a node receives a RREQ, it will check the information for first hop, hop count,
QoS requirements*/
if this node is not destination then
    if the first hop is the same as that in one of previously received RREQs
    or the hop count is larger than that of all previously received RREQs
    or the path contained in this RREQ does not satisfy QoS requirements then
        discard this RREQ
    else
        append its id to RREQ and rebroadcast it
    end if
else
    /*this node is the destination*/
    if this is the first RREQ that has been received from the source then
        start timer and wait for other RREQs from the same source
    else if timer not expire then
        if two completely node-disjoint paths have been found then
            send back RREPs immediately
        end if
    else
        discard this RREQ
    end if
end if

```

used to decide whether to perform the selection algorithm or not. Only when two paths contain different first hop node will destination further check rest of the nodes on the paths. Finally, after source receives RREPs, the path obtained from the first arriving RREP is considered as the primary path, and the path in late one is the alternate path.

The routes for real-time flows are computed in an on-demand way as described above. For best-effort traffic, route is first chosen based on the existing information in the routing table to speed up the route discovery and reduce the routing overhead.

Algorithm 3.1 summarizes the route discovery procedures.

3.3.1.5 Bandwidth Reservation

Before a qualified RREQ will be rebroadcast by the intermediate node, the bandwidth for real-time traffic will be reserved temporarily before route establishment. Also, even

established routes may stop being used or fail QoS requirements due to the channel dynamics, node mobility or even power failure. Thus efficient resource release mechanisms are needed to free the reserved resources at each node on the path if no reply is sent for RREQ or the established routes are broken or no longer used. This is done automatically by the use of soft state as introduced in Section 2.3.2. The reservation state of each nodes on established paths will be periodically refreshed by real-time traffic, and those not being refreshed for a pre-specified duration period will be purged. The reserved resources can thus be efficiently utilized in the networks.

3.3.1.6 Traffic Admission

After the source starts route discovery, it will expect to receive RREP before a certain time. Once two paths that satisfy QoS requirements are replied by destination before the time expires, real-time traffic will be admitted into the primary path and the alternate path will be maintained for future use. During the RREP propagation, one or two RREPs may fail to arrive at the source. If only one path is available when the timer runs out, the source will check whether this path is robust enough for real-time traffic by checking the length of this path. Real-time traffic can be admitted into this path if the length of the path is less than MAX_HOP and will keep finding the alternate path for future use. MAX_HOP is the result of a tighter delay bound divided by *one-hop-delay*, which is based on observation for the performance of real-time traffic in the simulations. With shorter distance between two nodes, the connection between them requires less transmissions over hops and is thus less prone to interruption. Meanwhile it will also take less time to recover from a route failure. If the route is not robust enough, the source may back-off and start a new route discovery procedure later or reject the flow request since the QoS for the real-time flow cannot be guaranteed.

3.3.2 Route Maintenance

3.3.2.1 QoS Violation Detection

QoS violation may occur due to link failure or excessive contention. The common approach used in existing ad hoc routing protocols for link failure detection is neighbor loss detection [PB94], i.e., when the beacon message from a node does not arrive at its neighbor in time. When neighbor loss is detected, a route breakage message is sent to the source to notify it of the breakage. However, the route breakage detection method by neighbor loss detection normally takes several seconds because the sending interval of beacon is usually set to large value to avoid flooding network, and it is not desirable for time sensitive real-time flows which have strict QoS requirements. In addition, the notifying process may incur more control overhead.

To quickly detect the link breakage, passive acknowledgment [JT87] is adopted in our proposal. This is illustrated in Figure 3.4. When node M receives a packet from L and unicasts it to N , node L can hear the next transmission. This single transmission not only forwards the packet to the next node in the route, but also acknowledges reception of the packet from the previous node. A packet is deemed lost if a node cannot hear the packet which was last transmitted by itself within *one-hop-delay*. Considering the channel dynamics and possible collision, a node may not hear the passive acknowledgement from its next hop temporarily. To solve this problem, only after *max-lost-num* (3 is used in our simulation) of packets are consecutively lost can the link to the next hop be determined broken. However, for the destination node D in the figure, such mechanism cannot be applied since that packet will not be propagated further by the destination node. An explicit acknowledgment is thus required from the destination node to confirm the successful reception of the packet.

If there is excessive contention in the network like due to burst of traffic, packets in the queue will have to wait for a longer time before being transmitted. This type of QoS

Figure 3.4: Passive acknowledgment

violation can also be detected using passive acknowledgement.

Using passive acknowledgment, QoS violation due to link breakage or excessive contention can be detected within $one-hop-delay * max-lost-num$. By exploiting the broadcast nature of wireless transmission, additional acknowledgement overhead is avoided. However, to use this passive acknowledgment, the nodes must work in promiscuous mode, that is, it must cooperate with the MAC layer to get the broadcast information even for transmission that is not destined to it.

3.3.2.2 Alternate Path Maintenance

After real-time traffic is admitted, the primary path will be refreshed periodically by ongoing traffic and QoS violation can be detected shortly using passive acknowledgement. For the alternate path, however, a small probe packet will need to be sent periodically from the source to the destination to both refresh the resource reservation state and detect link breakage. Since the alternate path is not used for real-time traffic transmission at the moment, the bandwidth reserved at the nodes on the path can be utilized by best-effort traffic to improve overall throughput. This will be discussed in Section 3.4.

3.3.2.3 Route Recovery

In our scheme, the route recovery will be invoked once QoS violation is detected on the path. Although this will add more overhead into the network, it is reasonable to do so

CHAPTER 3. ENHANCING QoS FOR REAL-TIME TRAFFIC IN MOBILE AD HOC NETWORKS

to provide the level of service required by the higher priority user.

When a QoS violation is detected, the node will inform the source node of the violation by sending a route error (RERR) message which contains the id of that flow. The intermediate nodes will release the resources reserved for that flow upon receiving RERR. For the nodes that are downstream of the broken link, the reserved resources will be automatically released after timeout period since there is no real-time traffic to refresh the soft state.

If a RERR is received from the primary path, the source will immediately switch real-time traffic to the alternate path which is maintained up-to-date and it is promoted to primary path. If the alternate path does not exist (this may occur when the route discovery for the alternate path fails to keep up with the fast-changing topology), real-time traffic will be dropped temporarily and will restart later; otherwise, real-time traffic will just suffer transient interruption for maximally a single trip time for RERR propagation from the destination to the source. If RERR is received from the alternate path, the source will just purge this path and start route recovery procedure.

Route recovery is invoked after a RERR is received. Since the real-time application is still in transmission, the route recovery packet might interfere with the ongoing transmission in the shared medium. Also, the new path is still preferred to be node-disjoint with the primary path as much as possible. With one path to the destination already known, some information is available at the source node such as the hop count and the intermediate nodes on the path. In QDMR, a new RREQ (we name it *type 2 RREQ*) is proposed to search for the new alternate path. In this new RREQ, the TTL is set to the hop count of the primary path plus one for the disjoint path selection. *Type 2 RREQ* will also carry the ids of all the nodes on the primary path as the payload to heuristically guide the propagation. This *type 2 RREQ* will propagate through the network following the rules defined in Section 3.3.1 except: if more than half of the nodes on the path it

has already traversed overlap with those on the primary path, this *type 2 RREQ* will be dropped. The new alternate path will be less reliable if it's more than 50% of the possibility that itself and the current active path will fail simultaneously. The remaining procedure stays the same as that defined in route discovery. In QDMR, most route discoveries are initiated by *type 2 RREQ* and such enhancement can dramatically reduce the flooding overhead.

3.4 Scheduling Policies

In previous sections, a QoS routing protocol with multipath support is proposed. Real-time traffic is admitted into the network if a single robust route or multiple routes with QoS requirements met are discovered. No admission control is performed at the intermediate nodes since most of the real time applications do not implement congestion control and the admission control for real-time traffic should work end-to-end. Meanwhile, traffic other than real-time traffic can coexist in the network. In this work only best-effort traffic and real-time traffic are concerned. The reserved resources for real-time traffic may not be available all the time due to the dynamics of networks. The best-effort traffic, if not controlled, will also impact the QoS for real-time traffic. To enhance the performance of real-time traffic after its admission, a local scheduling algorithm is proposed to provide differentiated services.

It is not feasible to implement a central control algorithm in highly dynamic MANETs where it is difficult and costly to track and update global information, and we argue that a distributed control algorithm is the most suitable for MANETs. In our proposal, we differentiate two classes of traffic through local packet scheduling. The **priority queueing with windows** [EGSM⁺95, SEGEG⁺95] scheme is adopted to manage the queues in our proposal because it's simple to be implemented with little overhead and requires no global information, making it suitable for MANETs. The original algorithm

CHAPTER 3. ENHANCING QOS FOR REAL-TIME TRAFFIC IN MOBILE AD HOC NETWORKS

is designed for scheduling the cells in ATM networks. We borrow the idea and modify it as follows:

- Traffic is classified into a certain number of classes with different priorities. The maximum number of packets that may be sent from each class queue i is defined as n_i .
- At the i th queue, packets are sent as long as the queue is not empty and the transmission does not exceed the maximum predefined number of packets.
- If the i th queue becomes empty before reaching its maximum allowable number of packets, send one packet from the $(i+1)$ th queue and return to check the i th queue. If it has received a new packet, send it, otherwise, serve the $(i + 1)$ th queue.
- When the i th queue finishes sending the maximum number of its packets, no packet will be further sent from that queue unless there is no available packets in all the other queues.

There are three types of traffic in our framework, real-time traffic, best-effort traffic and the control packets sent by the routing protocol. The traffic packets are classified into two queues as shown in Figure 3.5, one for best-effort traffic and another for real-time traffic. Since the control traffic of the routing protocol plays an important role in our framework for prompt reaction to the dynamics of the network, such traffic should be given the highest priority over other traffics. We implement this priority by always inserting the routing packet at the head of the real-time traffic queue. The queue works in a FIFO manner and will serve available routing packets first. By assigning different priorities to these two queues, real-time traffic and best-effort traffic can receive differentiated service which will favor real-time traffic but will not starve best-effort traffic since the minimum packet transmission in each round can be guaranteed.

Figure 3.5: Priority queueing

3.5 Performance Evaluation

In this section, our proposed QoS framework is evaluated. What we care the most is how the proposed QDMR and the scheduling algorithm working together can enhance the QoS for real-time traffic even for high mobility scenario. The most relevant metrics are the average end-to-end delay and the delivery ratio. The comparison with other QoS frameworks such as SWAN is not provided in this section because different assumptions and implementation in different layers make a fair comparison difficult to perform. For example, SWAN works independent of the routing protocol, but relies on the MAC layer to provide rate control; while our proposal requires the routing protocol to provide reliable routing with QoS requirements met and requires no modification to the MAC layer. In this section, the efficiency of our proposal is validated by measuring the QoS received for real-time traffic and the whole network performance.

3.5.1 Simulation Environment

We implement our proposed QoS routing protocol in the network simulator NS-2 [NS2]. The propagation path-loss model is free space and IEEE 802.11 DCF is used as the MAC layer protocol. The channel capacity is set to be 11 Mbps and every node in the simulation has an effective transmission range of 250 meters. Nodes in the simulation move according to the *random waypoint* model [BMJ⁺98]. Being set a pause time t , a node will keep stationary for the duration of t before makes the next movement to a randomly picked destination, repeating such procedure to the end. With pause time of 0s, a node will keep moving to the next destination without stop after reaching the destination.

We use constant bit-rate (CBR) traffic sources to simulate real-time traffic with bandwidth of 200 Kbps and an end-to-end delay bound of 0.1s. The payload is fixed to 512 bytes. Best-effort traffic is modeled as FTP applications using TCP with data packets of 128 bytes. The source-destination pairs of the flows are distributed randomly among 50 nodes that scattered in the area. Each flow starts the transmission randomly between 30 s and 180 s and continues to the end of simulation. In all simulations, a $1000m \times 500m$ rectangle area is chosen which provides enough space diversity and the opportunity for alternate routes.

3.5.2 Routing Performance

In this part, a total of 10 real-time flows are simulated in each scenario to evaluate the performance of our proposed routing protocol. Various mobility degrees are generated by using different pause time of 0s, 50s, 100s, 150s and 300s. Each simulation will run for 300s, so all the nodes with 0s pause time will keep moving while those with 300s pause time will stay stationary throughout the whole simulation duration. For each pause time, 5 randomly generated scenarios are used and the final result is calculated by averaging

the 5 readings. The minimum and the maximum speed of the nodes are set at 0 m/s and 20 m/s respectively. To evaluate the routing performance, we compare our proposed QDMR with DSR which is also an on-demand source routing protocol and only support best-effort service. The DSR implementation we use is included in the NS-2 with default parameters set. The operations of reply from cache, snoop source route, reply multiple replies are all switched on for all the simulations.

Figure 3.6 shows the average end-to-end delivery ratio of real-time traffic. From the figure, our proposed QDMR always achieves a high delivery ratio for admitted traffic, even in 0s pause time when all the nodes are always moving with average speed of 10 m/s. QDMR achieves more than 95% of packets successful delivered at 0s pause time while that achieved in DSR is only about 50%. In Figure 3.7, the admission ratio of real-time traffic is shown which is the ratio between the amount of real-time traffic admitted into the network and the amount of real-time traffic requested at the source nodes. As the node mobility increases (i.e. with smaller pause time), the admission ratio of real-time traffic decreases if using QDMR, since not all real-time flows can find qualified paths. But for DSR that does not implement admission control, the admitted traffic cannot be provided the QoS as required when the topology change is fast and higher loss ratio is expected as shown in Figure 3.7. From the results, we can understand the importance of admission control in terms of resource efficiency.

In Figure 3.6, when the network is stationary, DSR still fails to achieve 100% delivery ratio (about 2700 packets are dropped). Through analysis of the trace file of the simulation, we found that most of the packets are dropped because the queue is overflowed. In DSR, the shortest path is chosen in the route discovery, and the other nodes can also use the same routing information through snooping the source header of packets. Traffic tends to converge in some common links which is likely to be overloaded and then the queues overflow. In QDMR, two node-disjoint paths are established. If a packet is delayed

CHAPTER 3. ENHANCING QOS FOR REAL-TIME TRAFFIC IN MOBILE AD HOC NETWORKS

Figure 3.6: End-to-end delivery ratio of real-time traffic

Figure 3.7: Admission ratio of real-time traffic

CHAPTER 3. ENHANCING QoS FOR REAL-TIME TRAFFIC IN MOBILE AD HOC NETWORKS

in the queue of one intermediate node longer than a set threshold, such QoS violation will be detected and the sender will switch traffic to the alternate path upon receiving RRER. Load-balance can thus be achieved by re-routing traffic away from overloaded route and the global network resource efficiency is also improved.

Figure 3.8 illustrates the average end-to-end delay of delivered real-time traffic. QDMR in average yielded a small end-to-end delay even at the highest mobility which meets the delay bound requirement by the real-time traffic. Meanwhile, a much larger delay is achieved when using DSR. This can be attributed to two reasons: First, DSR takes more time to rediscover new route when link failure is detected, and during that time the data traffic is kept in the buffer which increases the delay. The second reason can be revealed from Figure 3.9 which shows the average hop distance of the discovered routes. As we can see, paths discovered by DSR are on average longer than that discovered by QDMR when there is mobility in the network. This is because the route information obtained from the cache in DSR tends to be stale when network is dynamic, while QDMR always maintains up-to-date routes for real-time traffic. While both protocols are able to find routes with similar length in stationary scenarios. With lower end-to-end delays and higher delivery ratio of the admitted traffic being achieved in QDMR, it is clear that QDMR can enhance QoS support for real-time traffic under high mobility scenarios by minimizing QoS violations.

Table 3.1 shows the node-disjoint degree of the paths found by QDMR with the pause time of 0s. The node-disjoint degree is calculated as the number of common nodes in two paths excluding the source and the destination. In a total of 5 simulations, QDMR is able to discover completely node-disjoint paths for most of the time and it proves the efficiency of our algorithm.

The persistence of QoS provisioning in dynamic networks is then evaluated by measuring the number of rejections for real-time flows as shown in Figure 3.10. Out of the

CHAPTER 3. ENHANCING QOS FOR REAL-TIME TRAFFIC IN MOBILE AD HOC NETWORKS

Figure 3.8: Average end-to-end delay

Figure 3.9: Average hop count of routes

Table 3.1: Node-disjoint degree of the paths

disjoint simulation	0	1	2	3	4
No. 0	275	6	2	1	0
No. 1	336	8	2	1	0
No. 2	292	4	2	0	0
No. 3	369	11	1	1	0
No. 4	445	6	0	0	0

Figure 3.10: Number of real-time traffic rejection

10 real-time flows in the simulation, for each pause time, the maximum number of rejection is about 18 which means on average every flow only suffers 2 rejections even in the scenario with the highest mobility. In stationary scenario, real-time traffic may still be rejected if qualified routes cannot be found at the start depending on the topology and traffic distribution. As we can see, QDMR is able to provide persistent QoS provisioning to real-time traffic for most of the time.

Figure 3.11: Normalized control overhead

Figure 3.11 shows the normalized control overhead with various degree of node mobility. The normalized control overhead is the result of dividing the number of all the overhead packets by the number of delivered packet. This metric reflects the routing efficiency in terms of resource efficiency. With a smaller normalized control overhead, more network bandwidth is utilized by useful traffic data. The trade-off for supporting real time traffic with enhanced QoS in QDMR is that a higher control overhead is incurred which is needed to maintain the route up-to-date with QoS requirements met. However, as shown in Figure 3.11, the normalized overhead incurred in QDMR is much less than that incurred in DSR which validates the routing efficiency of QDMR by providing reliable routing support. And it is obvious that QDMR can achieve better resource efficiency than that achieved by DSR.

3.5.3 Performance under Heterogeneous Traffic

In this part, both best-effort traffic and real-time traffic are used to evaluate the efficiency of our proposed QoS framework. In these simulations, 5 real-time traffic flows are simulated together with n best-effort traffic flows ($0 \leq n \leq 5$). All the nodes move with

Figure 3.12: End-to-end delivery ratio of real-time traffic

a pause time of 0s and all other settings remain the same as that in previous simulations. With the pause time set to 0s, all the nodes will keep moving during the whole simulation period that will cause topology and traffic pattern keep changing. We then use this simulation setting to evaluate the efficiency of our distributed scheduling algorithm.

Figure 3.12 and 3.13 show the performance of real-time traffic with various number of best-effort traffic flows. With more best-effort flows added into the network, the end-to-end delivery ratio of the admitted real-time traffic remains almost the same with only a small variation. The delay of the admitted real-time traffic also does not increase too much since real-time traffic receives a higher priority in our scheduling algorithm.

To enhance the QoS for real-time traffic, the trade-off, however, is that best-effort traffic will be controlled and no QoS is guaranteed. In Figure 3.14, the average end-to-end delay of best-effort traffic is shown that large delay will be experienced by best-effort traffic. Figure 3.15 then shows the number of best-effort traffic packets that are admitted into the network. We observe that the amount of admitted best-effort traffic increases in a controlled way with the number of best-effort flows increases. This is because best-effort traffic only utilizes the unused bandwidth in the network and will not affect the

Figure 3.13: End-to-end delay of real-time traffic

admitted real-time traffic, which is guaranteed by our proposed scheduling algorithm. When the number of best-effort flows increases from 3 to 5, the average end-to-end delay of real time traffic is not influenced much as shown in Figure 3.13. We also observe that the number of best-effort traffic packets drops slightly when there are 4 best-effort flows. This happens because there is no guarantee for QoS of best-effort traffic and instead it is controlled by TCP, the adaptive mechanism of which will self adjust to the dynamics of the network. When QDMR works together with scheduling algorithm, admitted real-time traffic can be provided good QoS with both small delay and high delivery ratio even with the coexistence of heterogeneous traffic.

3.6 Summary

In this chapter, we first identify the main challenges in enhancing QoS for real-time traffic in mobile ad hoc networks, i.e. minimize QoS violation by providing reliable routing support. We propose a QoS framework that supports both delay sensitive real-time traffic and best-effort traffic. A QoS routing protocol is first designed with multipath

CHAPTER 3. ENHANCING QoS FOR REAL-TIME TRAFFIC IN MOBILE AD HOC NETWORKS

Figure 3.14: Average end-to-end delay of best-effort traffic

Figure 3.15: Number of best-effort traffic packet

CHAPTER 3. ENHANCING QoS FOR REAL-TIME TRAFFIC IN MOBILE AD HOC NETWORKS

support to minimize the impact of interruption of ongoing transmission, using which traffic can be immediately switched to an alternate path upon detection of QoS violation such as link failure or excessive contention. The reliability of the routes for the flows is further enhanced through effective node-disjoint path discovery. To enhance the QoS of real-time traffic after the admission, differentiated services are provided through local packet scheduling. Simulation results show that our framework is able to enhance QoS for real-time traffic in terms of both end-to-end delay and delivery ratio, and persistent QoS provisioning is achievable even at high mobility. The work in this chapter is published in [WYN04].

Chapter 4

QoS-aware Routing in Multihop Wireless Networks with Infrastructure Support

4.1 Introduction

The convenience and rather low cost of deployment have made wireless networks more attractive for Internet service providers (ISPs) and many small companies. The easy availability of the wireless devices also helps nurture the market for commercial use. More handhelds are released with IEEE 802.11 compatible network interface cards (NICs) that allow users to connect to the access point (AP) or to perform peer-to-peer communications in ad hoc mode. Wireless Local Area Network (WLAN) is currently the major choice of last-hop wireless technologies for many applications due to its easy administration (using AP as a central entity), hassle-free set-up (cable-free) and low cost. However, the bandwidth that can be supported in wireless networks is quite limited compared to the wired network. The advertised 54 Mbps bandwidth for IEEE 802.11a/g is the peak rate that can be achieved and it drops greatly as the communication range increases. In WLANs, only the mobile hosts (MHs) that are within the transmission range of the APs can establish connection to the backbone. In reality, the effective coverage of the AP is limited in order to guarantee transmission quality, and more APs are needed to cover

CHAPTER 4. QoS-AWARE ROUTING IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

the service area. Meanwhile, mobile ad hoc networks (MANETs) have received intensive research in recent years where peer-to-peer connection between two out-of-range MHs is feasible through multihop links. By introducing multihop connections into WLANs, the coverage can be greatly extended without requiring additional APs. However, the dynamic nature of wireless networks is likely to cause congestion in such networks. For example, when a lot of students in the lecture room turn on their laptops for wireless connection, the burst of traffic will overload the AP that covers the room. Instead of deploying multiple APs for the additional coverage, recent works proposed to shift load to neighboring APs with multihop connection in order to use the resources more efficiently. In this chapter, we address the issues to improve routing efficiency in a multihop wireless network with infrastructure support (MWNI) which can be implemented by extending current WLANs with multihop transmissions.

In MWNI, the network coverage is extended with multihop wireless links, which enable the MHs to establish association with the APs that are out of direct transmission range. Like in WLANs, most user traffic in MWNI is to/from the Internet or Intranet which has to go through the APs that are connected to the backbone. In IEEE 802.11 standards, multiple orthogonal frequency channels are defined (3 for IEEE 802.11b/g and 12 for IEEE 802.11a). To reduce interference and maximize the usage of spectrum, neighboring APs usually operate at orthogonal channels. When take that into consideration for MWNI, to avoid extra delay incurred by channel switching and to reduce the overhead for channel synchronization, all MHs in the same flow are assumed to use the same channel as that used by the AP they associate with and thus form a spanning tree rooted at the AP. Each branch in the tree represents a valid route from the AP to the leaf MH. An example of MWNI is shown in Figure 4.1. If node *C* first associated with *AP1* finds that its QoS cannot be guaranteed since *AP1* is overloaded, it can switch its association to *AP2* which still has enough resources to satisfy the QoS requirements. In MWNI, the flexibility of network organization as that in ad hoc networks can

Figure 4.1: Multihop wireless network with infrastructure support (MWNI)

also be achieved, while at the same time the existence of infrastructure enables further performance optimization.

In view of channel diversity, one challenge in MWNI is that the channel assignment must be incorporated into routing protocol to guarantee the network connectivity. Without careful design, several simultaneous channel switchings may leave the network in partitions. Another challenge in MWNI is that the effectiveness of routing algorithm will greatly affect the network resource efficiency. In WMNI, the network capacity is determined by the topology since multihop flows will contend for the shared medium in common channel. The objective in this chapter is to propose an efficient and reliable routing protocol for MWNI that can provide enhanced QoS support and maximize the network resource utilization by exploring the available channel diversity. In this chapter, interference in MWNI is modeled based on the probabilistic distribution of the MHs with regard to the AP, and a new load metric is proposed to estimate the load at the AP considering contention in both time domain and spatial domain. A QoS-aware routing protocol (PDQR) is proposed using the probabilistic distribution based load metric. In the multihop part of the network, a mobile host will find multiple routes to the neighboring APs. The MH will choose the AP with the highest available bandwidth to be

CHAPTER 4. QoS-AWARE ROUTING IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

associated with. To adapt to the network dynamics, a distributed channel switching algorithm is proposed that can avoid excessive oscillations. In our proposal, only local information is needed for both load estimation and congestion detection, and the switching procedure is totally distributed. The performance of the proposed algorithm is evaluated through simulations and we will show that it is able to adapt to the dynamics of the wireless networks and efficiently utilize the network resources.

The rest of the chapter is organized as follows. In Section 4.2, related work in the area of supporting multiple channels in multihop wireless networks is reviewed. In Section 4.3, a new load metric based on probabilistic distribution of the MHs is proposed to estimate the load at the AP. In Section 4.4, a QoS-aware routing protocol with channel assignment algorithm is proposed that can efficiently use the available resources and enhance the QoS for the end users. Simulation results are presented in Section 4.5 and we give a summary of this chapter in Section 4.6.

4.2 Related Work

In recent years, there are several studies on the routing problem in multihop wireless networks where the dynamic switching of channels can be supported by the traditional IEEE 802.11 based network interface. Assuming that the existing single-channel MAC is used, the main focuses of those works are channel assignment and routing algorithms.

In [RGC04, RC05], Raniwala et al. propose an architecture and several algorithms to support multiple channels with multiple NICs in wireless mesh networks (WMNs). A centralized channel assignment algorithm is proposed in [RGC04] that assumes each node has multiple NICs that can effectively utilize the network bandwidth without incurring network partition. Their work is further extended to a distributed channel assignment algorithm proposed in [RC05]. A load-balancing routing protocol is also included which can adapt to traffic load changes as well as network failures.

CHAPTER 4. QoS-AWARE ROUTING IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

In [SV04], a multi-channel routing protocol called MCRR is proposed for multihop ad hoc networks. Single NIC and existing single-channel MAC can be supported by assigning the same channel to all nodes in the same flow. For nodes that are supporting more than one flow, periodic channel-switching must be performed that requires strict synchronization among the neighboring nodes. In [SV05], routing and channel assignment problems are studied in a multihop wireless network with infrastructure support where each node is assumed to be equipped with a single NIC. Although similar assumptions are used in their work and in ours, the focus on QoS enhancement differentiates our work from theirs. First, in [SV05], load-balance is achieved that each MH will associate with the AP with the least load. However, in our work, an MH will associate with an AP with more residual bandwidth which does not necessarily balance the load immediately. Second, the MH will switch channel to associate with a less-loaded AP in their work. In our work, fast reaction to the network dynamics is possible if QoS violation, such as high delay and low throughput, is detected locally at the MH, and the MH will switch association with other AP for better QoS.

4.3 Probabilistic Distribution based Load Estimation

Since all traffic will go through the APs in WMNIs, the APs will have full knowledge of the traffic that is distributed through itself to the MHs. However, as is pointed out in [SV05], the load measured at an AP cannot accurately reflect the practical channel contention if multihop transmission is available. To accurately estimate the actual load considering contention in both time domain and spatial domain, the geographic topology and traffic patterns must be known which are usually difficult to obtain in ad hoc networks. In this section, we will explore the existing traffic and topology patterns in MWNIs and propose a load metric that can effectively reflect the spatial contention in the spanning tree.

4.3.1 Assumptions

The following assumptions are used for our discussion of MWNIs in both this chapter and the next chapter:

- In MWNIs, we assume that most of the traffic is downlink traffic, i.e. from the APs to the MHs. This assumption is valid for scenarios where the backbone traffic constitutes the majority of the traffic in such networks, e.g. MWNIs.
- The nodes in MWNIs do not move frequently. Unlike MANETs, where the nodes may move frequently, such assumption usually does not stand in MWNIs. Considering the specific traffic pattern in MWNIs and possible applications in reality, like WLANs in lecture room, cafe, etc., users normally don't move frequently while occasional leaving and joining are expected.
- The routes used by the multihop flows in MWNIs have no *shortcut* (we follow the name used in [Li05]). For the example in Figure 4.2 (the circle represents the effective transmission range of the MH), there is a shortcut for the multihop flow in Figure 4.2(a) that the flow can take the shortcut by directly connecting from B to D instead of taking an extra hop through C . Such unnecessary forwarding will waste network resources and increase the end-to-end delay. In Section 4.4, our routing algorithm guarantees only the route like that in Figure 4.2(b) will be used.
- DCF MAC, which is proposed in IEEE 802.11 standard, is assumed to be used as the underlying MAC in our proposal. In packet-switched wired networks, only the backlogged flows for the same link will contend for the medium access. However, in wireless networks, all the packets are broadcast in the shared medium and two transmission in range may cause collision and packet loss. With DCF MAC, a node will sense the medium and only make transmission if the medium is idle.

Figure 4.2: A multihop flow

4.3.2 Preliminaries

A multihop flow in MWNI can be deemed as several link level flows where packets are transmitted from one end of the link to the other end of the link. To differentiate the link flows from the multihop flow which is from the source to the destination (end to end), we will use *flow* to exclusively denote the multihop flow and use *subflow* to denote the link flow in rest of our discussion. For example, a multihop flow F_2 in Figure 4.3(b) consists of 3 subflows, i.e. $f_{2,1}$ from AP to B , $f_{2,2}$ from B to C and $f_{2,3}$ from C to D .

Here we adopt the interference model that is used in [LLB00], where two subflows are defined as *contending subflows* (or we say one subflow contends with another subflow) if either end of one subflow falls within the transmission range of the nodes in either end of another subflow. If we say multiple subflows are contending flows, we actually mean that each of these subflows is contending with all of the other subflows and they form a contention group. By this definition, for all the subflows in Figure 4.3(b), every distinct pair of them are contending subflows and together they form a contention group.

Based on the definition of *contending subflows*, we further define contending flows in MWNI. Two flows in MWNI are contending flows if each of the first two subflows counted from the AP side (only first subflow if the total hop count of the flow is one) of one flow is contending with each of the first two subflows of another flow. In Figure 4.4, F_1 and F_2 (as well as F_3 and F_4) are contending flows according to the definition. The definition

CHAPTER 4. QoS-AWARE ROUTING IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

Figure 4.3: Interference in multihop wireless networks

Figure 4.4: Contending flows in MWNI

is supported by the following observations. First, only downlink traffic is considered in MWNI and all traffic starts from the AP. Second, the contention aggregates towards the AP and the load distribution in MWNI is skewed. Third, for a multihop flow without shortcut, spatial reuse in the flow is possible when the hop count is more than 3 according to the interference model we use in this thesis.

Figure 4.5: Spatial reuse around AP

4.3.3 Model the Interference in MWNIs

By drawing concentric circles centered at the AP with radius in multiples of R , the n^{th} tier is defined as the region that is enclosed by the perimeters of $O(0, nR)$ and $O(0, (n-1)R)$ as is illustrated in Figure 4.5, where R is the transmission range of the NIC, and $O(t, kR)$ is a circle with radius kR and the distance between its center and the AP is t . Here $O(0, nR)$ is also referred as the n^{th} concentric circle. For a multihop flow without shortcut in its path, the nodes on the flow are likely to be evenly distributed in different tiers starting from the 1^{st} tier. This assumption will hold if the node distribution is dense enough. With the routing algorithm proposed in Section 4.4, the route that the MH connected to the AP has no shortcut. For a h -hop flow starting from the AP, the tier t_s that the s^{th} subflow ($1 \leq s \leq h$) will locate in can be calculated from Equation (4.1).

$$t_s = \begin{cases} 1, & s = 1 \\ s - 1, & s > 1 \end{cases} \quad (4.1)$$

If we randomly pick two nodes N_i and N_j from the i^{th} tier and j^{th} tier ($i, j \geq 1$) respectively, the probability that the distance between these two nodes is less than or equal to kR ($P(|N_i, N_j| \leq kR)$) can be deduced using geometric methods if we assume that nodes are uniformly distributed, where k is a positive number and kR represents the

CHAPTER 4. QoS-AWARE ROUTING IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

Table 4.1: The probability $P(|N_i, N_j| \leq 2R)$ for nodes in different tiers

	1st	2nd	3rd	4th	5th	6th	7th
1st	1	0.719	0.169	0	0	0	0
2nd	0.719	0.452	0.283	0.073	0	0	0
3rd	0.169	0.283	0.256	0.181	0.049	0	0
4th	0	0.073	0.181	0.180	0.134	0.037	0
5th	0	0	0.049	0.134	0.139	0.107	0.030
6th	0	0	0	0.037	0.107	0.114	0.089
7th	0	0	0	0	0.030	0.089	0.096

interference range. The approach is to find the area in the j^{th} tier from where N_j being picked will interfere with N_i . If we first randomly pick node N_i and assume the distance between it and the AP is t ($i - 1 < t < i$), by drawing a circle $O(t, kR)$ with radius kR and centered at N_i , the intersection area $SA(t, j, k)$ between $O(t, kR)$ and the j^{th} tier, shown as the shaded area in Figure 4.5, represents the possible location of N_j which can be calculated with Equation (4.2). N_j picked from $SA(t, j, k)$ interferes with N_i since the distance between them is less than kR . The shaded area in Figure 4.5 illustrates the interfering area for $i = 4$ and $j = 3$. This probability $P(|N_i, N_j| \leq kR)$ can then be calculated with Equation (4.3). In the equation, $SC_{0,i}$ is the area of circle $O(0, iR)$.

$$SA(t, j, k) = SC_{0,i} \cap SC_{t,k} - SC_{0,j} \cap SC_{t,k} \quad (4.2)$$

$$P(|N_i, N_j| \leq kR) = \int_{i-1}^i \left(\frac{2\pi t}{SC_{0,i}} \cdot \frac{SA(t, j, k)}{SC_{0,j}} \right) dt \quad (4.3)$$

This equation can be solved using integral methods. Table 4.1 gives the calculated probabilities for $1 \leq i \leq 7$ and $1 \leq j \leq 7$. And k is set to 2 for the calculation, as in the IEEE 802.11 standard, where the interference range is about two times that of the transmission range. To simplify the presentation, we will use $P(i, j)$ to refer to the value decided from the i^{th} row and j^{th} column in the table.

CHAPTER 4. QoS-AWARE ROUTING IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

Table 4.2: Notations

R	transmission range of wireless devices which is fixed for a frequency
t_s	the geographic tier for the s^{th} subflow of a multihop flow
k	interfering factor
$O(t, k)$	a circle with radius k and the distance between its center and the AP is t
$SA(t, j, k)$	the intersection area between $O(t, k)$ and the j^{th} tier
$SC_{t,i}$	the area of $O(t, iR)$

4.3.4 Load Estimation

Since we assume neighboring APs will operate at orthogonal channels, the AP will be the bottleneck in the spanning tree because it is the bridge between the MHs and the backbone. The load at the AP can be estimated with the knowledge of each flow that goes through it. Interference will occur if the source of a transmission is located within the interference range of either the source or the destination of another subflow. For transmission starting from the AP, the source is always located in the 1st tier. For flow F_i with h_i hops from the AP, the probability that each of its subflows will possibly interfere with the AP can be checked from the first row of Table 4.1. Since the AP is assumed to be fixed, $P(1, 2) = P(2, 1) = 1$ will actually be used in our load estimation. For a h_i hop flow F_i with traffic rate b_i , a weighted load WL_i contributed to the AP should be concerned that each of its subflows interferes with the AP, which can be calculated with Equation (4.4). Considering all the flows that go through the AP, the actual load WL at the AP can be estimated with Equation (4.5).

$$WL_i = b_i \times \sum_{j=1}^{h_i} P(1, t_j) \quad (4.4)$$

$$WL = \sum_i \left(b_i \times \sum_{j=1}^{h_i} P(1, t_j) \right) \quad (4.5)$$

Table 4.2 gives the notations used in this section.

4.4 PDQR: A Probabilistic Distribution based QoS-aware Routing Protocol for MWNIs

In this section, a QoS-aware routing protocol is specially proposed for MWNIs using the probabilistic distribution based load metric proposed in previous section and we call this protocol PDQR. After starting up, each MH will associate with an AP based on the load estimation and a spanning tree is formed with the root at the AP. The AP will update its load information and periodically broadcast a beacon in the tree. The routing efficiency is further enhanced at the MHs which will actively monitor the QoS for the traffic locally. Upon detection of QoS violation like large end-to-end delay and small throughput, the MHs will perform channel switching and shift the load to neighboring APs with a lesser load. In MWNIs, the AP is responsible for the collection and dissemination of network state information in the spanning tree, and the MHs will only maintain minimal state information and choose the route based on local decision. In our protocol, only the sender will switch its channel to send any packet, and only the MHs operating at the exact same channel will be able to correctly interpret the information. For each message, the channel used by the sender will also be piggybacked which will be used by the receiver to send reply.

4.4.1 Route Establishment

In MWNIs, each MH need to associate itself with an AP to receive or send traffic. When an MH starts up, it will search for beacon broadcast by the APs and the neighboring MHs, and meanwhile broadcast a SCAN message to all the channels. The SCAN message will include the channel and the id of the MH. The AP will reply to the MH by unicasting a beacon after receiving the SCAN. The MHs that have AP information can also forward beacon to it with the information of the AP they are connected to. Such route information will be kept valid most of the time as the AP will periodically broadcast the beacons.

CHAPTER 4. QoS-AWARE ROUTING IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

Figure 4.6: Route establishment in MWNIs

After waiting for some time, the MH will choose an AP based on the routing metric that will be defined later in Section 4.3.4. The MH will switch its channel to the one used by the AP, then it will send an ASSO message to the AP to establish the association. The channel and the id of the MH will be included in the ASSO message. After receiving the ACCEPT message from the AP, the MH will update its route and switch the channel to the one used by the new AP. Then the MH will send an ESTAB message to the AP and a valid connection is established. After the association is established, the MH can request data from the AP. An example of route establishment procedure is shown in Figure 4.6, where the arrow head indicates the propagation direction and message closer to the head is sent earlier.

After associated with the AP, the MH will add a new entry to its routing table that marks an active route to the AP. This entry will include the destination id, the next hop node id, the hop count of the path and the load information of the AP. Upstream MHs will keep a route entry to this mobile host and will use this routing entry to forward traffic to it. A spanning tree is constructed gradually in this way and each branch represents a valid route to all the MHs on the branch. The route entry will be refreshed by the beacon broadcast periodically from the AP.

4.4.2 Routing Metric

In [SV05], load-balance is achieved by the MH choosing a new AP if the load of the new AP after the switching will still be less than that of the current AP. However, whether the route to the new AP can be discovered or not will depend on what is the existing channel assignment since all the nodes on the same route to the AP must use the same channel as the AP. Different routing metrics will lead to different topologies which will lead to different performance.

In Figure 4.7, there are three flows destined to A, B and E with traffic rate of 400 Kbps, 100 Kbps and 100 Kbps respectively. If the traffic from A starts first, then B and then E, the routes chosen by them will be as illustrated in Figure 4.7 if the load-balance metric in [SV05] is applied. Flow B and E will have to go through a longer path even if the nearer AP has enough bandwidth to support the traffic. Considering that a MH's choice of channel is determined by the channels used by its neighbors in order to maintain connectivity to the AP, there will be more route flips as the network traffic changes. Frequent flips will degrade the QoS for the traffic, and frequent channel switching in the current NIC will also cause additional delay [DPZ04]. In our work, we propose to use the residual bandwidth of the AP as the routing metric.

The AP will actively monitor the traffic for the MHs that associate with it and its load can be estimated using the algorithm proposed in Section 4.3.4. This information will be propagated to the MHs by piggybacking it in the beacon sent from the AP. Since the beacon is broadcast in all channels, a MH may receive beacon from its local AP (LA) or a foreign AP (FA). Upon receiving a beacon from a FA (if the MH is close to another spanning tree), the MH will update its routing table and record the route to this FA. Since the beacon carries the load information of the FA, it will be easy for the MHs to calculate the residual bandwidth of the AP as the capacity of the AP¹ is known and is

¹which is actually the capacity of wireless channel that is shared by all the flows in both time domain and spatial domain

Figure 4.7: Dynamic traffic in MWNI

assumed to be fixed. Also since more subflows will contend for the shared medium as the route gets longer, the residual bandwidth observed at the MH should be decreased with the hop count of the route from its associated AP to this MH with regard to the end-to-end performance. If the weighted residual bandwidth of the FA after the MH's joining is less than that of its LA, the MH will switch the channel and connect to the FA. For a MH connected to the AP with h_i hops route, Equation (4.6) calculates the residual bandwidth RC for the AP with weighted load measured as WL .

$$RC = \frac{C - WL}{h_i} \quad (4.6)$$

In [LBDC⁺01], it's pointed out that the bandwidth of multihop path will drop exponentially with the hop count. However, considering the application of MWNI, i.e. extension of existing WLANs, the length of route in MWNI will not be too long, so our assumption of linear relationship between bandwidth and hop count is an approximate estimation.

4.4.3 Route and Channel Switch

Since we have assumed that neighboring APs in MWNIs will operate at orthogonal channels, an MH will have to switch its channel in order to connect to a FA. Several issues have to be considered here: First, once an MH switches its channel, all its downstream children will have to switch their channels too, otherwise the connections between the child MHs and the current LA will be interrupted. Another issue is how to reduce oscillations in the network. After a new beacon is broadcast with updated information, several MHs may react and initiate the switch procedure, and it may cause temporary oscillations in the network. A distributed algorithm is therefore designed to cope with these problems.

If an MH decides to switch to an FA, it will wait for a chosen delay before it initiates the switch procedure. This delay D is calculated by Equation (4.7), where *UNIFORM* is a function to uniformly choose a value from an interval that starts from 0 to a given value $k * t_{AD}$, t_{AD} being the broadcast period of the beacon and k being an integer.

$$D = \text{UNIFORM}(k * t_{AD}) + l * t_d \quad (4.7)$$

- The first part of the equation staggers the time that the MHs will perform channel switch in the different beacon broadcasting intervals. This will allow the load information at the MHs to be updated after load distribution in the network is changed.
- The second part of the delay is added to constrain long-hop switches, where l is the length of the branch in which the MH is located and t_d is a time duration that determines how the long-hop flow will be penalized.
- In the duration of D , if either the state of the LA or the FA changes, e.g., some flows have switched out and the LA then has enough bandwidth or the FA accepts

Figure 4.8: Route and channel switch in MWNIs

flows from other trees and therefore has not enough bandwidth left, the switch procedure will be cancelled.

When an MH is ready to switch channels, it will send a reassociation message REASSO to FA and the following procedure is similar to association phase. The REASSO message differs from the ASSO message that as the ASSO request will always be granted while the REASSO request can be turned down by the FA. The FA will ignore the request if its load has exceeded its capacity. After an ACCEPT is received from the FA, the MH also needs to send a dissociation message DISSO to the LA to inform it of the dissociation. The MH will then switch the channel and an ESTAB message is sent to the FA to confirm the association. The FA then becomes the LA of the MH. The MH will then forward the message down to all its children using broadcast ACCEPT. The child MHs will then send the ASSO message to their new LA after channel switch is completed.

Figure 4.8 illustrates the switching procedure that C reassociates with AP_2 and dissociates with AP_1 . After C finishes the switching, D will also switch its association to AP_2 .

4.4.4 Enhancing Routing Efficiency

After the route is established, the following mechanism are provided to further enhance the routing efficiency. First, to allow the congested MHs to react fast to QoS violation, these MHs are given the highest priority to switch channels. We adopt metrics designed in [FGML02] such as the inter-packet delay difference (IDD) and the short-term throughput (STT) which can be monitored by each MH locally. If both a high IDD and a low STT are observed at the same time, we have high confidence that a congestion has occurred. Route switches will then be performed by the MHs if an MH has a fresh route to a FA with enough residual bandwidth. This mechanism was proven to incur very little overhead in the MHs through testbed experiments [FGML02]. Second, in the local tree, a shorter path to the LA may be discovered at a late time. If that happens, the MH can update its route by sending a new ASSO message to the LA. This also guarantees that no shortcut will exist in the routes using our protocol.

In our protocol, the propagation of the beacon will update the MHs of both route information and load information. To reduce the overhead, a large interval is chosen (3 seconds in our simulation) as the broadcast period. Also after receiving a beacon from the LA or a neighboring node, the MH will forward it only if the sender of the beacon is the previous node on this MH's path to the LA. If the AP is not overloaded, it should propagate beacons in all available channels so that MHs associated with FAs can learn about this alternate path. However, if an AP finds that it is overloaded, such propagation to other channels is unnecessary and the beacon will only be sent down in the local tree. Through the controlled propagation of beacons, unnecessary switching overhead is alleviated.

In our proposal, only local information is maintained at the MHs for both load estimation and congestion detection, and the switching procedure is totally distributed.

Figure 4.9: Simulation scenario

4.5 Performance Evaluation

4.5.1 Simulation Setup

In this section, the performance of the proposed QoS routing protocol PDQR is evaluated through simulation studies. Multi-channel infrastructure support is implemented in the network simulator NS-2 [NS2] with IEEE 802.11 DCF as the underlying MAC protocol. The effective transmission range of each node is 250m, and the data rate is set to 2 Mbps.

A simulation area of 1200m by 1200m is equally divided into 4 regions with an AP placed at the center of each region, which is also pre-assigned a non-overlapping channel as shown in Figure 4.9. There are 120 MHs evenly distributed in the area, with 30 of them in each region. To create an uneven traffic pattern, 10 MHs are randomly picked from the upper-right region and the bottom-left region (shaded area in Figure 4.9) as traffic sinks. After the sink MHs associate themselves with an AP, they will request traffic from the backbone and the AP will start forwarding traffic to them. CBR traffic is used in the simulations with various data rates.

Each simulation will run for 500 seconds, and each data point in the figures in the next section is the average of 5 runs unless otherwise specified.

4.5.2 Simulation Results

To simulate the dynamic traffic and to evaluate the efficiency of our proposal, 20 CBR traffic flows will start one by one in intervals of 10 seconds (i.e. all flows will be in transmission after 200 seconds and statistics will be collected from this point). We also implemented the load-balance algorithm (LB) with its weighted load metric that is proposed in [SV05] as the baseline to compare with our proposed PDQR.

We first evaluate the channel utilization of the multi-channel schemes. In Table 4.3, the aggregate throughput achieved in each AP is listed. If using routing protocol without multi-channel support, all the traffic sinks will be associated with AP1 or AP2 with the shortest paths. With the support for multiple channels, load distribution is achieved for LB and PDQR that as the network load increases, more traffic sinks will be associated with AP0 or AP3 with longer paths. Using LB, load-balance is achieved where the load is weighted by the route length from the MH to the AP as observed from the table. However, since PDQR uses the residual bandwidth as the metric, the load does not distribute evenly especially when there is still enough bandwidth available in the local channel, i.e. when the network load is less than 2500 Kbps in Table 4.3. From the table, traffic will aggregate to the nearer APs, i.e. AP1 and AP2 in this scenario. As the load increases, our algorithm is as efficient as the LB algorithm in distributing load among all available channels. Figure 4.10 shows the aggregate throughput in the network. We see that both schemes are able to utilize the available bandwidth in all channels.

The QoS of the end users is another issue that should be evaluated. Flows using orthogonal channels will not interfere with each other, but on the other hand, the channel diversity also constrains the choice of the route. In Figure 4.11, the average end-to-end delay is measured at the sinks. When the network is lightly loaded, most packets can be transmitted with small delay in both schemes. As the total network traffic increases, our proposed PDQR starts to outperform LB. The delay can be reduced up to 60% by using

CHAPTER 4. QoS-AWARE ROUTING IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

Table 4.3: per-AP aggregate throughput (Kbps)

Network Load (Kbps)	LB				PDQR			
	AP0	AP1	AP2	AP3	AP0	AP1	AP2	AP3
500	79.72	179.81	179.03	72.33	32.49	220.98	241.20	25.19
1000	154.74	333.08	375.72	157.44	50.10	433.23	480.17	61.30
1500	220.48	530.78	556.79	224.16	74.04	676.91	693.63	92.43
2000	316.64	749.41	749.42	312.31	196.19	851.26	851.93	228.78
2500	396.29	904.42	930.24	384.22	377.91	930.54	944.96	375.26
3000	470.42	987.59	961.21	451.85	462.97	1001.70	1032.66	474.83
3500	471.85	1018.44	1021.58	496.99	513.41	1095.09	1026.56	516.42
4000	494.02	1079.21	1052.53	488.15	449.98	1012.17	1006.03	446.64

Figure 4.10: Aggregate throughput with various network load

PDQR. This improvement can be explained as due to two factors: first, the MHs using LB will take a longer path than those used by PDQR. In LB, MHs make switching decisions locally so as to try to connect to the least-loaded AP even though a nearer AP still has enough bandwidth. Second, when the traffic in the network changes, the MH associated with a farther AP may switch back to connect to a nearer AP. With a longer path, such switching will have more impacts in the network and incur oscillation. In PDQR, the available bandwidth is first explored by the MHs before switching to other APs. Only

Figure 4.11: Average end-to-end delay with various network load

when the local AP is over-loaded, then the MHs will seek switching to neighboring APs.

With our distributed switch algorithm, the MH with a shorter path to the FA will have a higher probability to start to switch first. Resources is also more efficiently used in PDQR. As discussed in Section 4.4.2, the scheme LB will use more hops to forward the traffic especially when the network is dynamic, like the example in Figure 4.7 that flows start transmission one by one. Figure 4.12 further supports this argument that the MHs using LB take a longer path on average than those using PDQR even after the network traffic is stable after 200 seconds.

From the simulations, we also find that LB converges slower than our scheme especially when the network is dynamic. In Figure 4.13, the average number of channel switches by the sinks is shown which increases by 1 whenever an MH is associated with a new AP. Using LB, sinks will have to perform more channel switches before reaching stability. We further check the instantaneous aggregate throughput tracked from one of the simulation run where the flows start one by one at 10s interval and the total network traffic is 3500 Kbps. The result is shown in Figure 4.14. It can be seen that our scheme

CHAPTER 4. QoS-AWARE ROUTING IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

Figure 4.12: Average route length with various network load

Figure 4.13: Average number of channel switch with various network load

CHAPTER 4. QoS-AWARE ROUTING IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

Figure 4.14: Aggregate throughput over time

can converge more quickly than LB after the network traffic stabilizes, and thus achieves a higher throughput.

4.6 Summary

In this chapter, we studied the routing issues for multihop wireless networks with infrastructure support. Such networks can be extended from the existing IEEE 802.11 based WLANs with multihop transmission. Existing hardware can still be used in such networks as we assume a single IEEE 802.11 compatible NIC with single channel MAC support. A lot of existing wireless applications can benefit from such networks as the network coverage is extended without the need to deploy more APs and load-balance can also be achieved through multihop relaying. We further explore the routing and channel assignment problem to maximize the resource efficiency. A probabilistic distribution based load metric is proposed to explore the characteristics of both topology and traffic pattern in such networks. With effective load estimation, users in the network will connect to the APs that are providing more available bandwidth. In the face of dynamic traffic and unreliable wireless channels, QoS for the MHs is enhanced by exploring

CHAPTER 4. QoS-AWARE ROUTING IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

the available channel diversity. A distributed channel switching algorithm is proposed that can adapt quickly to dynamic traffic. The effectiveness of our proposal is validated through simulation studies. The work in this chapter is published in [WY06b].

Through the studies in this chapter, we learn that with careful design, the self-organization of MWNIs can be utilized to provide adaptation to the network dynamics. With the infrastructure support, global optimization is possible without incurring too much overhead in the network and overloading the resource-limited MHs, and it is very promising to support integrated services with various QoS requirements that require more sophisticated control and scheduling algorithms. The lessons learnt from this chapter inspire our work in the next chapter.

Chapter 5

A QoS Framework to Support Integrated Services (QFIS) in Multihop Wireless Networks with Infrastructure Support

5.1 Introduction

In Chapter 4, routing issues are studied in a multihop wireless network which is connected with the infrastructure (MWNI). MWNI can extend the existing WLANs to provide larger coverage and more flexibility in accommodating the dynamics of the network while still keeping the inter-operability by using a single half-duplex network interface. Comparing with MANETs where there are no infrastructure support and the topology is random, MWNI have a more predictable traffic pattern since most traffic in the network is between the mobile hosts and the backbone which is connected with the Internet or Intranet. With the infrastructure support, contention in such networks can be estimated with a more accurate model and with little overhead incurred at the resource-limited MHs. The self-organization property of ad hoc networks can also be retained which provides flexibility in adaption to the dynamics of the networks. The feasibility of MWNI is already explored in previous chapter and a QoS-aware routing protocol is proposed. In this chapter, we extend the previous study results and further consider providing QoS

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

support for integrated services in MWNIs that has a more practical meaning for the future commercial deployment.

In integrated-service networks, two types of traffic are usually considered. For real-time traffic, there is a strict minimum bandwidth requirement and a delay/jitter bound to be met. On the other hand, for best-effort traffic, there is no such guarantee. Considering the resource efficiency, the following objectives are pursued in our work: (1) allow more real-time flows to be admitted into the network subject to meeting the QoS requirements (which is equivalent to the minimization of the call-blocking ratio), (2) optimize the performance of best-effort traffic without losing fairness and (3) enhance QoS performance for real-time traffic, i.e. minimize the delivery delay and improve the delivery ratio. The first two objectives may contradict each other since resource reservation for real-time traffic is necessary to satisfy (1), which in turn will constrain the throughput of best-effort traffic. Objective (2) will also contradict with (3) since higher priority should be given to real-time traffic which will also limit the performance of best-effort traffic. Besides seeking trade-off among the above objectives, we will also exploit channel diversity and spatial reuse in our proposed QoS routing algorithm to provide an optimized network topology and to allow best-effort traffic to efficiently utilize the available bandwidth while the QoS for real-time traffic is still guaranteed.

In this chapter, a QoS framework (QFIS) consisting of various components is proposed to support integrated services in MWNIs. First, a QoS routing algorithm based on a flow contention graph (FCGQR) is proposed to find a route with the highest additional end-to-end achievable share (AEAS) for the flows based on the contention experienced by its bottleneck link. Second, based on the bandwidth of the route discovered, admission control for real-time traffic is performed at the AP which will keep the contention in the network below the level that can be supported by the channel capacity. Third, considering the coexistence of real-time traffic and best-effort traffic, scheduling policies should be

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

Figure 5.1: QoS framework for MWNIs

defined to guarantee the QoS for real-time traffic and to optimize the fairness among the best-effort flows. In this work, a hierarchical scheduling algorithm based on weighted fair queueing is implemented at the AP which further optimizes the performance for both real-time traffic and best-effort traffic. With the APs serving as the central scheduling point in each spanning tree, they can be easily updated to support our algorithms. The MHs can also greatly improve the QoS with little overhead incurred. Figure 5.1 illustrates the whole framework.

The organization of the rest of this chapter is as follows: In Section 5.2, related work in the literature is introduced. Preliminaries that are used in this chapter are introduced in Section 5.3. In Section 5.4, our algorithm to estimate the bottleneck bandwidth is presented. Our proposed QoS framework and its components are then explained in details in Section 5.5, Section 5.6 and Section 5.7, followed by the performance evaluation in Section 5.8. Finally, Section 5.9 summarizes the chapter.

5.2 Related Work

5.2.1 QoS Framework

FQMM [XSLC00] is the first QoS model proposed for small to medium size MANETs. It can be deemed as a hybrid scheme of IntServ [BCS94] and DiffServ [BBC⁺98] which are the QoS models for the Internet. In FQMM, only traffic with the highest priority is given per-flow granularity as in IntServ while other traffic is given per-class provisioning as in DiffServ, and the traffic is policed at the source node. This design can greatly reduce the overhead due to state information maintenance and also reduce the burden of traffic conditioning at the resource-limited mobile hosts, while still providing service differentiation for heterogeneous traffic. In our framework, a similar design philosophy is adopted where the scheduling policies are only implemented at the AP which is the entry point of all traffic and has enough resources to support complex scheduling algorithms. Mobile hosts only need to maintain real-time traffic information which constitutes only a small part of the traffic in the network. Moreover, FQMM only presents a general QoS model for MANETs, the details for its various components like resource metering, QoS routing and scheduling are not discussed in depth.

In SWAN [ACVS02], service differentiation provisioning for MANETs is provided using stateless approaches. Several design issues in their proposal are worth noting: first, they assume that the bandwidth for best-effort traffic can serve as a “buffer zone” and through dynamic rate control of best-effort traffic, end-to-end performance of real-time traffic can be guaranteed. Second, additive increase multiplicative decrease (AIMD) rate control mechanism is adapted to shape best-effort traffic which can effectively keep the delay of real-time traffic low and still achieve high network throughput. However, SWAN assumes most of the network capacity is utilized by best-effort traffic and only a small proportion is reserved for real-time traffic. In our work, one objective is to maximize the

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

number of real-time traffic that are admitted into the network, while another objective is to optimize throughput for best-effort traffic by exploiting channel diversity with our QoS routing protocol. Except for the basic assumption and the network environment, there still exists several major differences between SWAN and our work. First, SWAN relies on the underlying routing protocol to find routes and reserves resources through explicit signaling. However, the contention experienced by the probe message may not be the same as what the admitted traffic experiences because of inter-flow contention and intra-flow contention. Thus false admission is likely to occur which will waste the network resources and weaken the network stability. In our work, by cooperating with QoS routing, admission control works more efficiently by avoiding the overburden in the bottleneck region of the network. Such cooperation with the routing module will also benefit the discovery of a better topology which allows more efficient utilization of the network resources. Second, the distributed policing in SWAN requires support from the lower layers. Each mobile node needs to keep monitoring the MAC states and dynamically change the rate for sending best-effort traffic. We think that such a price may still be too high for mobile hosts and the requirement for the modification of MAC will prevent it from being widely deployed. Instead, our framework is designed in a more practical way that can be easily deployed in existing infrastructure wireless network since most functions are implemented at the AP, and all functions are implemented above the MAC layer thus making our proposal compatible with existing wireless devices. For those mobile hosts that do not implement FCGQR which provides multihop extension, they can still associate themselves with the AP within one hop. However, no guaranteed QoS is provided and their traffic is just delivered with best effort.

5.2.2 Packet Scheduling

Packet scheduling has been widely studied in the literature, and numerous algorithms have been proposed to approximate the ideal Generalized Processor Sharing (GPS) dis-

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

cipline, such as weighted fair queueing (WFQ) [DKS89, PG93], start-time fair queueing (SFQ) [GVC97], WF^2Q [BZ96], WF^2Q+ [BZ97], etc., and they are all able to provide a minimum throughput bound for backlogged flows. However, in multihop wireless networks, maintaining such strict fairness will lead to underutilization of the channel due to the spatial contention for the shared medium. In the literature, several centralized and distributed scheduling algorithms [LLB00, LCL04, LL00, HB01, Li05] have been proposed for multihop ad hoc networks to achieve various degrees of trade-off between the maximization of spatial reuse and maintaining fairness among flows. Although the centralized algorithms like that have been proposed in [HB01, LLB00] provide theoretical solution, they are difficult to be implemented in practical networks. For the distributed algorithms like that have been proposed in [LLB00, LCL04, LL00, Li05], on the other hand, require modification of the MAC to realize per-flow granularity scheduling with the auxiliary information maintained and exchanged at each node. Such implementation complexity at the lower layers is a huge burden for resource-limited mobile nodes and contradicts with our design criteria. Also, an important assumption in the above-mentioned packet scheduling algorithms is that all traffic to be scheduled has a weight, which is hard to be characterized for best-effort traffic because of their nature of burst and unpredictability.

In [Che99], a two-level hierarchical scheduling scheme is specially devised for the co-existence of QoS flow¹ and best-effort flows. The physical link capacity is divided between two logical scheduling servers which further schedule all real-time flows and all best-effort traffic separately. By applying weighted fair queueing at each level of server, the minimum bandwidth requirement for real-time flows can be satisfied without being impacted by best-effort traffic, meanwhile a fair share among the same type of traffic is also achievable at the same time. Their scheme is designed for traditional link-shared networks which do not consider the contention in the spatial domain, thus it cannot be applied in MWNIs

¹flow can be characterized with a bandwidth requirement, which can be deemed as a generalized form of real-time flow

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

especially when we consider the following characteristics of such networks: First, in their scheme, the link capacity is by default assumed to be fixed. However, the effective capacity in multihop wireless networks with shared medium is determined by the topology as is pointed out in [LBDC⁺01]. Without accurately assigning the weight to best-effort traffic, the end-to-end bandwidth for real-time traffic may be violated. Second, their scheduling algorithm is work-conserving² which will allow bursty traffic to be admitted into the network if fair MAC is used. Such bursty traffic will aggravate the contention in multihop wireless network with shared medium and will damage the end-to-end QoS performance for the admitted real-time traffic, e.g. end-to-end delay and delivery ratio. Third, their scheduling algorithm is only concerned with bandwidth requirement and does not consider other important QoS metrics such as end-to-end delay which is very sensitive for most real-time applications. In Section 5.7, a hierarchical packet scheduling algorithm is proposed based on a virtual sharing model which can not only better characterize the unique properties of MWNIs, but also provides enhancement to optimize other end-to-end QoS metrics, like end-to-end delay and delivery ratio, for real-time traffic.

5.3 Preliminaries

In this section, we discuss preliminaries that will be used in our QoS framework. The preliminaries in Section 4.3.1 and assumptions in Section 4.3.2 are also applied to the discuss of MWNIs in this chapter.

5.3.1 Characterization of Traffic

Each real-time flow can be characterized by a minimum bandwidth requirement which will be included in its traffic profile. For those real-time traffic that are supported by VBR (variable bit rate), its minimum bandwidth can be specified as the average bandwidth

²A scheduling policy is said to be work-conserving if it never idles when there are requests waiting to be scheduled

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

for guaranteeing the required quality. We consider end-to-end delay performance as a desirable QoS metric for real-time traffic, but no hard bound is defined for such metrics in our framework because the contention in MWNIs is not just in time domain (for the shared link), but also in spatial domain (for the shared medium). Thus end-to-end performance such as delay is greatly affected by the underlying MAC performance which is outside of the discussion in this thesis. Instead, an enhancement is provided in our proposed scheduling algorithm to minimize the average end-to-end delay performance for real-time traffic to within a tight bound. For best-effort traffic, there is no special QoS requirement so its minimum bandwidth requirement is assumed to be zero in the traffic profile. However, the relative priority among the best-effort flows can be defined to accommodate different specifications and to provide flexibility.

In our algorithm, we characterize each flow with a *weight* which is proportional to its minimum bandwidth requirement. For a flow F_i with a minimum bandwidth requirement of b_i , its weight w_i can be defined as b_i/r where r is the bandwidth that represents an available QoS level. So with a higher weight, the flow should be allocated more bandwidth. Since the performance of the flow such as delivery ratio, end-to-end delay, etc., is measured end-to-end, we define the weight of a subflow to be the same weight of the multihop flow it belongs to.

5.3.2 Network Model

Since we assume that neighboring APs in MWNIs operate at orthogonal channels, there is no interference between the spanning trees rooted at different APs, we can thus limit our discussion to a single spanning tree where all the MHs associate with the same AP and they operate in the same channel. For a tree with N flows, a flow F_i ($1 \leq i \leq N$) with weight w_i consists of multiple subflows $f_{i,j}$ ($1 \leq j \leq h_i$) where h_i is the hop count of the flow and j increases with the hop count to the AP, so $f_{i,1}$ is also called as the first subflow of F_i , $f_{i,2}$ is the second subflow of F_i and so on.

Figure 5.2: Subflow contention graph

We then model the contention in the network as an arbitrary undirected and weighted graph $G = (V, E)$, where V is the vertex set of G and $E \subseteq V \times V$ is the edge set of G . We then define the subflow contention graph G_{fl} to model the contention between the subflows where all the subflows constitute the vertex set and there is an edge between two subflows in the graph if they are contending subflows. For the scenario in Figure 4.4 (in Section 4.3.2), the according subflow contention graph is shown in Figure 5.2. Similarly, the flow contention graph G_F can be defined based on the definition of contention flow which is defined in Section 5.3. For example, G_F for the same topology in Figure 4.4 is given in Figure 5.3.

A clique cl in a graph is a subset of V such that for any pair of distinct nodes $u, v \in cl$, the edge $(u, v) \in E$ always stands. According to the definition, for the subflow contention graph as is shown in Figure 5.2, $\{f_{2,1}, f_{2,2}, f_{1,2}\}$ is a clique while $\{f_{2,1}, f_{2,2}, f_{4,2}\}$ is not a clique since $f_{2,2}$ does not contend with $f_{4,2}$. The weight of a clique is defined as the sum of the weight of all the nodes in the clique. The clique with the maximum weight in the graph is called *maximum clique* $\hat{cl}$ ³. Also we use $CL(V')$ to refer to the set of cliques for the set of nodes $V' \subseteq V$ in the graph, or simply CL if $V' = V$.

³In a weighted graph, a clique with the maximum cardinality is not definitely a maximum clique since different weights are attached to the nodes.

Figure 5.3: Flow contention graph

For a multihop flow, each subflow in it may experience different levels of contention which is location-dependent, and the end-to-end performance is constrained by the bottleneck subflow which encounters the highest contention and thus performs the worst among all the subflows of that flow. We further define that a subflow *constraint clique* $\hat{scl}_i$ for a flow F_i is a clique in the subflow contention graph G_{fl} (we call this the *subflow clique scl* to differentiate it from the clique found in the flow contention graph G_F which is called *flow clique fcl*) that contains any subflow $f_{i,j} \in F_i$ ($1 \leq i \leq N$, $1 \leq j \leq h_i$) and has the maximum weight.

5.4 Estimate the Bottleneck Bandwidth in MWNIs

In MWNIs, flows contend not only for the shared link, but also for the shared medium which is topology-dependent. So unlike in wired networks, a bottleneck region should first be identified in the wireless networks and all the subflows in the region are the corresponding bottleneck links for the flows. The contention between the subflows is modeled with the subflow flow contention graph in which a subflow clique represents a set of subflows that mutually contend with each other. So the problem of estimating bottleneck bandwidth is abstracted to find the subflow clique with the largest weight (such clique is called *maximum clique* in graph theory) in the graph which includes all the subflows covered in the bottleneck region. However, the problem of finding the maximum clique is known to be NP-hard [PS98]. In this section, we first summarize all the procedures to construct the subflow constraint clique for a flow from the subflow

contention graph and all the solution space will be explored after these procedures finish. All the procedures are then categorized based on different criteria, based on which an optimized algorithm is proposed which greatly relieves the computation complexity by reducing the search space from subflows to flows.

5.4.1 Procedures to Construct the Subflow Constraint Clique for a Flow

We first categorize all the flows in the network into two groups with respect to a flow F_x . In the first group F'_x , any flow F_m in F'_x ($1 \leq m \leq N, m \neq x$) are not contending flows with F_x . The second group $\overline{F'_x}$ represents all the other contending flows, including F_x . In Figure 4.4, the dotted line separates the two group of flows based on their contention relationships with F_3 . Here, $F'_3 = \{F_1, F_2\}$, and $\overline{F'_3} = \{F_3, F_4\}$.

Since traffic aggregates towards the AP in MWNIs, the first subflow starting from the AP to the first hop of the flow will always experience more contention than the other subflows of the flow, which is thus the bottleneck for the flow since it constrains the end-to-end performance of the flow. For each flow F_x in G_{fl} ($1 \leq x \leq N$), we can just check all the subflow cliques that include $f_{x,1}$, and the one with the maximum weight is the subflow constraint clique for the flow. For the cliques that contain $f_{x,1}$, the other subflows in the same clique must interfere with either the AP or the first hop of F_x or both according to the definition of contending subflows. In a multihop flow, there are at most 3 subflows in the same clique that will interfere in the transmission since the fourth hop of the flow will be out of the interference range of the first hop according to our interference model and the assumption that no shortcut exists in the flow. The search space can thus be characterized by Proposition. 1.

Proposition 1 *To construct the subflow constraint clique $\hat{scl}_x$ for a flow F_x ($1 \leq x \leq N$), only the contending subflows of $f_{x,1}$ should be considered, i.e. $\{f_{m,i} : \forall F_m \in F'_x \text{ and } 1 \leq i \leq 2\} \cup \{f_{n,j} : \forall F_n \in \overline{F'_x} \text{ and } 1 \leq j \leq 3\}$.*

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

Based on Proposition 1, the following procedures are performed, each searching for the maximum subflow clique in part of the subflow contention graph. The search procedure starts with an candidate subflow clique scl'_x that contain $f_{x,1}$, and we try to maximize the weight of scl'_x by adding subflows from the subflow set $\overline{scl'_x}$ to it. So $\overline{scl'_x}$ represents the search space subject to scl'_x and is initialized according to Proposition 1. The newly added subflows must contend with all the subflows that are previously added to scl'_x , and $\overline{scl'_x}$ always contains the candidate subflows that could be further added to it. A search procedure finishes only if no subflow in $\overline{scl'_x}$ can be added to scl'_x . The next procedure will start if can still find a different scl'_x . After all procedures finish, the subflow constraint clique can be identified as the subflow clique with the largest weight of all those found after each procedure. For the following discussion, we summarize all the procedures based on three criteria.

Criterion 1 *A second subflow in F'_x , i.e. $f_{m,2}$ ($F_m \in F'_x$), is contained in the candidate clique scl'_x .*

Considering the contention with both subflows $f_{x,1}$ and $f_{m,2}$, $\overline{scl'_x}$ is further reduced to $\{f_{m,i} : \forall F_m \in F'_x \text{ and } 1 \leq i \leq 2\} \cup \{f_{n,1} : \forall F_n \in \overline{F'_x}\}$. It is easy to prove that any first subflow $f_{j,1}$ ($1 \leq j \leq N$) will contend with any subflow in $\overline{scl'_x}$ due to the fact that in MWNIs: (1) all the first subflows share the same source, i.e. the AP; (2) the source of any second subflow is the 1-hop neighbor of the AP. So the candidate clique scl'_x could be enlarged to $\{f_{i,1} : 1 \leq i \leq N\}$ according to the clique definition and $\overline{scl'_x}$ becomes $\{f_{n,2} : \forall F_n \in F'_x\}$. By applying Proposition 2, the subflows in the maximum clique for the subflow set $\overline{scl'_x}$ could be added to scl'_x to obtain the maximum clique. It is noted that the flows including the subflows in the maximum clique for $\overline{scl'_x}$ will be contending flows according to the definition of contending flows. Following the above analysis, the maximum subflow clique $\hat{scl}$ that contains $f_{x,1}$ can be identified during the construction

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

of flow cliques for F'_x and its weight $W_{cl}(\hat{scl})$ can be calculated from Equation (5.1) and Equation (5.2) after all the procedures with this criterion finish, where $FCL(F'_x)$ is the set of flow cliques for the flow set F'_x and h_i is the hop count of the flow F_i .

$$W_{cl}(\hat{scl}) = \sum_{i=1}^N w_i + \max \left(\sum_{F_j \in fcl} w_j l'_j \right), \quad (5.1)$$

$$s.t. \quad \forall fcl \in FCL(F'_x)$$

$$l'_i = \begin{cases} 0, & h_i = 1 \\ 1, & h_i > 1 \end{cases} \quad (5.2)$$

In Equation (5.1), the first term calculates the sum of the weight of subflows in $\{f_{i,1} : 1 \leq i \leq N\}$, and the second term counts for the maximum clique for the set of second subflows that belong to F'_x .

Proposition 2 *In $G = (V, E)$, the maximum clique containing V' is the clique in G that includes all the nodes in V' and has the maximum weight. If every node in $V' \subset V$ has an edge with all the other nodes in the graph, i.e. $\forall v \in V'$ and $\forall w \in \{E - v\}$, there exists $(v, w) \in E$, then the maximum clique containing V' can be constructed as $V' \cup \hat{cl}(\overline{V'})$, where $\hat{cl}(\overline{V'})$ is the maximum clique in $CL(\overline{V'})$.*

Criterion 2 *A first subflow in F'_x , i.e. $f_{m,1}$ ($F_m \in F'_x$), is considered to be contained in scl'_x , but all the second subflows in F'_x are not included as been covered using Criterion 1.*

Following this criterion, $\overline{scl'_x} = \{f_{m,1} : \forall F_m \in F'_x\} \cup \{f_{n,i} : \forall F_n \in \overline{F'_x} \text{ and } 1 \leq i \leq 2\}$. Following the similar analysis as in Criterion 1, scl'_x could be enlarged to $\{f_{n,1} : 1 \leq n \leq N\}$ and $\overline{scl'_x} = \{f_{n,2} : \forall F_n \in \overline{F'_x}\}$. After all the procedures finish, the maximum clique $\hat{scl}$ containing $f_{x,1}$ could be obtained by including the subflows in the maximum clique for the subflow set $\overline{scl'_x}$ and the weight $W_{c2}(\hat{scl})$ can be calculated from Equation (5.3), where

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

$FCL(\overline{F'_x})$ is the set of flow cliques for the flow set $\overline{F'_x}$. The difference from Equation (5.1) lies that only the second subflows belonging to $\overline{F'_x}$ are considered in Equation (5.3).

$$W_{c2}(\hat{scl}) = \sum_{i=1}^N w_i + \max \left(\sum_{F_j \in fcl} w_j l'_j \right), \quad (5.3)$$

$s.t. \quad \forall fcl \in FCL(\overline{F'_x})$

Considering that F'_x and $\overline{F'_x}$ are non-overlapping, Equation (5.1) and Equation (5.3) can be combined together as Equation (5.4) where $FCL(F)$ is the set of flow cliques for all the flows in F . The maximum clique identified from Equation (5.4) includes all the first subflows and we denote it as the AP-constraint clique $\hat{scl}_{AP}$ since all the subflows in it interfere with the AP.

$$W_{c1 \cup c2}(\hat{scl}) = \sum_{i=1}^N w_i + \max \left(\sum_{F_j \in fcl} w_j l'_j \right), \quad (5.4)$$

$s.t. \quad \forall fcl \in FCL(F)$

Criterion 3 *With this criterion, we discuss the rest of the possible ways to construct the constraint clique for F_x , so only the subflows in $\overline{F'_x}$ will be included in the clique, i.e. $f_{n,j}$ ($F_n \in \overline{F'_x}$ and $1 \leq j \leq 3$). The procedures using Criterion 1 and Criterion 2 have already considered the cliques that include the first two subflows in $\overline{F'_x}$ and all the subflows in F'_x , and what has not been discussed is whether a third subflow in $\overline{F'_x}$ (if there exist $F_m \in \overline{F'_x}$ with $h_m > 2$) is included in scl' , i.e. $f_{n,3}$ ($F_n \in \overline{F'_x}$), which will be considered in the following procedures.*

Considering that the traffic in MWNIs is from the AP to the MHs, for two flows that are not contending flows, their second subflows cannot be contending flows, so the third subflow of one flow is unlikely to interfere with the subflows of another flow. So only the subflows of the contending flows for F_x , i.e. $f_{m,i}$ (F_m and F_x are contending flows, $1 \leq i \leq 3$), will be considered and the upper bound of the weight for the maximum

Table 5.1: Constraint clique construction for topology in Figure 5.2

criterion	maximum clique contains $f_{3,1}$	clique weight
1	$\{f_{3,1}, f_{1,1}, f_{1,2}, f_{4,1}, f_{2,1}, f_{2,2}\}$	6
2	$\{f_{1,1}, f_{3,1}, f_{3,2}, f_{2,1}, f_{4,1}, f_{4,2}\}$	6
3	$\{f_{3,1}, f_{3,2}, f_{3,3}, f_{4,1}, f_{4,2}, f_{4,3}\}$	6

clique containing $f_{x,1}$ can be obtained from Equation (5.5) and Equation (5.6), where fcl_x is a flow clique in $FCL(\overline{F'_x})$ and $F_x \in fcl_x$, and l''_i can be thought as the intra-flow interference factor that maximally 3 subflows of the same flow can contend with each other.

$$W_{c3}(\hat{s}cl) \leq \max \left(\sum_{F_i \in fcl_x} w_i l''_i \right), \quad (5.5)$$

$$s.t. \quad \forall fcl_x \in FCL(\overline{F'_x})$$

$$l''_i = \begin{cases} h_i, & h_i < 3 \\ 3, & h_i \geq 3 \end{cases} \quad (5.6)$$

All the subflow clique constructions that will include $f_{x,1}$ are examined and a maximum clique is identified after procedures with each criterion finish. The maximum clique with the largest weight is the subflow constraint clique for F_x which can be calculated during the flow clique construction with Equation (5.4) and Equation (5.5). For example, if all the flows in Figure 4.4 have the same weight $w_i = 1$, we can find the maximum cliques for F_3 that include the subflow $f_{3,1}$, as is indicated in Table 5.1 and the weight of the constraint clique for F_3 is 6.

5.4.2 Optimized Algorithm to Identify the Subflow Constraint Clique for the Flows

After going through all the procedures to find the subflow constraint clique from the set of subflows, we can see that by just constructing the flow cliques instead of subflow cliques, the weight of the subflow constraint clique for a flow can be calculated which will

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

Table 5.2: Notations used in Section. 5.4

F	the set of multihop flows starting from the same AP $F_1, F_2, \dots, F_N$
F_i	the i th flow in F
$f_{i,j}$	the j th subflow of flow F_i , j increases with the hop count to the AP
h_i	the hop count of flow F_i
w_i	the weight of flow F_i , as well as for all the subflows $f_{i,j} (1 \leq j \leq h_i)$
r	the bandwidth represents an available QoS level
C	the capacity of the channel
G	an arbitrary undirected and weighted graph
V	the vertex set of G
E	$E \subseteq V \times V$ is the edge set of G
cl	$cl \subseteq V$ is a clique in G , $\forall u, v \in cl (u \neq v)$, the edge $(u, v) \in E$
$\hat{cl}$	the clique with the maximum weight in G
$CL(V')/CL$	the set of cliques for the node set $V' \subseteq V$ in G ; if $V' = V$
G_F/G_{fl}	the flow/subflow contention graph, an edge exists between two flows/subflows if they are contending flows/subflows
fcl_i/scl_i	a flow/subflow clique, and $[F_i \in fcl_i]/[f_{i,1} \in scl_i]$
$\hat{scl}_i$	constraint clique for a flow F_i and $f_{i,1} \in scl_i$
$FCL(F)$	the set of flow cliques for the flow set F
$\hat{scl}_{AP}$	the AP-constraint clique, the weight of which represents the minimum contention experienced by the AP
F'_x	$\forall F_i \in F'_x$, F_i and F_x are not contending flows
$W(cl)$	the weight of a clique
W_B	the weight of any subflow clique is bounded by $\lceil C/r \rceil$

be used by FCGQR. By doing so, the computation complexity is dramatically reduced since the search space is reduced from subflows to flows. According to [HB01], a node can construct its local clique if and only if it can obtain the contention tree information from all its neighbors' neighbors. In our proposed algorithm, only the first 2-hop contention information is needed to decide if two flows are contending flows which is enough to construct the flow cliques. Together with the routing algorithm which will be introduced in the next section, the AP will maintain the information of contention flows and calculate

the weight of the subflow constraint clique for each flow that starts from it. The notations used in this section are given in Table 5.2.

5.5 FCGQR: Flow Contention Graph based QoS Routing to Support Integrated Services in MWNIs

Using the optimized algorithm proposed in previous section to estimate the bottleneck bandwidth based on flow contention graph, a QoS routing protocol (FCGQR) is proposed to support integrated services in MWNIs. The routing problem is addressed together with channel assignment, and the channel diversity is explored to maximize the effective network utilization for both real-time traffic and best-effort traffic. FCGQR relies on the load estimation to choose the route which will be updated periodically with the message exchanging. Based on the local information, the MH will perform association or reassociation with the AP. FCGQR shares most of the routing procedures of PDQR which is presented in Section 4.4, and the redundant statements will be omitted in this section.

5.5.1 Local State Update

In FCGQR, the flow information such as the minimum bandwidth requirement will be included in the routing messages. For the MHs, besides a routing table similar to that used in PDQR (Section 4.4) will be used to store the routes to all the APs it can connect to, the MHs will also maintain a flow table for those flows going through it or destined to it, where the hop count, weight and the first hop information of the flow as well as the weight of the constraint clique for this flow will be updated by the control messages received from its neighbors. By overhearing the beacons or control messages from its neighbors, the 1-hop neighbor MHs of the AP can identify the flows that are contending

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

with its local flows. For the 1-hop neighbor MHs of the AP, the contending flows information will be piggybacked in the beacon and other control messages and propagated which helps the AP to build the flow contention graph.

Similarly, the APs will also maintain a routing table and a flow table. By combining the information learned from its 1-hop neighbors with the information in the local flow table, the AP is able to generate the flow contention graph and construct the flow cliques. The weight of the constraint cliques for each flow can then be calculated as is described in Section 5.4. Since the flows going through the same 1-hop neighbor MH of the AP will all fall in the same constraint clique, the AP will piggyback the weight of all the constraint cliques indexed by its 1-hop neighbors in the beacon and send it to the MHs. Based on this, each MH can know the bottleneck bandwidth of the route which connects it with the LA.

5.5.2 Routing Metric

In MWNIs, the MH may connect to more than one AP with multihop transmission after switching the channel accordingly. Users may prefer to join a tree with less contention so as to enhance their QoS. We first propose a metric based on the contention experienced by the flow F_i which can be estimated by the weight of its subflow constraint clique $W(\hat{sc}l_i)$. An MH will choose the AP which provides the minimum $W(\hat{sc}l_i)$, and we denote this metric as LC , which favors the route with less contention. This metric is similar to that proposed in [SV05] and is able to achieve load-balance in the network.

Comparing with wired networks, the admitted traffic in multihop wireless networks will have more effect on the performance that can be achievable by other traffic considering the spatial contention. Using the LC metric, each flow is only concerned about its own QoS which might lead it to choose a longer path even if a shorter path can provide enough bandwidth. In Figure 5.4, for example, a real-time flow of 50 kbps (assume

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

Figure 5.4: Coexistence of real-time traffic and best-effort traffic in the network

$w = 1$) is established between A and $AP1$. Then B requests a best-effort flow (assume $w = 0$) from the infrastructure, and B will associate with $AP2$ if the routing metric LC is used. With 2 additional hops in the route, the achievable end-to-end throughput for best-effort traffic is affected by about a factor of 2 and it shows how the topology can affect the network efficiency.

To optimize the network efficiency, a new routing metric based on the additional end-to-end achievable share ($AEAS$) is proposed. Considering the intra-flow interferences, the maximum share of bandwidth that can still be allocated to flow F_i that goes through the MH after it is admitted into the network can be given by:

$$W_{AEAS} = \frac{W_B - W(\hat{scl})}{l_i''} \quad (5.7)$$

where l_i'' is the intra-flow interference factor that can be calculated from Equation (5.6) based on the hop count of F_i . Assuming the channel capacity is fixed in the network, the weight of any clique is bounded by C/r which is denoted as weight bound W_B in this chapter. Comparing with the metric LC , the metric $AEAS$ considers the available bandwidth that can still be allocated for further QoS improvement and at the same time can benefit other flows that will join the network later. This allows global optimization to be achieved.

5.5.3 Association

The MH will choose an AP to associate itself with based on the QoS it can receive. Although it is difficult to construct its constraint clique without knowing the topology before the association, we can actually calculate the upper bound of the weight of its constraint clique using Equation (5.8).

$$W(\hat{sc}l_x) \leq \max \left(W(\hat{sc}l_i) \right) + w_x l''_x, \quad (5.8)$$

$$s.t. \quad \forall F_i \in F$$

In the formula, F is the set of flows that are already admitted and F_x is the flow to be established for the MH after its association with the AP. For Equation (5.8), the first term considers the weight of the maximum clique in the tree rooted at that AP, and the second term calculates the maximum contention that can be introduced after the MH joins the network and F_x starts transmission. The MH will then choose an AP using the above proposed routing metric and ties are broken by giving preference to the shorter path. If the paths are of the same length, the MH will just pick one arbitrarily. The MH will then follow the procedures as introduced in Section 4.4.1.

5.5.4 Reassociation

Due to the dynamics of the wireless networks, the MH may need to reassociate itself with an FA that can provide better QoS to its traffic. For the MH that has 1-hop neighbors in another spanning tree, it will receive the beacons from the FA, the MH is then able to estimate the QoS it will receive after switching association to that FA through the calculation using Equation (5.8). By comparing the weight of its constraint clique in the new association with the FA and the current weight of its constraint clique in the LA, the MH will switch association to the FA if the FA has more available bandwidth.

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

If the MH initiating a reassociation has flows going through it to its children, the downstream MHs must also reassociate to the FA otherwise their association with the current AP will be broken. Since each MH has a flow table which records the information of the flows that destine to itself and its downstream MHs, it is enough to calculate the upper bound of the weight of the constraint cliques for those flows if switches to the FA. Equation (5.9) gives the new weight where F' is the set of flows that are recorded in the parent MH's flow table. The remaining procedure is almost the same as the case of a single MH initiating reassociation except that after the parent MH receives an ACCEPT, it will propagate the ACCEPT down to all the downstream MHs so that each of them will perform the same reassociation procedure as after the parent MH receives the ACCEPT.

$$W(\hat{s}cl_x) \leq \max \left(W(\hat{s}cl_i) \right) + \sum_{F_j \in F'} w_j l_j'', \quad (5.9)$$

$$s.t. \quad \forall F_i \in F$$

Figure 5.5 shows the control messages exchange for the procedures that G establishes a new association with AP_2 and C initiates a reassociation with AP_2 which changes C and D 's association from AP_1 to AP_2 after the reassociation procedure finishes.

The reassociation procedure follows most part of Section 4.4.3 except that in FCGQR the FA has enough information to perform admission control, which will be introduced in the next section, before the network load information is updated and it will avoid switching oscillations that are possible to happen in PDQR.

5.6 Admission Control

For a multihop flow, each one-hop subflow will contend for the shared link as well as for the shared medium. To meet those end-to-end QoS metrics, such as average delivery ratio, average end-to-end delay, delay jitter etc., each subflow should be guaranteed the minimum bandwidth to achieve them. In Section 5.4, the contention between multiple

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

Figure 5.5: Routing procedure in MWNIs

multihop flows can be modeled using flow contention graph. A constraint clique for a flow can then be identified which models the contention group that includes the bottleneck subflow of the flow. Assuming that the weight of the constraint clique $\hat{scl}_i$ for flow F_i is $W(\hat{scl}_i)$ and that the bandwidth is fairly allocated for all the subflows in the clique according to their weights, the maximum share of bandwidth B_i that can be allocated to F_i is calculated using Equation (5.10), which is the fair share for its bottleneck subflow. To satisfy the minimum bandwidth requirement of F_i , $B_i \geq b_i$ should be satisfied which can be solved as Equation (5.11) using Equation (5.10), in which $w_i = b_i/r$, and $W_B = C/r$ using the definition of the weight for the flow. The AP will then perform admission control based on the estimation of the weight of the constraint clique for a flow requesting traffic with certain QoS. The traffic is admitted only if Equation (5.11) is satisfied, otherwise the MH will send request later to check the bandwidth availability.

$$B_i = \frac{w_i C}{W(\hat{scl}_i)} \quad (5.10)$$

$$W(\hat{scl}_i) \leq W_B \quad (5.11)$$

In this work, we assume that the minimum bandwidth requirement for real-time traffic will always be met after it is admitted into the network. However, some real-time flows can adjust data rate to improve the quality received at the end users such as

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

encoding video with a higher resolution, or if the users are willing to pay a higher price to acquire more bandwidth. For the former case, if there is still bandwidth available after the minimum bandwidth requirement for all admitted real-time flows is satisfied, the remaining bandwidth can be allocated to the flows using the max-min fair share [BG87] algorithm. The traffic profiles at the scheduler will also be updated with the new weights calculated. We assume most sources of real-time traffic are cooperative and willing to send traffic conforming to the negotiated traffic profiles and rate-control mechanism such as leaky bucket can be planted at the admission controller to regulate those uncooperative real-time traffic. For the latter case, the traffic profile is updated with the maximum bandwidth requirement, and the same admission control procedure is performed at the AP and traffic will be admitted if Equation (5.11) can be satisfied. For best effort flows, they are by default assigned the same weight unless otherwise specified as will be discussed in the following section. To avoid starvation of best-effort traffic, a certain amount of bandwidth can be reserved.

5.7 Hierarchical Scheduling

5.7.1 Overview

A hierarchical scheduling scheme is proposed to achieve the following objectives: (1) to guarantee the minimum bandwidth requirement for real-time traffic which is specified in the respective traffic profile after it is admitted into the network, (2) to enhance delay performances for admitted real-time traffic and (3) to optimize the fairness among best-effort traffic. As is shown in Figure 5.6, we propose a two-level scheduling algorithm. On the first level, the available link capacity is divided between two logical scheduling servers which are denoted as *virtual servers* VS_{RT} and VS_{BE} . When the outgoing link is idle, one virtual server will be selected which will send real-time packets or best-effort packets into the network. With all real-time traffic and best-effort traffic being separately

Figure 5.6: Hierarchical scheduling for real-time flows and best-effort flows

scheduled by individual virtual servers, the minimum fair share for real-time traffic will not be affected by the best-effort traffic. Virtual servers can use existing packet fair queueing algorithms, like WFQ [DKS89, PG93], WF^2Q [BZ96], etc., to schedule the flows, thus the fairness among flows of the same type can be optimized. To enhance the delay performance for real-time traffic, a novel scheduling algorithm is proposed to heuristically adjust the share of bandwidth for the virtual server.

5.7.2 Design Issues

In our framework, real-time traffic is admitted only if there is enough bandwidth to meet its minimum bandwidth requirement. However, for best-effort traffic, they are usually transmitted with flow-controlled protocols like TCP which provides reliable delivery through retransmission. The slow-start and congestion avoidance mechanisms of such transmission protocol will make the sending traffic bursty and unpredictable. One important issue in our work is to guarantee that the contention incurred by admitting best-effort traffic will not affect the QoS for admitted real-time traffic. In [Che99], the capacity of the VS_{RT} is set as the bandwidth that is required to satisfy all real-time flows with their minimum bandwidth requirement, and the rest of the link capacity is all

CHAPTER 5. A QOS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

allocated to best-effort flows. However, considering that they have different assumptions, their algorithm cannot be directly applied to MWNIs. In their work, the link capacity is assumed to be fixed which is usually true for a wired link. However, this does not stand for a wireless link. In wireless networks, after a head-of-line (HOL) packet is scheduled, it has to be further scheduled by the MAC before being actually sent to the physical channel. Otherwise, simultaneous transmissions on the same channel from other nodes will interfere with it and prevent it from being correctly demodulated by the intended receiver. Most MAC protocols for wireless ad hoc networks are designed with collision avoidance mechanisms, e.g. CSMA/CA in IEEE 802.11 MAC [IEE99], in which a node can send the packet to the channel only upon sensing the channel to be free. For links that are contending for the same channel, i.e. within the interference range of each other, the achievable capacity for each link is to a large extent constrained by the topology (assuming that the MAC has no bias) as illustrated in Figure 5.7. Thus in wireless network, scheduling should be applied to multiple links that are contending for the same shared medium so as to fairly allocate bandwidth to all flows. This can be deemed as scheduling over a virtual link. Related algorithms such as [LLB00, LCL04, LL00, HB01, Li05] discussed in Section 5.2, try to approximate the centralized fair scheduling in a distributed environment, with cooperated scheduling required to be executed at every link. This is too complex to be used in MWNIs where MHs have limited power. In our framework, only the AP is in charge of the scheduling, thus the burden at the MHs is greatly alleviated.

Meanwhile, reducing delay to a small value for delivering data is also necessary to present satisfactory service to the end users. Generally, as the network load increases, the end-to-end delay will increase accordingly due to the queueing delays and the backoff delays during channel contention. After the load exceeds a certain level near the channel capacity, contention at some area (bottleneck area) will cause the queues of the links at

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

Figure 5.7: Multiple links contend for the shared medium in wireless networks

that area to gradually build up and the delay for the packets will increase in a much steeper slope. As the load increases after that point, the network throughput starts to grow at a slower rate since more packets are dropped as queues overflow and collisions occur. The process can be illustrated like in Figure 5.8 where the load near the vertical separating line approaches the channel capacity. If leave such congestion uncontrolled, the poor end-to-end performance will render the delivered real-time traffic useless. In our work, it is desirable to control the network load at the bottleneck area and keep it below the level of causing congestion. This can be achieved by sacrificing some bandwidth which otherwise will be used by best-effort traffic. Such bandwidth sacrifice is acceptable since the QoS for real-time traffic can be greatly enhanced and at the same time, best-effort traffic can still use most of the available bandwidth. Also, with the network being slightly underutilized, the whole system will also be more stable and such stability is especially important to be pursued in multihop wireless networks. Since the stability of such network relies on the cooperation among all members to be maintained, any dynamics may incur a lot of overhead to get the network re-stabilized. However, considering the limited bandwidth of such networks, additional load like these may greatly harm the network performance if the contention in some bottleneck area is already near the threshold of

Figure 5.8: End-to-end delay as function of traffic load

causing instability. In our work, the share for best-effort traffic is heuristically controlled by the scheduler to enhance the QoS for admitted real-time traffic.

5.7.3 Scheduling for Virtual Servers

There are two virtual servers in our hierarchical scheduling scheme: one for scheduling packets for all real-time flows (VS_{RT}) and another for scheduling packets for all best-effort flows (VS_{BE}). These virtual servers will be scheduled according to their weights which are assigned in proportion to the corresponding allocated bandwidth. For VS_{RT} , its weight W_{RT} is just the sum of the weights of all real-time flows that are queued at the server:

$$W_{RT} = \sum w(f), \quad f \in F_{RT} \quad (5.12)$$

Considering all the real-time subflows that interfere with the AP (i.e. contend with the subflows sent from the AP), cliques can be constructed based on their traffic profiles as described in Section 5.4. The one with the largest weight is the AP-constraint clique $\hat{scl}_{AP}$ which constrains the maximum achievable capacity of the AP⁴. The AP will share

⁴to be more accurate, should be the outgoing link of the AP

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

the channel capacity with the MHs that initiate subflows which are included in $\hat{scl}_{AP}$, and the weight of the AP-constraint clique $W(\hat{scl}_{AP})$ represents the minimum bandwidth that should be allocated to admitted real-time traffic assuming that maximum spatial reuse is achieved. The rest of the channel capacity $\hat{W}$ can then be allocated to best-effort flows based on Equation (5.13). However, considering multihop transmission, the actual weight W_{BE} that should be assigned for best-effort traffic at the AP will be less than $\hat{W}$, i.e. $W_{BE} \leq \hat{W}$.

$$\hat{W} = W_B - W(\hat{scl}_{AP}) \quad (5.13)$$

In our work, enhancing delay performance for real-time traffic is also our concern. In a multihop wireless network, the end-to-end delay for a packet will consist of the queueing delays at all the forwarding nodes and as well the backoff times in each node to contend for the medium. In MWNI, the AP is the entry point for all traffic flows and thus all traffic will contend for the same link there. The contention experienced by the packet will become less as it is propagated further downlink, thus the delay at the AP will contribute to a large extent the end-to-end delay for the packet. In MWNI, if the amount of best-effort traffic being admitted exceeds a certain threshold, the intensified contention with real-time traffic for both the shared link and the shared medium will cause the real-time packets to be detained much longer in the queue. It is therefore desirable to find the ideal share for best-effort traffic so that the contention will not jeopardize the QoS of the real-time traffic and at the same time the available network bandwidth can be efficiently utilized by the best-effort traffic. In our framework, an algorithm based on additive increase and multiplicative decrease (AIMD) is used to control the share that is allocated to best-effort traffic based on the average queue size of VS_{RT} , as shown in Algorithm 5.2 where d is multiplicative decrease factor that is set between 0 and 100, a is a fixed value to control the additive increase, and $\hat{W}_{BE}$ is an upper bound to prevent W_{BE} from exceeding $\hat{W}$. In our AIMD algorithm, $\hat{W}_{BE} = \hat{W}$.

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

Algorithm 5.2 AIMD control of W_{BE}

```

/*called every T second period*/
if  $avg_{RT} > max_{th}$  then
     $W_{BE} \leftarrow W_{BE}(1 - d/100)$ 
else if  $avg_{RT} \leq min_{th}$  then
     $W_{BE} \leftarrow W_{BE} + a$ 
    if  $W_{BE} > \hat{W}_{BE}$  then
         $W_{BE} \leftarrow \hat{W}_{BE}$ 
    end if
end if

```

Since real-time packets are scheduled one-by-one by the VS_{RT} , its queue size Q_{RT} can be taken as the number of real-time packets that are waiting for the scheduling at that moment. A low-pass filter same as the one used in RED [FJ93] is used to calculate its average queue size avg_{RT} . As shown in Equation (5.14) where α is a parameter to control how much the past history impacts the calculation. With a larger avg_{RT} , real-time packets will on average wait for a large time before being transmitted. In our algorithm, if the avg_{RT} exceeds a maximum threshold max_{th} , which indicates a possible *QoS-impaired* event being detected, a multiplicative decreasing of W_{BE} is then triggered to alleviate the contention. When the avg_{RT} drops below a minimum threshold min_{th} , W_{BE} is additively increased to allow more best-effort traffic to be scheduled to utilize the available bandwidth.

$$avg_{RT} = (1 - \alpha)avg_{RT} + \alpha Q_{RT} \quad (5.14)$$

With W_{RT} and W_{BE} being assigned to VS_{RT} and VS_{BE} respectively, the weighted fair queueing is used to schedule the servers. Each virtual server VS_i maintains the following states: (S_i, F_i, Q_i) , where S_i and F_i represent the virtual start and finish time as defined in [BZ97], and Q_i is the queue size of the virtual server which is equal to the number of packets that are queued at the corresponding server. The scheduling is implemented as follows:

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

- (i) When a packet with the size P_i arrives at the head of the queue of VS_i , the virtual start and finish time are updated according to the following

$$S_i = \begin{cases} F_i, & Q_i \neq 0 \\ \max(F_i, V), & Q_i = 0 \end{cases} \quad (5.15)$$

$$F_i = S_i + \frac{P_i}{W_i} \quad (5.16)$$

where Q_i is the queue size of the virtual server just before the moment the packet arrives, V is the virtual time maintained by the AP and W_i is the weight for the virtual server.

- (ii) When the AP senses that the channel is idle, Algorithm 5.3 will be executed to select the virtual server to schedule.

Algorithm 5.3 Schedule the virtual server when the outgoing link is idle

```

if  $Q_{RT} > 0 \ \&\& \ (Q_{BE} > 0 \ \&\& \ S_{BE} \leq V)$  then
    select  $VS_i$  with the minimum finish time to schedule (ties are broken arbitrarily)
else if  $Q_{RT} > 0$  then
    select  $VS_{RT}$  to schedule
else if  $S_{BE} \leq V \ \&\& \ Q_{BE} > 0$  then
    select  $VS_{BE}$  to schedule
else
    select none
end if

```

It should be noted that such a scheduling policy is non-work-conserving for best-effort traffic when real-time traffic is admitted since the rate control of best-effort traffic is needed to enhance the performance for real-time traffic. When no real-time traffic is admitted, $W_{BE} = W_B$ so the best-effort traffic can fully utilize the bandwidth.

5.7.4 Scheduling for Packets

Since we assume that admitted real-time flows will transmit traffic conforming to their traffic profiles, their fair share can always be guaranteed once there is enough bandwidth

being allocated to VS_{RT} . In our implementation, when the virtual server is idle, the packets from real-time flows will be scheduled in the FIFO order based on their arrival time at VS_{RT} .

The conforming assumption for real-time traffic cannot be applied to best-effort traffic which is controlled by the transport protocol to probe the maximum available bandwidth. The virtual server VS_{BE} will schedule all best-effort flows to optimize the fairness using weighted fair queueing. Packets from the same flow are stored in the same queue in the FIFO order. A timestamp $t_i(f)$ will be maintained for each queue and will specify the expected transmission completion time of the HOL packet. This is updated whenever a packet reaches the head of the queue according to the following

$$t_i(f) = \max(t_{i-1}(f), V_{BE}) + \frac{p_i(f)}{w(f)} \quad (5.17)$$

where $t_{i-1}(f)$ is the timestamp of the $(i-1)$ th packet being transmitted from the queue, V_{BE} is the virtual time [BZ97] of the virtual server, $p_i(f)$ is the size of the i th packet from the flow and $w(f)$ is the weight assigned to the flow. For best-effort flows, their weight is assumed to be 0 in the load estimation. However, in the scheduling algorithm, it is by default 1 to facilitate implementation. Different values can be set here to allow prioritized scheduling among the best-effort flows.

5.7.5 Enhanced AIMD Control (EAIMD)

The weight of the VS_{BE} determines the share of bandwidth allocated for best-effort traffic which is topology-dependent and hard to track considering the characteristics of best-effort traffic. In the hierarchical scheduling, W_{BE} is controlled by the AIMD algorithm to approximate the ideal share in the current topology. However, there may be a gap between the share of admitted best-effort traffic and its ideal share due to the mismatch between the AIMD control and the flow control of best-effort traffic. Such a gap can be

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

positive if the share of admitted best-effort traffic is larger than its ideal share and it is caused by the weight being increased before the actually admitted traffic fully utilizes the allocated share. The gap can be *negative* if the share of admitted best-effort traffic is less than what it was allowed due to the slow additive increase of the weight after a multiplicative decrease. With a large *positive* gap, the burst of admitted best-effort traffic will harm the QoS for admitted real-time traffic. On the other hand, with a large *negative* gap, the available network bandwidth cannot be fully utilized. Such large gaps will cause the performance of the AIMD control to oscillate between overloading and underutilization. The amplitude of the gap is determined by additive increase factor a , multiplicative decrease factor d and $\hat{W}_{BE}$.

To enhance the stability of the AIMD control, we propose to control $\hat{W}_{BE}$ so as to prevent W_{BE} from deviating too much from the ideal value. This helps to reduce both the *positive* gap and the *negative* gap, and thus enables the network to converge over time. When the AIMD algorithm is executed by the AP, *QoS-impaired* event will be detected if the average queue size of V_{RT} exceeds max_{th} . The queue size Q_{RT} at that time is used as a reference. The enhanced AIMD (EAIMD) is depicted in Algorithm 5.4:

Algorithm 5.4 Enhanced AIMD

```

/*called upon QoS-impaired detected*/
if  $Q_{RT}(t) > \beta$  then
     $\hat{W}_{BE} = \hat{W}_{BE} - s/100 \times \hat{W}$ 
end if

```

where β is a parameter that determines the minimum queueing delay of real-time packets at the AP and s is a factor to shape the maximum weight of the best-effort traffic. When the $(\beta + 1)$ th packet arrives at $V_{S_{RT}}$ with its queue size of β at that moment, it has to wait until all the β packets queued are all sent out before it will be transmitted, $T_q^{\beta+1} \geq \sum_{i=1}^{\beta} t_{pr}^i$ where $T_q^{\beta+1}$ is the queueing delay for the $(\beta + 1)$ th packet and t_{pr}^i is

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

the propagation time for the i th packet. Since the multiplicative decrease of W_{BE} will be invoked when avg_{RT} exceeds the maximum threshold, the queue size of VS_{RT} should be controlled without deviating too much beyond max_{th} . By setting $\beta \gg max_{th}$, if $Q_{RT} > \beta$ and $avg_{RT} > max_{th}$ are detected together, it is of high probability that the $\hat{W}$ is larger than the ideal share which will allow an excessive burst of best-effort traffic being admitted and harm the QoS for admitted real-time traffic.

Unlike AIMD, in which additive increase factor a and multiplicative decrease factor d work in opposite direction to control W_{BE} , EAIMD is triggered only when $\hat{W}$ is larger than the ideal share for best-effort traffic and eventually keeps W_{BE} below the ideal level by shaping its upper bound $\hat{W}_{BE}$. The stability of AIMD is thus enhanced by the one-way change in EAIMD. There might be a *negative* gap after executing EAIMD because it is hard to find the ideal share for best-effort traffic. The parameter s is thus set conservatively to allow efficient utilization of the network resources.

This algorithm assumes that the QoS of real-time traffic is harmed by excessive admitted best-effort traffic. In wireless networks, a link failure due to node movement or the dynamics of the channel (multipath, fading, etc.) will also impact the QoS for real-time traffic. Such events are usually detected by the routing protocols (the dynamics of the channel is also indicated as temporary link failure at the routing protocol), and they can be easily filtered out without affecting our algorithm.

5.7.6 Parameter Analysis

In our heuristic algorithm, the achievable performance greatly depends on the choice of the parameters. There are a few parameters that are not discussed in detail in the above section, and a brief analysis is provided here to guide the choosing of these parameters in the implementation.

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

- α and max_{th} α controls how the history impacts the average queue size. An appropriate value should be set to filter out transient QoS violation. Similar analysis as used in RED [FJ93] is applied in our algorithm.

Assuming that the queue of VS_{RT} is initially empty, i.e. an average queue size of zero, and a burst of L packets arrives. After the L th packet arrives at the queue, the average queue size avg_L is

$$\begin{aligned} avg_L &= \sum_{i=1}^L i\alpha(1-\alpha)^{L-i} \\ &= L + 1 + \frac{(1-\alpha)^{L+1} - 1}{\alpha} \end{aligned}$$

Given max_{th} and that the burst of L real-time packets are allowed to arrive at the AP, α should be set to satisfy the equation $avg_L < max_{th}$. To guarantee the delay performance of real-time traffic, a small value for max_{th} is desirable to allow the detection of possible *QoS-impaired* event. Given $max_{th} = 0.5$ and $L = 3$, for example, it is necessary to choose $\alpha \leq 0.0884$.

- min_{th} After *QoS-impaired* is detected, the bandwidth share for best-effort traffic is reduced which will grant more chances for the detained real-time packets in the queue to be sent out. Assuming that the average queue size is max_{th} at that moment, and that the queue becomes empty and remains empty before the arrival of next packet if the congestion is alleviated, the average queue size avg_M after the M th arrival packet is

$$avg_M = max_{th}(1-\alpha)^M \tag{5.18}$$

which is deduced from Equation (5.14). The value of M provides an estimation for the network state. If consecutive M real-time packets arrive at VS_{RT} which has an empty queue, it is highly possible that the network still has available bandwidth to accommodate more traffic which supports the equation $min_{th} \leq avg_M$. In our work, M is equal to L which is the number of burst of packets as discussed above.

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

Table 5.3: Notations used in Section. 5.7

$w(f)$	the weight assigned to flow f in proportion to its minimum bandwidth requirement
F_{RT}	the set of real-time flows
F_{BE}	the set of best-effort flows
VS_{RT}	the virtual server to schedule all real-time flows
VS_{BE}	the virtual server to schedule all best-effort flows
W_{RT}	the weight assigned to VS_{RT} which is in proportion to its allocated bandwidth
W_{BE}	the weight assigned to VS_{BE} which is in proportion to its allocated bandwidth
$\hat{W}$	the maximum weight could be assigned to VS_{BE}
$\hat{W}_{BE}$	upper bound of W_{BE}
$W(\hat{cl}_{AP})$	the weight of the AP-constraint clique
a	additive increase rate
d	multiplicative decrease factor
α	queue weight
max_{th}	maximum threshold for queue
min_{th}	minimum threshold for queue
β	threshold to execute shaping control of W_{BE}
s	shaping factor in EAIMD

The notations used in this section are given in Table 5.3.

5.8 Performance Evaluation

5.8.1 Simulation Setup

In this section, the performance of the proposed QoS framework is evaluated through extensive simulation studies. Multi-channel support is implemented in the network simulator NS-2 [NS2] with IEEE 802.11 DCF as the underlying MAC protocol. The effective transmission range and the interference range are set to 250m, and the data rate of the channel is set to 2 Mbps.

A simulation area of 1200m by 1200m is equally divided into 4 regions with one AP placed at the center of each region, and each AP is pre-assigned a non-overlapping channel, as is shown in Figure 5.9. There are totally 120 MHs with 30 of them evenly

Figure 5.9: Placement of the APs

distributed in each region. After the MHs associate themselves with the APs, they will request traffic from the backbone according to their traffic profile and the AP will start forwarding traffic to them.

All real-time traffic used in this section is sent in constant bit rate (CBR) using UDP with a packet size of 512 bytes and the weight of the real-time flow is calculated based on its data rate which is assumed to be the minimum bandwidth required to support acceptable service. Newreno TCP [FH99] with default parameters set in NS-2 is used to send best-effort traffic with a packet size of 128 bytes.

5.8.2 Efficiency of Admission Control

In this section, we will first evaluate the efficiency of our proposed admission control scheme considering scenarios with only real-time traffic in the network. In this scenario, there are 20 real-time flows which is CBR traffic with the same data rate. Among all the MHs, 10 of them in region II and another 10 in region III are selected as the traffic sinks. Various QoS requirements are simulated by varying the data rate of real-time traffic and updating the traffic profiles accordingly. Each simulation will last 500s and the results

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

are averaged over 5 runs. Only the performance of admitted flows is measured unless otherwise mentioned.

The value of W_B is important to avoid excessive real-time traffic being admitted by the admission controller and it should reflect the effective channel capacity in the network. We first provide a simple analysis on the maximum throughput of IEEE 802.11 MAC based on [LDGN04]. The average time to transmit a packet in the MAC is:

$$T = DIFS + (CW_{min})/2 \times T_{slot} + 3 \times SIFS + T_{RTS} + T_{CTS} + T_{DATA} + T_{ACK} \quad (5.19)$$

where $DIFS$, $SIFS$ and T_{slot} are fixed values as defined in the standard. $(CW_{min})/2 \times T_{slot}$ represents the average backoff time before transmission to the physical channel starts. T_{RTS} , T_{CTS} , T_{DATA} , T_{ACK} represent the transmission time for the corresponding physical layer frame. For a packet with L bytes, the maximum throughput (in Mbps) is calculated as $(L \times 8)/T$. Assuming the average packet size is 512 bytes and data rate of the channel is 2 Mbps, the maximum throughput is about 56% of the raw channel rate. Based on this estimate, W_B can be calculated since the values for C and r are all fixed for a certain system.

We first check the QoS performance for the admitted real-time traffic. As shown in Figure 5.10, admitted real-time traffic can achieve a high delivery ratio with both routing metrics if admission control is performed at the APs. With an average of above 98% of admitted packets being successfully delivered with various QoS requirements, the effectiveness of admission control that the flows being admitted will be given their required bandwidth is validated. Meanwhile, if all flows are admitted into the network without admission control (for the metric LC only), the performance will deteriorate greatly after the load exceeds a certain point, which is at about 2500 Kbps in our simulation. When the load keeps increasing, a larger delay will occur for delivering data as shown in Figure 5.11. At the same time, the aggregate network throughput increases slowly after that point as shown in Figure 5.12.

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

Figure 5.10: Efficiency of admission control: average delivery ratio of real-time traffic

Figure 5.11: Efficiency of admission control: average end-to-end delay of real-time traffic

5.8.3 Efficiency of Routing Algorithm

We now check how each individual AP will perform in the same scenarios so as to validate the efficiency of our proposed QoS routing algorithm. The three tables 5.4, 5.5 and 5.6 presented here record the traffic that is successfully admitted into the network and delivered to the MHs by each AP using FCGQR with different schemes. As is shown

Figure 5.12: Efficiency of admission control: aggregate throughput of real-time traffic

from all three tables, FCGQR can adapt to the traffic dynamics by balancing the load among all available channels as the network load increases. In our simulation scenarios, all the traffic sinks are close to AP_1 or AP_2 , so if the sink MHs associate with AP_0 or AP_3 , they have to take a longer path. For the load incurred due to multihop transmission, it should also be estimated to reflect the actual load in the network as FCGQR does. Such estimation does not necessarily equal to the end-to-end throughput that is measured in the simulation, which then explains why the admitted traffic shown in the tables does not seem to be balanced among all the APs. The difference of two routing metrics being proposed can be observed from Table 5.5 and Table 5.6. As shown in Table 5.6, when the network load is light, less load will be shifted to the APs that are farther since the metric *AEAS* measures the end-to-end achievable bandwidth which is shaped by the factor of hop count.

5.8.4 Efficiency of Routing Metrics

In the previous simulation with only real-time traffic in the network, the metrics *LC* and *AEAS* achieve similar overall performance. To study in depth how the different routing

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

Table 5.4: per-AP performance: use the metric LC without admission control

Real-time Traffic (Kbps)	Admitted Traffic (Kbps)				Delivered Traffic (Kbps)			
	AP_0	AP_1	AP_2	AP_3	AP_0	AP_1	AP_2	AP_3
500	88.26	152.02	157.49	93.58	88.24	152.02	157.49	88.66
1000	176.72	304.26	315.33	187.25	176.72	304.25	315.32	177.35
1500	265.45	471.26	472.74	265.11	265.44	471.26	472.72	250.35
2000	368.89	633.99	655.58	388.85	368.86	633.96	655.55	368.43
2500	507.90	786.71	787.40	457.15	495.18	786.69	787.33	431.68
3000	606.80	939.48	939.96	545.22	559.61	939.40	923.36	507.81
3500	704.96	1092.71	1091.54	634.57	578.23	1032.95	992.23	551.89
4000	736.79	1269.97	1312.92	778.35	600.91	1071.33	993.98	573.16

Table 5.5: per-AP performance: use the metric LC with admission control

Real-time Traffic (Kbps)	Admitted Traffic (Kbps)				Delivered Traffic (Kbps)			
	AP_0	AP_1	AP_2	AP_3	AP_0	AP_1	AP_2	AP_3
500	93.49	157.40	152.01	88.16	93.49	157.40	152.01	88.16
1000	186.88	314.61	315.19	166.96	186.87	314.61	315.18	166.95
1500	280.68	472.51	471.97	250.55	280.64	472.49	471.96	250.54
2000	409.63	655.54	656.09	327.70	389.01	655.49	656.07	327.66
2500	381.64	838.69	863.78	456.70	381.60	838.65	863.76	456.00
3000	454.46	1010.52	910.69	505.59	454.42	1010.44	910.63	502.79
3500	493.05	985.78	1058.65	528.34	487.98	985.63	1058.55	528.25
4000	450.50	942.76	1024.06	532.48	439.66	942.70	1024.00	489.83

metrics may impact the performance to support integrated services, we further study the scenarios with the coexistence of best-effort traffic and real-time traffic. Besides the 20 real-time flows that are used in the above scenarios, 20 best-effort flows are added. In each region I-IV in the simulation area, as shown in Figure 5.9, 5 MHs are chosen as the best-effort traffic sinks which are all one hop away from the AP located at the centre of the region. Upon establishing the association with the APs, the traffic flows will start and will run to the end of the simulation. To fairly compare the performance of different

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

Table 5.6: per-AP performance: use the metric *AEAS* with admission control

Real-time Traffic (Kbps)	Admitted Traffic (Kbps)				Delivered Traffic (Kbps)			
	AP_0	AP_1	AP_2	AP_3	AP_0	AP_1	AP_2	AP_3
500	34.36	211.62	201.29	44.46	34.36	211.61	201.29	44.46
1000	69.29	422.23	402.76	88.56	69.28	422.21	402.74	88.56
1500	176.75	545.70	574.26	177.00	176.74	545.67	574.24	176.98
2000	429.40	717.89	553.85	348.63	429.35	717.86	553.83	328.05
2500	457.20	863.95	811.81	405.76	432.81	863.93	811.76	405.72
3000	479.60	985.08	910.08	530.25	479.53	985.00	910.03	524.4
3500	528.66	915.77	1056.49	528.72	485.49	915.67	1056.41	528.62
4000	450.63	941.46	1023.50	532.93	440.12	941.41	1023.44	490.53

routing metrics, admission control is performed at the APs. From the results presented in Section 5.8.2, it is observed the network throughput ceases to increase after reaching about 3000 Kbps. So only the simulation results with total real-time traffic requirement between 500 Kbps and 3000 Kbps are presented in this section.

In Figure 5.13, the aggregate throughput of best-effort traffic and real-time traffic is shown. First, it should be noted that the performance of the real-time traffic is not affected too much by the coexisting best effort traffic. This is because UDP sources are not subject to the same adaptation as TCP and also no end-to-end reliability is provided by UDP, while TCP is flow-controlled and will only utilize the available bandwidth and reliability is provided through retransmission of the lost packets. We can observe that the performance of real-time traffic using different routing metrics is almost the same. However, the performance of best-effort traffic achieved by using the routing metric *AEAS* is much better than that achieved by using the metric *LC*, which has up to 100% improvement in the throughput.

To better understand the performance difference for best-effort traffic using different routing metrics, in Table 5.7 and Table 5.8, the admitted best-effort traffic and real-time traffic from each AP is measured using both metrics *LC* and *AEAS* respectively which

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

Figure 5.13: Aggregate traffic throughput with different routing metrics

will further demonstrate how the routing performance affects the resource efficiency in the network. We can observe from the tables that real-time traffic distributes more evenly among all the channels if using the metric LC ; on the other hand, best-effort traffic admitted from each AP is not that balanced. In this simulation, the distribution of real-time traffic affects the distribution of admitted best-effort traffic. This can be explained from two aspects: First, by using the metric LC , more real-time traffic will be served through longer paths, e.g. from AP_0 and AP_3 as shown in Table 5.7. This is because using the metric LC will balance the admitted real-time traffic among all available channels which even requires the MHs to associate with a farther AP but less loaded. The additional multihop transmission uses more bandwidth, the admitted best-effort traffic is thus affected since best-effort traffic only uses the available bandwidth. Second, best-effort traffic will take longer paths to be served using the metric LC , as shown in Figure 5.14. By checking the trace files for the simulation using the metric LC , we found that the best-effort traffic sinks in region II and III tend to associate with AP_0 or AP_3 which will use longer paths for those MHs in region I and IV. Further analysis reveals the reason: all the real-time traffic sinks are in region II and III, it is of higher

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

Table 5.7: per-AP admitted traffic: use the metric LC with admission control

Real-time Traffic (Kbps)	Best-effort Traffic (Kbps)				Real-time Traffic (Kbps)			
	AP_0	AP_1	AP_2	AP_3	AP_0	AP_1	AP_2	AP_3
500	223.77	90.90	114.24	108.22	108.68	137.71	148.20	98.16
1000	135.00	48.62	36.82	90.90	216.00	314.75	294.10	157.59
1500	99.14	40.21	32.10	75.41	325.37	472.16	442.71	236.27
2000	81.94	41.67	24.08	50.22	429.62	635.08	614.45	369.50
2500	54.49	29.23	30.78	56.33	432.16	838.43	788.73	482.85
3000	70.55	1.97	23.27	47.41	454.57	970.97	877.67	485.82

probability that AP_1 and AP_2 are more loaded which “lures” the best-effort traffic sinks in region II and III to associate with AP_0 and AP_3 .

Table 5.8: per-AP admitted traffic: use the metric $AEAS$ with admission control

Real-time Traffic (Kbps)	Best-effort Traffic (Kbps)				Real-time Traffic (Kbps)			
	AP_0	AP_1	AP_2	AP_3	AP_0	AP_1	AP_2	AP_3
500	322.67	272.66	275.58	308.94	24.52	211.63	206.82	49.75
1000	292.82	214.59	225.75	287.23	78.41	413.73	393.33	98.45
1500	246.46	176.85	175.33	238.14	162.52	546.25	575.12	192.31
2000	138.33	167.70	115.82	148.69	348.06	593.70	779.18	327.93
2500	93.78	93.75	53.12	101.25	406.37	812.22	889.51	431.59
3000	34.61	2.10	34.79	53.48	575.96	1000.72	878.27	515.81

Using the metric $AEAS$ is able to find a global optimized topology that will associate the MHs with the AP which can provide more end-to-end achievable bandwidth. We can see from Figure 5.14 that the length of routes used by best-effort flows using $AEAS$ is close to 1, so the best-effort sinks prefer nearer APs which in the end provide more end-to-end throughput. At the same time, the flexibility of adaptation to the dynamic traffic is achieved that real-time traffic will be distributed to available channels when the local channel is overloaded. Through the simulation and analysis of the results, it

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

Figure 5.14: Average route length of best-effort flows with different routing metrics

is obvious that the topology has a large impact on the network resource efficiency due to the shared-medium nature of wireless networks. This further supports our argument that the routing efficiency should be counted for QoS provisioning.

5.8.5 Efficiency of Hierarchical Scheduling

In this section, the efficiency of the proposed hierarchical scheduling (HS) algorithm is evaluated. We first study how the parameters a and d used in the algorithm affect the performance which will provide us with some idea on choosing the parameters in our simulations. In our implementation, a is normalized as $\hat{W}$ multiplied by a fixed factor which is referred as a for the following discussion. The scenario with only one AP is considered first which is a 400m by 400m area with the AP located in the centre. This means that all the MHs are within direct transmission range. There are 4 real-time flows with data rate of 50 Kbps using CBR and 4 best-effort flows using TCP. Each simulation runs for 200s in this scenario.

We are interested in the throughput of best-effort traffic and the QoS of real-time traffic which is measured by the average end-to-end delay. In our discussions, *Original*

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

Figure 5.15: Average end-to-end delay of real-time traffic with various AIMD parameters

is used to denote the default queueing algorithm set in NS-2 which serves packets in the order of FIFO. From Figure 5.15, it can be concluded that as the multiplicative decrease factor d increases, the average delay of real-time traffic will decrease. However, d ceases to affect the delay when it exceeds 40. So a 40% multiplicative decrease of best-effort traffic share is enough to recover the QoS performance of real-time traffic. The figure also suggests that a conservative value should be set for the additive increase rate a , otherwise admitted best-effort traffic will soon increase over the threshold of harming the QoS of admitted real-time traffic. This is shown in Figure 5.15 where the delay of real-time traffic will increase in a large amplitude when a is greater than 4. It can also be observed that using HS is able to achieve much better delay performance comparing with *Original*.

For the throughput of best-effort traffic, the parameter a seems to be less sensitive, while the decrement rate d has a more obvious effect, as can be observed from Figure 5.16. It is not difficult to understand that with a smaller d , best-effort traffic will get more share, thus more traffic be admitted. Looking at Figure 5.15 and Figure 5.16 together, less delay of real-time traffic comes with less throughput of best-effort traffic. However,

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

Figure 5.16: Aggregated throughput of best-effort traffic with various AIMD parameters

if proper parameters are chosen, such loss of throughput of best-effort traffic can be controlled to be within an acceptable level while still providing satisfactory service for real-time traffic.

In Section 5.7.5, we discussed another important factor that may affect the performance which is the large gap between the actual share of bandwidth being assigned to best-effort traffic and the ideal share. EAIMD is then proposed to enhance the stability. We evaluate the algorithm using the same single AP scenario with a and d fixed to 2 and 40 respectively. This choice of a and d is based on the observations from Figure 5.15 and Figure 5.16. We then evaluate how EAIMD works by assigning the value 20 to the shaping factor s . In Figure 5.17, the instantaneous weight being assigned to the best-effort traffic is monitored. Compared with AIMD, EAIMD is able to converge to a stable state after some time. The only difference between the AIMD and the EAIMD is the shaping control of the upper bound of the weight that can be assigned to best-effort traffic. Such stability also benefits best-effort traffic since the throughput loss due to frequent multiplicative decrement is also reduced, as shown in Figure 5.18. Together with Figure 5.18, two points can be observed. First, the admitted best-effort traffic follows

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

Figure 5.17: Weight assigned to best-effort traffic over time

closely with the weight (observed in Figure 5.17) that is dynamically controlled by our proposed EAIMD algorithm. Second, the peak throughput achieved with AIMD is very close to that achieved from *Original* where no central control is imposed on best-effort traffic. This also explains the oscillation of the performance with AIMD.

To further investigate how the shaping factor s may affect the performance, multihop

Figure 5.18: Aggregated throughput of best-effort traffic over time

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

flows are used in the following scenario so that the effective bandwidth in the network is shaped by multihop transmission. A single AP locates at the center of an area of 1000m by 1000m with 36 MHs evenly distributed in it. Taking reference from the AP, 2 1-hop MHs and 2 2-hop MHs are selected as the sinks for best-effort flows, and 2 2-hop MHs are selected as the sinks for real-time traffic with data rate of 50 kbps. Various values of s between 0 and 50 are used for the following simulations where $s = 0$ represents no shaping control, i.e. AIMD.

From Figure 5.19, it is observed that the delay of real-time traffic will increase as s decreases since the burst of best-effort traffic will greatly harm the performance of real-time traffic. It is further observed from Figure 5.20 that such burst contributes little to improve the network utilization as the admitted best effort traffic is about the same when s is set to values between 5 and 30, which however, is still higher than that achieved without using shaping control ($s = 0$) due to the oscillation of AIMD. Concluding from the above analysis, it is desirable to control admitted best-effort traffic below the ideal share to enhance the performance of admitted real-time traffic, which also supports our motivation for the shaping control. When s exceeds a certain value (e.g. 30 in this simulation), no more improvement of the delay for real-time traffic can be observed which can be reasoned that the contention from best-effort traffic has little effect on the performance of real-time traffic after it is well controlled below the level corresponding to the ideal share. It can be concluded that s should be set conservatively to allow efficient utilization of the network resource.

In Figure 5.21, the fairness among all admitted best-effort flows is measured with the well-known Jain's fairness index [JCH84] which can be calculated by:

$$FI(x) = \frac{(\sum x_i)^2}{n(\sum x_i^2)} \quad (5.20)$$

where x_i is the amount of admitted traffic from the best-effort flow F_i and n is the number of admitted best-effort flows. An FI of 1 means all the best-effort flows get the exact

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

Figure 5.19: Average end-to-end delay of real-time traffic with various shaping factors

Figure 5.20: Aggregate throughput of best-effort traffic with various shaping factors

Figure 5.21: Fairness index of best-effort traffic with various shaping factors

same share of bandwidth. It can be seen from the figure that s has negligible effect on the fairness among all best-effort flows. Comparing with the *Original* scheme that does not implement fair queueing, our proposed hierarchical scheduling based on weighted fair queueing can optimize the fair share for all best-effort flows.

5.8.6 Efficiency of QoS Framework

Through the above simulation studies, the components of our proposed QoS framework are individually evaluated and the simulation results effectively support our design principles. In the following simulations, the efficiency of the QoS framework is evaluated to check how it can optimize the QoS of real-time traffic with various requirements and maximize the network efficiency. The same scenario used in Section 5.8.4 is used for this simulation and the traffic consists of 20 real-time flows and 20 best-effort flows. The same data rate is assumed for all real-time flows and the enhanced QoS requirement is simulated by increasing the data rate. In the following simulations, our proposed QoS framework which is FCGQR using the routing metric *AEAS* and with the hierarchical scheduling algorithm (HS) as is denoted “FCGQR+AEAS with HS” in the figures, which

is compared with “FCGQR+AEAS w/o HS” that does not implement the hierarchical scheduling algorithm.

The QoS of admitted real-time traffic is shown in Figure 5.22 and Figure 5.23. With all components working closely together, both the average end-to-end delay and the average delivery ratio are not impacted very much by the varying QoS requirement. However, without hierarchical scheduling (i.e. best-effort traffic is not controlled), the performance of admitted real-time traffic is unpredictable, and an average of 5% admitted real-time traffic is lost. There is also 10 times more delay incurred due to the contention with best-effort traffic. It can also be observed that after the network resource is almost all allocated for real-time traffic (after the backlogged real-time traffic exceeds 3000 Kbps), there is little throughput improvement for best-effort traffic using scheme without HS over scheme with HS, as is shown in Figure 5.24. However, the QoS of real-time traffic in both delay and delivery ratio is still worse than that achieved with our proposed QoS framework, as shown in Figure 5.22 and Figure 5.23. This shows the importance of keeping network stable by leaving some bandwidth unused as a “buffer zone”. In our QoS framework, real-time traffic is regulated with admission control while best-effort traffic is policed with scheduling at the AP, which optimizes the efficiency of the network with dynamic traffic.

5.9 Summary

In this chapter, two design issues are addressed in an integrated-service wireless network: first, how to accurately estimate the actual contention in the network so that a global optimized topology can be discovered which will further improve the resource efficiency in the network. Second, considering the coexistence of heterogeneous traffic with different QoS requirements, how to optimize the QoS for those traffic without compromising

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

Figure 5.22: Efficiency of QFIS: average end-to-end delay of real-time traffic

Figure 5.23: Efficiency of QFIS: average delivery ratio of real-time traffic

CHAPTER 5. A QoS FRAMEWORK TO SUPPORT INTEGRATED SERVICES (QFIS) IN MULTIHOP WIRELESS NETWORKS WITH INFRASTRUCTURE SUPPORT

Figure 5.24: Efficiency of QFIS: average network throughput

the network utilization. A QoS framework is then proposed to support integrated services in multihop wireless networks with infrastructure support which consists of three major components. First, a QoS routing algorithm is designed, by exploiting all available channels, which is adaptive to the traffic dynamics and is able to discover a global optimized topology. Second, based on the estimation of the contention in the current topology, admission control is performed at the APs based on the traffic profiles. Finally, a hierarchical scheduling algorithm is implemented at the APs to further enhance the QoS for admitted traffic. Through exhaustive simulation studies, our design principles are justified and the QoS framework is shown to be able to achieve our objectives. The work in this chapter is published in [WY06a] and [WY08].

Chapter 6

Conclusion and Future work

In this chapter, the main contributions in this thesis are summarized and the future work is presented.

6.1 Main Research Contributions

In this thesis, we study the problem of providing Quality of Service (QoS) in wireless ad hoc networks. The following challenges are addressed in our proposals considering the unique characteristics of such networks: first, considering the limited bandwidth in the network, resource efficiency should be achieved when designing routing algorithms. Accurate load estimation is required before designing the routing metrics, and it should estimate the contention in both time domain and spatial domain, and the latter is especially challenging in multihop wireless networks. Channel assignment should also be considered together with the routing algorithm for wireless mesh networks (WMNs) where multiple access points (APs) or routers may operate at different channels. Global optimization should be achieved in such network. Second, the dynamics of such network will incur frequent QoS violation. To alleviate the disruption thus caused, fast recovery is needed. Third, due to the lack of central management entity in wireless ad hoc networks, all algorithms should be designed in a distributed manner which makes the issue of reducing overhead more important. Our main contributions are as follows:

- A QoS framework to enhance QoS for Real-time Traffic in Mobile Ad Hoc Networks:** In MANETs, node movement is unpredictable and the topology keeps changing randomly, both making it difficult to track and maintain global state information. In this part, we address the QoS provisioning for real-time traffic in Mobile Ad Hoc Networks (MANETs) through a QoS routing protocol with multipath support (QDMR). The reliability is further enhanced by node-disjoint paths discovery. First, by efficiently discovering and maintaining an alternate path with enough resources, traffic disruption due to link failure in the current path can be alleviated that the traffic can be immediately switched to the alternate path. Second, by using passive acknowledgement, QoS violation due to link failure or excessive contention can be quickly detected without incurring too much overhead. To further enhance the QoS of admitted real-time traffic, a packet scheduling algorithm is implemented locally at each node to provide differentiated services. Through simulation studies, our QoS framework is able to satisfy QoS requirements of real-time traffic in MANETs that achieves both low end-to-end transmission delay and high delivery ratio, even under high mobility scenario.
- A QoS-aware routing protocol for multihop wireless networks with infrastructure support:** A feasible application of wireless ad hoc networks in reality is based on the existing WLANs with the extension of multihop transmission. The network coverage in such multihop wireless network with infrastructure support (MWNI) can be greatly extended with minimum deployment cost. In this part, a QoS-aware routing protocol (PDMR) is proposed to enhance the QoS and maximize the resource efficiency in MWNI. By exploiting the characteristics of both topology and traffic pattern in MWNI, a new load metric is proposed based on the probabilistic distribution of the MHs with regard to the APs. This metric is more accurate than existing proposals because it considers possible spatial reuse

CHAPTER 6. CONCLUSION AND FUTURE WORK

and explores the channel diversity in such networks. Based on the load estimation at the APs, the MHs will associate with the APs which can provide more residual bandwidth, and it allows better network resource utilization and thus enhances the QoS of the MHs. In PDMR, a distributed channel switch algorithm is also proposed to avoid switching oscillation with little overhead incurred. The performance of PDMR is evaluated through simulation study, the results show that our proposal can achieve stable performance with dynamic traffic and efficient resource utilization.

- **A QoS framework to support integrated services in MWNIs:** From the study in Chapter 4, MWNIs combine the flexibility of MANETs and the infrastructure support of WLANs that make it very promising to support integrated services. In this part, a QoS framework (QFIS) is proposed to support integrated services in MWNIs that can meet various QoS requirements of traffic. QFIS differs from the QoS framework in Chapter 3 in three respects: First, QFIS explores the existing traffic and topology patterns in MWNIs which make it possible for global performance optimization, while the dynamics in MANETs make it too difficult and too costly to perform such global optimization. Second, instead of providing reliable routes like QDMR does, QFIS provides routes that can achieve a global optimized topology which in turn enhances the QoS for all the traffic. Third, with the infrastructure support, QFIS is not only able to enhance the performance of real-time traffic, but also enhance the performance of best-effort traffic.

There are three main components in QFIS: First, a new QoS routing protocol (FCGQR) that discovers global optimized topology based on the effective estimate of the load in the bottleneck areas in the network, which is calculated based on the network interference which is modeled by flow contention graph. A novel routing

CHAPTER 6. CONCLUSION AND FUTURE WORK

metric (AEAS) is specially designed to improve the global resource efficiency that the MH will associate with the AP that can provide more additional end-to-end available resource.

Second, admission control is performed at the AP so that a flow will be admitted only if the network has enough bandwidth to support its minimum bandwidth requirement.

Third, a two-level hierarchical scheduling algorithm based on weighted fair queueing is implemented at the APs to further enhance the QoS of admitted real-time traffic and to optimize the fairness among all best-effort flows. On the first level, the available link bandwidth is divided between two virtual servers each of which will schedule packets of the same type. Since the available link bandwidth is determined by the topology considering contention in both time domain and spatial domain, only the AP will perform such scheduling which requires a lot of resources. The actual share to be allocated to best-effort traffic is controlled by an additive increase multiplicative decrease (AIMD) algorithm based on the performance of admitted real-time traffic, and its stability is further enhanced with a shaping control mechanism (EAIMD). On the second level, each virtual server performs weighted fair queueing algorithm so the fairness among all backlogged flows of the same type is optimized.

The efficiency of individual components as well as the whole QoS framework is evaluated through exhaustive simulation studies. The results show the following objectives can be achieved with our proposed QoS framework: (1) allow more real-time flows to be admitted into the network, (2) optimize the performance of best-effort traffic without losing fairness and (3) enhance QoS performance of admitted real-time traffic.

6.2 Future Work

The future work is then discussed as follows:

6.2.1 Predictive Routing

In QDMR, we choose routes with the criteria of disjoint degree and QoS. But we didn't look into the stability of the route. The worst case may exist where two paths are to break shortly after their establishment because some nodes on the paths are moving in opposite direction to the routes. In this case, frequent re-route discovery will be invoked and more resources will be wasted. When determine a route, its stability should be considered, e.g. how long the route can be kept unbroken. Some information about the nodes can be utilized to predict its stability, such as velocity, scheduling (when node will start movement or keep stationary). To reduce overhead, such information can be piggybacked in the *type 1* and/or the *type 2* RREQ, and the intermediate nodes receiving the RREQ can calculate the stability of the link with its upstream node and decide whether to relay it. The information can also benefit our traffic switching scheme in multipath routing. With this optimization, the reliability of the route will be enhanced and the overhead will be reduced which also boosts the performance. To obtain the extra information about the node, various mechanisms can be adopted, e.g. GPS based method or through local metering. When we study these issues, the trade-off between the overhead thus incurred and the improvement of the performance should be carefully considered.

6.2.2 Support Peer-to-peer Communications in MWNIs

In our proposed PDQR and QFIS in this thesis, we assume only downlink traffic from the AP to the MHs is supported. This stands for most scenarios if the users request traffic from the Internet or Intranet that connected with the backbone. In future work, we will consider support for peer-to-peer communications between the MHs which has

also become popular in the current Internet applications. The routing between two MHs will have more choices and we categorize different routing strategies into two types based on the location of the intermediate node. First, like in our PDQR and FCGQR, an MH must connect to its local AP first before connecting to other destinations even in the same tree. Our routing algorithm can be directly applied to this case. Second, the MH will choose the best route which is more like the routing in MANETs. For example, if two MHs are located in two non-neighboring spanning trees, going through the AP may be the shortest path. But if two MHs are in the same tree or they are not far away but are associated with different APs, pure ad hoc routing will be more efficient for this scenario. To find an optimal route, having global information at the MHs is needed which is usually not feasible since MWNIs are supposed to support a large number of users. One possible solution is that each AP maintains only a small amount of information of the MHs associated with other APs such as the AP id and the hop count from it to its home AP (HA), which is feasible through periodic information exchange between the APs. When an MH wants to communicate with an MH associated with other AP, it will first enquire the HA. Based on the information of the destination maintained at the HA, the source MH will then decide whether to go through the HA or to initiate a route discovery by broadcasting route request in all channels and expecting that the destination can hear it (just like on-demand routing in MANETs). For the hierarchical scheduling algorithm, it will have no effect if the connection does not go through the AP. We should then consider to apply the algorithms developed for MANETs in this case.

6.2.3 Support Multiple Transceivers in MWNIs

In Chapters 4 and 5, we assume only one transceiver is equipped with each MH which stands for most recent mobile devices. Research work [RGC04, KV05] has suggested the capacity improvement by deploying multiple transceivers at the routers or the MHs.

CHAPTER 6. CONCLUSION AND FUTURE WORK

And as they argue, with the production cost keeps dropping, future network interface card (NIC) may encapsulate multiple transceivers and simultaneous transmission can be performed at the MH or the AP. With various assumption of the number of transceivers at the MH and the AP, algorithms must be correspondingly designed to take the advantage of multiple transceivers. We can consider extending our work in Chapter 4 and 5 for such multiple transceiver support. First, we assume only the APs will support multiple transceivers while the MHs only support one. By viewing each transceiver as an individual AP, our proposed algorithms can be directly applied without any modification. Second, the MHs or the APs can deploy n transceivers where $n \geq 1$. In this case, beside the channel assignment, the transceiver assignment (i.e. use which transceiver to send or receive traffic data) must also be considered in designing the routing algorithm that should maximize spatial reuse without losing connectivity. For the packet scheduling, it should be implemented individually for each transceiver in the same device since different channels should be used to avoid local interference.

References

- [ACVS02] G. S. Ahn, A. T. Campbell, A. Veres, and L. H. Sun. SWAN: Service Differentiation in Stateless Wireless Ad Hoc Networks. In *INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, volume 2, pages 457–466 vol.2, 2002.
- [BBC⁺98] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss. An Architecture for Differentiated Services. *RFC 2475*, Dec. 1998.
- [BCS94] R. Braden, D. Clark, and S. Shenker. Integrated Services in the Internet Architecture - an Overview. *IETF RFC1663*, Jun. 1994.
- [BCSW98] S. Basagni, I. Chlamtac, V.R. Syrotiuk, and B.A. Woodward. A Distance Routing Effect Algorithm for Mobility (DREAM). *Proceedings of the MOBICOM*, pp. 76-84, 1998.
- [BG87] D. Bertsekas and R. Gallager. *Data Networks*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1987.
- [BMJ⁺98] J. Broch, D. Maltz, D. Johnson, Y.-C. Hu, and J. Jetcheva. A Performance Comparison of Multihop Wireless Ad Hoc Network Routing Protocols. *Proc. ACM/IEEE Int'l Conf. Mobile Computing and Networking (MobiCom)*, Oct. 1998.
- [BZ96] J. C. R. Bennett and H. Zhang. WF^2Q : Worst-case Fair Weighted Fair Queueing. In *Proc. INFOCOM '96*, volume 1, pages 120–128 vol.1, 1996.

REFERENCES

- [BZ97] J. C. R. Bennett and H. Zhang. Hierarchical Packet Fair Queueing Algorithms. *IEEE/ACM Transactions on Networking*, 5(5):675–689, 1997.
- [BZB⁺98] R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin. Resource ReSer-Vation Protocol (RSVP) - Version 1 Functional Specification. *RFC 2205*, Sep. 1998.
- [CG97] C. Chiang and M. Gerla. Routing and Multicast in Multihop Mobile Wire-less Networks. *Proceedings of IEEE ICUPC'97, San Diego, CA*, pp. 546-551, Oct. 1997.
- [Che99] S. Chen. *Routing Support for Providing Guaranteed End-to-End Quality-of-Service*. PhD thesis, Univ. of IL at Urbana-Champaign, 1999.
- [CM01] S. Chakrabarti and A. Mishra. QoS Issues in Ad Hoc Wireless Networks. *Communications Magazine, IEEE*, 39(2):142–148, 2001.
- [CN99] S. Chen and K. Nahrstedt. Distributed Quality-of-Service Routing in Ad-Hoc Networks. *IEEE Journal on Selected Areas in Communications*, 17(8):1488–1505, 1999.
- [CNRS98] E. Crawley, R. Nair, B. Rajagopalan, and H. Sandick. A Framework for QoS-Based Routing in the Internet. *RFC 2386*, <http://www.ietf.org/rfc/rfc.2384.txt>, Aug. 1998.
- [DKS89] A. Demers, S. Keshav, and S. Shenker. Analysis and Simulation of a Fair Queueing Algorithm. In *SIGCOMM '89: Symposium proceedings on Communications architectures & protocols*, pages 1–12, New York, NY, USA, 1989. ACM Press.
- [DPZ04] R. Draves, J. Padhye, and B. Zill. Routing in Multi-Radio, Multi-Hop Wireless Mesh Networks. *MobiCom '04: Proceedings of the 10th annual international conference on Mobile computing and networking*, pages 114–128, 2004.

REFERENCES

- [EGSM⁺95] H. El-Gebaly, A. Sabaa, J. Muzio, F. El-Guibaly, and D. Shapak. Performance Evaluation of a Window-Based Approach for ATM Cell Scheduling. *CCECE/CCGEI'95*, 1995.
- [FGML02] Z. Fu, B. Greenstein, X. Meng, and S. Lu. Design and Implementation of a TCP-Friendly Transport Protocol for Ad Hoc Wireless Networks. *ICNP '02: Proceedings of the 10th IEEE International Conference on Network Protocols*, pages 216–225, 2002.
- [FH99] S. Floyd and T. Henderson. The NewReno Modification to TCP's Fast Recovery Algorithm. *IETF RFC 2582*, Apr. 1999.
- [FJ93] S. Floyd and V. Jacobson. Random Early Detection Gateways for Congestion Avoidance. *IEEE/ACM Transactions on Networking*, Vol. 1, No. 4, Aug. 1993.
- [GVC97] P. Goyal, H. M. Vin, and H. Cheng. Start-time Fair Queueing: a Scheduling Algorithm for Integrated Services Packet Switching Networks. *IEEE/ACM Trans. Netw.*, 5(5):690–704, October 1997.
- [HB01] X. L. Huang and B. Bensaou. On Max-min Fairness and Scheduling in Wireless Ad-Hoc Networks: Analytical Framework and Implementation. In *MobiHoc '01: Proceedings of the 2nd ACM international symposium on Mobile ad hoc networking and computing*, pages 221–231, New York, NY, USA, 2001. ACM Press.
- [HP97] Z. J. Haas and M. R. Pearlman. The Zone Routing Protocol (ZRP) for Ad Hoc Networks. *IETF MANET, Internet Draft*, Dec. 1997.
- [IEE99] Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications. IEEE Standard 802.11, June 1999.
- [IEE05] IEEE Std. 802.11e-2005, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications. Amendment 8: Medium Access Control (MAC) Quality of Service Enhancements. IEEE Standard 802.11e, 2005.

REFERENCES

- [JCH84] R. Jain, D. Chiu, , and W. Hawe. A Quantitative Measure of Fairness and Discrimination for Resource Allocation in Shared Computer System. *Technical Report*, Dec. 1984.
- [JMHJ02] D. B. Johnson, D. A. Maltz, Y-C. Hu, and J. G. Jetcheva. The Dynamic Source Routing Protocol for Mobile Ad Hoc Networks (DSR). *Internet Draft*, draft-ietf-manet-dsr-07.txt, work in process, April 2002.
- [JT87] J. Jubin and J. Tornow. The DARPA Packet Radio Network Protocols. *Proceedings of the IEEE*, 75(1): 21-32, 1987.
- [KV98] Y. B. Ko and N. H. Vaidya. Location-aided Routing (LAR) in Mobile Ad Hoc Networks. *MOBICOM*, pp. 66-75, 1998.
- [KV05] P. Kyasanur and N. H. Vaidya. Capacity of Multi-Channel Wireless Networks: Impact of Number of Channels and Interfaces. *MobiCom '05: Proceedings of the 11th annual international conference on Mobile computing and networking*, pages 43–57, 2005.
- [LAZC00] S. B. Lee, G. S. Ahn, X. Zhang, and A. T. Campbell. INSIGNIA: An IP-Based Quality of Service Framework for Mobile Ad Hoc Networks. *Journal of Parallel and Distributed Computing, Special issue on Wireless and Mobile Computing and Communications*, Vol. 60, No. 4, pp. 374-406, April 2000.
- [LBDC⁺01] J. Li, C. Blake, D. S. J. De Couto, H. I. Lee, and R. Morris. Capacity of Ad Hoc Wireless Networks. In *MobiCom '01: Proceedings of the 7th annual international conference on Mobile computing and networking*, pages 61–69, New York, NY, USA, 2001. ACM Press.
- [LBS97] S. Lu, V. Bharghavan, and R. Srikant. Fair Scheduling in Wireless Packet Networks. *in Proceedings, ACM SIGCOMM'97*, San Francisco, 1997.
- [LCL04] H. Luo, J. Cheng, and S. Lu. Self-coordinating Localized Fair Queueing in Wireless Ad Hoc Networks. *IEEE Transactions on Mobile Computing*, 3(1):86–98, 2004.

REFERENCES

- [LDGN04] Z. Li, A. Das, A. K. Gupta, and S. Nandi. Full Auto Rate MAC Protocol for Wireless Ad Hoc Networks. In *IEE Proceedings Communication*, 2004.
- [LG01] S. J. Lee and M. Gerla. Split Multipath Routing with Maximally Disjoint Paths in Ad Hoc Networks. In *ICC 2001: IEEE International Conference on Communications*, volume 10, pages 3201–3205 vol.10, 2001.
- [Li05] B. Li. End-to-End Fair Bandwidth Allocation in Multi-Hop Wireless Ad Hoc Networks. In *Proceedings of IEEE International Conference on Distributed Computing Systems*, pages 471–480, 2005.
- [LL00] H. Luo and S. Lu. A Topology-independent Fair Queueing Model in Ad Hoc Wireless Networks. In *Proceedings of IEEE International Conference on Network Protocols*, pages 325–335, 2000.
- [LLB00] H. Luo, S. Lu, and V. Bharghavan. A New Model for Packet Scheduling in Multihop Wireless Networks. In *MobiCom '00: Proceedings of the 6th annual international conference on Mobile computing and networking*, pages 76–86, New York, NY, USA, 2000. ACM Press.
- [LLP⁺01] R. Leung, J. Liu, E. Poon, A.-L. C. Chan, and B. Li. MP-DSR: A QoS-aware Multi-path Dynamic Source Routing Protocol for Wireless Ad-Hoc Networks. *Proceedings of the 26th IEEE Annual Conference on Local Computer Networks (LCN 2001)*, Tampa, Florida, Nov. 2001.
- [MD01] M. K. Marina and S. R. Das. On-Demand Multipath Distance Vector Routing in Ad Hoc Networks. In *Ninth International Conference on Network Protocols*, pages 14–23, 2001.
- [MGLA96] S. Murthy and J. J. Garcia-Luna-Aceves. An Efficient Routing Protocol for Wireless Networks. *ACM/Baltzer Journal on Mobile Networks and Applications, MANET 1, 2*, pp. 183-197, October 1996.
- [NCD01] A. Nasipuri, R. Castaneda, and S. R. Das. Performance of Multipath Routing for On-Demand Protocols in Mobile Ad Hoc Networks. *Mob. Netw. Appl.*, 6(4):339–349, August 2001.

REFERENCES

- [NS2] The Network Simulator ns-2. <http://www.isi.edu/nsnam/ns/>.
- [PB94] C. E. Perkins and P. Bhagwat. Highly Dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for Mobile Computers. *Computer Communication Review* 24, 4, pp. 234-244, October 1994.
- [PC98] V. Park and S. Corson. Temporally-ordered Routing Algorithm. *Internet Draft*, Aug. 1998.
- [PG93] A. K. Parekh and R. G. Gallager. A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: the Single-node Case. *IEEE/ACM Trans. Netw.*, 1(3):344–357, June 1993.
- [PGHC99] G. Pei, M. Gerla, X. Hong, and C.-C. Chiang. A Wireless Hierarchical Routing Protocol with Group Mobility (WHIRL). *Technical Report 990006*, 25, 1999.
- [PR99] C. E. Perkins and E. M. Royer. Ad-hoc On-Demand Distance Vector Routing. *Proceedings of the Second IEEE Workshop on Mobile Computing Systems and Applications*, pages 90–100, February 1999.
- [PS98] C. H. Papadimitriou and K. Steiglitz. *Combinatorial Optimization: Algorithms and Complexity*. Dover Publications, July 1998.
- [RC05] A. Raniwala and T. C. Chiueh. Architecture and Algorithms for an IEEE 802.11-based Multi-Channel Wireless Mesh Network. In *INFOCOM 2005: Proceedings of the 24th Annual Joint Conference of the IEEE Computer and Communications Societies.*, volume 3, pages 2223–2234, 2005.
- [RGC04] A. Raniwala, K. Gopalan, and T. C. Chiueh. Centralized Channel Assignment and Routing Algorithms for Multi-Channel Wireless Mesh Networks. *SIGMOBILE Mob. Comput. Commun. Rev.*, 8(2):50–65, April 2004.
- [SEGEG⁺95] A. Sabaa, H. El-Gebaly, F. El-Guibaly, J. Muzio, and D. Shpak. Implementation of a Window-Based Scheduler in an ATM Switch. *CCECE/CCGEI'95*, 1995.

REFERENCES

- [SSB99] R. Sivakumar, P. Sinha, and V. Bharghavan. CEDAR: a Core-Extraction Distributed Ad Hoc Routing Algorithm. *IEEE Journal on Selected Areas in Communications*, 17(8):1454–1465, 1999.
- [SV95] M. Shreedhar and G. Varghese. Efficient Fair Queuing using Deficit Round Robin. in *Proceedings, ACM SIGCOMM'95*, Berkeley, 1995.
- [SV04] J. So and N. H. Vaidya. A Routing Protocol for Utilizing Multiple Channels in Multi-Hop Wireless Networks with a Single Transceiver. Oct. 2004.
- [SV05] J. So and N. H. Vaidya. Routing and Channel Assignment in Multi-Channel Multi-Hop Wireless Networks with Single Network Interface. in *The Second International Conference on Quality of Service in Heterogeneous Wired/Wireless Networks (QShine)*, Aug 2005.
- [WH01] K. Wu and J. Harms. QoS Support in Mobile Ad Hoc Networks. *Crossing Boundaries - an interdisciplinary journal*, Vol. 1, No. 1, 2001.
- [WY06a] H. Wang and K. C. Yow. A QoS Framework to Support Integrated Services in Multihop Wireless Networks with Infrastructure Support. *IPCCC*, April 2006.
- [WY06b] H. Wang and K. C. Yow. QoS Routing in Multi-channel Multihop Wireless Networks with Infrastructure Support. *INTERSENSE*, May 2006.
- [WY08] H. Wang and K. C. Yow. A QoS Framework to Support Integrated Services in Multihop Wireless Networks with Infrastructure Support. *Computer Communications*, 2008.
- [WYN04] H. Wang, K. C. Yow, and S. Nandi. Providing Quality of Service to Real-time Traffic through Multipath Discovery. *ICMU*, pages 152–157, January 2004.
- [XS01] S. Xu and T. Saadawi. Does the IEEE 802.11 MAC Protocol Work Well in Multihop Wireless Ad hoc Networks? *Communications Magazine*, June 2001.

REFERENCES

- [XSLC00] H. Xiao, W. K. G. Seah, A. Lo, and K. C. Chua. A Flexible Quality of Service Model for Mobile Ad-Hoc Networks. In *Proceedings of IEEE 51st Vehicular Technology Conference*, volume 1, pages 445–449 vol.1, 2000.