ABSTRACT

SERVICE ORIENTED NETWORKING FOR MULTIMEDIA APPLICATIONS IN BROADBAND WIRELESS NETWORKS

by
Ehsan Haghani Dogaheh

Extensive efforts have been focused on deploying broadband wireless networks. Providing mobile users with high speed network connectivity will let them run various multimedia applications on their wireless devices. In order to successfully deploy and operate broadband wireless networks, it is crucial to design efficient methods for supporting various services and applications in broadband wireless networks. Moreover, the existing "access-oriented" networking solutions are not able to fully address all the issues of supporting various applications with different quality of service requirements. Thus, "service-oriented" networking has been recently proposed and has gained much attention.

This dissertation discusses the challenges and possible solutions for supporting multimedia applications in broadband wireless networks. The service requirements of different multimedia applications such as video streaming and Voice over IP (VoIP) are studied and some novel service-oriented networking solutions for supporting these applications in broadband wireless networks are proposed. The performance of these solutions is examined in WiMAX networks which are the promising technology for broadband wireless access in the near future. WiMAX networks are based on the IEEE 802.16 standards which have defined different Quality of Service (QoS) classes to support a broad range of applications with varying service requirements to mobile and stationary users.

The growth of multimedia traffic that requires special quality of service from the network will impose new constraints on network designers who should wisely allocate the limited resources to users based on their required quality of service. An efficient resource management and network design depends upon gaining accurate information about the traffic profile of user applications. In this dissertation, the access level traffic profile of VoIP applications are studied first, and then a realistic distribution model for VoIP traffic is proposed. Based on this model, an algorithm to allocate resources for VoIP applications in WiMAX networks is investigated. Later, the challenges and possible solutions for transmitting MPEG video streams in wireless networks are discussed. The MPEG traffic model adopted by the WiMAX Forum is introduced and different application-oriented solutions for enhancing the performance of wireless networks with respect to MPEG video streaming applications are explained. An analytical framework to verify the performance of the proposed solutions is discoursed, and it is shown that the proposed solutions will improve the efficiency of VoIP applications and the quality of streaming applications over wireless networks. Finally, conclusions are drawn and future works are discussed.

SERVICE ORIENTED NETWORKING FOR MULTIMEDIA APPLICATIONS IN BROADBAND WIRELESS NETWORKS

by
Ehsan Haghani Dogaheh

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Electrical Engineering

Department of Electrical and Computer Engineering

January 2010

APPROVAL PAGE

SERVICE ORIENTED NETWORKING FOR MULTIMEDIA APPLICATIONS IN BROADBAND WIRELESS NETWORKS

Ehsan Haghani

Nirwan Ansari, Dissertation Advisor — Date
Professor, Department of Electrical and Computer Engineering, NJIT

Doru Calin, Committee Member — Date
Technical Manager, Bell Labs, Alcatel-Lucent

Roberto Rojas-Cessa, Committee Member — Date
Associate Professor, Department of Electrical and Computer Engineering, NJIT

Yun-Qing Shi, Committee Member — Date
Professor, Department of Electrical and Computer Engineering, NJIT

Osvaldo Simeone, Committee Member — Date
Assistant Professor, Department of Electrical and Computer Engineering, NJIT

BIOGRAPHICAL SKETCH

Author:	Ehsan Haghani Dogaheh
Degree:	Doctor of Philosophy
Date:	Janauary 2010

Undergraduate and Graduate Education:

- Doctor of Philosophy in Electrical Engineering,
  New Jersey Institute of Technology, Newark, NJ, 2010
- Master of Science in Communications Engineering,
  Chalmers University of Technology, Gothenburg, Sweden, 2005
- Bachelor of Science in Electrical Engineering,
  Sharif University of Technology, Tehran, Iran, 2003

Major: Electrical Engineering

Presentations and Publications:

E. Haghani, S. Parekh, D. Calin, E. Kim, and N. Ansari, "A Quality-driven Cross-layer Solution for MPEG Video Streaming over WiMAX Networks," *IEEE Transactions on Multimedia*, Volume 11, Issue 6, Pages 1140 - 1147 , Oct. 2009.

E. Haghani and N. Ansari, "An Analytical Framework for Video Streaming Application-Oriented Solutions in Wireless Networks," *To be submitted to IEEE Transactions on Multimedia.*

E. Haghani and N. Ansari, "Application-Oriented Networking: A Solution for Improving QoE in Wireless Networks," *submitted to IEEE Network.*

N. Ansari, P. Sakarindr, E. Haghani, C. Zhang, A. K. Jain, and Y. Q. Shi, "Evaluating Electronic Voting Systems Equipped with Voter-Verified Paper Records," *IEEE Security and Privacy*, Volume 6, Issue 3, Pages 30-39, May 2008.

E. Haghani, N. Ansari, S. Parekh, and D. Calin, "Traffic-Aware Video Streaming in Broadband Wireless Networks," *submitted to IEEE Wireless Communications and Networking Conference (WCNC), 2010.*

E. Haghani and N. Ansari, "VoIP Traffic Scheduling in WiMAX Networks," *IEEE Global Telecommunications Conference(Globecom)*, New Orleans, LA, December 1, 2008.

E. Haghani, S. De, and N. Ansari, "On modeling VoIP traffic in broadband networks," *IEEE Global Telecommunications Conference(Globecom)*, Washington, DC, November 30, 2007.

S. Motahari, E. Haghani, and S. Valaee, "Spatio-Temporal Schedulers in IEEE 802.16," *IEEE Global Telecommunications Conference(Globecom)*, St. Louis, MO, November 30, 2005.

To my beloved parents,
lovely wife,
and dear sister.

ACKNOWLEDGMENT

This dissertation would not have been possible without the support and inspiration of many people. First of all, I would like to express my sincere gratitude to my advisor, Prof. Nirwan Ansari, for his insightful guidance, enthusiastic encouragement, and constructive feedback throughout my Ph.D. studies. He gave me the freedom to work on what I liked the most while providing me with the needed assistance at the right moment. Furthermore, he taught me how to become a better researcher and helped me become a better writer. Working with him was indeed a rewarding experience and I look forward to collaborating with him in future.

Next, I would like to thank my dissertation committee members, Dr. Doru Calin, Dr. Roberto Rojas-Cessa, Dr. Yun-Qing Shi, and Dr. Osvaldo Simeone for reading this dissertation and spending their valuable time in my proposal and dissertation defenses. Their constructive comments have greatly helped in improving the quality of this work.

I am very grateful to Dr. Shyam Parekh, Dr. Doru Calin, Ms. Eunyoung Kim, and Dr. Kenneth Budka in the Bell Laboratories at Alcatel-Lucent for providing me with the terrific opportunity to work with them as an intern and for their unprecedented support, productive suggestions and invaluable contribution to my research.

I am so thankful to my fellow graduate students in the Advanced Networking Lab at NJIT for their feedback on my research and discussions on various exciting topics. I am also very thankful to all my friends at NJIT whom I spent my time with, for their genuine friendship and making my life at NJIT memorable.

I would like to thank my parents. I am forever indebted to them for their devotion, encouragements and unconditional help. Lastly, though not least, I am so grateful and obliged to my wife for her love, patience, and persistent support in every aspect of my life.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Chapter Page

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS
(Continued)

Chapter Page

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES
(Continued)

Figure Page

CHAPTER 1

INTRODUCTION

1.1 Background Information

Broadband wireless networks have been very attractive for providing broadband access because of quick and cost-effective deployments. In addition to the features of high data rate and large coverage, broadband wireless networks also promise to rapidly provide broadband access to mobile users. Till now, most existing efforts have been focused on the basic "access" capability. However, to successfully deploy and operate broadband wireless networks, a crucial issue must be addressed: "How to support a variety of services and applications within the broadband wireless networks?" For instance, there is an increasing need to efficiently support applications such as voice over IP (VoIP), video streaming, music downloading, and IPTV.

Apparently, the existing "access-oriented" design may not be the good answer because of the following reasons. (1) Efficiency: The existing design may not be able to fully exploit the potential of wireless networks because a customer always needs to connect to an access point, which is very likely becoming a bottleneck. (2) Quality of service: For many existing designs that are developed to gain Internet access, the customer may not be able to obtain the desired quality-of-service (QoS) because all services are provided by the Internet, which can only provide best-effort services in practice. (3) Incentive of the service provider: One critical issue that has been largely ignored is that the access-oriented design may not be fair to the service provider who develops the infrastructure, because the service provider can only earn the access fee, which is usually paid monthly and is relatively low as compared to the deployment cost.

To address the above issues, the service-oriented design on the network layer has been proposed recently and has gained increasing interests from both the research and industrial communities. With such a momentum, there is an urgent need to better understand the service-oriented broadband wireless network architecture and propose novel solutions to enhance the efficiency of broadband wireless networks with respect to different applications.

1.2 Objective

This dissertation discusses the challenges and possible solutions for supporting multimedia applications in broadband wireless networks. The service requirements of different multimedia applications such as video streaming and VoIP are studied and some novel service-oriented networking solutions for supporting these applications in broadband wireless networks are proposed. Moreover, this dissertation examines the performance of the proposed solutions in WiMAX, which is the promising technology for broadband wireless access in the near future. It will be shown that the proposed solutions will improve the efficiency of VoIP applications and the quality of video streaming applications over wireless networks.

1.3 Organization

The outline of the rest of this dissertation is as follows. After a brief introduction to Broadband Wireless Networks in Chapter 2, Chapter 3 presents a survey on VoIP traffic modeling. In this chapter, the access level traffic profile of VoIP applications is studied and a realistic distribution model for VoIP traffic is proposed. Based on the proposed model, an algorithm for resource allocation in wireless networks is introduced. It is shown that using that scheduling algorithm will enhance the delay and utilization performance of the network. Chapter 4 proposes a new algorithm for scheduling the uplink VoIP packets generated at the end user of a WiMAX subscriber station. The scheduling algorithm is inspired by the realistic VoIP traffic model introduced in Chapter 3.

To study the main characteristics of the MPEG video traffic, Chapter 5 explains the structure of MPEG video traffic and introduces a traffic model for video streaming application. Chapter 6 presents some challenges of streaming video traffic in wireless networks. Chapter 7 explains a cross-layer design to enhance the quality of MPEG video streaming for the end users in WiMAX networks. The proposed solution uses the characteristics of MPEG traffic to give priority to the more important frames and protect them against dropping. Chapter 8 introduces a queue management strategy in which the BS can intelligently drop less effective frames when congestion happens. It is shown that incorporating the intelligent dropping scheme along with the proposed cross-layer solution can enhance the quality of video streaming applications in wireless networks. An analytical frame to verify the performance of the streaming solutions is discoursed in Chapter 9. Finally, concluding remarks and discussions about future work are given in Chapter 10.

CHAPTER 2

BACKGROUND ON BROADBAND WIRELESS NETWORKS

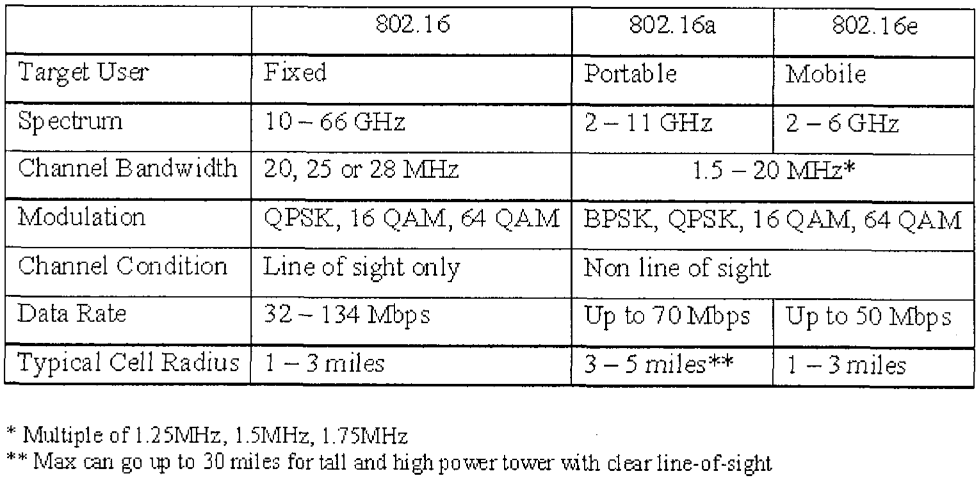
2.1 Objective

The increasing demand for broadband wireless access has called for the design and implementation of different wireless technologies such as WiMAX and LTE. These technologies are considered suitable for Metropolitan Area Network (MAN) or Wide Area Network (WAN), and some consider them also as solutions for the "last-mile" problem. This chapter provides a general introduction to the physical layer (PHY), medium access control layer (MAC) and system architecture of these technologies, in particular, WiMAX networks. However, most of the introduced features are common to other emerging technologies such as LTE.

2.2 Introduction

A Broadband Wireless Network (BWN) has a system similar to mobile phone network; it has a base station (BS) located inside the cell to provide connection to subscriber stations (SSs). Some BSs will connect to the core network directly through the Internet service provider where other BS, will act as a wireless backhaul and relay packets between the SSs and the Internet. The SSs can be fixed, portable or mobile terminals.

Worldwide Interoperability for Microwave Access (WiMAX) is based on the IEEE standard group 802.16 to provide broadband wireless access to fixed, portable and mobile systems. IEEE 802.16-2001 is the first standard for fixed broadband wireless access for the 10 - 66 GHz spectrum. It defines the air interface and the medium access control layer where Line of Sight (LoS) is required between the BS and SS. During the time of creating the standard for 802.16-2001, some other standards were also started. IEEE standard 802.16a designed for 2 - 11GHz spectrum due



	802.16	802.16a	802.16e
Target User	Fixed	Portable	Mobile
Spectrum	10 – 66 GHz	2 – 11 GHz	2 – 6 GHz
Channel Bandwidth	20, 25 or 28 MHz	1.5 – 20 MHz*	
Modulation	QPSK, 16 QAM, 64 QAM	BPSK, QPSK, 16 QAM, 64 QAM	
Channel Condition	Line of sight only	Non line of sight	
Data Rate	32 – 134 Mbps	Up to 70 Mbps	Up to 50 Mbps
Typical Cell Radius	1 – 3 miles	3 – 5 miles**	1 – 3 miles

* Multiple of 1.25MHz, 1.5MHz, 1.75MHz
** Max can go up to 30 miles for tall and high power tower with clear line-of-sight

Figure 2.1 IEEE standard 802.16.[2]

to the demand for lower frequency band where Line of sight is not required as lower frequency with longer wavelength have higher tolerance for any physical obstructions. All of these IEEE 802.16 standards were included in the 802.16-2004 (802.16 REVd during drafting), and released on October 1, 2004 [1]. To provide coverage for mobile users, the IEEE standard 802.16e was released on February 28, 2006 [3]. It is similar to 802.16a except that it operates between 2 - 6 GHz spectrums, and it provides hand-off for true mobility. Other revisions of the IEEE 802.16 standards such as 802.16j, 802.16h, and 802.16m are work in progress.

2.3 Physical Layer

2.3.1 Physical Layer Standard

WiMAX supports 3 types of users based on the IEEE standard 802.16, 802.16a, and 802.16e. Fig. 2.1 shows each standard's target user type, its working spectrum, channel bandwidth, supported modulation type, required channel condition, data transfer rate, and the typical cell radius.

2.3.2 Physical Layer Air Interface

WiMAX has two sets of physical layer specifications, one for LoS and the other for non-Line-of-Sight (NLoS). LoS service operates between 10 - 66 GHz, and nLoS service operates between 2 - 11 GHz; both services will share the same physical layer air interface. There are three types of air interface: wirelessMAN-SC, wirelessMAN-OFDM, and wirelessMAN-OFDMA. WirelessMAN-SC is a single carrier air interface modulation for the 10-66 GHz spectrum. WirelessMAN-OFDM uses Orthogonal Frequency Division Multiplexing (OFDM) and WirelessMAN-OFDMA uses Orthogonal Frequency Division Multiplexing Access (OFDMA) where the former with 256 point transform produces 256 subcarriers and the latter with 2048 point transform produces 2048 subcarriers. Both of them work in the 2 - 11GHz spectrum. There is also a wirelessMAN-SC2, a single carrier air interface modulation for the 2 - 11 GHz spectrum, but OFDM modulation is selected as the standard for the portable users and OFDMA for the mobile users.

2.3.3 Physical Layer Duplex Mode

Time-Division Duplexing (TDD) and Frequency-Division Duplexing (FDD) are the main type of duplexing mode supported by the IEEE 802.16. TDD is where both uplink and downlink use the same channel but transmit in different time slot (subframe). Fig. 2.2 shows the frame structure of TDD, where the number of downlink and uplink subframes is adjustable, and each subframe contains 4 QAM symbols. FDD is where both uplink and downlink use different channel, and both channels can transmit at the same time.

2.3.4 Physical Layer Features

Adaptive Modulation: Adaptive modulation selects the suitable modulation scheme among different types of available modulation techniques from BPSK to 64-QAM

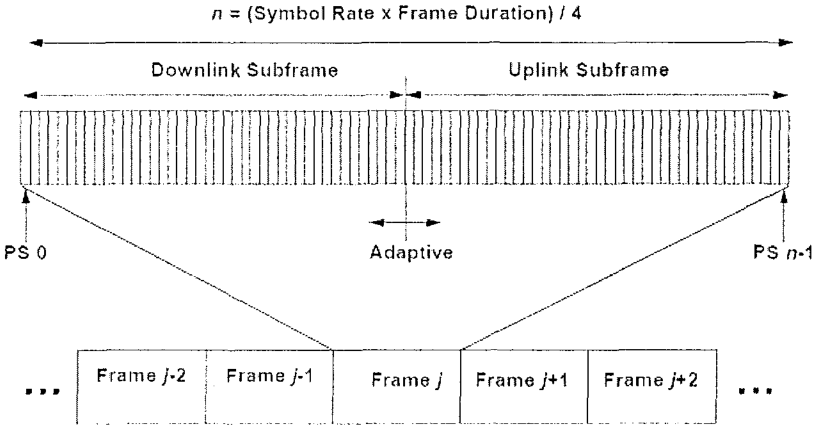


Figure 2.2 TDD Frame structure.

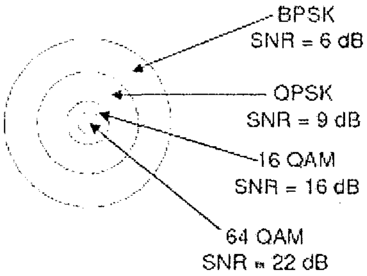


Figure 2.3 Adaptive Modulation and cell radius.

depending on the Signal to Noise Ratio (SNR). If the SS is close to the BS, the SNR will be higher, and a more efficient modulation technique likes 64-QAM or 16-QAM can be used. However, as the SS gets further away from the BS, the SNR will decrease and the modulation is switched to a less complex modulation technique like QPSK or BPSK. Fig. 2.3 shows the cell radius, and the required SNR level for each type of modulation. Adaptive modulation is also used to combat wireless channel impairments caused by interference, fading, etc.

Smart Antenna System: WiMAX supports Multi Input Multi Output (MIMO) antennae which can increase the system capacity and reduce interference. By having more than one signal to choose from, MIMO allows the wireless device to reach much greater ranges, with better signal strength (and hence better speed) than conventional wireless devices.

Variable Channel Size: For portable and mobile systems, the channel bandwidth can vary between 1.5MHz and 20MHz as long as the channel bandwidth is a multiple of 1.25MHz, 1.5MHz or 1.75MHz. This will consume the provided spectrum more efficiently.

Error Correction: WiMAX has Forward Error Correction (FEC) using Reed-Solomon GF (256) code with convolutional encoding and interleaving algorithm which can detect and correct errors caused by fadings or burst errors. It improves the overall system performance and also helps decrease the SNR requirements since it corrects errors caused by interferences or fadings.

Power Control: Similar to CDMA networks, the BS will send power control information to SSs to control their power to a predefined level. This will reduce the power consumption and interferences in the network.

2.4 MAC Layer

WiMAX has adopted the MAC layer of the IEEE 802.16 standards. This section introduces the main characteristics of the MAC layer in the IEEE 802.16 [1].

Medium access control carries connection-oriented service flows (SF). Each service flow is mapped to a MAC connection with a unique connection ID (CID). The 802.16 protocol was designed for point to multipoint (PMP) broadband wireless access applications. It accommodates hundreds of terminals per channel, and multiple end users per terminal, and provides each user with high bit rates [42].

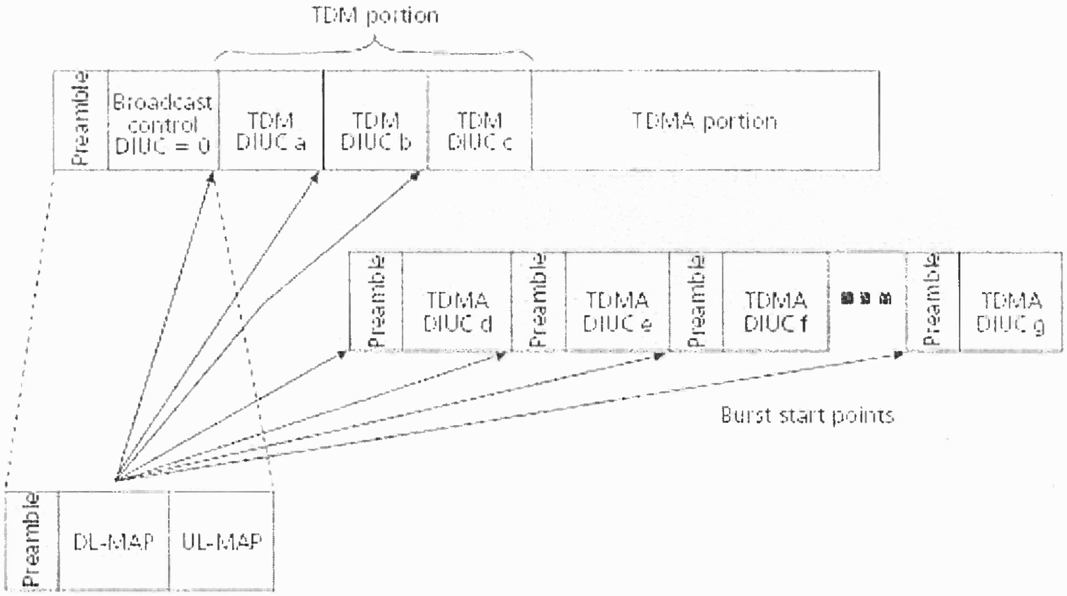


Figure 2.4 Downlink Frame Structure [17].

There are two defined types of MAC layer headers: a generic header and a Bandwidth Request (BR) MAC header, which is used by the SS to request more bandwidth. There are different types of subheaders including Automatic Repeat Request (ARQ), Grant Management subheader and fragmentation and packing subheaders [17]. The 802.16 MAC layer provides fragmentation and packing of MAC Service Data Units (SDUs) for efficient bandwidth allocation. ARQ processing performs maintenance of MAC SDU blocks that have errors or have been lost. Grant Management message is used to communicate bandwidth allocation between the SS and the BS.

The downlink frame is shown in Fig. 2.4 which is extracted from Reference [17]. The frame begins with a broadcast control section containing the downlink map (DL-MAP) for the downlink frame being sent currently and the uplink map (UL-MAP) for the next frame.

UL-MAP grants bandwidth to specific SSs. DL-MAP lets all SSs know when to listen for their own data in the current frame, and the UL-MAP informs SSs of their future transmission opportunities. The opportunities are based on each SS's dynamic

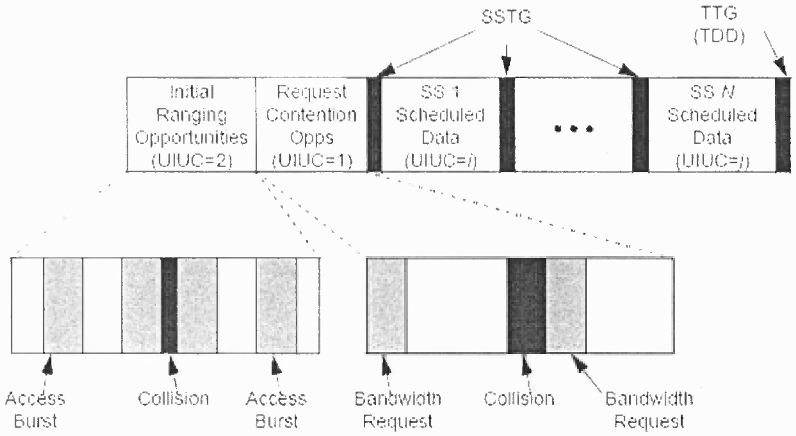

SSTG
TTG (TDD)
Initial Ranging Opportunities (UIUC=2)
Request Contention Opps (UIUC=1)
SS 1 Scheduled Data (UIUC=i)
• • •
SS N Scheduled Data (UIUC=j)
Access Burst
Collision
Access Burst
Bandwidth Request
Collision
Bandwidth Request

Figure 2.5 Uplink Frame Structure [17].

bandwidth request or on the service agreements made before. These opportunities may be pre-allocated for particular subscribers or be available for contention to all SSs.

A TDM portion usually follows these maps. This portion carries data for SSs. Burst profile for each SS is negotiated by the Downlink Interval Usage Code (DIUC). The intervals are in order of decreasing modulation robustness. The interval used by each SS may be different from that of the other SSs and is variable in time. For better supporting the half-duplex systems and to prevent them losing synchronization, the DL frame may also contain a TDMA portion. In the TDMA portion, the intervals are separated by some preambles which have information about the following interval. Thus, the users with half duplex systems are not required to be synchronized all the time, and preambles let them synchronize on time.

The uplink frame is shown in Fig. 2.5 which is extracted from Reference [17]. Initial maintenance opportunities are used to determine network delay and to request power or profile changes. Moreover, it is also used for ranging, and new users send their registration request in this period. In the request opportunities portion, SSs request bandwidth in response to a polling from the BS.

The uplink frame is composed of transmissions from different SSs. The SS's data is scheduled according to the BS's discretion and this scheduling is indicated in the UL-MAP, following the opportunities for bandwidth request. The BS may also earmark some intervals for initial maintenance and bandwidth request. Guard times, which are used as SS transition gaps, are between different intervals to help the BS synchronize to transmissions of different SSs. Collisions are possible to occur at these intervals.

In order to communicate in the network, each SS needs to successfully pass the network entry process corresponding to the desired BS. This includes downlink channel synchronization, initial ranging, capabilities negotiation, authentication message exchange, registration and IP connectivity stages. Then, the SS creates one or more connections to send data to the BS [17].

When a SS wants to enter the network, it first looks for a channel in the defined frequency list. Then, it tries to synchronize at the physical layer level with the DL channel. After receiving the DL and UL-MAP for a frame, it begins the initial ranging process by sending a ranging request MAC message on the initial ranging interval at the minimum transmission power level. It will gradually increase the transmission power till it receives a response from the BS which includes timing and power correction information for the SS. Capability request message is also sent at this time by the SS to BS and contains information about supported modulation levels, coding schemes, coding rates, and duplexing modes. A SS may be accepted or denied based on that message. If it is admitted at this stage and also accepted at the authentication stage, which is the next step, the SS is registered in the network. Then, it starts DHCP to obtain an IP address.

The IEEE 802.16 needs an advanced Radio Link Control (RLC) for its advanced physical layer technology. In addition to the traditional RLC functions of power control and ranging, it should control the capability of the Physical Layer (PHY) to transit from one burst profile to another.

2.5 MAC Scheduler

The MAC scheduler is responsible for allocating radio resources to clients efficiently based on the QoS requirements, PHY layer conditions, and any other scheduler criteria. The scheduler prevents collisions which are caused by simultaneous transmission, by regulating the transmissions' orders and medium access permutations among users.

2.5.1 Scheduling Mechanisms

The purpose of the MAC layer scheduler is to provide traffic flows with efficient medium access to transmit data of different applications such as voice and video with various QoS requirements over varying wireless channels. In general, the MAC layer schedulers should possess the following features to facilitate broadband wireless communications.

Efficient Data Scheduler: The scheduler is designed to efficiently allocate available resources to traffic flows over varying wireless channels. Each traffic flow is characterized by its corresponding QoS parameters used by the BS to schedule packet transmissions in the MAC layer. The scheduler is responsible to serve traffic flows according to their defined QoS requirements despite rapid changes in the rate of bursty traffic or variations of the wireless channel conditions. As a channel indicator, the Channel Quality Indicator Channel (CQICH) provides the base station with instantaneous channel information and helps it select the optimum modulation and coding scheme for each transmission. Different WiMAX specifications such as Adaptive Modulation and Coding (AMC) and Hybrid ARQ (HARQ) protect data transmission over the lossy PHY channel. Fig. 2.6 shows the input parameters of a general MAC scheduler.

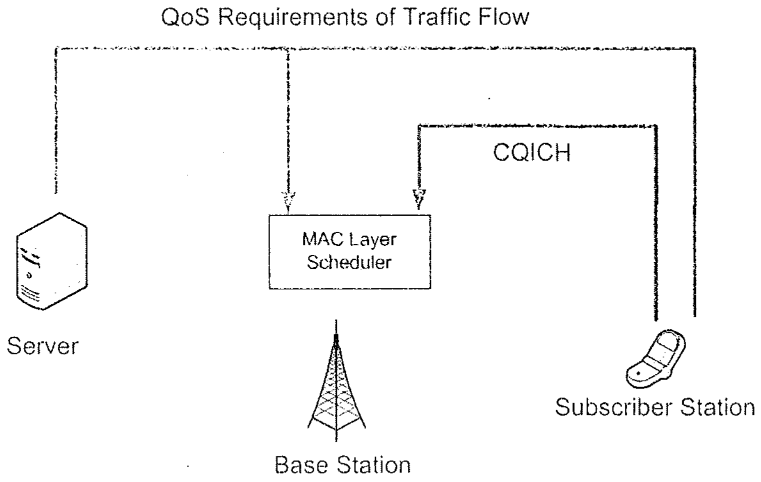


Figure 2.6 Scheduler Inputs.

Scheduling for both UL and DL: The scheduler should allocate resources to both UL and DL traffic flows. For efficient scheduling of the DL traffic, the BS uses the feedback from the SS, QoS parameters of the DL traffic flow, and the status of DL queue of the traffic flow located at the BS. To provide required QoS to UL traffic flow and to efficiently provide the UL traffic flows with medium access, the BS receives adequate and prompt feedback from the SS about its queue status and bandwidth requirements. Owing to different QoS service classes, various bandwidth request and polling mechanisms have been designed to deliver required information to the BS to efficiently allocate resources to UL traffic flows. The polling and bandwidth request mechanisms are chosen based on the UL service flow. Since the co-channel interference among adjacent cells is prohibited by the OFDM scheme, UL scheduling can better fulfill QoS requirements.

Dynamic Scheduling: The BS broadcasts the resource allocation information for both UL and DL in the MAP message at the beginning of each MAC frame. Thus, the scheduling is performed on frame-by-frame basis and the BS can choose the optimum scheduling strategy according to the instantaneous status of the network traffic and wireless channel conditions in each frame. Scheduling in both time and frequency domains on a per frame basis facilitates efficient resource allocation for supporting QoS in the dynamic wireless network.

QoS Provisioning: The MAC scheduler supports true QoS for all traffic flows. In a connection oriented fashion, each traffic flow has a service flow ID which includes a set of QoS parameters used to determine the QoS requirements of the traffic flow. The BS uses the QoS parameters of the traffic flow to schedule both UL and DL traffic and to provide them with required QoS. The QoS parameters define the possible bandwidth request and polling mechanisms of UL traffic flow. Furthermore, per frame scheduling ensures the efficient performance of the UL scheduler.

Time-Frequency Scheduling: The scheduler incorporates the OFDM technology to allocate resources in both time and frequency dimensions. The scheduler may pseudo randomly allocate a traffic flow with different subchannels across the bandwidth to achieve the highest frequency diversity gain. The scheduler can also assign a subscriber station to its strongest available subchannels, and thus maximizing the transmission rate. The scheduler identifies the strongest subchannels by using information provided by the Channel Quality Indicator (CQI). Moreover, by applying time-frequency resource allocation, the scheduler can more efficiently allocate available resources to users.

2.5.2 Uplink Scheduling

The scheduler receives bandwidth requests from subscriber stations and grants the requests upon availability of resources and according to the QoS requirements of the UL service flow. There are different mechanisms for the subscriber station to request bandwidth from the scheduler. Existing mechanisms include:

- Polling

- Piggybacking
- Bandwidth request through ranging message

The subscriber station can use any of the available requesting mechanisms for a traffic flow according to its QoS service class. In a connection oriented fashion, the scheduler efficiently allocates resources to UL traffic flows similar to DL flows, and thus supports QoS for UL traffic as well.

2.5.3 Scheduling Metrics

In general, the MAC scheduler calculates a metric M for all traffic flows to determine their serving order. The metric M_i may depend on different parameters of the traffic flow i such as QoS requirement, delay, throughput, etc.

$$M_i = f(QoS_i, Delay_i, Throughput_i, other\ parameter_i) \tag{2.1}$$

Based on the scheduler objectives, different metrics and scheduling algorithms have been proposed [50]. Although it is up to the network designers to adopt the appropriate scheduling scheme according to their network properties, some of the general scheduling schemes used in different networks along with their attributes are introduced below:

1. Round Robin Scheduler: Serves the traffic flows in a round robin fashion until all flows are served. Although it is simple to implement, it is usually not capable of fulfilling different objectives such as utilization, QoS, and throughput maximization [7, 44].

2. Round Robin with Priorities: Classifies the traffic flows according to their priority classes according to their QoS requirements, and serves them in a round robin manner until all flows are served [13, 45].

3. Max Carrier/Interference (C/I) Scheduler: Arranges the traffic flows in decreasing order based on their C/I ratio values. This method has the tendency to maximize the throughput, but it may not be fair to the subscriber stations with weak PHY layer wireless channels [53, 60].

4. Proportional Fair Scheduler (PFS): This scheme provides a better compromise between many conflicting objectives, such as throughput maximization, fairness to all flows, and ease of implementation. This scheme applies a greedy approach to maximize the throughput [6, 33].

After determining the serving order of traffic flows, the scheduler performs the following tasks sequentially to finalize the resource allocation procedure. The scheduler may carry out heuristic solutions in each of the following tasks according to the specific network design and specifications.

- Determine the amount of traffic to be transmitted by a chosen flow.
- Allocate adequate resources in the DL or UL subframes.
- Create the DL-MAP and UL-MAP.
- Generate and process the TDD frame in accordance to the MAP.

In the general deployment case, the MAC scheduler should be capable of supporting different QoS classes defined by the IEEE 802.16e air interface, which has also been adopted by the WiMAX Forum system profile. Figs. 2.7 and 2.8, which are extracted from Reference [62], show general schematics of the DL and UL schedulers, respectively.

2.5.4 QoS Support

The IEEE 802.16 defines five QoS service classes: Unsolicited Grant Scheme (UGS), Real Time Polling Service (rtPS), Extended Real Time Polling Service (ertPS), Non

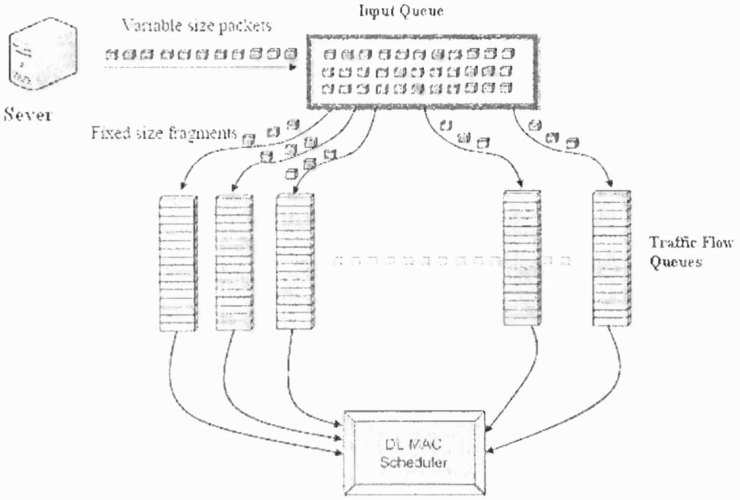


Figure 2.7 Downlink Packet Scheduler [62].

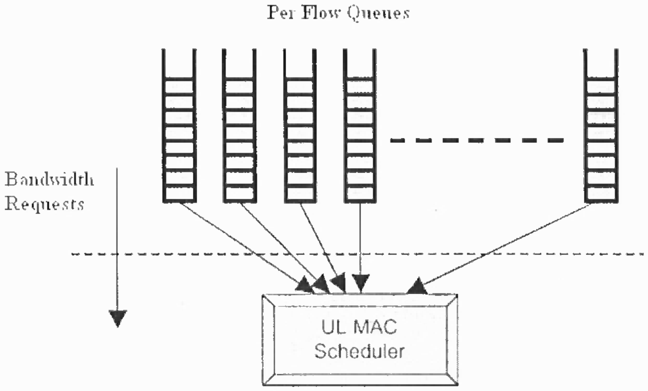


Figure 2.8 Uplink Packet Scheduler [62].

Real Time Polling Service (nrtPS), and Best Effort Service (BE). Each of these has its own QoS parameters such as minimum throughput requirement and delay/jitter constraints.

UGS: This service class provides a fixed periodic bandwidth allocation. Once the connection is setup, there is no need to send any other requests. This service is designed for constant bit rate (CBR) real-time traffic such as E1/T1 circuit emulation. The main QoS parameters are maximum sustained rate (MST), maximum latency and tolerated jitter (the maximum delay variation).

rtPS: This service class is for variable bit rate (VBR) realtime traffic such as MPEG compressed video. Unlike UGS, rtPS bandwidth requirements vary and so the BS needs to regularly poll each MS to determine the appropriate allocations. The QoS parameters are similar to the UGS, but minimum reserved traffic rate and maximum sustained traffic rate need to be specified separately. For UGS and ertPS services, these two parameters are the same, if present.

ertPS: This service is designed to support VoIP with silence suppression. No traffic is sent during silent periods. The ertPS service is similar to UGS in that the BS allocates the maximum sustained rate in the active mode, but no bandwidth is allocated during the silent period. There is a need to have the BS poll the MS during the silent period to determine if the silent period has ended. The QoS parameters are the same as those in UGS.

nrtPS: This service class is for non-real-time VBR traffic with no delay guarantee. Only minimum rate is guaranteed. File Transfer Protocol (FTP) traffic is an example of applications using this service class.

BE: Most of data traffic falls into this category. This service class guarantees neither delay nor throughput. The bandwidth will be granted to the MS if and only if there is a left-over bandwidth from other classes. In practice, most implementations allow specifying minimum reserved traffic rate and maximum sustained traffic rate, even for this class.

It is worth noting that for non-real-time traffic, traffic priority is also one of the QoS parameters that can differentiate among different connections or subscribers within the same service class. By considering bandwidth request mechanisms for uplink, it is understood that UGS, ertPS and rtPS are real-time traffic. UGS uses a static allocation. ertPS is a combination of UGS and rtPS. Both UGS and ertPS can reserve the bandwidth during setup. Unlike UGS, ertPS allows all kinds of bandwidth request including contention resolution. However, rtPS cannot participate in contention resolution. For other traffic classes (non real-time traffic), nrtPS and BE, several types of bandwidth requests are allowed such as piggybacking, bandwidth stealing, unicast polling and contention resolution. Reference [54] provides a survey on scheduling in the IEEE 802.16e networks.

Table 2.1 shows the defined QoS service classes of WiMAX and their admission control parameters.

Finally, it is noted that the admission control process is used to determine how many traffic flows of each class with what QoS parameters can be admitted in the network at a given time. This function is outside the scope of the MAC scheduler. A generic MAC scheduler may classify flows based on their service classes and schedule users within classes, based on the following criterion:

- Delay constraints of real-time traffic flows.
- Throughput requirements.
- Jitter constraints.

Table 2.1 WiMAX QoS Service Classes Summary

Class	Applications	Admission Control Parameters
UGS	CBR real-time periodic traffic like T1 connection	Maximum Sustained Traffic Rate, Minimum reserved Traffic Rate, Maximum Latency, Tolerated Jitter
ertPS	VoIP with silence suppression/Video conference (real-time variable-size periodic data)	Maximum Sustained Traffic Rate, Minimum reserved Traffic Rate, Maximum Latency
rtPS	Real-time Video (real-time variable size data on periodic basis)	Minimum reserved Traffic Rate, Maximum Sustained Traffic Rate, Maximum Latency
nrtPS	FTP (variable size data)	Minimum reserved Traffic Rate, Maximum Sustained Traffic Rate,
BE	Web browsing traffic	Minimum reserved Traffic Rate, Maximum Sustained Traffic Rate,

- Power constraints of power-limited subscriber stations.

2.6 WiMAX Architecture

The mobile WiMAX End-to-End network architecture is based on an All-IP platform. Owing to the All-IP architecture, it is possible to use a common network core without the need to maintain both packet and circuit core networks separately which reduces the maintenance overhead of supporting two separate networks; this results in lower cost, high scalability, and easier deployment since the networking functionalities are mainly becoming software-based services.

In order to successfully deploy efficient and operational WiMAX networks, it is crucial to support beyond the IEEE 802.16 (PHY/MAC) air interface specifications. In particular, it is needed to design a core set of networking functionalities as a complimentary part of the End-to-End WiMAX system architecture. Some of the basic concepts that have been applied to the WiMAX architecture are introduced next.

1. The network architecture is designed based on a packet-switched framework which is compatible to the original specifications introduced in the IEEE 802.16 standard, IETF RFCs, and Ethernet standards.

2. The architecture permits separating the access architecture (and supported topologies) from IP connectivity services. Network elements of the connectivity system are transparent to the IEEE 802.16 radio specifications.

3. The network architecture is flexible enough to support a broad range of deployment options including:

   - Small-scale (sparse) to large-scale (dense) radio coverage .
   - Efficiency for rural, suburban, and urban radio propagation environments.

- Operating in licensed and/or licensed-exempt frequency bands.
- Possibility of hierarchical, flat, multi-hop or mesh topologies.
- Supporting fixed, nomadic, portable and mobile users.

2.7 Summary

With the high demand for wireless broadband access and increasing need for ever-widening range of applications that encompass fixed, nomadic, portable and mobile data access as well as fixed and mobile voice services, WiMAX is committed to meeting the requirements of all these applications.

This chapter presented some important features of WiMAX in both the physical and MAC layers. WiMAX has many advantages that allow it to provide NLoS access, with essential features such as OFDM technology, error correction and adaptive modulation. Furthermore, WiMAX has many other features such as ARQ, diversity and space-time coding that provides many invaluable solutions.

In this chapter, the All-IP WiMAX network architecture and its connectivity with IP based networks were introduced. The importance of designing a core set of networking functions for successful and efficient transferring of data from the base stations to the IP networks were also explained.

CHAPTER 3

MODELING VOIP TRAFFIC IN BROADBAND NETWORKS

3.1 Objective

With the general trend towards ubiquitous access to networks, more users will prefer to make voice calls through the Internet. Voice over IP (VoIP) as the application which facilitates voice calls through the Internet will increasingly occupy more traffic. The growth of delay sensitive traffic that requires special quality of service from the network will impose new constraints on network designers who should wisely allocate the limited resources to users based on their required quality of service. An efficient resource allocation depends upon gaining accurate information about the traffic profile of user applications. In this chapter, the access level traffic profile of VoIP applications is studied and a realistic distribution model for VoIP traffic is proposed. Based on the proposed model, an algorithm for resource allocation in networks is introduced. It is shown by using that algorithm will enhance the delay and utilization performance of the network.

3.2 Introduction

Voice over IP (VoIP) is a rapidly growing service which is providing voice communications over packet-switched networks. With extensive growth of the Internet and growing demand for provisioning various applications and services, more service providers are trying to provide people with new applications and technologies to make their voice calls through the Internet.

With the advent of new technologies, people can access the Internet through different types of connections. These different technologies may impose various challenges on network design, but, in the terms of their functionalities, they all have to provide customers with network resources to run their applications. Regardless of the data link and physical layer protocols used in these technologies, the IP layer traffic generated by VoIP applications does exhibit similarities. These similarities become more visible when users are employing a broadband access, demanding services for their applications through a high speed connection. For example, a user who makes voice calls through a 3G wireless network expects the same voice quality as that of a DSL user. Towards the necessity of a thorough study of VoIP traffic, it is tried to investigate some of the key factors of VoIP traffic in this chapter.

There are many IP based applications that generate VoIP packet traffic in the Internet. Some of the main characteristics of the these traffics are studied in this chapter. Although there are various VoIP commercial applications that provide voice connections between PCs and phones, they all generate IP packets in the Internet. Some of the main commercial applications are Skype, MSN, Yahoo messenger, etc. These applications could be run on any PC or wireless device, and it is up to the service providers to provision their customers with enough network resources to make their voice calls.

Knowledge of VoIP traffic characteristics becomes more crucial especially when the service providers are encountering scarcity in network resources and they are required to allocate their resources as efficiently as possible. This demand has motivated many researchers to study and model voice traffic. By reviewing the literatures, it is found that many articles on modeling call arrival rate and call duration [15, 16]. In this research, it is interested to capture traffic characteristics of a single VoIP connection at the end user. Understanding the main features of the VoIP traffic at the end user will help to anticipate the packet generation time which can be capitalized to improve the network efficiency.

There have been much research effort in modeling the VoIP traffic. The main traffic model adopted for voice traffic at the end user is the *ON-OFF* model [43]. This model is inspired by the nature of voice which is composed of periods of silence

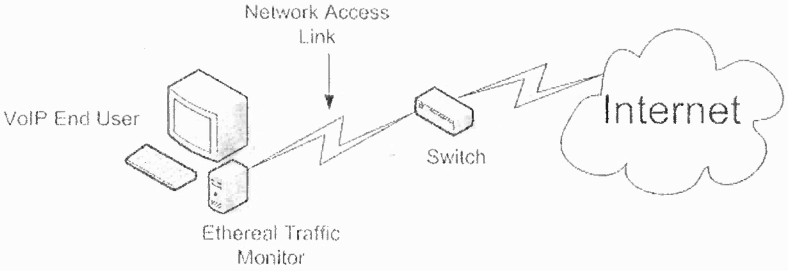


Figure 3.1 Traffic Monitoring System.

and sound. In this model, the source generates equal-size packets separated equally in time during the *ON* period and either does not generate any packet or generate smaller packets during the *OFF* period . Based on the voice characteristics, the duration of each period is assumed to be predetermined in this model[43].

Although the ON-OFF model has been used for studying the behavior of VoIP applications in networks [32, 16], the modeling of VoIP packets generated at end users requires further investigation. This is becoming more obvious when it is noted that VoIP applications do not necessarily perform the same procedure for generating voice packets [11]. It is also worth noting that the impact of transport layer protocols on VoIP packet streams is not considered in the ON-OFF model. Since VoIP applications might use different transport layer protocols such as TCP, UDP, or SCTP, the procedure of generating IP packets would be different, and the last-mile networks that deal with IP flows need further information to anticipate the behavior of VoIP traffic.

In order to model the traffic profile at the end user more accurately, it is decided to run VoIP applications and monitor the packets traversed through networks. By monitoring the uplink traffic at each user, it is possible to generate real traces for VoIP traffic, that incorporates the impact of all protocols above the IP layer. The resulted traces would be helpful to gain better insight on the VoIP traffic profile which would guide us towards more detailed and accurate traffic modeling of VoIP applications.

This chapter proposes a realistic model for VoIP traffic traversed in the uplink based on real traffic traces. In particular, the model describes the *inter-packet time* of VoIP traffic. The inter-packet time is the time between two consecutive packets sent to the network from the VoIP applications. The outline of the rest of the chapter is as follows. In Section 3.3, the methodology for generating the traffic traces is explained. Section 3.4 discusses the characteristics of the generated traces. In Section 3.5, a VoIP traffic model is introduced, and its parameters are elaborated. Section 3.6 presents simulation results and performance comparison of conventional TDMA networks with that of networks that exploit the proposed model in resource allocation for VoIP traffic. Concluding remarks are given in Section 3.7.

3.3 Methodology for Trace Generation

As mentioned in Section 3.2, in order to understand the traffic characteristics of VoIP connections at end users, it is needed to capture the VoIP packets. The *ethereal* is run at each side of the VoIP connection. By using ethereal, it is possible to capture the traffic at each layer, and the captured data would be helpful to generate real-time traffic traces. An overview of the trace generating system is shown in Fig. 3.1.

This method of generating traffic traces has also been used in other research works [57, 55]. Nevertheless, this method will be applied to derive a legitimate model for VoIP traffic. For this reason, hundreds of traffic traces are generated, and by studying them it was possible to conceive the common characteristics of those traces. Additionally, different voice calls between different source-destination pairs with different VoIP applications have been monitored.

3.4 Trace Results

VoIP connections can be initiated by different applications, and either side of a connection might have access to different kinds of networks and use different devices.

In order to generate legitimate traces, different kinds of VoIP connections via different applications are established. Also, many voice calls to PCs and phones located in different parts of the world are made. As depicted in Fig. 3.1, Ethereal is run on a PC which is connected to the network via different technologies. The VoIP traces especially generated from DSL and 100 Mbps LAN connections are studied as they provide the end users with high speed network access. Skype, MSN, and Yahoo Messenger are used as the applications for generating VoIP traffic. These applications can be easily installed on any PC or mobile device.

Some particular characteristics of the VoIP traffic are observed by studying the resulted traces. The size of packets and inter-packet time in the uplink for each voice connection are examined. A summary of the results is described next.

3.4.1 Packet Size

By capturing the packets generated by VoIP applications it is found that the packet sizes are not varying much during the time of a conversation. Although different voice connections might result in different packet sizes, each connection will bond to a relatively fixed size for the majority of its packets. The resulted probability mass function (PMF) of the VoIP packet size is shown in Fig. 3.2.

As mentioned in Reference [49], G.711 and G.723.1 are two of the standard speech codecs used in VoIP applications. These standards generate equal size packets. The size of the packets is a function of the available bandwidth. On the other hand, transport layer protocols may change the size of the data segment, but the key factor resulted from the traces and standards is the fact that the majority of packets of a connection bond to a fixed size.

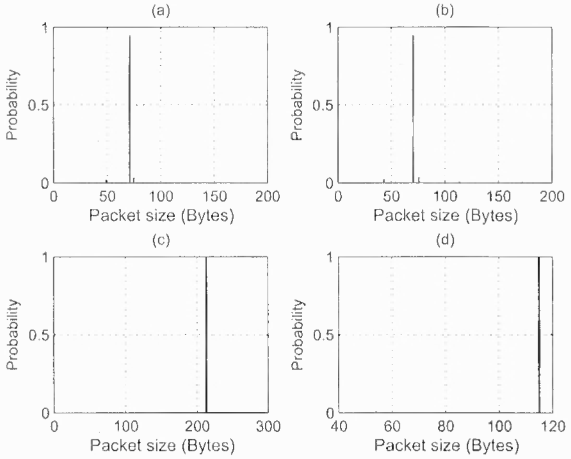


Figure 3.2 PMF of VoIP packet size: a) Skype; b) another instance of Skype; c) Yahoo Messenger; d)MSN.

3.4.2 Inter-Packet Time

In order to gain real insight about the behavior of the VoIP packets, it is crucial to know their variations in a timely fashion. The uplink packets made by VoIP applications are monitored in different scenarios and the time between subsequent packets which is called *inter-packet* time is measured. The inter-packet time of VoIP packets is measured for different destinations with different voice applications. The measurement tests were run during the call duration which is in the order of a few minutes. Some of the resulted distributions for inter-packet time of VoIP packets are shown in Fig. 3.3.

The inter-packet time of VoIP traffic has been measured in some other research works as well [49, 55, 11]. However, these works presented similar patterns for the traffic profile, but they had not neither proposed a distribution model for inter-packet time nor used this distribution in resource allocation.

As shown in Fig. 3.3, the inter-packet time for different calls will result in different distributions. Nevertheless, all the distributions can be accurately modeled, which will be discuss next.

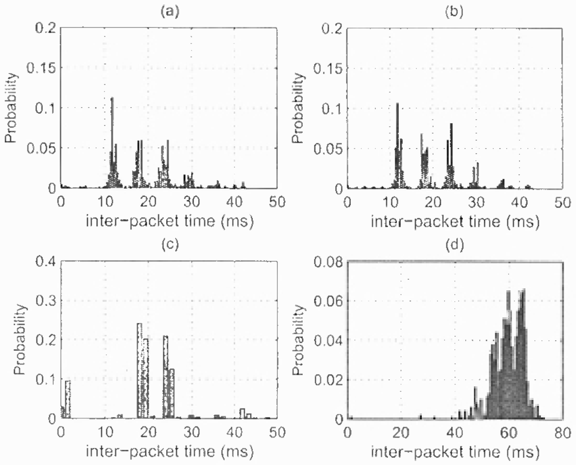


Figure 3.3 PMF of VoIP inter-packet time from real traces: a) Skype; b) another instance of Skype; c) Yahoo Messenger; d) MSN.

3.5 Distribution Model

From the resulted traces shown in Section 3.4, it is desired to introduce a model which captures the behavior of VoIP traffic at the uplink of end users. Considering the resulted traces, it is found that more than 95% of generated packets had the same packet size. Based on the route bandwidth, any VoIP application may deploy various voice coding standards. Thus, the size of packets generated by a VoIP application might not be the same for different voice calls[49].

The captured traffic traces is use to model the inter-packet time in the VoIP packets. As depicted in Fig. 3.3, for any voice connection, the inter-packet time is mainly located close to a few distinct values which are referred to as *taps* in this chapter. Inspired by this observation, an inter-packet distribution is defined as shown in Fig. 3.4. The distribution of the inter-packet time of uplink packets can be written as (3.1).

$$P_{\Delta}(\Delta) = \sum_{i=1}^{N} p_i(\Delta_i) \tag{3.1}$$

In (3.1), $P_{\Delta}(\Delta)$ is the PMF of the inter-packet time and each of the distinct values is called a *tap*. Each tap indicates the probability of having an inter-packet time equal to Δ_i or

$$p_i = Pr\{\Delta = \Delta_i\}$$

The distribution of inter-packet time for any VoIP connection can be modeled with two matrices:

- The inter-packet time matrix Δ, which is $1 \times N$, shows the locations of taps.

$$\Delta = [\Delta_1, \Delta_2, ..., \Delta_N]$$

- The probability matrix P, which is $1 \times N$, shows the value of PMF for each tap.

$$P = [p_1, p_2, ..., p_N]$$

N is the number of taps in the model. For each voice connection, the parameter N is constant while different calls might have different numbers of taps. As mentioned before, the size of the packets, S, is also constant in this model. Therefore, one can understand the parameters of the multi-tap traffic model by studying (P, Δ, S).

Based on the PMF model matrices, the probability of sending the next packet in Δ_n^{ms} is

$$\Pr\{\Delta < \Delta_n\} = \sum_{i=1}^{n-1} p_i \tag{3.2}$$

Thus, the probability of having an inter-packet time longer than Δ_N is

$$\Pr\{\Delta > \Delta_N\} = 1 - \sum_{i=1}^{N} p_i \tag{3.3}$$

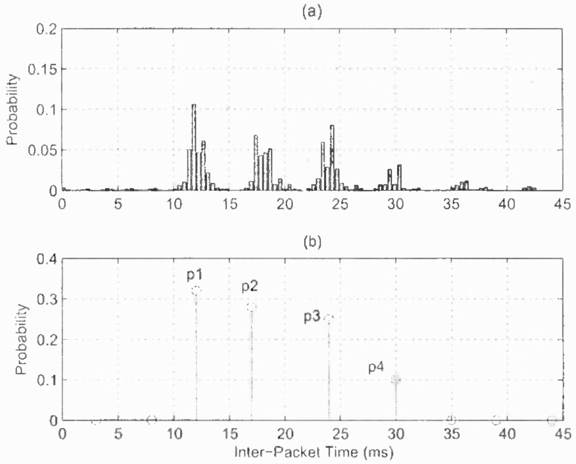


Figure 3.4 Distribution of inter-packet time: (a) real distribution from the trace; (b) resulted model for distribution.

Based on the PMF, the cumulative distribution function (CDF) of the inter-packet time can be described by the following matrix C.

$$C = [C_1, C_2, ..., C_N]$$

$$C_j = \Pr\{\Delta < \Delta_j\} = \sum_{i=1}^{j-1} p_i \tag{3.4}$$

As an example, for a voice call trace and its deduced model shown in Fig. 3.4, the parameters of the traffic model are $N = 4$, $\Delta = [12, 17, 24, 30]$ (ms), and $P = [0.32, 0.28, 0.25, 0.10]$.

For any voice call, the number of taps, N, and the value of PMF at each tap might be different, but as will be shown later, finding the parameters of the model, will be quite fast as compared to the call duration. It allows the resource allocators to make optimum decisions in the minimal time. As mentioned in Section 3.4, in all of the generated traces, the packet sizes and inter-packet time values are close to some fixed values that are functions of the network bandwidth. Since these are the parameters of the model, revealing accurate model parameters in the least time

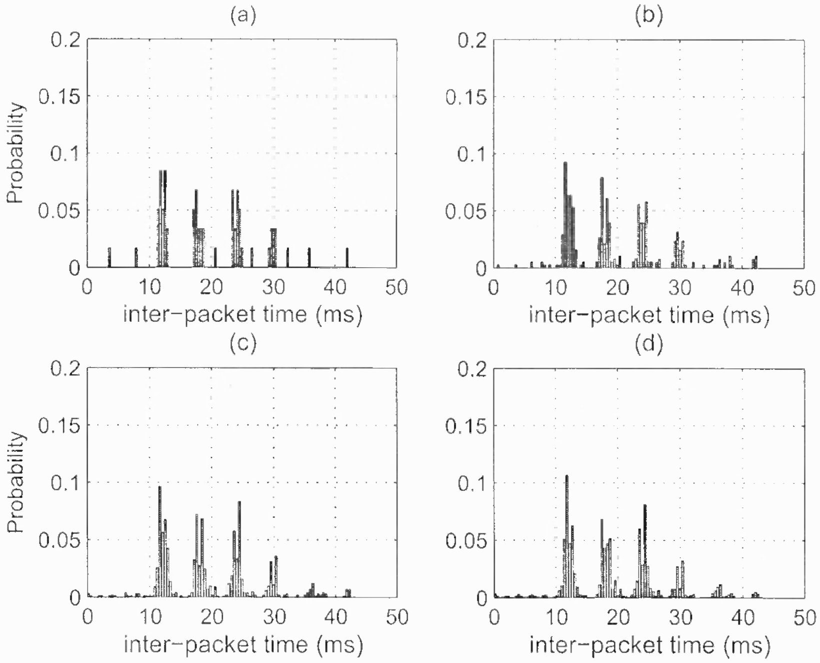


Figure 3.5 Convergence of PMF for inter-packet time from traces with duration of T: a) T=1 sec; b) T=7 sec; c) T=20 sec; d) T=60 sec.

will significantly improve the performance of the network. In the case of a change in network status as the codecs will update the packetization procedure, the model parameters will also be updated within a short period of time.

Fig. 3.5 shows the resulted PMF of inter-packet time calculated for the same trace at different times. As depicted in Fig. 3.5 the location of taps can be computed even in a short period of time like 1*sec*.

3.6 System Simulations

In this section, an algorithm for resource allocation is proposed and the impact of the proposed algorithm on the network performance is simulated. Since VoIP traffic is delay sensitive, the delay efficiency and bandwidth utility are especially noted. For this purpose, a user running a VoIP application and requiring network access to transmit its packets in the uplink towards the destination is considered. It is assumed that parameters of the distribution model have been converged and the network connection is in the steady state mode. The real traces explained in Section 3.3 are used for simulating the VoIP packet stream. As mentioned in Section 3.4, the traces were generated in broadband networks with available uplink bit rates of more than $1Mbps$. Thus, the resulted inter-packet times shown in Fig. 3.3 and functions of the VoIP application rather than media access control (MAC) protocol. Therefore, the generated traces could be used for simulating the packet stream in any broadband network.

The MAC layer of a broadband TDMA system consisting of frames each with the length equal to Γ is considered. Each frame is composed of a constant number of time slots, each with a length of τ. In the simulations, it is assumed that τ is equal to the time that a user needs to transmit a VoIP packet. As VoIP traffic is delay sensitive, it is desired that the user can transmit the VoIP packet with the minimum delay. Therefore, the scheduling algorithm has to assign a time slot to the user as soon as it has a packet to send. Regarding the distribution of inter-packet time discussed in Section 3.4, the time difference between two consecutive packets is a random variable. Furthermore, as mentioned in Section 3.4, the size of VoIP packets is relatively small, and it will incur a waste of bandwidth and introduce additional delay if the user requests a time slot for any single packet. Accordingly, it is assumed that the user cannot send requests for bandwidth; however, it was permitted to piggyback the bandwidth request with the data packets. Hence, if the user has more than one packet in its queue or the packet size is bigger than normal, it will piggyback the request for extra time slots in the data packet. Thus, the out of order or big packets will not waste any bandwidth as the scheduler will know their existence prior to any resource allocations.

There is a probability that the scheduler reserves a time slot for the user but the user does not have any packet to send, and as a result that reserved time slot would be wasted. This probability is higher if reservations are made frequently. On the other hand, if the scheduler assigns time slots less frequently, the transmission delay of the VoIP packet will increase. Therefore, the scheduler has to optimize the reservation in order to reduce both the packet delay and unused time slots. Two different scheduling algorithms are compared: a conventional TDMA access method, and a novel method based on the traffic model discussed in Section 3.5.

- *Conventional Method*

  In this method, the user will have periodical access to network for sending its uplink VoIP packets towards the destination. A variable U is defined as the period of access in terms of frames. For example, $U = 1$ means the user will have access to the network in each consecutive frame, and $U = 2$ means it will gain access in every second frame. In the simulations, the impact of $1 \leq U \leq 4$ is examined.

- *Novel Method*

  In this method, it is supposed that the scheduler knows the parameters defined in Section 3.5. Therefore, the scheduler has enough knowledge to model and estimate the inter-packet time matrix elements Δ_i for $i = 1, 2, ..., N$. Based on this information, the scheduler wisely reserves a time slot for the user to transmit its data at the time of Δ_1. It is also assumed that if $\Delta_1 < \Gamma$, the packet would be sent in the very next frame, and in other cases the packet would be sent in the frame, located Δ_1 away from the pervious packet. If the packet has not been generated till Δ_1, the scheduler will reserve another access in the frame located Δ_2 away from the previous packet. In this case, the first reservation would be wasted as it could have been assigned to other users. The scheduler will continually reserve time slots for the user at each Δ_i which is an element of the inter-packet time matrix Δ unless the source has a packet to send or the time elapses Δ_N from the previous packet. As shown in Fig. 3.3, more than 95% of the packets will be generated within Δ_N of the previous packet, but for those few remainders the scheduler will reserve a time slot for the VoIP user in every consecutive frame located further than Δ_N till the user sends its packet or the connection becomes timed out.

The average number of frames a packet should wait till the user secures a permission for transmission is measured. The average number of wasted reserved time slots caused by the false estimation of the next packet generation time is also computed. The number of wasted time slots can be used as an indicator for the efficiency of the network bandwidth utilization. The frame duration is changed from $5ms$ to $40ms$, and the impact of U in the conventional method is observed.

Fig. 3.6 shows the average delay time that a packet waits before transmission versus frame duration Γ for the novel method and conventional method with reservation frequency $1 \leq U \leq 4$. Fig. 3.7 demonstrates the average number of time slots wasted for each packet before it grants the reservation. As depicted in Figs. 3.6 and 3.7, it is observed that although the conventional method will waste slightly less time slots at longer frame lengths, its delay performance is far worse which is not acceptable for delay sensitive applications such as VoIP. It is also worth noting that for any frame length, the difference between the average number of wasted reservations for the conventional and novel methods would be less than 1 time slot which is negligible. Thus, the effectiveness of the proposed model is demonstrated.

3.7 Summary

In this chapter, the VoIP traffic behavior at end users is studied. VoIP is an important application for next generation broadband access networks. Thus, understanding the characteristics of VoIP traffic is crucial for designing efficient networks. In order to determine an accurate model for VoIP traffic transmitted in the uplink, the uplink

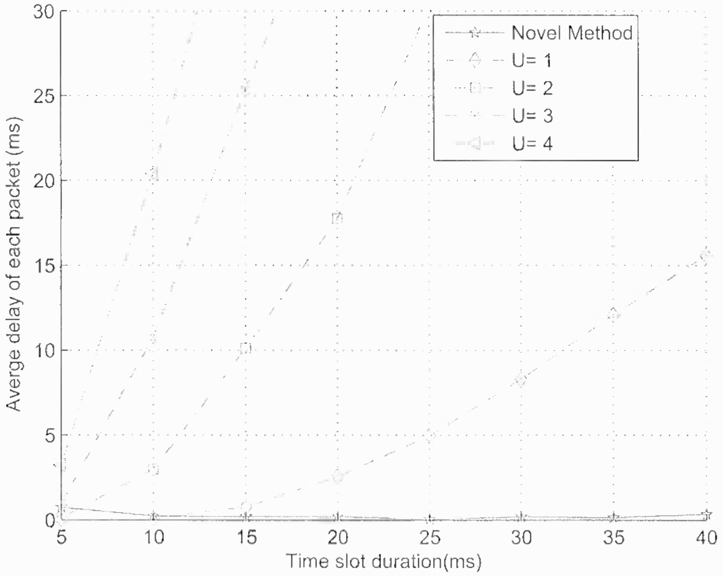


Figure 3.6 Delay comparison.

VoIP traffic generated by different applications at the end user is captured, and an accurate model based on the resulted traces is defined.

In this work, the packet size and inter-packet time of VoIP traffic is modeled. It is found that the packet sizes do not vary much during the conversation time. This work also proposes a multi-tap model to capture the features of the inter-packet time. Based on this model, an algorithm for resource allocation in TDMA networks is proposed. It is observed that with the accurate anticipation of the packet generation time, the average number of missed bandwidth reservations will be less than that of conventional methods for shorter frame lengths, and comparable to that of conventional methods for longer frame lengths. It was also shown that as the frame duration increases, the average delay that a packet waits at the source before transmission in the uplink using this algorithm will be less than that of conventional resource allocation methods.

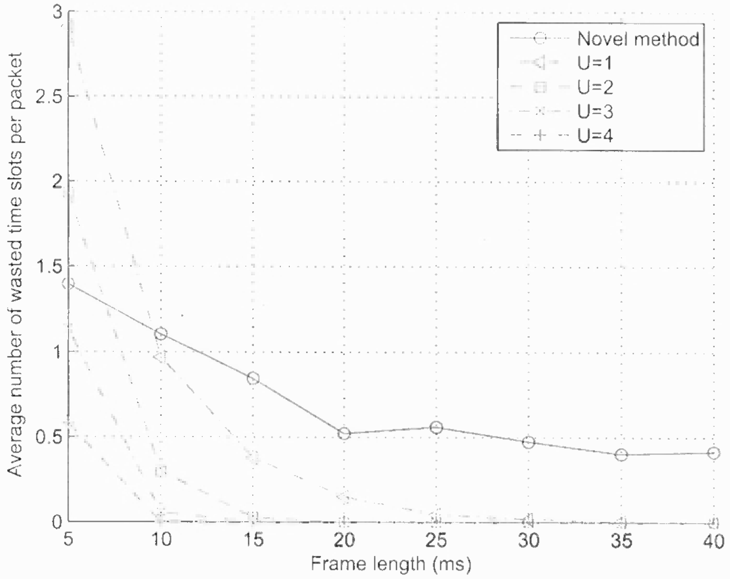


Figure 3.7 Average wasted bandwidth.

CHAPTER 4

VOIP TRAFFIC SCHEDULING IN WIMAX NETWORKS

4.1 Objective

Popularity of Voice over IP (VoIP) applications such as Skype, Google Talk, and MSN Messenger along with emerging deployment of WiMAX networks is making VoIP over WiMAX an attractive market and a driving force for both carriers and equipment suppliers in capturing and spurring the next wave of telecommunications innovation, though challenges remain. Optimization of the VoIP call capacity over WiMAX networks is one such crucial challenge and remains an open research issue. While conventional scheduling methods have not considered the traffic characteristics of VoIP, in this chapter, a traffic aware scheduling algorithm for VoIP applications in WiMAX networks is propose. The performance of the proposed method is studied and compared with that of some conventional methods. The tradeoff between delay and bandwidth efficiency is discussed, and it is shown that using the scheduling method enhances the efficiency of VoIP over WiMAX.

4.2 Introduction

WiMAX is the promising technology for broadband wireless access for the near future. The excessive demand for providing mobile users with broadband wireless access has attracted tremendous investment from the telecommunications industry in the development and deployment of WiMAX networks. Voice over IP (VoIP) over WiMAX will be one of the killer applications for rapid deployment of WiMAX networks. The legitimate desire for bundling voice and data will increase the portion of voice traffic in the WiMAX networks. Hence, VoIP, as the current technology for making voice calls through packet switch networks, will be a key application in WiMAX networks.

The scarcity of available bandwidth in wireless networks has called for efficient resource management. The IEEE 802.16 standard, which has been adopted by WiMAX, has defined different scheduling services and QoS mechanisms, but the details of traffic scheduling are intentionally left open for vendors' innovation to design the best scheduling method suitable for their networks. To design the optimum scheduling algorithm, it is crucial to understand the traffic features and service requirements of different applications consuming network resources. An imprecise model of the traffic leads to waste of bandwidth and lower efficiency. The main VoIP traffic model used in research literature is the ON-OFF model [28]. In this model, it is assumed that the source generates equal-size packets separated equally in time during the ON period and either does not generate any packet or generate smaller packets during the OFF period. Although natural voice might conform to the ON-OFF model, experimental traces of VoIP packets, incorporating the impact of the application codes, transport layer, and IP layer, do not exhibit the characteristics of an ON-OFF traffic[25, 8, 41].

The importance of efficient resource management has prompted a keen interest in the research community on scheduling VoIP traffic in WiMAX networks. Reference [38] proposes a scheduling method based on the ON-OFF model. Reference [65] also uses the ON-OFF model to perform a predictive scheduling of VoIP traffic in IEEE 802.16 systems. An analysis of the voice packet transmission in IEEE 802.16 is presented in [67]. They have studied the performance of conventional service scheduling methods on VoIP traffic in IEEE 802.16 systems. They also assumed that the packet generation process can be modeled by the ON-OFF model. A qualitative experimental study of VoIP traffic in a WiMAX testbed is reported in [46].

In this chapter, a new algorithm for scheduling the uplink VoIP packets generated at the end user of a WiMAX subscriber station [24] is proposed. The algorithm is inspired by the realistic VoIP traffic model introduced in [25]. The outline of the rest of the chapter is as follows: Section 4.3 provides an overview on service scheduling methods defined in the IEEE 802.16, and illustrates the proposed algorithm for scheduling VoIP traffic in WiMAX networks. Section 4.4 presents simulation results and performance comparison of conventional scheduling methods with that of the proposed scheduling algorithm. Concluding remarks are given in Section 4.5.

4.3 VoIP Traffic Scheduling in WiMAX

WiMAX networks are planned to support different types of traffic. While the network can be designed to work as the backhaul for broadband communications, it can be tailored to provide wireless access to mobile users. Supporting various types of traffic requires flexibility in design and functionality. Owing to this demand, there are many available options in the IEEE 802.16 standard which are supposed to be chosen by vendors based on their network requirements.

Traffic scheduling is one of the important issues that is intentionally left outside the scope of the IEEE 802.16 standard. It is up to vendors to make the best decision based on their network traffic. The IEEE 802.16 has defined different types of scheduling services for various types of traffic. This section elaborates on scheduling services suitable for VoIP traffic. VoIP traffic is real-time and delay sensitive, and it is required to allocate network resources to this traffic within a limited period of time. One of the important scheduling points in the IEEE 802.16 is the scheduling of the Up Link (UL) subframe by the base station (BS). To schedule the UL subframe, the BS receives the requests from subscriber stations (SSs), and after processing them, it creates the UL MAP message of the next UL subframe and distributes it to SSs. After receiving the UL MAP message, each SS will know the time and amount of bandwidth reserved for its very next UL subframe. This process requires bandwidth negotiations between the BS and each SS. Based on the type of traffic and policy of the network, there are different approaches to perform the bandwidth request and grant procedures which are explained next.

4.3.1 IEEE 802.16 Service Scheduling

There are three types of scheduling services defined by the IEEE 802.16 that are capable of supporting real-time traffic such as VoIP. They are the unsolicited grant service (UGS), real-time polling service (rtPS), and extended real-time polling service (ertPS). These scheduling services are briefly elaborated here.

- UGS: In this service, the BS periodically allocates a fixed amount of bandwidth resources to the subscriber station and the SS does not need to send bandwidth request.

- rtPS: In this service, the BS periodically polls the SS about its uplink bandwidth request and allocates bandwidth to it in the next uplink subframe.

- ertPS: It basically works similarly to UGS but the SS has the opportunity to request the BS to allocate different amount of bandwidth whenever the SS needs to change the transmission rate.

By studying these scheduling service methods, it is understood that UGS will have the best delay performance, but the allocated bandwidth is "wasted" when the user does not have enough traffic to transmit in a UL subframe. This fact will be observed in the simulations discussed in Sec. 4.4. As mentioned in Chapter 3, VoIP packets are usually small and the required bitrate is also low. Thus, the mentioned problem is most likely experienced by users generating only VoIP traffic. The rtPS scheduling method has better bandwidth efficiency as a user requests bandwidth (BW) based on its queue. However, the BW polling procedure requires some BW allocation by itself. Moreover, in this method, packets will encounter a deterministic delay proportional to the frame length. Therefore, the delay performance of the network will be degraded. The ertPS scheduling method has issues similar to UGS and rtPS.

By considering the issues encountered by these three scheduling services, it is understood that in order to reduce the delay for the UL packets at SS, the BS has to allocate BW to users more frequently and this can increase the bandwidth loss, and thus decrease the BW efficiency. Thus, there is a trade off between the delay and bandwidth loss in these scheduling methods. This issue will be investigated in Sec. 4.4.

4.3.2 Multi-Tap Scheduling

Providing users with a high quality voice connection, without wasting the invaluable and limited bandwidth, is a crucial task for service providers. Therefore, it is necessary for them to choose an optimum scheduler to enhance the performance of their networks. Thus, a scheduling method based on the traffic model discussed in Chapter 3 is proposed. It is referred to as *multi-tap* scheduling which uses the information of the traffic model to perform a more efficient scheduling. In this method, it is supposed that the SS has captured the characteristics of the VoIP traffic, and thus knows the parameters of the VoIP traffic model discussed in Chapter 3, i.e., the inter-packet time matrix, $\Delta_{1\times N}$, and the probability matrix, $P_{1\times N}$. Therefore, it knows the PMF of its VoIP traffic. It is also assumed that the SS knows the nominal size of the packets, S, as well.

By making these assumptions, the average inter-packet time, $\bar{\Delta}$, can be calculated as

$$\bar{\Delta} = \sum_{i=1}^{N} P_i.\Delta_i \tag{4.1}$$

Two parameter are defined next. R_{avg} is the average bitrate required by the SS to transmit its VoIP packets to the BS, and R_{max} is the maximum bitrate the BS can allocate to SS. R_{max} is determined by the service level agreement (SLA) as well as the status of the network. It is obvious that in order to keep the delay bounded, it is necessary to have $R_{avg} \leq R_{max}$. Otherwise, the voice call cannot be admitted into the network. R_{avg} is calculated as

$$R_{avg} = \frac{S}{\bar{\Delta}} \quad (bps) \tag{4.2}$$

Another parameter used in the scheduling algorithm is the *availability factor*, ρ. This parameter indicates the availability of bandwidth for the VoIP traffic. The larger the parameter, the higher bitrate BS can allocate to the VoIP traffic. ρ is calculated based on R_{avg} and R_{max} as follows.

$$\rho = 1 - \frac{R_{avg}}{R_{\max}} \tag{4.3}$$

In the scheduling method, the SS predicts the generation time of the next packet whenever it gets the opportunity to transmit to the BS. The SS uses the matrix Δ for this prediction. It supposed that the packet generation will happen only at time intervals equal to Δ_i, for $i = 1, 2, ..., N$, from the generation time of the previous packet. In the algorithm, it is assumed that the SS has already transmitted the parameters of the traffic model, i.e., Δ, P, and S, to the BS. Therefore, the BS knows the value of the i^{th} element in the inter-packet time matrix, i.e., Δ_i. It is worth that the transmission of these parameters imposes only a negligible overhead since it is done once during a call time of a few minutes. The scheduling algorithm consists of two parts: request and grant. The request part is run in the SS while the grant part is run in the BS, as will be explained next.

Request Algorithm As mentioned before, the SS has to tell the BS the index of the tap that it predicts the next packet will be ready for transmission. The SS may transmit any information to the BS only if the BS has allocated to the SS part of

```
$g$ = generation time of the last packet
$t$ = current time
$d \leftarrow (t - g)$
if $d \leq \Delta_N$
    $index \leftarrow arg \min_i (\Delta_i - d \geq 0)$
else
    $index \leftarrow (N + 1)$
end
send index to BS
```

Figure 4.1 Request Algorithm.

the UL subframe. Therefore, the SS predicts and piggybacks the index in the packet. When the SS is transmitting a packet to the BS, it calculates the time difference between the last packet generation time and the current time. This time difference is denoted as d. The SS finds the smallest Δ_i which is greater or equal to d. Then, it sets $index = i$ or sets $index = N+1$ if $d > \Delta_N$. The request algorithm is summarized in Fig. 4.1.

Note that this request algorithm is used for the prediction scenario only. The SS has to piggyback a separate BW request in the packet if it has traffic already queued up in its buffer. Owing to the time sensitivity, the BS allocates enough resources for this request as soon as possible.

Grant Algorithm The BS has to reserve appropriate time slots based on the received *index* parameter and the network constraints. As mentioned earlier, due to the trade off between the delay and bandwidth loss, the BS has to select an operating point which satisfies both the delay and bandwidth constraints. The purpose of the grant algorithm is to reserve enough BW for the SS to transmit its traffic to the

BS. In order to choose the time slot, the BS uses the parameters of the VoIP traffic model to determine the values of R_{avg} and ρ. By using the values of *index*, which is sent by the SS, and ρ, the BS calculates the next transmission time of the SS by using the grant algorithm summarized in Fig. 4.2. Intuitively, the BS considers the availability factor, ρ, and the requested transmission time received from the VoIP user via the *index* parameter. It assigns a transmission time at least Δ_{index} away from the previous packet transmission time. The extra delay in transmission might be imposed due to the lack of BW which is translated to the small availability factor. In this algorithm, a smaller availability factor results in longer delay.

As an example, the network parameters mentioned in Chapter 3 is considered. By using Equation. (3.4), it is understood that the CDF matrix is $C = [0.32, 0.6, 0.85, 0.95]$. It is also assumed that SS has sent a message and indicated that $index = 1$ and the BS has calculated $\rho = 0.5$ for a VoIP user. By using the grant algorithm, the BS finds that the smallest i which satisfies $\rho \geq 1 - c_i$ is $i = 2$. Therefore, it calculates $k = max(index, i) = 2$, and allocates an available time slot, which is at least $\Delta_2 = 17ms$ away from the previous packet transmission time, to the subscriber station.

It is worth noting that the SS might not have any traffic to send at the allocated time slot. In this case, the BW is 'wasted'. However, the SS uses this BW to send a new *index* calculated from the request algorithm to the BS. The BS uses the *index* to assign another BW in future UL subframes for the SS. The performance of this scheduling algorithm will be evaluated in the next section.

4.4 Simulation Results

In this section, the performance of the scheduling methods discussed in Sec. 4.3 is elaborated and the delay and BW performance of these methods are studied through some simulations. Some real VoIP traffic traces generated in [25] are employed

```
$index = last\ requested\ index$
$t = last\ allocation\ time$
$if\ \ index \leq N$
  $if\ \rho \leq 1 - C_N$
    $i \leftarrow N$
  $else$
    $i \leftarrow arg \min_i\ (\rho \geq 1 - c_i)$
  $end$
  $k = max(index, i)$
  $allocate\ BW\ equal\ to\ S\ at\ least\ \Delta_k\ from\ t$
$else$
  $allocate\ BW\ equal\ to\ S\ as\ soon\ as\ possible$
$end$
$t \leftarrow allocated\ time$
$put\ allocated\ time\ in\ UL\ MAP$
```

Figure 4.2 Grant Algorithm.

and it is assumed that the subscriber stations only generate VoIP traffic during the simulations. This is a rational assumption since many mobile WiMAX enabled handsets will use VoIP to make voice calls. In the simulations, the physical and MAC layers of a WiMAX network with parameters described in Table 4.1 are simulated. Since it is desired to compare the performance of the uplink scheduling methods in the MAC layer of the WiMAX, an ideal physical layer without any losses is assumed. Therefore, the presented results are considered the best achievable ones with respect to the MAC layer.

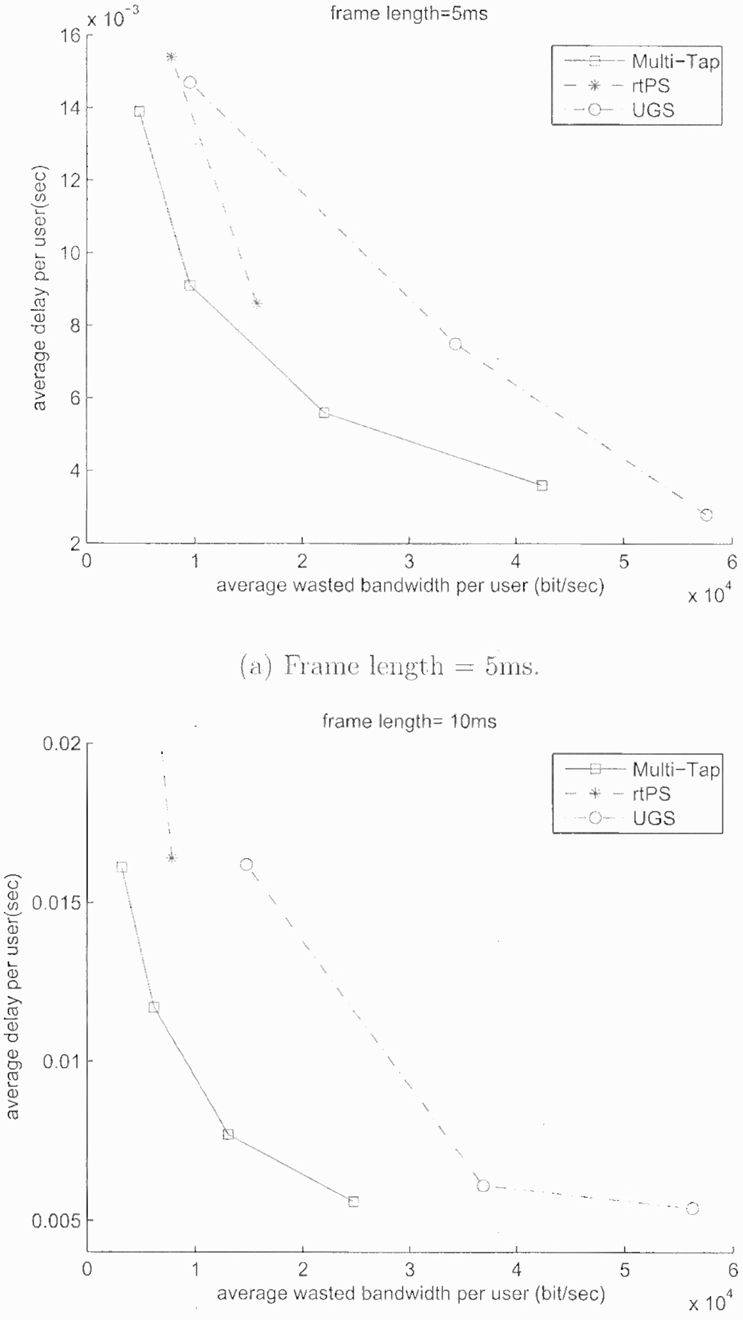
In the simulations, two different values for the MAC layer frame length are chosen, and the average delay that each packet experience before transmission in the

Table 4.1 Simulation Parameters

WiMAX Parameters	Value
MAC frame length	$5ms$, $10ms$
Bits per timeslot	192
Duplexing	TDD
Channel bandwidth	5 MHz
Uplink modulation	16 QAM
Uplink control slots	4 timeslots
Uplink data symbols $5ms$	21
Uplink data symbols $10ms$	45

UL will be observed. the BW loss is also measured. Simulations are conducted based on the UGS, rtPS and multi-tap methods. For the UGS, the scenarios that either one, two, or three time-slots are allocated to the VoIP user in each frame are considered. For the rtPS, the scenarios that the polling is done in either every one or other frame are shown. It is noted that decreasing the resource allocation or polling frequency will result in higher delay. Nevertheless, it might result in saving some BW. For the multi-tap model, different values of the availability factor, i.e., ρ are considered. It is worth noting that increasing the number of users in the network will decrease the maximum rate that the BS can allocate to each user and it results in a smaller availability factor.

The achievable operating points are sketched in the simulation results. The operating points whose average delay is less than 20ms are shown. As depicted in Figs. 4.3(a) and 4.3, reducing the average delay requires more BW loss in both UGS and multi-tap scheduling methods. In rtPS, both BW loss and average delays are constant and proportional to the frame length. As discussed in Sec. 4.3, in the rtPS



(a) Frame length = 5ms.

(b) Frame length = 10ms.

Figure 4.3 Delay vs. BW waste.

method, the BS periodically asks the SS about its available traffic in its queue and allocates BW to the SS if it has already had some traffic to send. The only BW loss in the rtPS algorithm is due to the polling overhead. The BS has to allocate 6 bytes to the SS in the UL subframe whenever it wants to poll the SS about its queue size. In the simulations, it is supposed that the BS polls the SS once in either every one or other frame. It is also worth noting that the former will achieve the minimum reachable average delay.

The performance of scheduling methods on traffic of a user with R_{avg}=29Kbps is observed. As shown in Figs. 4.3(a) and 4.3, rtPS cannot reach an average delay less than almost 1.5 times the frame length. It is also shown that it is possible to decrease the average delay in UGS, by allocating more BW to the user. However, the waste of BW will increase. Finally, it is shown that the multi-tap scheduling algorithm reaches the best performance with respect to the delay and BW loss. The average delay of the packets using the multi-tap method decreases as the larger value for ρ is used by the BS. However, increasing ρ results in more waste of BW. By using the multi-tap scheduling algorithm, the service providers can provide the VoIP SSs with less delay while they save more BW as compared to other scheduling methods. This results in higher capacity and efficiency in networks.

4.5 Summary

User preference for bundled services over the same network has resulted in a high demand for ubiquitous network access. This demand will prompt WiMAX system providers, as the promising technology providers for broadband wireless access, to specially provisioning their customers with reliable and qualified voice connections via VoIP applications. In order to build an efficient WiMAX network, service providers have to acquire real insight about the behavior of the VoIP traffic in WiMAX networks. This chapter incorporates a realistic traffic model that has been derived based on real VoIP traffic traces. Available QoS classes for VoIP in WiMAX networks have been described, and based on these classes some MAC layer uplink traffic scheduling methods are examined. This chapter has also proposed a heuristic scheduling algorithm for VoIP traffic based on the traffic model, and elaborated on the trade off between bandwidth efficiency and delay in each of the scheduling methods. It is demonstrated that using the proposed scheduling method can lead to a better delay and bandwidth efficiency.

CHAPTER 5

MPEG VIDEO TRAFFIC MODEL

5.1 Objective

The availability of various video applications requiring different services, from low bitrate video conferencing to high bitrate movie streaming, has called for development of different video codecs. Understanding and modeling the main characteristics of the video traffic is needed to create efficient methods for transporting video from different applications with different codecs over various networks. In this chapter, the main features of video traffic are studied and a traffic model for the video traffic in explained.

5.2 Introduction

The transmission of digital video over broadband communication networks is an important service. However, the existence of different video generating applications along with the availability of numerous networking technologies with different service features have made this an extremely challenging problem. Providing the required quality of service to the end users is a difficult problem requiring in depth understanding of the video traffic characteristics.

The variety of video applications and networking technologies has called for implementation of a variety of video coder and decoder (codec) standards. These standards may be deployed in a broad set of applications ranging from low bitrate video conferencing to high bitrate movie streaming. However, the efficiency of these codecs is different for different applications. Valid traffic models accounting for the key video characteristics are required to investigate how best to transport video from different applications and codecs over different networks.

5.3 Video Streaming Traffic Model

Moving Picture Expert Group (MPEG) has a series of advanced video compressing standards, of which MPEG-2 and MPEG-4 are the most pervasive ones, and the latter is the latest and the most advanced one. All these standards rely on removing the redundant information of each frame by predicting the changes between subsequent frames. The idea of prediction is based on the fact that consecutive scenes have few differences and the information in their pictures is highly correlated. By coding the small differences between the scenes, much less data needs to be transported and thus achieving data compression.

MPEG-4 encodes the input video into a sequence of frames called Group of Picture (GoP). The number of frames in each GoP is typically constant. It is possible that an MPEG frame is fragmented into multiple IP packets when transmitted over an IP network. The MPEG-4 encoder divides each scene of the video into a number of consecutive GoPs. The number of GoPs generating a scene is a function of the scene complexity and compression ratio. There has been extensive research work on modeling the MPEG video traffic. Reference [35] has separated the video traffic into I, P, and B frames. The authors have modeled each type of frames separately and have also provided a model for the combined traffic. Reference [40] has also suggested a traffic model by separating the MPEG frame into three different types. The WiMAX Forum has adopted the traffic model proposed in these papers, and has given the parameters determined from empirical video traces for different video applications such as video conferencing and video streaming [62]. A rather comprehensive work on modeling MPEG video traffic is given in [4]. A unified traffic model for MPEG-4 and H.264 is introduced in [14]

As mentioned above, the MPEG coded videos are composed of three different frame types, i.e., I, P and B. I (Intra coded) frames are single still images used as the reference frame in each GoP. I frames are used for synchronization of all frames in a

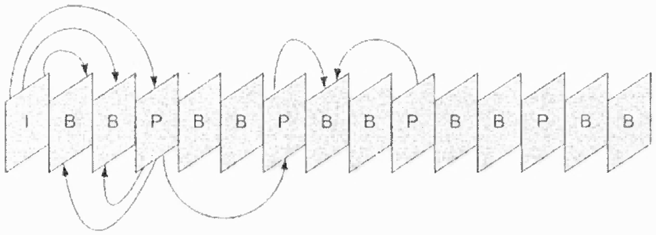


Figure 5.1 GoP Pattern.

GoP. If a GoP is lost or corrupted, the next GoP will be built based on its I frame which is coded without using any other frames. P (Predicted) frames are built by predicting the changes from the closest match in the preceding I or P frames of their GoP. However, B (Bi-predictive) frames use previous I or P frame and the next P frame to predict the changes in the picture. Thus, the B frames are used to predict both the backward and forward changes in the motion. Based on these definitions, it is understood that these frames are interrelated, and some P and B frames are derived from each I frame in a GoP. Similarly, some B frames are also derived from each P frame. Fig. 5.1 shows the schematic of a general GoP that begins with an I frame and interdependencies among frames of the GoP. Therefore, loss of I or P frames will affect some other frames in their corresponding GoP, and this will degrade the perceived quality. Chapter 7 will elaborate on this problem.

Although not required by the standard, MPEG encoders usually use a fixed pattern of frames in GoPs. The GoP pattern indicates the number of frames in each GoP, and their permutation order. In a regular GoP pattern, a GoP begins with an I frame and the number of B frames between I and P frames or between two P frames is constant. Such regular GoPs can be defined by two parameters: the I-to-I distance 'N', and the I-to-P or P-to-P distance 'M'. A schematic illustrating the decomposition of video scenes into GoPs and formation of a GoP with $N = 15$ and $M = 3$ is depicted in Fig. 5.2.

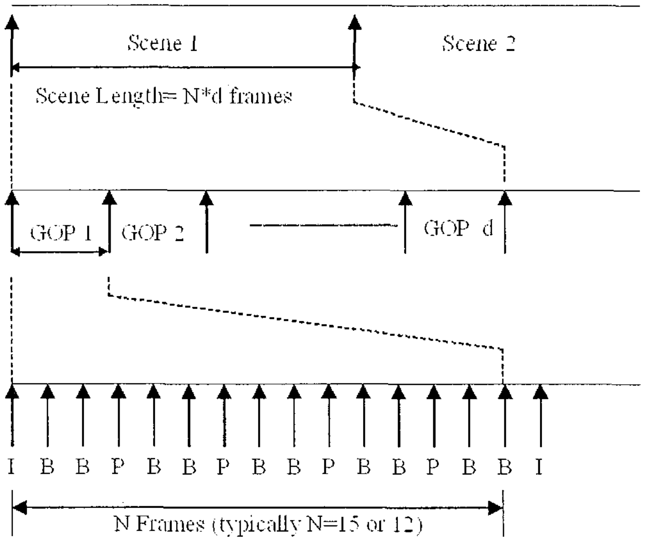

Scene 1
Scene 2
Scene Length= N*d frames
GOP 1
GOP 2
GOP d
I B B P B B P B B P B B P B B I
N Frames (typically N=15 or 12)

Figure 5.2 Source Model.

The need for simulating the network performance has introduced different traffic models for different video applications such as video conferencing and video streaming. In this dissertation, a traffic model for video streaming adopted by the WiMAX Forum [62] is used. In this model, I, P and B frames are modeled separately, and a fixed pattern, similar to what is shown in Fig. 5.2, is used in building GoPs. As shown in Fig. 5.2, the number of GoPs in each scene is denoted by d, and it is modeled as a geometrically distributed random variable. The MPEG-4 traffic model will be explained next.

The I frames are modeled as Variable Bit Rate (VBR) traffic. Based on the real MPEG traffic traces, the I frames have exhibited different behavior at different time scales. At the shorter time scales of a few seconds, the bitrate varies a little around a mean level. However, the mean level varies tremendously at larger time scales. The change of the mean levels in the large time scales is called the *scene* variation [37]. A scene is a short part of the movie that does not contain sudden changes in the view while it can possibly include some zooming or object movement. In the adopted traffic model, the concept of scene has been incorporated in the model, which results in more accurate performance prediction.

As explained above, the variations of the size of I frames have two scales: 1) the small variations within a scene duration; 2) the large variations among different scenes. Thus, the model considers two independent components for defining the size of the n^{th} I frame of the video stream, $X_I(n)$, located at the k^{th} scene.

$$X_I(n) = \bar{X}_I(k) + \Delta_I(n) \tag{5.1}$$

$\bar{X}_I(k)$ is the mean activity of scene k and represents the large variations, and thus it may vary greatly from scene to scene. $\bar{X}_I(k)$ is constant for all I frames in scene k while it will be different for other scenes. $\bar{X}_I(k)$ is modeled by a log-normally distributed random variable [35]. $\Delta_I(n)$ represents the small variations of the I frames around the mean level of each scene. The $\Delta_I(n)$ is modeled by an order two autoregressive process, AR(2).

$$\Delta_I(n) = a_1\Delta_I(n-1) + a_2\Delta_I(n-2) + \varepsilon(n) \tag{5.2}$$

The a_1 and a_2 are assumed to be constant for each video stream and $\varepsilon(n)$ is a normal random variable with zero mean and constant variance for each stream [35]. The parameters defining the random variables depend on the content of the video; however, a constant set of parameters will be used in the simulations which is similar to what is adopted by the WiMAX Forum.

The sizes of P and B frames are modeled by log-normal distributions with parameters (μ_P, σ_P) and (μ_B, σ_B). The correlation between P frames (and similarly B frames) is negligible as compared to that of I frames, and thus the model considers them as independent random variables [35]. The MPEG model parameters of different proposed video applications are presented in Table 5.1. More details about this MPEG model is given in Reference [62].

By considering the parameters displayed in Table 5.1, it is understood that if 30 frames are generated per second, $N = 15$, and $M = 3$, two GoPs would be generated per second. Thus, the model generates two I frames per second, eight P frames per second, and twenty B frames per second. Hence, the average bitrate for each frame type is as follows: $\bar{R}_I = 273Kbps$, $\bar{R}_P = 588Kbps$, and $\bar{R}_B = 1094Kbps$. Thus, the overall average bitrate for each video stream is $\bar{R}_{tot} = 1955Kbps$. It is observed that the average bitrate of B frames is higher than those of I and P frames. Although the average size of a B frame is less than that of other types, its bitrate is higher since there are more B frames in each GoP, as discussed earlier.

5.4 Summary

This chapter introduced the general characteristics of the MPEG video traffic. The structure of the MPEG traffic was illustrated, and the interdependencies among video frames were discussed. The importance of exploring genuine traffic models for video applications was explained, and different approaches for modeling the video traffic were studied. A comprehensive traffic model proposed by the WiMAX forum was described in detail, and its proposed values for different parameters was reported. Throughout the rest of this dissertation, the general characteristics of MPEG video will be applied in designing more efficient video streaming strategies, and different video traces will be generated according to the WiMAX forum traffic model to study and verify the performance of the video streaming strategies via rigorous simulations.

Table 5.1 MPEG-4 Model Parameters [62]

Model Parameter	Video Conferencing		Movie Streaming		TV Broadcasting	
Display size	176x144	320x240	176x144	320x240	176x144	320x240
I frame size (bytes)	Log-normal $(\mu = 6210, \sigma = 1798)$	Log-normal $(\mu = 18793, \sigma = 5441)$	Log-normal $(\mu = 5640, \sigma = 2632)$	Log-normal $(\mu = 17068, \sigma = 7965)$	Log-normal $(\mu = 19504, \sigma = 2213)$	Log-normal $(\mu = 59025, \sigma = 6697)$
P frame size (bytes)	Log-normal $(\mu = 2826, \sigma = 1131)$	Log-normal $(\mu = 8552, \sigma = 3422)$	Log-normal $(\mu = 3037, \sigma = 2315)$	Log-normal $(\mu = 9190, \sigma = 7005)$	Log-normal $(\mu = 9891, \sigma = 2310)$	Log-normal $(\mu = 29933, \sigma = 6990)$
B frame size (bytes)	Log-normal $(\mu = 1998, \sigma = 716)$	Log-normal $(\mu = 6048, \sigma = 2168)$	Log-normal $(\mu = 2260, \sigma = 1759)$	Log-normal $(\mu = 6839, \sigma = 5323)$	Log-normal $(\mu = 6496, \sigma = 1896)$	Log-normal $(\mu = 19658, \sigma = 5737)$
Mean BW for compressed stream (Mbps)	0.54	1.65	0.58	1.74	1.1	3.35

CHAPTER 6

APPLICATION-ORIENTED NETWORKING FOR VIDEO STREAMING APPLICATIONS

6.1 Objective

The growth of multimedia traffic that requires special quality of service from the network is imposing new constraints on network designers who should wisely allocate the limited resources to users based on their required quality of service. An efficient resource management and network design, with the goal of providing customers with satisfactory perceived quality, depends upon gaining accurate information about the service requirements of multimedia applications. This chapter discusses the main challenges of streaming video content over wireless networks, and introduces some application-oriented solutions which can improve the quality of experience at the end users.

6.2 Introduction

Extensive deployment of multimedia services such as Video on Demand (VoD) or IPTV, facilitates the boom of various video streaming applications. Furthermore, rapid expansion of broadband wireless networks is providing mobile users with broadband access through which they can run various applications. It is expected that multimedia applications and, in particular, video streaming applications will grow along with the population of many online video servers. Therefore, an excessive number of people will use their wireless devices to access numerous video streaming contents in the Internet.

Video streaming in wireless networks is challenging owing to the stringent service requirements of video traffic and impairments of wireless channels. However, the increasing demand for broadband wireless access has called for the design and implementation of different wireless technologies such as WiMAX and LTE. These technologies are supposed to provide users with network connectivity to run different applications with various Quality of Service (QoS) requirements. Hence, network designers of broadband wireless technologies should provision adequate resources to support QoS requirements of video streaming applications with high bitrate and low latency. Moreover, supporting QoS requires proper resource allocation which makes video streaming in wireless networks a complicated problem.

One of the issues for streaming real-time video over wireless networks is sustaining the satisfactory perceived video quality even when congestion happens or the wireless channel become less reliable. This aspect of video streaming, commonly referred to as *Quality of Experience (QoE)*, is more crucial than providing required QoS. Different techniques have been proposed to support required QoS in the networks with limited resources; nevertheless, sustaining the satisfactory perceived quality is more complicated than supporting required QoS. This is attributed to the fact that users' perceived quality depends on streaming quality at the application layer; however, most of the QoS strategies run at the lower layers of the protocol stack such as network or MAC layers. Consequently, service providers have to allocate resources according to application requirements, and cross-layer approaches seem to be a viable solution.

So far, extensive efforts have been focused on providing adequate resources to multimedia applications through *Access-Oriented* solutions. In access-oriented solutions, the resource allocator does not generally consider the traffic content and its impact on QoE. Supporting QoS via different approaches such as Differentiated Services (DiffServ) or Integrated Services (IntServ) may guarantee some networking parameters like as delay, jitter, or loss, but they cannot optimize QoE with respect to available resources. Access-oriented solutions treat data packets of each application similarly; however, different packets may have different impacts on the QoE at the end user, and thus retain different importance [22]. In order to optimize the QoE, it is wise to serve various packets differently according to their impact on QoE.

To enable networking elements to abstract the impact of multimedia data packets on QoE at the end user, a cross-layer approach, which includes the application layer information in the IP packets, is a feasible solution. This approach is referred to as *Application-Oriented* solution.

This chapter discusses some of the challenges service providers are confronting in supporting video streaming applications in wireless networks with acceptable QoE. Different examples of incorporating the application-oriented approach are explained, and it will be illustrated that the QoE at the end user will be improved by applying this approach.

The main characteristics of the video traffic explained in Chapter 5 are helpful to better understand application-oriented solutions for video streaming applications.

6.3 MPEG Streaming Challenges

Video streaming over IP networks is explosively growing as an increasing number of users retrieve various available video resources, and a variety of video streaming applications with various service requirements are operating in the networks. Although the main purpose of video streaming is providing the end users with quality pictures, a variety of demands has called for numerous streaming solutions. It is possible to categorize streaming solutions with respect to *architecture, delay sensitivity*, and *bitrate.*

Architecture: video sources are accessible via different communication architectures. While some of the streaming contents, such as Video on Demand (VoD), are transmitted from a video server to a client in the unicast fashion, some others such as IPTV are transmitted in the multicast fashion. In P2P streaming applications, on the other hand, multiple servers may cooperate to transmit video contents to a plurality of clients.

Delay sensitivity: video streaming applications have different delay requirements while real-time applications such as IPTV or conferencing are extremely sensitive to transmission delay, and some non-real time applications such as VoD exhibit less delay sensitivity.

Bitrate: owing to the content of video and streaming application, video traffic may present a wide range of bitrates from IPTV applications which require relatively high bitrate for transmission to conferencing applications with potentially low bitrates.

Ever increasing deployment of streaming applications with different characteristics requires discreet provisioning by service providers. Challenges of video streaming are more complicated in wireless networks where the resources are scarce and the demand is growing rapidly. Some of the major challenges that service providers are confronting in wireless networks are explained next.

6.3.1 Distribution Delay

Distribution delay represents the time packets traverse the network to reach the client. It includes network propagation, switching, and queuing delay. Wireless networks may experience longer queuing delay at the access points because access points have to share limited resources among all users. Queuing delay may become unacceptably large when congestion occurs in the network or wireless channel becomes less reliable. In cellular networks, base stations are responsible for allocating adequate resources to streaming traffic flows such that video packets become available at the end user before the decoding deadline for uninterrupted play of the video. Packets that cannot meet the delivery deadline should be dropped by the base station as their transmission will only waste some of the network resources.

6.3.2 Variable Bitrate Traffic

In order to provide the streaming traffic with adequate resources, network elements have to gain knowledge about the service requirements of traffic flows. Service requirements are determined in the Service Level Agreement (SLA) which is negotiated during the Call Admission Control (CAC) between the end user and service provider when the user requests for service initiation. In wireless networks, QoS is negotiated between the base station and end user. To fulfill the QoS requirement, the base station has to reserve some of the resources for each video streaming traffic flow. The SLA determines the amount of services the BS has to guarantee such as minimum bitrate or maximum delay. It is worth noting that during congestion periods the BS is committed to provide the user with only guaranteed services. In this case, the BS will allocate resources equal to the minimum reserved bitrate for the video traffic flow. It is also noted that there is a trade-off between network utility and QoS because reserving higher bitrate for each traffic flow will decrease the network utility as the BS can admit less number of users in the network while reserving lower bitrates will result in higher delay, and thus lower quality.

As discussed above, the video traffic exhibits variable bitrate. However, base stations should reserve a minimum bitrate which will be allocated to the user during the congestion periods. Much research effort has been carried out to determine the minimum bitrate which guarantees the queuing delay, and some approaches such as equivalent capacity have been proposed [64]. However, choosing the optimum reserved bitrate which can guarantee the minimum perceived video quality still remains an open problem.

6.3.3 Packet Loss

A packet is lost if it does not arrive at the decoder of the end user within a certain period of time. Packet loss may be caused by:

Delay: As discussed earlier, packets should be available to the decoder of the end user at the time of decoding. Otherwise, the packet is lost. Different reasons may delay the arrival of the packet. In wireless networks, packets may be delayed at the queue of the base station due to lack of resources for transmission or unreliable wireless channels. The base station will drop the packet if it exceeds the predefined maximum tolerable delay. Hence, the BS should keep track of queuing delay of each packet. Maximum delay tolerance is defined at the SLA within the CAC period.

Congestion: During congestion periods, queues of network elements build up and they may overflow. An arriving packet may be dropped at the ingress point of a network element if its queue is full. In wireless networks, congestion happens if the amount of incoming traffic exceeds outgoing available resources, and the BS will drop incoming packets if it does not have enough queueing space. Therefore, queue overflow or long queuing delay may result in packet loss which degrades the video quality at the end user [39].

6.4 Application-Oriented Solutions

The main characteristics of the MPEG video traffic and challenges mentioned above have inspired us to propose some application-oriented solutions to improve the quality of perceived video at the end users. In application-oriented solutions, network elements and especially base stations of wireless networks gain knowledge about the content of packets before scheduling those packets for transmission. Since processing contents of all packets is a costly procedure, it is preferred to provide the base stations with application layer information via slightest processing effort. In the following, some of the application-oriented solutions in combating streaming challenges are explained.

6.4.1 Cross-layer Information Exchange

As mentioned in Chapter 5, MPEG video streaming traffic is composed of I, P and B frames. Although B frames generate the most amount of traffic, they have the least impact on the video quality. It is wise then for the BS to distinguish the frame type of each packet and serve them according to their frame type. There are different approaches for informing the BS about the frame type of each packet. Direct access to the data part of each packet is a trivial solution but it is time consuming and increases the computational complexity at the BS. Another solution is establishing a control channel between the base station and video server in addition to the data channel. This solution increases the traffic in the network, and in particular the BS. It would be less complicated and more efficient if the information becomes readily available to the BS among its routine procedures. An efficient solution will thus be including the frame type of each packet in its header part. Since base stations have to process the IP header of each incoming packet, including this information in the header part will not impose extra load to the network or the base station. For that reason, it is possible to include the frame type of each packet in the Type of Service (ToS) part of each packet. The Explicit Congestion Notification (ECN) section provides two bits which can determine the three types of each MPEG packet, i.e., I frame, P frame, or B frame. In addition to the IP header section, it is possible to include frame type information in consecutive header sections belonging to upper layers such as UDP and RTP. For example, the source port section of the UDP header can be used to indicate frame type information. It is worth noting that IP, UDP, and RTP headers have known size, and therefore the BS can derive required information from specified locations of headers. Fig. 6.1 shows the schematic of MPEG packets. It is understood that these approaches do not add any overhead to the packets, and are thus readily implementable.

IPv4 Header (20)	UDP Header (8)	RTP Header (12)	MPEG Data

Figure 6.1 MPEG Packet.

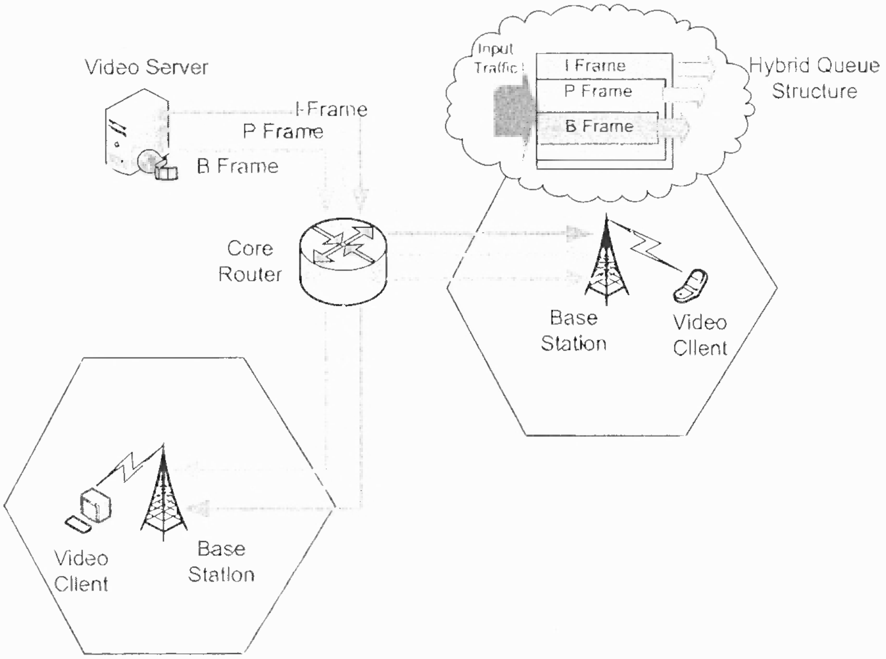


Figure 6.2 Network Architecture.

6.4.2 Multi-level Service Classification

In this solution, the video server incorporates one of the cross-layer information exchange solutions to inform the BS about the frame type of each packet, and MPEG frames are mapped into three different traffic flows with different QoS parameters [26]. The QoS of each traffic flow is determined by the user's video application or the video server, and is sent to the BS during the CAC process. A schematic explaining the concept of *Multi-level Service Classification* is shown in Fig. 6.2. It is noted that the proposed traffic classification does not require any changes at the BS. However, the video servers have to initiate three different traffic flows for each frame type upon receiving a request for video transmission. This solution will be elaborated in Chapter 7.

In order to sustain the video quality, it is crucial to send as many frames as possible to the end user. Furthermore, because of the hierarchical structure of GoPs and the inter-dependency of frames, it is understood that the network should protect the I frames with the highest priority, and then the P frames from loss to prohibit the *propagation of errors* effect.

When a traffic flow of a streaming application is admitted into the wireless network, the BS has to reserve a minimum bitrate. In the multi-level service classification method, the BS guarantees different bitrates for each traffic flow and serves each flow according to its frame type. Owing to the importance of each frame type, the BS has to protect the more important frames against dropping by reserving relatively higher bitrate to their traffic flow. It is worth noting that suitable bitrate values should be reserved for each of the traffic flows to provide the user with at least the minimum acceptable perceived quality.

6.4.3 Intelligent Queue Management

This solution deploys the cross-layer information exchange and multi-level service classification schemes. Thus, the BS can indicate the frame type of each packet in its header. Furthermore, it is supposed that the BS can derive the GoP number of each packet and its frame number as well. As mentioned earlier, the video server can exchange this information by the BS via different approaches such as indicating this information in the headers of upper layer protocols.

When a traffic flow is admitted into the wireless network, it will have its corresponding downlink (DL) queue at the BS. Thus, the DL traffic from the video server to the user is enqueued at that DL queue. By incorporating the multi-level service classification scheme in this solution, it is assumed that the MPEG traffic is divided into three different traffic flows based on their frame type. Therefore, video streaming packets will be enqueued at three DL queues corresponding to their frame types. However, in the Intelligent queue management solution, it is considered that hybrid priority queuing is deployed at the BS. To better understand hybrid priority queuing, a virtual queuing structure depicted in Fig. 6.2 is assumed. At the arrival of each packet, the BS determines its frame type, GoP, and frame number by retrieving this information from the header of the packet. If the queue has enough space, the BS will put the packet in its corresponding subqueue according to its GoP and frame number, but if the maximum queue size is reached, the BS checks the frame type of the incoming packet. If there is any lower priority packet in the queue, the BS will drop as many lower priority frames as needed from the queue to enqueue the arriving packet, and protect it against dropping. Otherwise, the BS will drop the incoming packet if there is not enough lower priority packets available in the queue suitable for dropping. The BS will also drop all the dependent packets that will arrive subsequently in the future and require the dropped packet for decoding, as the end user will not be able to decode them without receiving the dropped frame. Frames in each GoP are prioritized as follows:

1. I frame.

2. P frames in descending order. In each GoP, the higher order P frames have less priority because less number of frames is derived from them.

3. B frames.

As explained earlier, the BS should drop the packets awaiting in the queue for more than the maximum tolerable delay which is equal to the caching time at the end user. With that said, the BS regularly monitors the delay time of the first packet of each sub-queue, and if the maximum delay is reached, it will drop that packet from the queue. It is noted that the BS has to also drop all the lower priority packets that are depending on an already dropped packet for decoding. This solution will be explained in Chapter 8.

6.4.4 Rate Allocation

At the call admission control process, it is necessary to negotiate service requirements of each traffic flow as the BS has to decide whether it can provide the user with the requested service requirements. Each user has to ask for the QoS needed for running its applications. Although requesting higher service rates may result in better QoE, the BS may not grant the request if it cannot provide the user with requested services. In general, the BS prefers to guarantee the least amount of bandwidth (BW) such that it can admit more number of users in the network, and thus increasing the network utility. Finding the bitrate which guarantees the minimum acceptable QoE is a complicated problem which is addressed here and in Chapter 9.

Although it is possible for the video server to estimate the optimum bitrates for offline and non-real time streaming applications such as VoD, it is more difficult for real-time applications in which the server cannot fully predict the required bitrate of a VBR traffic flow. The main challenge with respect to optimizing the QoE is modeling user satisfaction and measuring QoE. Many research works [36] have been proposed to address rate allocation for VBR traffic to limit the queue size or queuing delay, and different groups such as Video Quality Experts Group (VQEG) have proposed some criteria to assess the video quality (http://www.its.bldrdoc.gov/vqeg/).

Based on the MPEG traffic model discussed earlier and the multi-level service allocation scheme, it is of interest to find suitable bitrates for each of the traffic flows which carry different types of frames. Intuitively, a higher bitrate for the flow corresponding to I frames is requested to minimize the dropping probability of I frames. Therefore, a bitrate which is higher than the average bitrate of I frames is conservatively requested. It is worth noting that if the I frame traffic flow cannot consume all the allocated bandwidth, the BS will spare the remaining part to other traffic flows according to the intelligent queuing solution. To sustain a minimum video quality, minimum reserved bitrates for the P and B frames are requested as well. However, the requested bitrates are less than the average bitrates of these flows, and the bitrate for P frames is higher than that of the B frames. Thus, B frames are more likely to be dropped.

Results of many ongoing research works in modeling the QoE and especially exploring thresholds of acceptable video packet loss for each frame type can be employed to enhance the performance of proposed application-oriented solutions. However, it is shown that the QoE at the end users can be improved even by heuristic rate allocation method discussed above. An analytical framework for choosing optimum bitrate will be introduced in Chapter 9.

6.5 Summary

In this chapter, some challenges of streaming video traffic in wireless networks are presented. By incorporating the characteristics of the MPEG traffic, some application-oriented solutions to improve the quality of experience of video streaming applications at the end users are introduced. In a cross-layer fashion, it is elaborated that by informing the network elements about the impact of the video packets on the perceived quality, network elements can serve streaming packets according to their importance. The proposed application-oriented solutions are simple and easily applicable, and are thus likely to be deployed in future networks to improve quality of experience.

CHAPTER 7

MPEG VIDEO STREAMING OVER WIMAX NETWORKS

7.1 Objective

Extensive efforts have been focused on deploying broadband wireless networks. Providing mobile users with high speed network connectivity will let them run various multimedia applications on their wireless devices. Satisfying users with different quality of service requirements while optimizing resource allocation is a challenging problem. This chapter discusses the challenges and possible solutions for transmitting MPEG video streams over WiMAX networks. It will use the MPEG traffic model suggested by the WiMAX Forum and explained in Chapter 5. A cross-layer solution for enhancing the performance of WiMAX networks with respect to MPEG video streaming applications is explained. The proposed solution uses the characteristics of MPEG traffic to give priority to the more important frames and protect them against dropping. Besides, it is simple and compatible with the IEEE 802.16 standards, and thus readily deployable. It is shown that the proposed solutions will improve the video quality over WiMAX networks.

7.2 Introduction

The excessive demand for ubiquitous broadband wireless access has attracted tremendous investment from the telecommunications industry in the development and deployment of WiMAX networks. The WiMAX technology is promising to provide broadband wireless access to mobile users in the near future. It is expected that video streaming will be a very attractive application for the rapid deployment of WiMAX networks. The stringent QoS requirements including high bitrate and low latency are some of the challenges the service providers and network designers are confronting. Furthermore, the popularity of many online video servers such as YouTube, will encourage an increasing number of users to watch video clips on their mobile devices.

The scarcity of available bandwidth in wireless networks has called for efficient resource management. WiMAX networks are based on the IEEE 802.16 standards which have defined different QoS classes to support a broad range of applications with varying service requirements. The IEEE 802.16 standards provide true QoS classes for different types of applications. As a result, in WiMAX networks, each traffic flow is mapped into an appropriate service class based on its service requirements and the user's Service Level Agreements (SLA). Selecting appropriate service classes with proper parameters to support the required QoS while not wasting the scarce resources is the key challenge that is addressed in this chapter. The traffic characteristics of video streaming applications is studied, and it will be shown that an application driven, traffic aware service classification will provide the WiMAX Subscriber Stations (SSs) with better video quality.

The importance of efficient resource management has prompted a keen interest in the research community on supporting video streaming applications in wireless networks. Reference [68] has reviewed the challenges of video streaming in wireless networks. It has also proposed a network adaptive rate control and cross-layer design for enhancing the overall received video quality. Many other research works have also considered feedback based video rate control [5, 31, 58, 27]. In most of the rate adaptive methods, the server receives some information such as the available bandwidth, loss rate, buffer size at the receiver, or the end-to-end delay to adapt to the optimum video coding rate. One of the main drawbacks for the rate adaptive methods is caused by the channel variations in the wireless networks. Owing to the fast variation in the wireless physical channels, the adaptation methods may not be able to track the fast changes in radio channel conditions and adapt to the optimum rate accordingly. Furthermore, selecting the appropriate rate increases the computational complexity at the video server which can result in overloading the video streaming servers. Moreover, sending feedback is not a feasible option in some multicasting applications such as IPTV or MobileTV. In such cases, a video server transmits the same content to multiple receivers with different physical channels. The heterogeneity of receivers in these applications makes it very complicated for the server to attain the flexibility and sustain the efficiency.

In order to reduce the complexity at the server side and support various types of clients, Scalable Video Coding (SVC) has been introduced in Reference [47]. The goal of this method is to encode high quality video streams into some groups of bit streams including one base sublayer and multiple enhancement sublayers. All clients register to receive the base sublayer. The addition of enhancement sublayers improves the video quality. Thus, clients select the number of enhancement sublayers to receive based on their network connection and the availability of resources [63].

The challenging and crucial problem of video streaming over WiMAX networks has attracted many researchers. Reference [30] presents an adaptive SVC approach for streaming video on demand to the subscriber stations. Different WiMAX network architectures and cross-layer solutions for supporting the broadcasting and multicasting applications have been introduced [52, 61]. They take advantage of the broadcasting capabilities of WiMAX wireless medium and leverage this feature to deliver video multicasting applications such as IPTV to WiMAX customers. A channel based, rate adaptive solution for video streaming in WiMAX network has been introduced in Reference [34]. Reference [51] illustrates an active frame dropping approach for streaming real-time video over IEEE 802.16 networks. In this active dropping approach, the base station drops a frame if it does not have enough confidence about successfully delivering a video frame within the application delay limit. The concept of active dropping at the video server for video streaming has been employed in different research works. A frame discarding solution based on the packet life-time is introduced in Reference [23]. In this method, frames that cannot meet the deadline are dropped by the video server or the intermediate routers. Reference [29] explains a priority based frame dropping algorithm. In response to a temporary bandwidth reduction, the video server selectively drops least effective frames. In this work, the authors have considered an MPEG video streaming system in which the video server determines the priority of each frame based on the frame type.

This chapter proposes a cross-layer design to enhance the quality of MPEG video streaming for the end users in WiMAX networks. The proposed solution uses the characteristics of MPEG traffic to give priority to the more important frames and protect them against dropping. Unlike the approach proposed in [29], the complexity of video servers is not increased since the frames are dropped at the BS. Moreover, the complexity of the BS is not increased as well. In fact, the proposed scheme is capable of being used by any WiMAX certified BS. In this method, real-time feedback is not sent to the video server. Thus, a video server will be able to support multiple clients simultaneously, and it makes this solution flexible for multicasting applications as well.

The outline of the rest of the chapter is as follows. Section 7.3 provides an overview of the quality of service support in WiMAX networks. Section 7.4 explains a proposed solution for video streaming in WiMAX. The simulation results are presented in Section 7.5. Concluding remarks are given in Section 7.6.

7.3 Quality of Service in WiMAX

WiMAX networks are planned to support traffic from many different applications. While the WiMAX technology can be designed to support the backhaul connectivity for broadband communications, it can also be tailored to provide wireless access to mobile users. Supporting different types of traffic requires flexibility in design and functionality. Owing to this requirement, there are many available options in the

IEEE 802.16 standard that are to be chosen by vendors and service providers in their product design or configuration. IEEE 802.16 has defined different types of QoS classes. In this section, the QoS classes suitable for video streaming applications are elaborated.

Video streaming applications generate variable bit rate traffic which is real-time and delay sensitive. Such applications require the network to allocate network resources to handle the corresponding traffic within a limited period of time. By considering these service requirements, it is observed that the real-time Polling Service (rtPS) class is suitable for supporting video streaming traffic. In the rtPS class, each traffic flow is characterized by a few parameters such as the minimum reserved traffic rate and the maximum sustained traffic rate. An rtPS traffic flow will not get admitted into the network if the BS cannot guarantee it the requested minimum reserved bitrate. The BS periodically polls the rtPS queues and assigns resources based on the bandwidth requests and connection parameters.

7.4 Multi-level Service Classification

This section introduces a novel solution for enhancing the quality of the video streams at the WiMAX subscriber stations. As mentioned in Section 7.3, the video streaming traffic is mapped into the rtPS class. In order to classify the video traffic as a rtPS class, it is necessary to determine the minimum reserved bandwidth. As described in Chapter 5, the average bitrate of a video stream that is used in this research is about 2 Mbps, and as shown in Table 1, the average size of each frame depends on the frame type, and its average value varies from 6.8 Kbytes for B-frames to 17 Kbytes for I-frames. It is worth noting that each video frame may be fragmented into some IP packets and MAC layer Data Units (MDU). Thus, the loss of an MDU can result in the loss of a frame. In regular WiMAX networks, a video stream is mapped into an rtPS class, and in the case of MAC layer congestion or poor physical

(PHY) layer conditions, the MDUs are dropped either at the BS or lost at the air interface. Considering the fact that each frame is fragmented into multiple MDUs, it is observed that the probability of a frame dropping increases when the network becomes more congested or the wireless link between the SS and BS becomes less reliable. The main characteristics of the MPEG video traffic discussed earlier, has inspired us to propose a cross-layer, content-aware traffic classification method called *Multi-level Service Classification.*

As mentioned in Chapter 5, MPEG video streaming traffic is composed of I, P and B-frames. Although B-frames generate the most amount of traffic, they have the least impact on the video quality. In the proposed method, the video server indicates the type of each frame in the Type of Service (ToS) field of the IP header. Therefore, the BS can distinguish the frame type of each packet. At the BS, MPEG frames are mapped into three different rtPS classes with different minimum reserved bandwidth parameters. The parameters of each traffic flow are determined by the SS or the video server, and are then sent to the BS during the call admission control process. Hence, each video stream is projected into three traffic flows. A schematic explaining the concept of Multi-level Service Classification is shown in Figure 7.1. It is noted that the proposed traffic classification is fully compatible with the current WiMAX certified products and does not require any changes at the BS or SS.

In order to sustain the video quality, it is crucial to send as many frames as possible to the end user. Furthermore, it is important to protect the more valuable frames, i.e., the I-frames, against dropping. As expected, the loss of frames will affect the video quality. Since all frames of GoPs are built over the base frame which is the I-frame, the loss of an I-frame will propagate throughout the GoP and all other frames will be corrupted. Similarly, the loss of a P-frame also affects the proceeding P-frame and some B-frames in each GoP. However, the loss of B-frames will not propagate

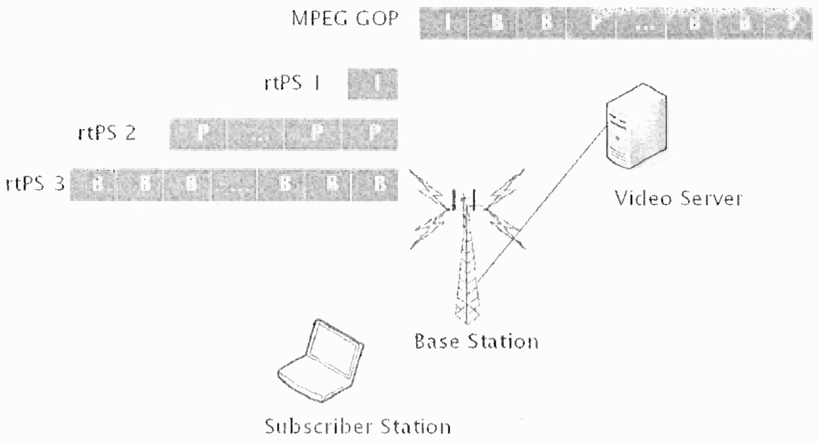


Figure 7.1 Multi-level Service Classification.

and will result in less quality degradation. Therefore, it is more important to protect the I and then P-frames from dropping.

When a traffic flow of a streaming application is admitted in the WiMAX network, it has the corresponding downlink (DL) queue in the BS. Thus, the DL traffic from the video server to a WiMAX SS is enqueued at its corresponding DL queue in the BS. If the traffic flow is admitted as an rtPS flow, the BS has to guarantee the requested minimum reserved bitrate for that. The DL queue will overflow if the input traffic rate exceeds the guaranteed reserved bitrate and the BS cannot allocate more resources to that flow due to either congestion or lack of bandwidth availability. In any case, the waiting time in the queues will impose some delay to the traffic flow. Full queues will drop the incoming traffic, and this will degrade the video quality. On the other hand, requesting higher bitrate for flows will decrease the chance of admittance in the network. Hence, increasing the minimum reserved bitrate will make the BS admit less number of flows in the network, and this will decrease the overall network utility. Thus, there is a trade-off between the video quality and network utility; this is optimized by choosing the optimum minimum reserved bitrate value for each video stream.

As explained earlier, the I, P and B-frames require different bitrates and the B-frames require the largest bandwidth while they have less effect on the video quality. Inspired by these observations, larger *minimum* reserved bitrate is allocated to the rtPS flow corresponding to the I-frames in the multi-level classification method. In order to minimize the probability of dropping I-frames, a bitrate which is higher than the average bitrate of I-frames is conservatively requested. It is worth noting that if the I-frame traffic flow cannot consume all the allocated bandwidth, the BS will spare the remaining part to other traffic flows. In order to sustain a minimum video quality, minimum reserved bitrates for the P and B-frames are also requested. However, the requested bitrates are less than the average bitrates of these flows. higher bitrates for P-frames are requested as compared to those of B-frames, and therefore increasing the probability of dropping B-frames. These threshold values will be more elaborated in Section 7.5 where the simulation results are discussed.

7.5 Simulation Results

In this section, the performance of the rtPS flow classification methods discussed in Section 7.4 is examined. The number of video frames received by the clients and the number of video frames dropped by the BS are studied through some simulations. Video streaming traffic is generated based on the model parameters mentioned in Table 7.2 and the OPNET simulator and its WiMAX package is used to perform the simulations. In order to achieve more accurate results, the PHY and MAC layers of WiMAX are fully simulated. The parameters used in the simulations are described in Table 7.1.

In the simulations, the MPEG model parameters introduced in Chapter 5 are used and they are presented in Table 7.2. By considering the parameters displayed in Table 7.2, it is understood that there are two GoPs per second. Thus, the model generates two I-frames per second, eight P-frames per second, and twenty B-frames

Table 7.1 Simulation Parameters

WiMAX Parameter	**Value**
MAC Frame Length	5ms
Symbol Duration	102.86 μs
Frequency band	5GHz
Bandwidth	20 MHz
Duplexity	TDD
Modulation and Coding	Adaptive
Number of Subcarriers	2048
DL Usage Mode	PUSC

per second. Hence, the average bitrate for each frame type is as follows: $\bar{R}_I = 273Kbps$, $\bar{R}_P = 588Kbps$, and $\bar{R}_B = 1094Kbps$. Thus, the overall average bitrate for each video stream is $\bar{R}_{tot} = 1955Kbps$. It is observed that the average bitrate of B-frames is higher than those of I and P-frames. Although the average size of a B-frame is less than that of other types, the bitrate associated with the B-frames is higher since there are more B-frames in each GoP, as discussed earlier.

In the simulations, a WiMAX network comprising of a BS and eleven SSs is considered. Each of the SSs receives an MPEG-4 video stream from one video server. An overview of the network architecture is shown in Figure 7.2. In the conventional scheme, each SS asks for one rtPS traffic flow with the minimum reserved bitrate equal to 884 Kbps. While in the multi-level traffic classification scheme, each SS asks for three rtPS flows corresponding to three different MPEG frame types. The SS will request 384 Kbps for I-frames, 300 Kbps for P-frames, and 200 Kbps for B-frames. The simulation is run under highly traffic loaded conditions of the network. Each subscriber station asks for 884 Kbps reserved bandwidth while each video stream requires almost 2 Mbps. Hence, the BS may drop some MDUs due to lack of resources. As mentioned in Table 7.1, the adaptive modulation and coding scheme (MCS) is used in simulations, and the BS chooses different MCSs based on the PHY channel status.

Simulations are conducted for both of the traffic classification schemes, and their performances are compared with respect to the number of MPEG frames received correctly at the end users. Simulations are performed under similar PHY characteristics and video traffic for each SS in both the schemes to understand the impact of the multi-level service classification scheme on the video quality. Figures 7.3(a) and 7.3(b) show that in both schemes, the BS has almost dropped/sent the same amount of traffic. Thus, any difference at the video quality at the SSs is due to the different classification schemes. Figures 7.4(a) and 7.4(b) represent the number of frames received by the users with the worst physical channels and the best physical channels,

Table 7.2 MPEG Model Parameters

Model Parameter	**Value**
Display size	320x240
GoP pattern	$N = 15$, $M = 3$
Frame rate	30 frames per second
Scene size parameter (d)	Exponential with p=0.1
I-frame size (bytes)	Log-normal ($\mu = 17068$, $\sigma = 7965$)
P-frame size (bytes)	Log-normal ($\mu = 9190$, $\sigma = 7005$)
B-frame size (bytes)	Log-normal ($\mu = 6839$, $\sigma = 5323$)
AR coefficients	$a_1 = 0.39$, $a_2 = 0.15$, $\sigma_\varepsilon = 4.36$

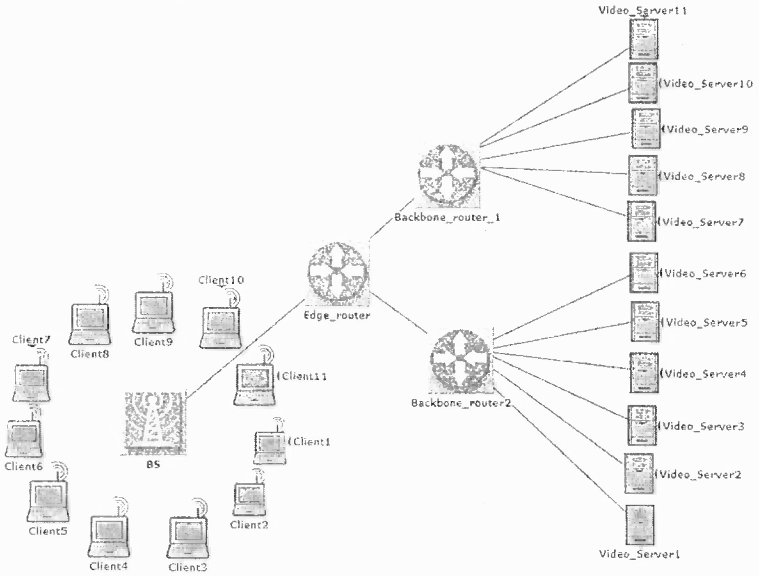
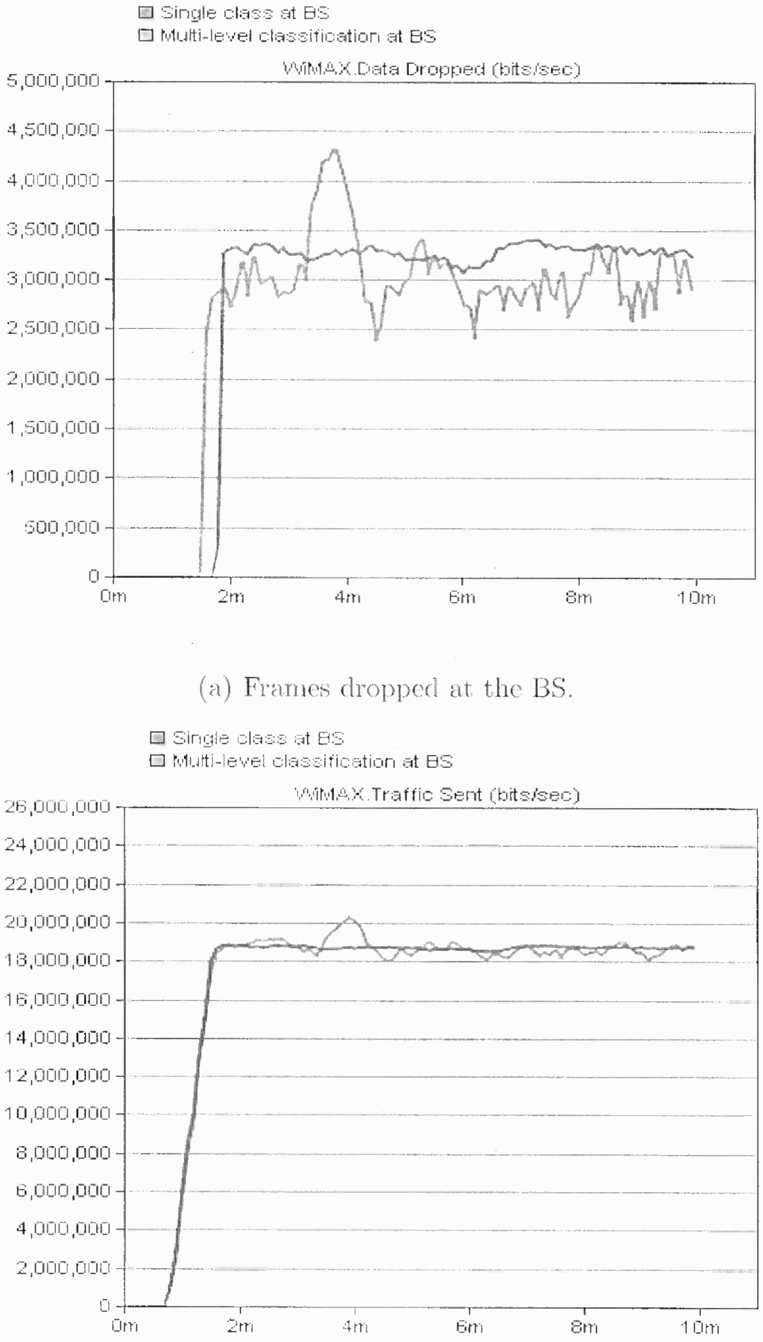


Figure 7.2 Network Architecture.

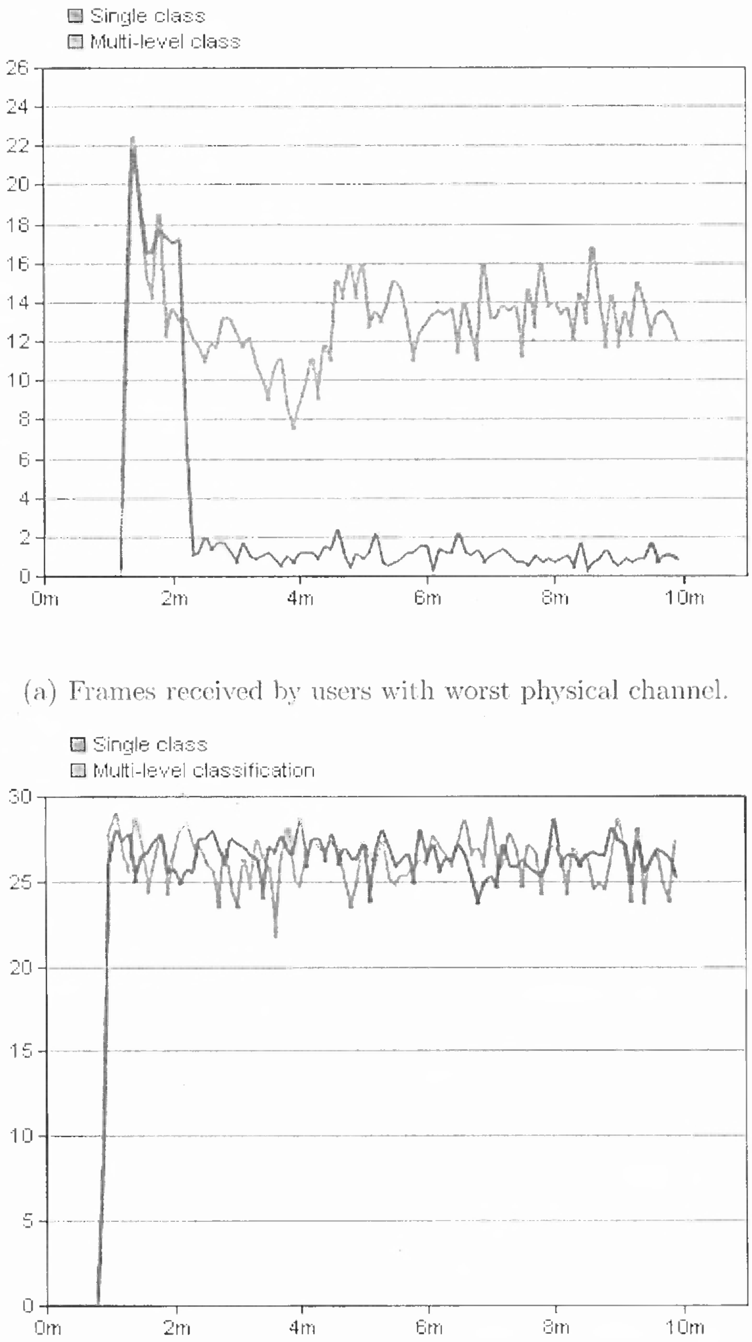
respectively. It is shown that although the SSs with the best PHY channels have received almost the same number of frames, the SSs with the worst PHY channels in the single class scheme have received fewer number of frames as compared to that for the SSs with the worst PHY channels in the multi-level service classification scheme. Figures 7.5(a) and 7.5(b) show the number of frames received for each type of MPEG frames by the user with the worst and the best PHY channels, respectively. It is noted that mainly B-frames are dropped in this scheme, while the I-frames and most of the P-frames are received at the SSs. It is also worth noting that in the single service class scheme, all types of frames are subject to dropping since no preference among the frame types is made. Figure 7.6 presents the average number of frames received by all users in both schemes. It is shown that the multi-level classification scheme outperforms the single class scheme in terms of the average number of frames received by all users in the network.



(a) Frames dropped at the BS.

(b) Traffic sent by BS.

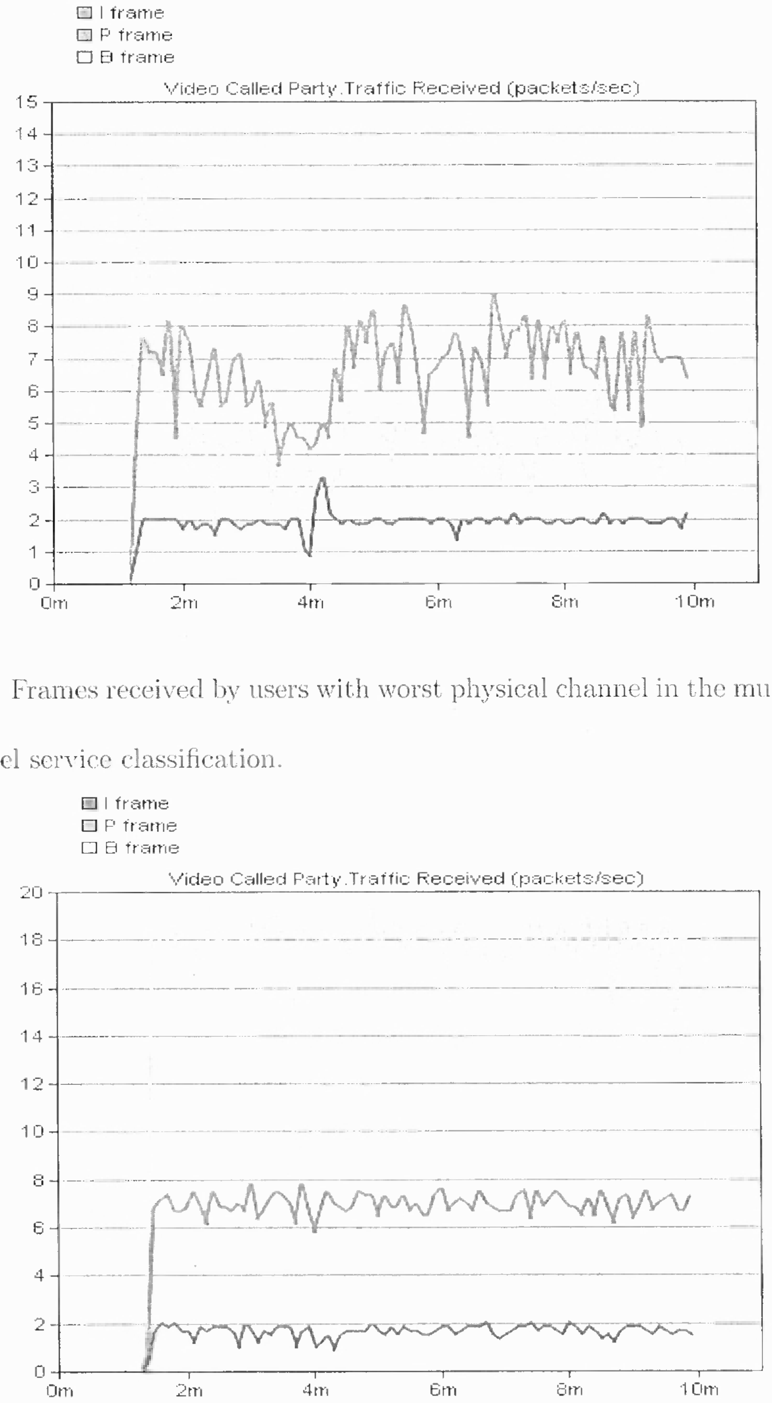
Figure 7.3 BS performance.



(a) Frames received by users with worst physical channel.

(b) Frames received by users with best physical channel.

Figure 7.4 Client performance.



(a) Frames received by users with worst physical channel in the multi-level service classification.

(b) Frames received by users with best physical channel in the multi-level service classification.

Figure 7.5 Multi-level classification performance.

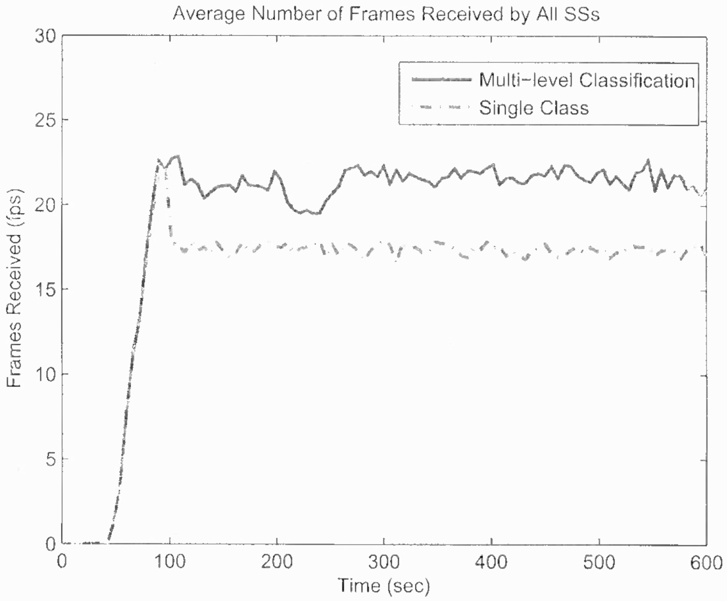


Figure 7.6 Average number of frames received by all SSs.

7.6 Summary

This chapter has proposed some novel solutions to increase the performance of MPEG video transmission over WiMAX networks. A cross-layer approach which relies on the characteristics of the MPEG frames and the elaborated QoS classification features at the WiMAX MAC layer is introduced. It has explained the challenges of transmitting video traffic over wireless networks, and discussed some of the WiMAX networks constraints and design tradeoffs, which can dramatically impact the quality of video. The main characteristics of the MPEG traffic and the MPEG model which categorizes the traffic frames into three types: I, P and B-frames have been illustrated. It is shown that by providing the BS with information about the type of video frame, it can map I, P and B-frames into three different rtPS service classes with different service requirements. It is shown that by incorporating the proposed traffic classification scheme at the BS, the overall number of frames delivered to each SS is increased; this enhances the video quality at the end users. It is a simple and reliable scheme which can be readily deployed in WiMAX networks.

CHAPTER 8

TRAFFIC-AWARE VIDEO STREAMING IN BROADBAND WIRELESS NETWORKS

8.1 Objective

With increasing implementation of broadband wireless networks and extensive deployment of multimedia services such as Video on Demand (VoD) or IPTV, the demand for video streaming applications will be increased and more people will use their wireless devices to reach numerous video contents available in the Internet. Streaming real-time video in wireless networks is a challenging problem due to the stringent service requirements of video traffic and impairments of wireless channels. Providing the required Quality of Service (Qos) through efficient resource allocation is a complicated problem that service providers are confronting. This chapter proposes a traffic-aware, cross-layer solution for enhancing the perceived video quality at the end user in wireless networks. This solution incorporates the characteristics of the MPEG traffic to give more priority to the more important frames and to protect them against dropping when available resources of the network are not sufficient for providing the desired QoS to the traffic flow. It is shown that the proposed solution will improve the perceived video quality over the broadband wireless networks.

8.2 Introduction

The growing interest for deployment of broadband wireless networks is providing the mobile users with a broadband access through which they can run various applications. It is expected that multimedia applications and in particular video streaming applications will grow due to the population of many online video servers. Therefore, an excessive number of people will use their wireless devices to access to the numerous video streaming contents on the Internet.

The increasing demand for broadband wireless access has called for the design and implementation of different wireless technologies such as WiMAX and LTE. These technologies are supposed to provide the users with network connectivity to run different applications with various Quality of Service (QoS) requirements. Hence, the network designers of the broadband wireless technologies should provision adequate arrangements to support different QoS requirements. The stringent QoS requirements of video streaming applications including high bitrate and low latency are some of the challenges the service providers are confronting. Moreover, supporting QoS requires proper resource allocation which makes video streaming a complicated problem.

Efficient allocation of scarce resources in wireless networks is a crucial and challenging problem. The importance of this problem has attracted a keen interest in the research community. In this section, some of the research work devoted to the problem of video streaming in wireless networks are reviewed. Rate adaptive video streaming is one of the solutions highly discussed in the literature [31, 27, 68]. Most of the rate adaptive solutions assume that the video server receives some feedback information such as the end-to-end delay or the loss rate from the end client. By using the feedback information, the video server can choose the optimum coding option. The rate adaptive solutions may not be feasible in wireless networks where the adaptive methods may not be able to track the fast changes in the channel. Moreover, this solution increases the computational complexity at the video server which may result in overloading the server in large networks. Furthermore, sending feedback information may not be possible in some video streaming applications such as IPTV in which the video server has to multicast the same content to different clients.

One of the issues for streaming real-time video over wireless networks is sustaining the satisfactory video quality even when congestion happens or the wireless channels become less reliable. Different techniques have been proposed to achieve an acceptable video quality with respect to the limited network resources; nevertheless, in some cases, it is inevitable to prevent the video packets from being lost due to transmission errors over wireless channels or dropped due to overflow of the queues at the Base Station (BS). The effect of packet loss on the video quality has been studied by many researchers. Liang *et al.* [39] analyzed the effect of bursty errors which occur frequently in wireless networks. Meanwhile, there has been extensive effort to find solutions to mitigate the loss effects. Feamster and Balakrishnan [19] introduced a selective re-transmission mechanism to recover more important packets of the video streams. Examples of error moderating schemes include, rate adaptive coding [31], forward error correction schemes [56], and scalable video coding [63].

In this chapter, a cross-layer solution for streaming video over wireless networks is proposed. Unlike the rate adaptive solution, the computational complexity of the video servers is not increased. The proposed solution is, therefore, appropriate for video multicasting applications. MPEG video streaming is considered and the characteristics of the MPEG traffic is used to enhance the video quality perceived by the end user. In the traffic aware video streaming mechanism, the video server includes the necessary information, which the Base Station (BS) requires to handle the video traffic, in the IP header of the video packets. The BS can determine the importance of each packet and its effectiveness on the perceived video quality at the end user. The BS can therefore protect the more important packets against dropping in the wireless medium, and thus enhances the video quality. The proposed solution may increase the processing load of the BS but, as will be discussed later, the availability of the required information in the IP header of each packet will minimize the extra processing load because the IP header of all packets are usually processed by the BS for other networking purposes.

The validity of the proposed solutions is examined by simulating a broadband wireless network which has implemented the traffic aware video streaming. Without loss of generality, the WiMAX technology is considered as the platform for evaluating the performance of the proposed mechanism. WiMAX is a promising technology expected to deliver broadband wireless connectivity to mobile users in near future. It must be noted that although the impact of the solution on the quality of the video stream transmitted over a WiMAX network is presented, similar quality enhancement is expected on other emerging wireless technologies comprised of base stations and subscriber stations. Video streaming over WiMAX networks will be a popular application for users and an attractive market for service providers. There are many research works trying to tailor the general solutions for WiMAX networks. Kim *et al.* [34] introduced a channel adaptive scheme while Hillestad *et al.* [30] proposed the scalable video coding to transmit video traffic over WiMAX. Unlike most of the proposed solutions that either increase the complexity of video server or are not applicable to multicasting applications, the cross-layer solution is scalable and suitable for both unicasting and multicasting applications of real-time video streaming while it does not increase the complexity of the video servers.

The outline of the rest of this chapter is as follows. In Section 8.3, a cross-layer solution for video streaming in broadband wireless networks is proposed. The simulation results are presented in Section 8.4. Concluding remarks are given in Section 8.5.

8.3 Intelligent Queue Management

This section introduces a novel solution for enhancing the performance of video streaming in broadband wireless networks. This solution incorporates the characteristics of the video traffic discussed in Chapter 5. MPEG video is a variable bitrate traffic which is delay sensitive. In WiMAX, this kind of traffic is mapped into real-time Polling Service (rtPS) class based on the QoS requirements. As a result, the BS has to guarantee a minimum reserved bitrate for the rtPS traffic flow at the call admission process. The BS periodically polls the rtPS queue and assigns resources based on the bandwidth request. It worth noting that increasing the guaranteed minimum bitrate will decrease the queuing delay and the probability of packet loss happened due to queue overflow; however, it also decreases the network utility since the BS can admit fewer users into the network. Choosing the optimum bitrate which provides the end users with at least the minimum satisfactory video quality is an open problem. In this section, it is supposed that the BS has already chosen a minimum bitrate for the video stream, and it will be shown that the video quality is enhanced by incorporating the proposed novel solution.

As discussed in Chapter 5, the loss of the video frames will degrade the video quality, and this is more severe if an I frame is lost. Owing to the hierarchical structure of GoPs and the inter-dependency of frames, it is understood that the network should protect the I frames and then the P frames from loss to prohibit the *propagation of errors* effect. Considering the fact that each video frame may be fragmented into multiple IP packet or MAC layer Data Units (MDU), it is perceived that the probability of a frame loss is increased when the network becomes more congested. The main characteristics of the MPEG video traffic has inspired us to propose a cross-layer, traffic-aware queuing solution called *Intelligent Queue Management*, to be discussed next.

In the proposed solution, the video server indicates the type of each frame in the Type of Service (ToS) field of the IP header, and thus the BS can determine the frame type of each packet. The BS can also understand the GoP number of each packet and its frame number. It is worth to recall that in the traffic model, the video encoder generates constant number of frames per second, and thus the BS can distinguish the frame number of each packet and its GoP number by inspecting the sequence number and ToS fields of the IP header. As mentioned earlier, this solution protects the more important frames against dropping when congestion happens at the BS.

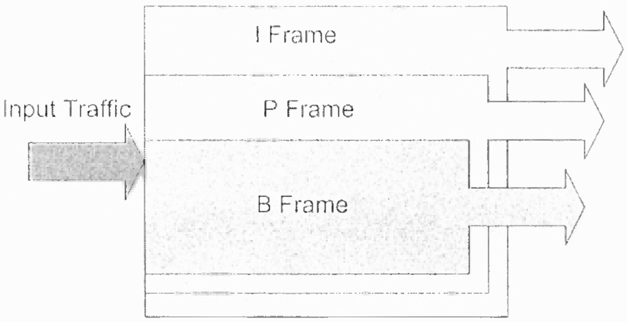


Figure 8.1 Intelligent Queue Schematic.

In WiMAX, the MPEG video streaming traffic is enqueued as an rtPS class flow at the BS. As mentioned in Chapter 5, the traffic bitrate is about 1955 kbps, but the BS may not be able to guarantee this bitrate for the traffic flow. It is noted that most of the bitrate is consumed by the B frames while they have least impact on the video quality. Hence, the BS may guarantee a lower bitrate for this traffic flow to increase the number of admitted users in the network. If congestion happens, the BS would drop the least important frames while it tries to adhere to the minimum video quality by transmitting the more important frames on time. It is also worth noting that during a congestion period, the BS provides the user with only the minimum bitrate which is guaranteed at the call admission process. In this case, the queue size of the corresponding video stream and the queuing delay will be increased. The BS considers a maximum possible size for queuing each traffic flow, the excessive incoming traffic beyond the maximum queue size will result in frame loss. Long queuing delay may also result in frame dropping because of the maximum caching time at the end user. The video traffic encountering delay time more than the maximum caching time cannot be used by the end user, and thus the BS should keep track of queuing delay of each packet. In WiMAX, the maximum tolerable delay is also determined at the call admission process. Therefore, it is understood that the overflow of the queue or long queuing delay may result in frame loss which degrades the video quality at the end user.

In order to sustain the video quality, it is crucial to send as many frames as possible to the end user. Moreover, it is necessary to protect the more important frames, i.e., the I frames and then the P frames, against dropping. In the proposed solution, a priority queuing at the BS is considered. To better understand the queue structure, a virtual queuing structure as depicted in Fig. 8.1 is assumed. The arriving packets will be enqueued at the BS based on their frame type. The packets are ordered in each queue according to their GoP and frame number. At the arrival of each packet, the BS determines its frame type, GoP and frame number by checking its IP header. If the queue has enough space, the BS will put the packet in its corresponding sub-queue, but if the maximum queue size is reached, the BS checks the frame type of the incoming packet. If there is any lower priority packet in the queue, the BS will drop as many lower priority frames as needed from the queue to enqueue the arriving packet and protect it against dropping. It is noted that loss of a fragmented packet will result in the loss of the whole frame, and thus the BS will prefer to drop the frames whose fragments have not still been sent in part to the end user. The BS will drop the incoming packet if there is not enough lower priority packets available in the queue suitable for dropping. The BS will also drop all the dependent packets that will arrive subsequently in the future and require the dropped packet for decoding. This is because the end user will not be able to decode them without receiving the dropped frame. The procedure executed upon arrival of a new packet by the BS is shown in Fig. 8.2. The priorities of the frames in each GoP are as follows:

1. I frame.

2. P frames in descending order. In each GoP, the higher order P frames have less priority because less number of frames is derived from them.

3. B frames.

As explained earlier, the BS should drop the packets awaiting in the queue for more than the maximum tolerable delay which is equal to the caching time at the end user. With that said, the BS regularly monitors the delay time of the first packet of each sub-queue and if the maximum delay is reached, it will drop that packet from the queue. It is worth noting that the BS has to drop all the lower priority packets that are dependent on an already dropped packet for decoding. For example, the drop of the third P frame, which is the tenth frame of the GoP, will result in dropping of all B and P frames located from the 8th to 15th frame of that GoP, as depicted in Fig. 5.2. The delay monitoring procedure regularly performed by the BS is shown in Fig. 8.3.

It is worth noting that to avoid long delay the queues should be served properly in time. This may not be possible during the congestion period when the BS can provide each queue with only the minimum reserved bitrate. Inspired by this fact, the available bitrate is divided between three sub-queues, and a higher minimum reserved bitrate is conservatively allocated to the I frame sub-queue. Therefore, it is possible to avoid the long delays for more important sub-frames. It is also note that if the I frame traffic cannot consume all the allocated bandwidth, the BS will spare the remaining part to other traffic flows. A higher bitrate is chosen for the P frame sub-queue as compared to that of B frame, and thus the probability of dropping the P frames is reduced. The chosen bitrates for each of the sub-queues are more elaborated in Section 8.4 where the simulation results are discussed.

In order to allocate resources to the video stream, the BS polls each of the sub-queues separately. The polling period depends on the minimum reserved bitrate of each sub-queue and the higher the bitrate, the more frequent polling. During the congestion time the BS can only provide the user with the minimum reserved bitrate. In this case, the minimum amount of traffic associated to the minimum reserved bitrate is transmitted to the end user upon polling the sub-queue. The

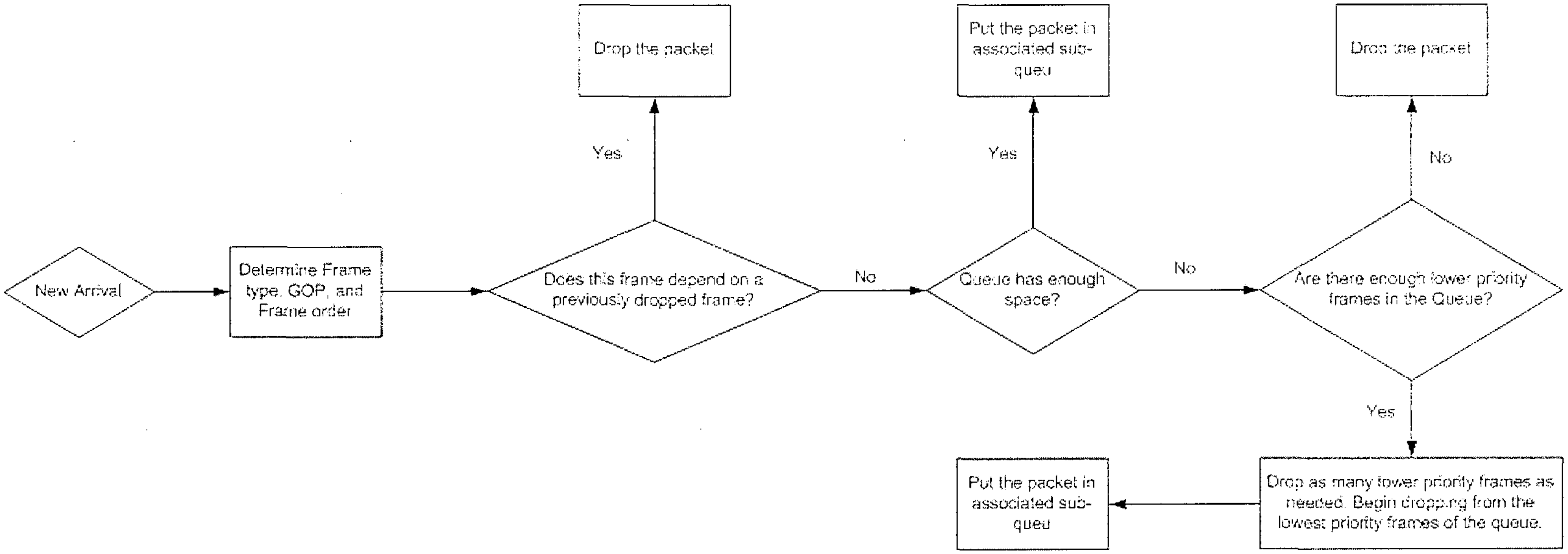

New Arrival
Determine Frame type, GOP, and Frame order
Does this frame depend on a previously dropped frame?
Yes
Drop the packet
No
Queue has enough space?
Yes
Put the packet in associated sub-queu
No
Are there enough lower priority frames in the Queue?
No
Drop the packet
Yes
Drop as many lower priority frames as needed. Begin dropping from the lowest priority frames of the queue.
Put the packet in associated sub-queu

Figure 8.2 Frame Arrival Management.

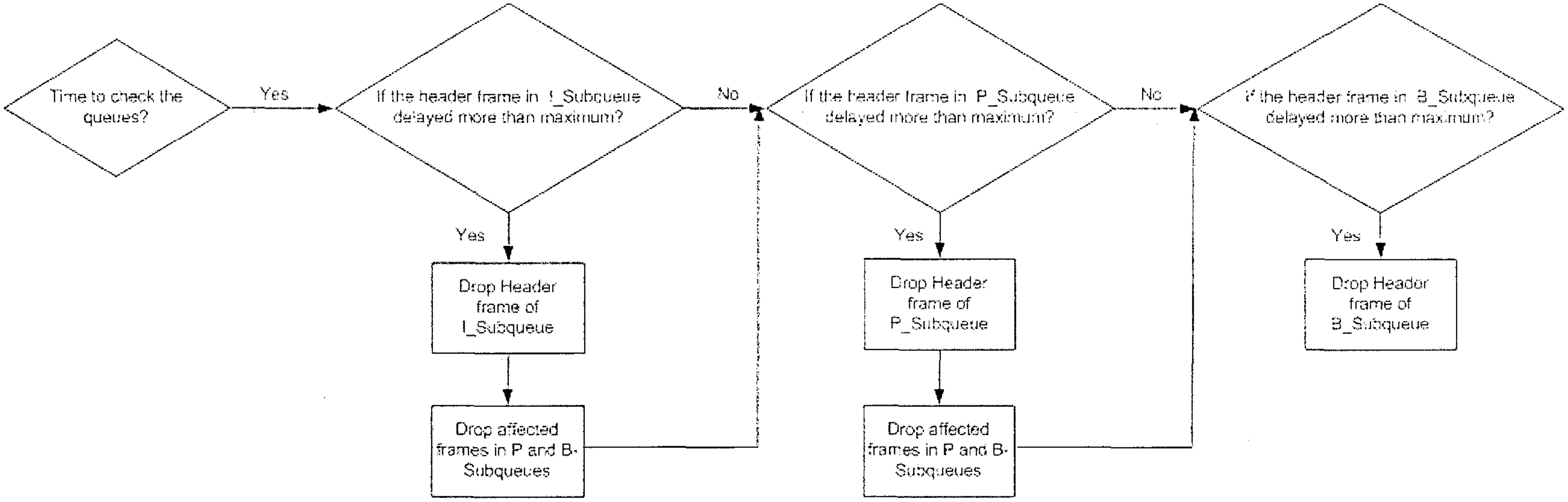

Time to check the queues?
Yes
If the header frame in I_Subqueue delayed more than maximum?
No
If the header frame in P_Subqueue delayed more than maximum?
No
If the header frame in B_Subqueue delayed more than maximum?
Yes
Drop Header frame of I_Subqueue
Drop affected frames in P and B-Subqueues
Yes
Drop Header frame of P_Subqueue
Drop affected frames in P and B-Subqueues
Yes
Drop Header frame of B_Subqueue

Figure 8.3 Delay Management Procedure.

BS will not allocate any resources to the sub-queue if there is no traffic available in that sub-queue. Frames are sorted according to their frame number and GoP number in each sub-queue. Therefore, the departure permutation of the frames from each sub-queue is regulated. In the intelligent queue management, the loss of the higher priority frames will result in the loss of the dependent and lower-order frames of their GoPs. Moreover, the sub-queues with higher priority are guaranteed to receive higher bitrates while they have smaller average bitrate. Therefore, it is not necessary to synchronize the permutation of departing frames among different sub-queues. Moreover, the end user caches the arrival frames, and thus out of sequence arrivals will be adjusted accordingly.

8.4 Simulation Results

In this section, the performance of the proposed intelligent queue management scheme discussed in Section 8.3 is examined. The number of frames received by clients is studied through some comprehensive system level simulations. The traffic is generated based on the model explained in Chapter 5. In the simulations, the performance of the MAC layer of WiMAX is only examined, and the PHY layer effects are not considered. The network is simulated under congestion circumstances in which the BS can only provide each traffic flow with the minimum reserved bandwidth. The conventional WiMAX network in which the video stream is classified as the rtPS service class is also simulated. In the convetional network simulations, the traffic is enqueued in a regular FIFO queue at the BS. Furthermore, in the convetional scheme, packets are dropped regardless of their frame types when the queue becomes full or the delay becomes long. It is supposed that the maximum possible queue size is 6MB in both schemes and the maximum tolerable delay is 10 seconds. Therefore, the frames, which are delayed for more than 10 seconds in the queue, will be dropped.

It is supposed that the BS has reserved 900 Kbps for the traffic flow which is less than 1955 Kbps required for streaming the video. Since the departure rate of the queue will be less than the arrival rate, the queue will overflow, and packet loss in both queuing schemes is expected. In the intelligent queue management scheme, it is assumed that the BS divides the reserved bitrate among the sub-queues and allocates 360 Kbps for the I frames, 340 Kbps for the P frames, and 200 Kbps for the B frames sub-queues, respectively. It is noted that the BS has assigned a value more than the average rate of the I frames to them. This protects I frames against dropping. As mentioned earlier, the BS can spare the bandwidth to other traffic flows if the I frames cannot consume all the allocated bandwidth. It is expected to observe the loss of the P fames and B frames due to shortage of the pre-allocated bandwidth to these flows, and it is expected to observe a higher number of dropped B frames since they receive the least share of the bandwidth.

Simulations are conducted for both of the intelligent queue management and conventional queuing schemes, and their performances are compared with respect to the number of MPEG frames received and decoded correctly at the end user. Obviously, receiving a higher number of decodable frames will result in better perceived video quality. This can be a useful criterion for comparing the performances of different networking schemes while considering the importance (or priority) of the different frame types. Fig. 8.4 shows the number of frames of each GoP decoded by the end user in the intelligent queue management scheme. It is observed that all I frames have been delivered, and on average, three P frames and two B frames were also delivered. In the intelligent queuing scheme, by receiving all of the I frames and most of the P frames, the end user can adhere to the minimum perceived video quality during the congestion periods. Fig. 8.5 shows the comparison on the number of frames of each GoP received and decoded by the end user when applying the intelligent queuing and conventional queuing schemes. It is noticed that the intelligent queuing

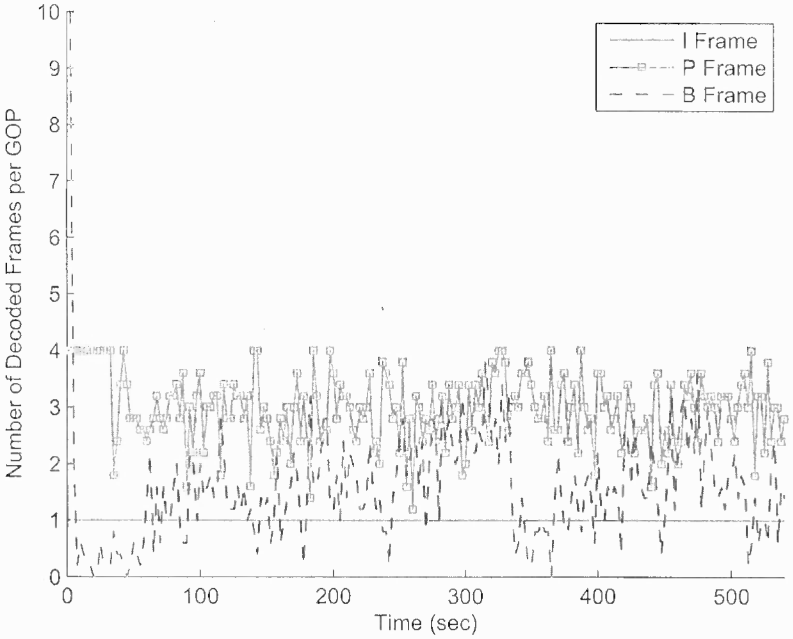


Figure 8.4 Intelligent Queue Management Performance.

scheme delivers more frames to the end user, and thus it provides better video quality. However, in the conventional queueing scheme, the end user sometimes will not be able to decode any frames of some GoPs due to the loss of the higher priority frames. Owing to the sever disruptions in the video streaming, the conventional queuing scheme is not a feasible solution for the networks with the possibility of congestion or shortage of resources.

8.5 Summary

This chapter has proposed a novel solution for enhancing the quality of MPEG streaming over broadband wireless networks. Inspired by the characteristics of the MPEG traffic model and the effect of frame loss on the video quality of the end user a cross-layer scheme has been designed. This scheme provides the BS with the frame information of each incoming video streaming packet. In the proposed solution, the BS can deduce the frame type and frame order of each incoming packet by examining

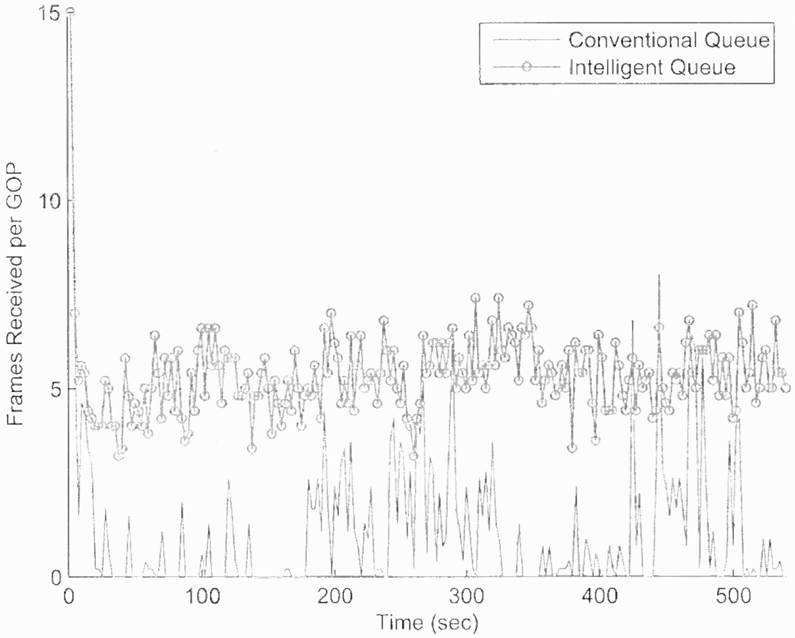


Figure 8.5 Intelligent Queue and Conventional Queue Comparison.

its IP header. A queue management strategy is proposed. In the intelligent queue solution, the BS can drop less effective frames when congestion happens. It is shown through some simulations that by incorporating the proposed cross-layer solution and intelligent queuing scheme, it is possible to deliver more video frames to the end users. Moreover, the proposed design protects the most important frames, i.e., I frames and P frames, against dropping, and thus provides better perceived video quality.

CHAPTER 9

RATE ALLOCATION AND GUARANTEED QOS FOR VIDEO TRAFFIC

9.1 Objective

Video streaming applications require stringent QoS requirements which should be furnished by service providers. In wireless networks, it is up to the base stations to provide the streaming applications with required services by reserving enough bandwidth and allocating enough resources to fulfill the service requirement. Since service providers are interested in admitting more users into the their networks to maximize the network utility, they are interested to reserve the least amount of bandwidth which satisfies the minimum QoS requirements. In streaming applications, it is necessary to understand the minimum QoS requirements needed to provide the users with acceptable quality of experience. This chapter introduces an analytical framework through which service providers and streaming applications can realize the achievable probability of frame loss as a function of guaranteed bitrate. With this knowledge, service providers may choose the optimum bitrate to maximize the network utility while providing the users with acceptable picture quality.

9.2 Introduction

At the call admission control process, it is necessary to negotiate service requirements of each traffic flow as the BS has to decide whether it can provide the user with the requested service requirements. Each user has to ask for the QoS needed for running its applications. Although requesting higher service rates may result in better QoE, the BS may not grant the request if it cannot provide the user with requested services. In general, the BS prefers to guarantee the least amount of bandwidth (BW) such that it can admit more number of users in the network, and thus increasing the network utility. Finding the bitrate which guarantees the minimum acceptable QoE is a complicated problem which is addressed here.

Although it is possible for the video server to estimate the optimum bitrates for offline and non-real time streaming applications such as VoD, it is more difficult for real-time applications in which the server cannot fully predict the required bitrate of a VBR traffic flow. The main challenge with respect to optimizing the QoE is modeling user satisfaction and measuring QoE. Many research works have been proposed to address rate allocation for VBR traffic to limit the queue size or queuing delay [36], and different groups such as Video Quality Experts Group (VQEG) have proposed some criteria to assess the video quality (http://www.its.bldrdoc.gov/vqeg/).

Based on the MPEG traffic model discussed earlier and the multi-level service allocation scheme, it is of interest to find suitable bitrates for each of the traffic flows which carry different types of frames. Intuitively, a higher bitrate for the flow corresponding to I frames is requested to minimize the probability of dropping I frames. Therefore, a bitrate that is higher than the average bitrate of I frames is conservatively requested. It is noticed that if the I frame traffic flow cannot consume all the allocated bandwidth, the BS will spare the remaining part to other traffic flows according to the intelligent queuing solution. To sustain a minimum video quality, the minimum reserved bitrates are also request for both the P and B frames. However, the requested bitrates are less than the average bitrates of these flows, and the bitrate for P frames is higher than that of the B frames. Thus, B frames are more likely to be dropped.

Results of many ongoing research works in modeling the QoE and especially exploring thresholds of acceptable video packet loss for each frame type can be employed to enhance the performance of proposed application-oriented solutions.

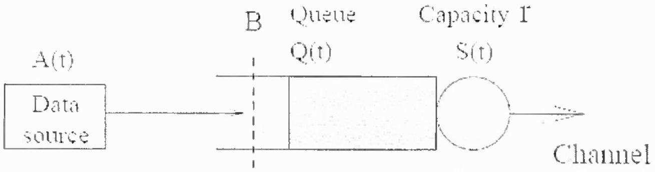


Figure 9.1 A general queueing system.

9.3 Background on Effective Bandwidth

QoS guarantees have been studied in many research works [20]. A large number of papers have considered QoS in the physical layer by modeling physical layer characteristics such as bit error rate with respect to the SNR [21]. While studying QoS guarantees can be applied to circuit switch applications such as cellular voice, providing QoS to the traffic flows in the packet switch networks requires understanding the impact of higher layers on the QoS. Some papers have considered QoS guarantees in the MAC layer [21, 10]. In modeling the MAC layer, the general queue system shown in Fig. 9.1 is used. This queue system represents a First-In-First-Out(FIFO) queue and models the delay and loss of incoming packets.

Queuing analysis relies on the knowledge of incoming traffic characteristics and outgoing service characteristics. In this queuing system, the arrival process, $\{A(t), t \geq 0\}$, represents the amount of arrival traffic (in bits) during the time interval $[0, t)$. The actual service of the queued bits over the interval $[0, t)$ is represented by $S(t) = \int_0^t r(\tau)d\tau$. The instantaneous size of the queue is also shown by $Q(t)$.

In this chapter, the theory of *Effective Bandwidth*, which is discussed in various literature sources [9, 48, 18, 64], is applied to analyze the performance of the proposed application-oriented solutions discussed in Chapter 6. It is of interest to study the delay and loss performance of the proposed solutions by applying the tools introduced in the theory of effective bandwidth. In order to apply that theory, the parameters used in this theory are first introduced. In this chapter, the generic parameter notations of Reference [64] are used.

The *asymptotic log-moment generating function* of $A(t)$ is defined as

$$\Lambda(u) = \lim_{t \to \infty} \frac{1}{t} \log E\left[e^{uA(t)}\right]. \tag{9.1}$$

If this functions exists for all $u \geq 0$, the *effective bandwidth function* of $A(t)$ is defined as [48]

$$\alpha(u) = \frac{\Lambda(u)}{u}, \quad \forall u \geq 0. \tag{9.2}$$

Under appropriate conditions, it can be shown that $\alpha(u)$ is continuous and increasing in u [9].

It was explained earlier that at the call admission control process, the BS will not admit a new video traffic flow into the network unless it can guarantee a minimum bitrate for it. It was discussed that the BS will only allocate the minimum reserved bitrate to the traffic flow during congestion periods when it does not have enough resources to serve all users in the network. This chapter analyzes the performance of the proposed solution in the worst case in which the BS serves the video traffic flow with the guaranteed bitrate. In this case, the service rate, r, is considered constant. It is shown that the probability of $Q(t)$ exceeding the B is [48]

$$\sup_t \Pr\{Q(t) \geq B\} \approx \gamma e^{-\theta B} \tag{9.3}$$

Both γ and θ are functions of the actual service bitrate r. θ is the *QoS Component* with respect to the size of the queue and it is the solution of $\alpha(\theta) = r$ which means θ is the inverse function of the effective bandwidth $\alpha(u)$.

$$\theta = \alpha^{-1}(r) \tag{9.4}$$

$\gamma = Pr\{Q(t) \geq 0\}$ is the probability of non-empty queue at the random time instant t [12]. Since the worst case performance is considered here, it is assumed that the queue is not empty at the congestion periods and thus $\gamma = 1$.

If the delay performance of the queue is of interest, the corresponding QoS exponent is $\theta^{(d)}$ which is defined in Reference [66].

$$\theta^{(d)} = \theta \times r. \tag{9.5}$$

Therefore, the probability of $D(t)$ exceeding a delay bound D_{max} is calculated from Equations (9.3), and (9.5).

$$\sup_t Pr\{D(t) \geq D_{max}\} \approx \gamma e^{-\theta r D_{max}} \tag{9.6}$$

9.3.1 Finite Buffer Effects

In the queuing system explained above, it is assumed that the maximum buffer size is unlimited. However, if the buffer is finite the incoming packets would get dropped when the queue size reaches it maximum value $B_{\max}$. In this case, the probability of packet loss due to buffer overflow is less than $\Pr\{Q(t) \geq B_{\max}\}$. It is because the dropped packets are removed from the system. Since the probabilities of dropping are calculated under the worst case here, Equation (9.3) is used to calculate Pr_{loss} which is the probability of packet loss caused by buffer overflow in the finite buffer queuing system.

$$\begin{aligned} Pr_{loss} &\leq \sup_t \Pr\{Q(t) \geq B_{\max}\} \\ &\approx e^{-\theta B_{\max}} \end{aligned} \tag{9.7}$$

Dropping of packets affects the delay performance of the queuing system. If part of incoming traffic is dropped at the entrance of the queue with the probability of Pr_{loss}, the buffer limited queuing system is equivalent to an infinite buffer queuing system whose source traffic is $A(t)\,(1 - Pr_{loss})$. It is assumed that QoS component of this system is θ'. By using Equation 9.2, it is understood that

$$\lim_{t\to\infty} \frac{1}{\theta' t} \log E\left[e^{\theta'(1-Pr_{loss})A(t)}\right] = r \Rightarrow (1 - Pr_{loss})\,\theta' = \alpha^{-1}\left(\frac{r}{1 - Pr_{loss}}\right) \qquad (9.8)$$

Since $\alpha(u)$ is a continuous and increasing function of u and $Pr_{loss} < 1$, it is deduced that $(1 - Pr_{loss})\,\theta' \geq \theta$.

The probability of excessive delay in the finite buffer queuing system, Pr_{delay} is calculated by using Equation 9.6 as follows:

$$\begin{aligned} P_{delay} &= \sup_t Pr\{D(t) \geq D_{max}\} \approx e^{-(1-Pr_{loss})\theta' r D_{\max}} \\ &\leq (1 - Pr_{loss})\, e^{-\theta' r D_{\max}} \leq (1 - Pr_{loss})\, e^{-\theta r D_{\max}} \\ &= (1 - Pr_{loss})\, Pr_{delay}^{infinite} \end{aligned} \qquad (9.9)$$

9.4 Loss Analysis

In this section, the theory of effective bandwidth is used to analyze the loss performance of the multi-level service classification and intelligent dropping schemes, and to compare them with that of conventional video streaming methods. As discussed before, the BS has to consider a limited queue space for each traffic flow and when the queue size reaches its maximum limit, any extra incoming traffic will result in a packet loss. The probabilities of packet loss in each of the video streaming schemes in studied next.

9.4.1 Multi-level Service Classification

In this method, as explained in Chapter 7, the MPEG video stream is separated into three different traffic flows according to the frame type of the video packets. The BS reserves r_I, r_P, and r_B as the minimum bitrate of corresponding traffic flows.

By using Equation (9.7), the probability of queue overflow and thus packet loss of each traffic flow in the Multi-level Service Classification method is

$$Pr^{Ml}_{loss}(I) = \sup_t \Pr\{Q_I(t) \geq B_I^{\max}\} \approx e^{-\theta_I B_I^{\max}} \tag{9.10}$$

$$Pr^{Ml}_{loss}(P) = \sup_t \Pr\{Q_P(t) \geq B_P^{\max}\} \approx e^{-\theta_P B_P^{\max}} \tag{9.11}$$

$$Pr^{Ml}_{loss}(B) = \sup_t \Pr\{Q_B(t) \geq B_B^{\max}\} \approx e^{-\theta_B B_B^{\max}} \tag{9.12}$$

where $Pr^{Ml}_{loss}(I)$ represents the loss probability of the I-frame traffic flow in the Multi-level service classification scheme, and similar notations are used for the P and B-frame traffic flows. $Q_I(t)$, $Q_P(t)$, and $Q_B(t)$ are the queue of each traffic flow at time instant t. B_I^{max}, B_P^{max}, and B_B^{max} are correspondingly the maximum buffer sizes assigned to each of the traffic flows of a video stream.

9.4.2 Intelligent Dropping

It was discussed in the Chapter 8 that in the Intelligent Dropping scheme, the BS separates the packets of a video stream into three traffic flows according to their frame type, but it considers one buffer with the maximum size of B_{max} to be shared by all of the traffic flows of a video stream. To make it comparable with the Multi-level Service Classification Scheme, it is assumed that the maximum size of the queue in the Intelligent Dropping scheme is equal to the summation of maximum sizes of all three queues of a video stream in the Multi-level Service Classification scheme

$$B_{\max} = B_I^{\max} + B_P^{\max} + B_B^{\max}. \tag{9.13}$$

Owing to the priority queuing system in Intelligent Dropping method, the I-frame traffic flow has the highest priority, and thus can fully fill up the buffer. By using Equation (9.7), the probability of queue overflow and thus packet loss of the

I-frame traffic flow in the Intelligent Dropping scheme is

$$Pr_{loss}^{ID}(I) = \sup_t \Pr\{Q_I(t) \geq B_{\max}\} \approx e^{-\theta_I B_{\max}}. \tag{9.14}$$

In order to calculate the loss probability of P-frames in the Intelligent Dropping scheme, it is necessary to note that because of the priority queuing system, in this scheme, P-frames have lower priority than the I-frames, and thus the maximum size of the P-frame sub-queue at any time, t, is $Q_P^{\max}(t) = B^{\max} - Q_I(t)$. By using Equation (9.7), the probability of queue overflow and thus packet loss of the P-frame traffic flow in the Intelligent Dropping scheme is

$$\begin{aligned} Pr_{loss}^{ID}(P) &= \sup_t \Pr\{Q_P(t) \geq B_{\max} - Q_I(t)\} \\ &= \sup_t \Pr\{Q_P(t) \geq B_{\max} - Q_I(t) \mid Q_I(t) = Q_I\}. \Pr\{Q_I(t) = Q_I\} \\ &= \int_0^{B_{\max}} (1 - F_{Q_P}(B_{\max} - q)).f_{Q_I}(q)dq \end{aligned} \tag{9.15}$$

where F_{Q_P} is the Cumulative Distribution Function(CDF) of the size of P-frame sub-queue in the Intelligent Dropping scheme. From Equation (9.7), it is deduced that

$$\begin{aligned} F_{Q_P}(B_{\max} - q) &= Pr\{Q_P \leq (B_{\max} - q)\} \\ &\approx 1 - e^{-\theta_P(B_{\max} - q)}. \end{aligned} \tag{9.16}$$

f_{Q_I} is the Probability Distribution Function(PDF) of the size of I-frame sub-queue in the Intelligent Dropping scheme, and it is thus calculated by taking derivative of the CDF of the size of I-frame sub-queue.

$$\begin{aligned} f_{Q_I}(q) &= \frac{d}{dq}(F_{Q_I}) \\ &\approx \frac{d}{dq}(1 - e^{-\theta_I q}) = \theta_I e^{-\theta_I q} \end{aligned} \tag{9.17}$$

By substituting Equations (9.16) and (9.17) in Equation (9.15), the probability of queue overflow and thus packet loss of the P-frame traffic flow in the Intelligent Dropping scheme is

$$\begin{aligned} Pr_{loss}^{ID}(P) &= \int_0^{B_{\max}} e^{-\theta_P(B_{\max}-q)}.\theta_I.e^{-\theta_I q}dq \\ &= \theta_I.e^{-\theta_P B_{\max}} \int_0^{B_{\max}} e^{q(\theta_P-\theta_I)}dq \\ &= \left(\frac{\theta_I}{\theta_P-\theta_I}\right)\left(e^{-\theta_I B_{\max}} - e^{-\theta_P B_{\max}}\right). \end{aligned} \tag{9.18}$$

In the special case when $\theta_P = \theta_I$, the values of numerator and denominator are both zero in Equation (9.18), and to calculate the value of $Pr_{loss}^{ID}(P)$ in this case, l'Hôpital's rule is used.

$$\begin{aligned} \lim_{\theta_P \to \theta_I} Pr_{loss}^{ID}(P) &= \left(\frac{\frac{\partial}{\partial(\theta_P)}\left(\theta_I\left(e^{-\theta_I B_{\max}} - e^{-\theta_P B_{\max}}\right)\right)}{\frac{\partial}{\partial(\theta_P)}\left(\theta_P - \theta_I\right)}\right) \\ &= B_{\max}\theta_I e^{-\theta_I B_{\max}} \end{aligned} \tag{9.19}$$

To calculate the loss probability of B-frames in the Intelligent Dropping scheme, it is worth noting that because of the priority queuing system, B-frames have lower priority than both I-frames P-frames. Therefore, the maximum size of a B-frame sub-queue at any time, t, is $Q_P^{\max}(t) = B_{\max} - Q_I(t) - Q_P(t)$. By using Equation (9.7), the probability of queue overflow and thus packet loss of the B-frame traffic flow in the Intelligent Dropping scheme is

$$Pr_{loss}^{ID}(B) = \sup_t \Pr\{Q_B(t) \geq B^{\max} - Q_I(t) - Q_P(t)\} \tag{9.20}$$

$$\begin{aligned} &= \sup_t \Pr\{Q_P(t) \geq B^{\max} - Q_I(t) - Q_P(t) \mid Q_I(t) = Q_I, Q_P(t) = Q_P\} \\ &\quad \times \Pr\{Q_I(t) = Q_I, Q_P(t) = Q_P\} \end{aligned} \tag{9.21}$$

$$= \int_0^{B_{\max}} \int_0^{B_{\max}-q_I} (1 - F_{Q_B}(B_{\max} - q_I - q_P))f_{Q_P}(q_P)f_{Q_I}(q_I)dq_P dq_I \tag{9.22}$$

Similar to Equations (9.16 , 9.17), $F_{Q_B}(B_{\max} - q_I - q_P) = 1 - e^{-\theta_B(B_{\max}-q_I-q_P)}$ and $f_{Q_P}(q_P) = q_P e^{-\theta_P q_P}$. The probability of queue overflow and thus packet loss of the B-frame traffic flow in the Intelligent Dropping scheme is calculated.

$$\begin{aligned}
Pr_{loss}^{ID}(B) &= \int_0^{B_{\max}} \int_0^{B_{\max}-q_I} e^{-\theta_B(B_{\max}-q_I-q_P)} \theta_I e^{-\theta_I q_I} \theta_P e^{-\theta_P q_P} dq_P dq_I \\
&= \int_0^{B_{\max}} \theta_I \theta_P e^{-\theta_I q_I} \int_0^{B_{\max}-q_I} e^{-\theta_B(B_{\max}-q_I)} e^{-q_P(\theta_P-\theta_B)} dq_P dq_I \\
&= \int_0^{B_{\max}} \theta_I \theta_P e^{-\theta_I q_I} e^{-\theta_B(B_{\max}-q_I)} \int_0^{B_{\max}-q_I} e^{-q_p(\theta_P-\theta_B)} dq_P dq_I \\
&= \int_0^{B_{\max}} \theta_I \theta_P e^{-\theta_B B_{\max}} e^{q_I(\theta_B-\theta_I)} \left(\frac{1}{\theta_B-\theta_p}\right) \left(e^{(\theta_B-\theta_P)(B_{\max}-q_I)} - 1\right) dq_I \\
&= \left(\frac{\theta_I \theta_P}{\theta_B-\theta_p}\right) e^{-\theta_B B_{\max}} \int_0^{B_{\max}} \left(e^{(\theta_B-\theta_P)B_{\max}} e^{q_I(\theta_P-\theta_I)} - e^{q_I(\theta_B-\theta_I)}\right) dq_I \\
&= \left(\frac{\theta_I \theta_P}{(\theta_B-\theta_p)(\theta_P-\theta_I)}\right) \left(e^{-\theta_I B_{\max}} - e^{-\theta_P B_{\max}}\right) \\
&\quad - \left(\frac{\theta_I \theta_P}{(\theta_B-\theta_p)(\theta_B-\theta_I)}\right) \left(e^{-\theta_I B_{\max}} - e^{-\theta_B B_{\max}}\right) \qquad (9.23)
\end{aligned}$$

In a special case when $\theta_B = \theta_P = \theta_I$, the probability of packet loss of the B-frame traffic flow in the Intelligent Dropping scheme is known by calculating

$$\lim_{\substack{\theta_B \to \theta_I \\ \theta_P \to \theta_I}} Pr_{loss}^{ID}(B) = \lim_{\theta_B \to \theta_I} \left(\lim_{\theta_P \to \theta_I} Pr_{loss}^{ID}(B) \right). \qquad (9.24)$$

In this case, the values of the numerators and denominators in Equation (9.23) are zero, and to calculate the value of $Pr_{loss}^{ID}(B)$, l'Hôpital's rule is used again.

$$\begin{aligned}
\lim_{\theta_P \to \theta_I} Pr_{loss}^{ID}(B) &= \lim_{\theta_P \to \theta_I} \left(\left(\frac{\theta_I \theta_P}{(\theta_B-\theta_p)(\theta_P-\theta_I)}\right) \left(e^{-\theta_I B_{\max}} - e^{-\theta_P B_{\max}}\right) \right) - \\
&\quad \lim_{\theta_P \to \theta_I} \left(\left(\frac{\theta_I \theta_P}{(\theta_B-\theta_p)(\theta_B-\theta_I)}\right) \left(e^{-\theta_I B_{\max}} - e^{-\theta_B B_{\max}}\right) \right) \\
&\overset{l'Hopital}{=} \frac{\theta_I^2 B_{\max} e^{-\theta_I B_{\max}}}{\theta_B-\theta_I} - \frac{\theta_I^2 \left(e^{-\theta_I B_{\max}} - e^{-\theta_B B_{\max}}\right)}{(\theta_B-\theta_I)^2} \qquad (9.25)
\end{aligned}$$

By using Equations (9.24) and (9.25), the probability of packet loss of the B-frame traffic flow in the special case is calculated as follows.

$$\begin{aligned}\lim_{\substack{\theta_B \to \theta_I \\ \theta_P \to \theta_I}} Pr_{loss}^{ID}(B) &= \lim_{\theta_B \to \theta_I} \left(\frac{\theta_I^2 B_{\max} e^{-\theta_I B_{\max}}}{\theta_B - \theta_I} - \frac{\theta_I^2 \left(e^{-\theta_I B_{\max}} - e^{-\theta_B B_{\max}} \right)}{\left(\theta_B - \theta_I \right)^2} \right) \\ &= \lim_{\theta_B \to \theta_I} \left(\frac{\left(\theta_B - \theta_I \right) \left(\theta_I^2 B_{\max} e^{-\theta_I B_{\max}} \right) - \theta_I^2 \left(e^{-\theta_I B_{\max}} - e^{-\theta_B B_{\max}} \right)}{\left(\theta_B - \theta_I \right)^2} \right) \\ &\overset{l'Hopital}{=} \lim_{\theta_B \to \theta_I} \left(\frac{\theta_I^2 B_{\max} e^{-\theta_I B_{\max}} - \theta_I^2 B_{\max} e^{-\theta_B B_{\max}}}{2 \left(\theta_B - \theta_I \right)} \right) \\ &\overset{l'Hopital}{=} \frac{1}{2} \theta_I^2 B_{\max}^2 e^{-\theta_I B_{\max}} \qquad (9.26)\end{aligned}$$

9.4.3 Conventional Queueing

In the conventional queueing scheme, only one traffic flow for video streaming is established between the video server and the client. The BS does not evaluate the contents of the video packets and buffers them in a queue with the maximum size of $B_{\max}$. By using Equation (9.7), the probability of queue overflow and thus packet loss of video packets in the conventional queueing scheme is

$$Pr_{loss}^{CQ} = \sup_t \Pr\{Q(t) \geq B_{\max}\} \approx e^{-\theta B_{\max}}. \qquad (9.27)$$

9.5 Delay Analysis

In this section, the theory of effective bandwidth is used to analyze the delay performance of the multi-level service classification and intelligent dropping schemes, and to compare them with that of the conventional video streaming method. As discussed before, the BS has to drop a packet from the queue if its waiting time inside the queue exceeds a maximum limit. The probabilities of excessive delay and thus packet loss in each of the video streaming schemes in studied next.

9.5.1 Multi-level Service Classification

In this scheme, as explained in Section 9.4.1, the BS reserves r_I, r_P, and r_B as the minimum bitrate of corresponding traffic flows. The loss probability of a packet due to excessive queuing delay in the Multi-level Service Classification method is contingent to the admittance to the queue. By using Equation (9.9), excessive delay probabilities of each traffic flow are calculated as follows.

$$\begin{aligned} Pr^{Ml}_{delay|adm}(I) &= \sup_t Pr\{D_I(t) \geq D_I^{\max}|adm^{Ml}(I)\}.Pr\{adm^{Ml}(I)\} \\ &\approx e^{-\theta_I r_I D_I^{\max}} \left(1 - e^{-\theta_I B_I^{\max}}\right)^2. \end{aligned} \tag{9.28}$$

$$\begin{aligned} Pr^{Ml}_{delay|adm}(P) &= \sup_t Pr\{D_P(t) \geq D_P^{\max}|adm^{Ml}(P)\}.Pr\{adm^{Ml}(P)\} \\ &\approx e^{-\theta_P r_P D_P^{\max}} \left(1 - e^{-\theta_P B_P^{\max}}\right)^2. \end{aligned} \tag{9.29}$$

$$\begin{aligned} Pr^{Ml}_{delay|adm}(B) &= \sup_t Pr\{D_B(t) \geq D_B^{\max}|adm^{Ml}(B)\}.Pr\{adm^{Ml}(B)\} \\ &\approx e^{-\theta_B r_B D_B^{\max}} \left(1 - e^{-\theta_B B_B^{\max}}\right)^2. \end{aligned} \tag{9.30}$$

Here, $Pr^{Ml}_{delay|adm}$ represents the excessive delay probability of a video packet in the Multi-level service classification scheme and $Pr\{adm^{Ml}\} = 1 - Pr^{Ml}_{loss}$ is the probability of admittance of a video packet in its corresponding queue in this scheme. $D_I(t)$, $D_P(t)$, and $D_B(t)$ are the maximum waiting times of each traffic queue at the time instant t. D_I^{max}, D_P^{max}, and D_B^{max} are correspondingly the maximum allowed delays for each of the traffic flows of a video stream.

9.5.2 Intelligent Dropping

As discussed in Section 9.4.2, in the Intelligent Dropping scheme, the BS separates the packets of a video stream into three traffic flows according to their frame type, but it considers one buffer with the maximum size of B_{max} to be shared by all of the traffic flows of a video stream. According to Equation (9.13), the maximum size of the queue is equal to the summation of the maximum sizes of all three queues of a video stream.

Because of the priority queuing system in Intelligent Dropping method, the I-frame traffic flow has the highest priority and thus can fully fill up the buffer. However, if the buffer size reaches its maximum value, $B_{\max}$, the incoming packets will be dropped. The loss probability of an I-frame packet due to excessive queuing delay in the Intelligent Dropping scheme is contingent to the admittance to the queue and encountered excessive delay. By using the Equations (9.9) and (9.14),

$$\begin{aligned} Pr^{ID}_{delay|adm}(I) &= \sup_t Pr\{D_I(t) \geq D_I^{\max}|adm^{ID}(I)\}.Pr\{adm^{ID}(I)\} \\ &\approx e^{-\theta_I r_I D_I^{\max}}\left(1 - Pr^{ID}_{loss}(I)\right)^2 = e^{-\theta_I r_I D_I^{\max}}\left(1 - e^{-\theta_I B_{\max}}\right)^2 \quad (9.31) \end{aligned}$$

where $Pr\{adm^{ID}(I)\} = 1 - Pr^{ID}_{loss}(I)$ is the probability of admittance of an I-frame video packet in its corresponding queue in this scheme. Similarly, by using Equations (9.9) and (9.18), the loss probability of a P-frame video packet because of excessive delay in the Intelligent Dropping scheme is calculated as follows.

$$\begin{aligned} Pr^{ID}_{delay|adm}(P) &= \sup_t Pr\{D_P(t) \geq D_P^{\max}|adm^{ID}(P)\}.Pr\{adm^{ID}(P)\} \\ &\approx e^{-\theta_P r_P D_P^{\max}}\left(1 - Pr^{ID}_{loss}(P)\right)^2 \quad (9.32) \\ &= e^{-\theta_P r_P D_P^{\max}}\left[1 - \left(\frac{\theta_I}{\theta_P - \theta_I}\right)\left(e^{-\theta_I B_{\max}} - e^{-\theta_P B_{\max}}\right)\right]^2 \quad (9.33) \end{aligned}$$

In the special case when $\theta_P = \theta_I$, by substituting Equation (9.19) in Equation (9.32), the value of $Pr^{ID}_{delay}(P)$ is calculated.

$$\lim_{\theta_P \to \theta_I} Pr^{ID}_{delay|adm}(P) = e^{-\theta_I r_P D_P^{\max}}\left[1 - B_{\max}\theta_I e^{-\theta_I B_{\max}}\right]^2 \quad (9.34)$$

By using Equations (9.9) and (9.23), the loss probability of a B-frame because of excessive delay in the Intelligent Dropping scheme is calculated as follows.

$$\begin{aligned} Pr^{ID}_{delay|adm}(B) &= \sup_t Pr\{D_B(t) \geq D_B^{\max}|adm^{ID}(B)\}.Pr\{adm^{ID}(B)\} \\ &\approx e^{-\theta_B r_B D_B^{\max}}\left(1 - Pr^{ID}_{loss}(B)\right)^2 \qquad (9.35) \\ &= e^{-\theta_B r_B D_B^{\max}}[1 - \left(\frac{\theta_I \theta_P}{(\theta_B - \theta_p)(\theta_P - \theta_I)}\right)\left(e^{-\theta_I B_{\max}} - e^{-\theta_P B_{\max}}\right) \\ &\quad + \left(\frac{\theta_I \theta_P}{(\theta_B - \theta_p)(\theta_B - \theta_I)}\right)\left(e^{-\theta_I B_{\max}} - e^{-\theta_B B_{\max}}\right)]^2 \qquad (9.36) \end{aligned}$$

In the special case when $\theta_B = \theta_P = \theta_I$, the probability of excessive delay of the B-frame packets in the Intelligent Dropping scheme is known by substituting Equation (9.26) in Equation (9.35).

$$\lim_{\substack{\theta_B \to \theta_I \\ \theta_P \to \theta_I}} Pr^{ID}_{delay|adm}(B) = e^{-\theta_I r_B D_B^{\max}}\left[1 - \frac{1}{2}\theta_I^2 B_{\max}^2 e^{-\theta_I B_{\max}}\right]^2 \qquad (9.37)$$

9.5.3 Conventional Queueing

As discussed before, in the conventional queueing scheme, only one traffic flow for video streaming is established between the video server and the client. The BS does not evaluate the contents of the video packets and buffers them in a queue with the maximum size of $B_{\max}$. Buffered packets will be dropped from the queue if their waiting times exceed the maximum allowed delay, $D_{\max}$. By using Equations (9.9) and (9.27), the loss probability of a video packet because of excessive delay in the conventional queueing scheme is

$$\begin{aligned} Pr^{CQ}_{delay|adm} &= \sup_t \Pr\{D(t) \geq D_{\max}|adm^{CQ}\}.Pr\{adm^{CQ}\} \\ &\approx e^{-r\theta D_{\max}}\left(1 - e^{-\theta B_{\max}}\right)^2. \qquad (9.38) \end{aligned}$$

9.6 Loss and Delay Combination Analysis

In the Multi-level Service Classification and Conventional Queuing schemes, each packet may be dropped as a result of either full queue or long delays; it is worth to estimate the overall probability of dropping a video packet in the previously discussed queuing schemes. If the probability of full queue loss is Pr_{loss} and the probability of excessive delay is Pr_{delay}, the overall probability of packet dropping, Pr_{Drop} is

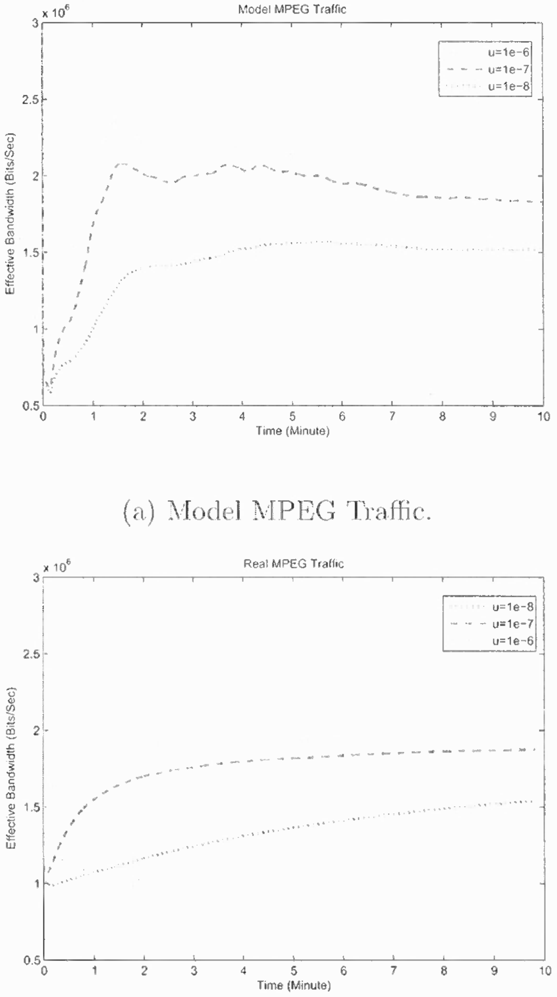
$$Pr_{Drop} \approx Pr_{loss} + Pr_{delay|adm} = 1 - (1 - Pr_{loss})\,(1 - Pr_{delay}) \qquad (9.39)$$

9.7 Performance Comparison

In this section, the performance of the proposed queuing schemes are studied via the introduced analytical framework and compared with the simulation results. The reliability of the analytical framework is also evaluated through comparisons of its results with those of simulations. As discussed earlier in this chapter, the main purpose of the analytical framework is to give the base stations and video servers the insight about streaming performance under the worst case scenario when the BS can provide the traffic flow with the minimum reserved bandwidth. By incorporating the results of this chapter, the BS and video server can negotiate for the optimum guaranteed QoS in which the minimum video quality is guaranteed.

9.7.1 Effective Bandwidth

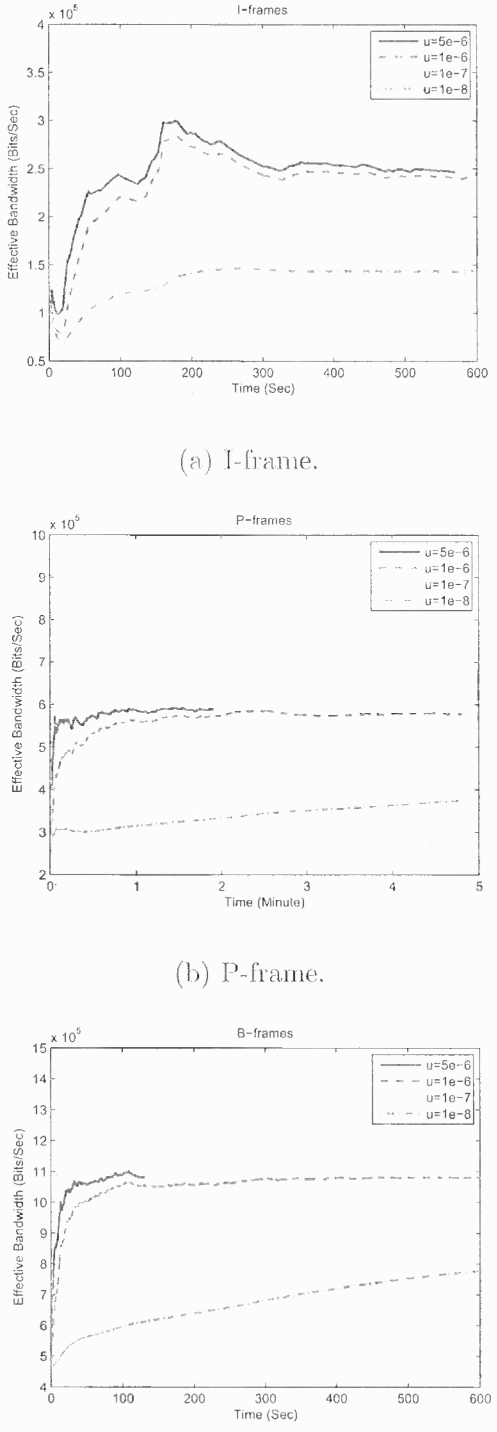
The calculation of effective bandwidth defined in Equation (9.2) requires knowledge about the source traffic. To understand the possible values of effective bandwidth, α, and relations between the QoS component, θ, and service bitrate, r, sample video streams are used. A sample MPEG traffic generated by the traffic modeled discussed in Chapter 5 and a real MPEG video traffic trace from Reference [59] with mean bitrate of $1.95Mbps$ are studied, and Fig. 9.2 shows the variations of effective bandwidth over the duration of video streaming which was equal to 10 minutes. It is observed that both model traffic and real traffic converge to a service bitrate value as a function of QoS component $u = \theta$. For the rest of this chapter, video traffics generated by different models will be used. Fig. 9.3 shows the variation of effective bandwidth for each of the I, P, and B frame traffic generated by the video model. Table 9.1 shows the resulted values of effective bandwidth for each traffic flow as a function of θ.



(a) Model MPEG Traffic.

(b) Real MPEG Traffic.

Figure 9.2 Effective Bandwidth Variation.



(a) I-frame.

(b) P-frame.

(c) B-frame.

Figure 9.3 Effective Bandwidth Variation per Frame Type.

Table 9.1 Effective Bandwidth Results

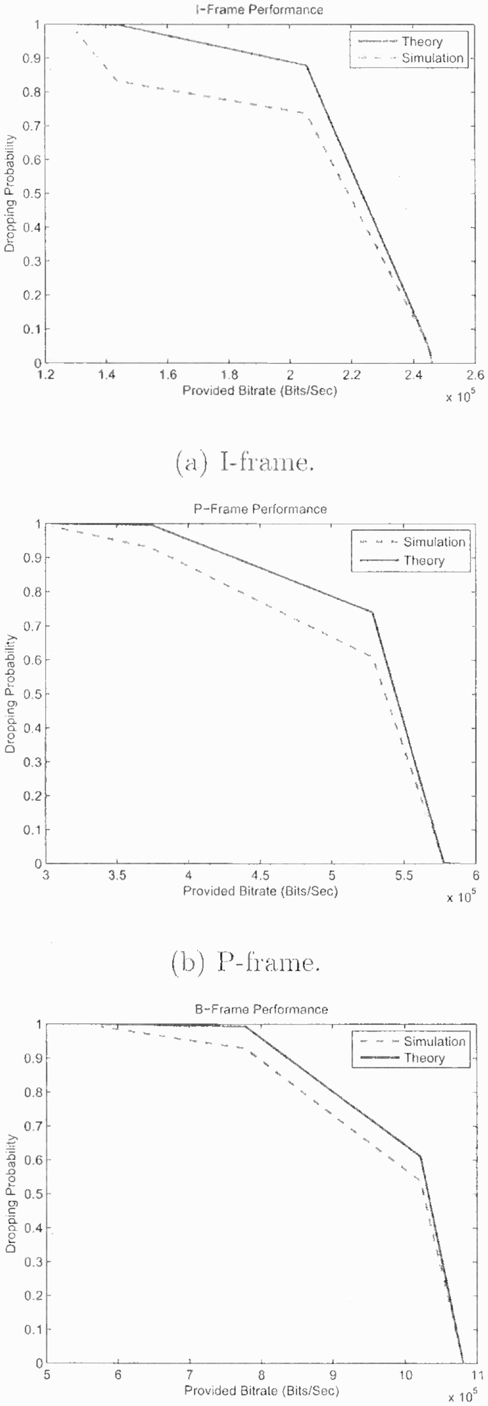
θ	5e-6	1e-6	1e-7	1e-8	1e-9
Effective BW for All (Bits/Sec)	1.94e6	1.93e6	1.876e6	1.54e6	1.071e5
I-Frame Effective BW (Bits/Sec)	2.46e5	2.447e5	2.054e5	1.438e5	1.303e5
P-Frame Effective BW (Bits/Sec)	5.888e5	5.776e5	5.283e5	3.740e5	3.042e5
B-Frame Effective BW (Bits/Sec)	1.081e6	1.080e6	1.022e6	7.788e5	5.757e5

9.7.2 Multi-level Service Classification Performance

The performance of the Multi-level service classification is studied by using analytical framework introduced earlier in this chapter. The performance is also verified by comparing the analytical results with those of the simulation results which are obtained by simulating MPEG traffic generated by the video model. In the simulations, it is supposed that $B_I^{\max} = B_P^{\max} = B_B^{\max} = 1MB$ and $D_I^{\max} = D_P^{\max} = D_B^{\max} = 12Sec$. Fig. 9.4 shows the dropping probability of the frame type in the Multi-level Service classification when different bitrates are guaranteed for each traffic flow. The difference between analytical results and simulation results is less for the B-frames as there are more B-frame packets during 10 minutes of simulation.

9.7.3 Intelligent Dropping Performance

In order to evaluate the performance of the introduced analytical framework of Intelligent Dropping Scheme, simulations are performed over video traffic generated by the traffic model. In the simulations, it is assumed that $B^{\max} = 3MB$ and $D_I^{\max} = D_P^{\max} = D_B^{\max} = 12Sec$. In this scheme, the bitrates of higher priority traffic flows impact on the performance of the lower priority frames. It is thus possible to run various simulations for various scenarios by independently choosing different bitrates for each traffic flow and observing the performance of the analytical framework in each scenario. The results of the special case scenarios discussed earlier in this chapter are shown here. Fig. 9.5 presents the dropping probability of each traffic flow when equal QoS components are considered for all traffic flows.

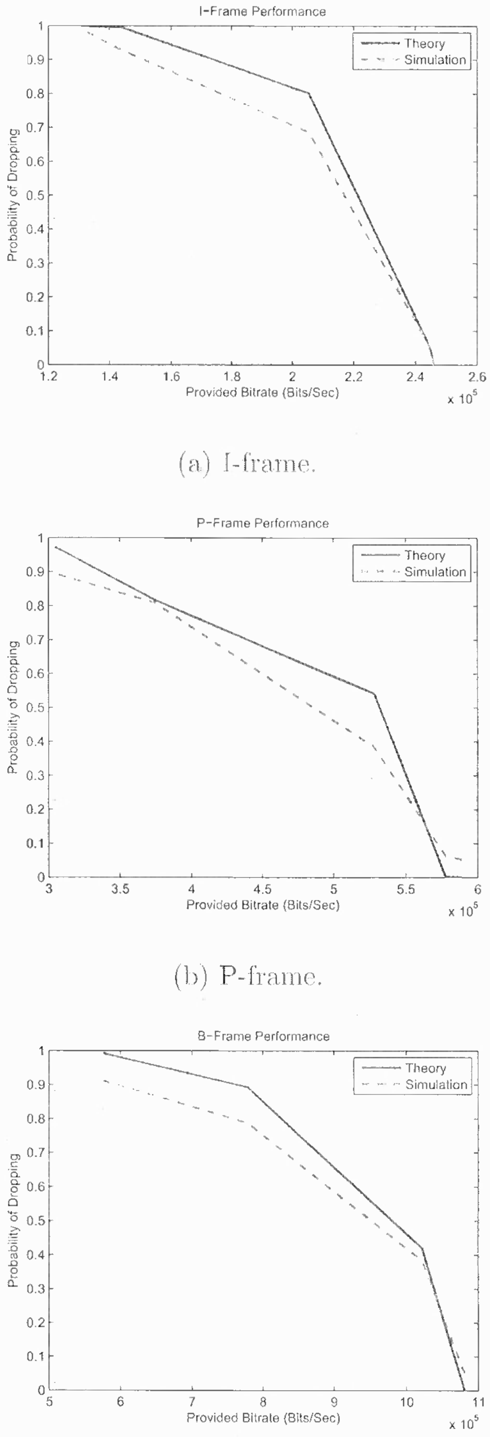


(a) I-frame.

(b) P-frame.

(c) B-frame.

Figure 9.4 Dropping Probability in Multi-level Service Classification Scheme.



(a) I-frame.

(b) P-frame.

(c) B-frame.

Figure 9.5 Dropping Probability in Intelligent Dropping Scheme.

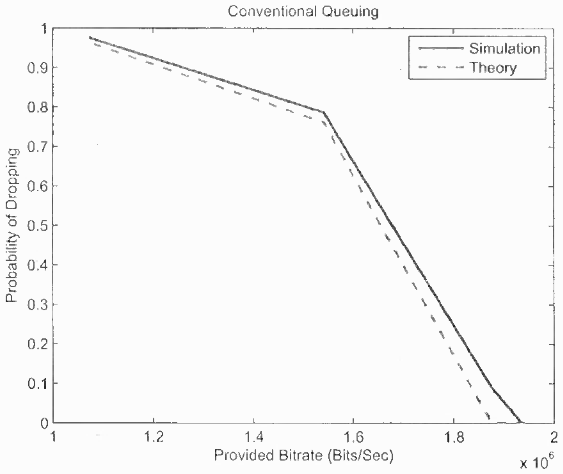


Figure 9.6 Dropping Probability in Conventional Queuing Scheme.

9.7.4 Conventional Queueing Performance

The performance of the analytical framework for Conventional Queueing is also verified by comparing its results with those of simulations. In the simulations, it is assumed that $B^{\max} = 3MB$ and $D^{\max} = 12Sec$. Fig. 9.6 presents the dropping probability of video packets in the Conventional Queueing scheme.

9.8 Summary

In this chapter, an analytical framework for performance analysis of different video streaming schemes is established. The proposed framework applies the theory of effective bandwidth to calculate the probability of frame dropping in each of the video streaming schemes. By incorporating this framework, the base stations and video servers will gain the knowledge about the performance of the video streaming in each scheme under provided bitrate. They can use this knowledge to choose the optimum bitrate that the BS should provide to each video stream to guarantee the minimum video quality during congestion periods. Understanding the maximum frame dropping rate which is translated into the minimum acceptable video quality is another research work which can complement the analytical framework introduced in this chapter.

CHAPTER 10

CONCLUSION AND FUTURE WORK

This dissertation presents novel solutions for application driven network design in broadband wireless networks. Providing different users with high speed network connectivity will let them run various multimedia applications on their wireless devices. In order to successfully deploy and operate broadband wireless networks, it is crucial to investigate efficient methods for supporting various services and applications in broadband wireless networks. Thus, there is a high demand to comprehend the traffic characteristics and service requirements of different applications.

In Chapter 3, a realistic traffic model for VoIP traffic traversed in the uplink based on real traffic traces is presented. In particular, the model describes the inter-packet time of VoIP traffic. The proposed model incorporates the impact of all protocols above the IP layer and it help us gain better insight on the VoIP traffic profile. Thus, it guides us towards more detailed and accurate traffic modeling for VoIP applications.

Chapter 4 discusses that with the extensive demand for ubiquitous network access more users will prefer to make their voice call through the same network connection. This demand will make the WiMAX system providers, as the promising technology providers for broadband wireless access, to specially consider providing their customers with reliable and qualified voice connections via VoIP applications. It is explained that in order to build an efficient WiMAX network, service providers have to obtain real insight about the behavior of the VoIP traffic in WiMAX networks. Available QoS classes for VoIP in WiMAX networks are described and based on those classes, some MAC layer traffic scheduling methods are examined. Based on the traffic model introduced in Chapter 3, the multi-tap scheduling algorithm for VoIP traffic is proposed. Moreover, the trade off between bandwidth efficiency and delay efficiency in each of the possible scheduling methods is also elaborated. It is shown that based on the status of network resources, using the multi-tap scheduling method can lead to a better delay and bandwidth efficiency.

Video streaming applications are of high interest in broadband wireless networks. The challenges of transmitting video traffic over wireless networks are discussed in Chapter 7. The MPEG-4 traffic model suggested by the WiMAX Forum is described and a cross-layer solution for enhancing the performance of WiMAX networks with respect to MPEG video streaming applications is explained. In the multi-level service classification method, MPEG frames are mapped into three different rtPS classes with different minimum reserved bandwidth parameters at the BS. It is discussed that to sustain the video quality, it is crucial to send as many frames as possible to the end user. Furthermore, it is important to protect the more valuable frames, i.e., the I frames, against dropping. The proposed solution allocates higher minimum reserved bitrate to the rtPS flow corresponding to the I frames and thus decreases the probability of dropping of more valuable frames. In brief, the multi-level service classification method uses the characteristics of MPEG traffic to give priority to the more important frames and protect them against dropping. Besides, it is simple and compatible with the IEEE 802.16 standards, and thus readily deployable. It is shown that the proposed solutions will improve the video quality over WiMAX networks.

As a follow-up to the results reported in Chapter 7, the scheme that incorporates selective dropping of different MPEG frames (i.e., different types of frames are dropped at different queue occupancy thresholds) at the BS is illustrated in Chapter 8. This chapter proposed a queue management strategy in which the BS can intelligently drop less effective frames when congestion happens. In the proposed queueing architecture, a fixed buffer is shared by all three traffic flows of a video stream. The proposed queue architecture applies a priority scheme to protect the more important frames against dropping. In an intelligent dropping fashion, the queue may drop the packets with least priority from the queue to make space for an incoming packet with higher priority. The queue also drops all the packets depending on the dropped packet for decoding. The proposed intelligent dropping scheme can deliver more decodable frames to the user while it protects more effective frames against dropping.

In Chapter 9, an analytical framework for performance analysis of different video streaming schemes is established. The proposed framework applies the theory of effective bandwidth to calculate the probability of frame dropping in all video streaming solutions. By incorporating this framework, the base stations and video servers will gain the knowledge about the performance of the video streaming in each scheme under provided bitrates. They can use this knowledge to choose the optimum bitrate which the BS should provide to each video stream to guarantee the minimum video quality during congestion periods

Understanding the maximum frame dropping rate which is translated into the minimum acceptable video quality is another research work which can complement the analytical framework introduced in this dissertation. A future work in this area can include both objective and subjective investigations on modeling the impact of the video frame loss on the perceived video quality at the end users. Despite extensive research works addressing this problem, acquiring an analytical model for studying the impacts of networking and transmission deficits on the streaming quality is one of the open problems to be addressed in future research efforts.

REFERENCES

[1] IEEE 802.16-2004 WirelessMAN standard for wireless metropolitan area networks, 2004.

[2] Telephonys complete guide to WiMAX. *Telephony Magazine*, May 2004.

[3] IEEE 802.16e air interface for fixed and mobile broadband wireless access systems, Feb 2006.

[4] N. Ansari, H. Liu, Y. Shi, and H. Zhao. On modeling MPEG video traffics. *IEEE Transactions on Broadcasting*, 48(4):337–347, Dec. 2002.

[5] K. Balachandran, D. Calin, E. Kim, and K. M. Rege. Proactive content rate selection for enhanced streaming media quality. In *IEEE Sarnoff Symposium*, pages 1–6, April 2008.

[6] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushyana, and S. Viterbi. CDMA/HDR: a bandwidth efficient high speed wireless data service for nomadic users. *IEEE Communications Magazine*, 38(7):70–77, Jul 2000.

[7] P. Bhagwat, P. Bhattacharya, A. Krishna, and S. Tripathi. Enhancing throughput over wireless LANs using channel state dependent packet scheduling. In *Proceedings IEEE INFOCOM'96*, volume 3, pages 1133 – 1140, Mar. 1996.

[8] D. Bonfiglio, M. Mellia, M. Meo, D. Rossi, and P. Tofanelli. Revealing Skype Traffic: when randomness plays with you. In *Proceedings ACM SIGCOMM 2007*, Aug. 2007.

[9] C.-S. Chang. Stability, queue length and delay of deterministic and stochastic queueing networks. *IEEE Transactions on Automatic Control*, 39:913–931, 1994.

[10] W.-T. Chen, S.-H. Chen, and J.-C. Liu. An efficient QoS guaranteed MAC protocol in wireless ATM networks. Los Alamitos, CA, USA, Jan. 2001.

[11] W.-H. Chiang, W.-C. Xiao, and C.-F. Chou. A performance study of VoIP applications: MSN vs. Skype. In *Proceedings The First Multimedia Communications Workshop (MULTICOMM)*, pages 13 – 18, Jun. 2006.

[12] G. Choudhury, D. Lucantoni, and W. Whitt. Squeezing the most out of ATM. *IEEE Transactions on Communications*, 44(2), Feb 1996.

[13] C. Cicconetti, L. Lenzini, E. Mingozzi, and C. Eklund. Quality of service support in IEEE 802.16 networks. *IEEE Network*, 20(2):50–55, March-April 2006.

[14] M. Dai, Y. Zhang, and D. Loguinov. A unified traffic model for MPEG-4 and H.264 video traces. *IEEE Transactions on Multimedia*, 11(5):1010–1023, Aug. 2009.

[15] J. Daigle and J. Langford. Models for analysis of packet voice communications systems. *IEEE Journal on Selected Areas in Communications*, 4:847 – 855, 1986.

[16] T. D. Dang, B. Sonkoly, and S. Molnar. Fractal analysis and modeling of VoIP traffic. In *Proceedings 11th International Telecommunications Network Strategy and Planning Symposium*, pages 123 – 130, Jun. 2004.

[17] C. Eklund, R. Marks, K. Stanwood, and S. Wang. IEEE standard 802.16: a technical overview of the wirelessman air interface for broadband wireless access. *IEEE Communications Magazine*, 40(6):98–107, June 2002.

[18] J. Evans and D. Everitt. Effective bandwidth-based admission control for multiservice CDMA cellular networks. *IEEE Transactions on Vehicular Technology*, 48(1):36–46, Jan. 1999.

[19] N. Feamster and H. Balakrishnan. Packet loss recovery for streaming video. In *In 12th International Packet Video Workshop*, 2002.

[20] P. Ferguson and G. Huston. *Quality of Service: Delivering QoS on the Internet and in Corporate Networks.* Wiley, New York, NY, 1998.

[21] A. G. Gotsis, N. T. Koutsokeras, and P. Constantinou. PHY- and MAC-aware resource allocation and packet scheduling for single-cell OFDMA packet networks. *Journal of Communications*, 3(4):59–70, 2008.

[22] J. Greengrass, J. Evans, and A. Begen. Not all packets are equal, part 2: The impact of network packet loss on video quality. *IEEE Internet Computing,*, 13(2):74–82, March-April 2009.

[23] A. Gurtov and R. Ludwig. Lifetime packet discard for efficient real-time transport over cellular links. *ACM Mobile Computing and Communications Review*, 7(4):32–45, Oct. 2003.

[24] E. Haghani and N. Ansari. VoIP traffic scheduling in WiMAX networks. In *IEEE Global Telecommunications Conference, 2008. IEEE GLOBECOM 2008.*, pages 1–5, Dec. 2008.

[25] E. Haghani, S. De, and N. Ansari. On modeling VoIP traffic in broadband networks. In *Proc. IEEE Global Telecommunications Conference(Globecom'07)*, pages 1922–1926, Washington DC, Nov. 2007.

[26] E. Haghani, S. Parekh, D. Calin, E. Kim, and N. Ansari. A quality-driven cross-layer solution for mpeg video streaming over WiMAX networks. *IEEE Transactions on Multimedia*, 11(6):1140–1147, Oct. 2009.

[27] M. Hassan and M. Krunz. Video streaming over wireless packet networks: An occupancy-based rate adaptation perspective. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(8):1017–1027, Aug. 2007.

[28] H. Heffes and D. Lucantoni. A markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance. 4(6):856–68, Sept. 1986.

[29] M. Hemy, U. Hengartner, P. Steenkiste, and T. Gross. MPEG system streams in best-effort networks. *IEEE Packet Video*, 1999.

[30] O. I. Hillestad, A. Perkis, V. Genc, S. Murphy, and J. Murphy. Adaptive H.264/MPEG-4 SVC video over IEEE 802.16 broadband wireless networks. In *Proceedings of Packet Video*, pages 26–35, Nov. 2007.

[31] C.-Y. Hsu, A. Ortega, and M. Khansari. Rate control for robust video transmission over burst-error wireless channels. 17(5):756 – 773, May 1999.

[32] W. Jiang and H. Schulzrinne. Analysis of on-off patterns in VoIP and their effect on voice traffic aggregation. In *Proceedings Ninth International Conference on Computer Communications and Networks*, pages 82 – 87, Oct. 2000.

[33] H. Kim and Y. Han. A proportional fair scheduling for multicarrier transmission systems. *IEEE Communications Letters*, 9(3):210–212, March 2005.

[34] H.-S. Kim, H.-M. Nam, J.-Y. Jeong, S.-H. Kim, and S.-J. Ko. Measurement based channel-adaptive video streaming for mobile devices over mobile WiMAX. *IEEE Transactions on Consumer Electronics*, 54:171 – 178, Feb. 2008.

[35] M. Krunz and S. K. Tripathi. On the characterization of VBR MPEG streams. In *ACM SIGMETRICS Performance Evaluation Review*, volume 25, pages 192 – 202, Jun. 1997.

[36] W. Kumwilaisak, Y. Hou, Q. Zhang, W. Zhu, C.-C. Kuo, and Y.-Q. Zhang. A cross-layer quality-of-service mapping architecture for video delivery in wireless networks. *IEEE Journal on Selected Areas in Communications*, 21(10):1685–1698, Dec. 2003.

[37] A. A. Lazar, G. Pacifici, and D. E. Pendarakis. Modeling video sources for real-time scheduling. In *Proceedings of the IEEE GLOBECOM*, pages 835–839, Nov. 1993.

[38] H. Lee, T. Kwon, and D. H. Cho. An enhanced uplink scheduling algorithm based on voice activity for VoIP services in IEEE 802.16d/e system. *IEEE Communications Letters*, 9:44–51, 2005.

[39] Y. J. Liang, J. G. Apostolopoulos, and B. Girod. Analysis of packet loss for compressed video: Effect of burst losses and correlation between error frames. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(7):861–874, Jul. 2008.

[40] A. Matrawy, I. Lambadaris, and C. Huang. MPEG4 traffic modeling using the transform expand sample methodology. In *Proceedings of the IEEE 4th International Workshop on Networked Appliances*, pages 249–256, Jan. 2002.

[41] M. Menth, A. Binzenhfer, and S. Mühleck. Source models for speech traffic revisited. Technical Report 426, University of Wüzburg, Germany, May 2007.

[42] G. Nair, J. Chou, T. Madejski, K. Perycz, D. Putzolu, and J. Sydir. IEEE 802.16 medium access control and service provisioning. *Intel Technology Journal*, 8, Aug. 2004.

[43] P. Pragtong, T. J. Erke, and K. M. Ahmed. Analysis and modeling of VoIP conversation traffic in the real network. In *Proceedings Fifth International Conference on Information, Communications and Signal Processing*, pages 388 – 392, Dec. 2005.

[44] A. Sayenko, O. Alanen, and T. Hämäläinen. Scheduling solution for the IEEE 802.16 base station. *Computer Networks*, 52(1):96–115, 2008.

[45] A. Sayenko, O. Alanen, J. Karhula, and T. Hämäläinen. Ensuring the QoS requirements in 802.16 scheduling. In *MSWiM '06: Proceedings of the 9th ACM international symposium on modeling analysis and simulation of wireless and mobile systems*, pages 108–117, 2006.

[46] N. Scalabrino, F. D. Pellegrini, R. Riggio, A. Maestrini, C. Costa, and I. Chlamtac. Measuring the quality of VoIP traffic on a WiMAX testbed. In *Proceedings TRIDENTCOM 2007*, May 2007.

[47] H. Schwarz, D. Marpe, and T. Wiegand. Overview of the scalable video coding extension of the H.264/AVC standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(9):1103–1120, Sep. 2007.

[48] C. shang Chang and J. A. Thomas. Effective bandwidth in high speed digital networks. *IEEE Journal on Selected Areas in Communications*, 13:1091–1100, 1999.

[49] S. Sharafeddine, A. Riedl, J. Glasmann, and J. Totzke. On traffic characteristics and bandwidth requirements of voice over IP applications. In *Proceedings Eighth IEEE International Symposium on Computers and Communication (ISCC)*, pages 1324 – 1330, Jul. 2003.

[50] M. Shariat, A. Quddus, S. Ghorashi, and R. Tafazolli. Scheduling as an important cross-layer operation for emerging broadband wireless systems. *IEEE Communications Surveys & Tutorials*, 11(2):74–86, 2^{nd} Quarter 2009.

[51] J. She, F. Hou, and P.-H. Ho. An application-driven MAC-layer buffer management with active dropping for real-time video streaming in 802.16 networks. In *Proceedings of 21st International Conference on Advanced Information Networking and Applications AINA*, pages 451 – 458, May 2007.

[52] J. She, X. Yu, F. Hou, P.-H. Ho, and E.-H. Yang. A framework of cross-layer superposition coded multicast for robust IPTV services over

WiMAX. In *Proceedings of IEEE Wireless Communications and Networking Conference(WCNC)*, pages 3139 – 3144, April 2008.

[53] V. Singh and V. Sharma. Efficient and fair scheduling of uplink and downlink in IEEE 802.16 ofdma networks. In *IEEE Wireless Communications and Networking Conference, 2006. WCNC 2006.*, volume 2, pages 984–990, April 2006.

[54] C. So-In, R. Jain, and A. K. A. Tamimi. Scheduling in IEEE 802.16e mobile WiMAX networks: key issues and a survey. *IEEE Journal on Selected Areas in Communications*, 27(2):156–171, 2009.

[55] A. Sukhov, P. Calyam, W. Daly, and A. Illin. Towards an analytical model for characterizing behavior of high-speed VVoIP applications. In *TERENA Networking Conference (TNC)*, Jun. 2005.

[56] W. T. Tan and A. Zakhor. Video multicast using layered FEC and scalable compression. *IEEE Trans. Circuits Syst. Video Technol*, 11:373–386, 2001.

[57] H. Toral-Cruz and D. Torres-Roman. Traffic analysis for IP telephony. In *Proceedings 2nd International Conference on Electrical and Electronics Engineering*, pages 136 – 139, Sept. 2005.

[58] P. van Beek and M. Demircin. Delay-constrained rate adaptation for robust video transmission over home networks. In *IEEE International Conference on Image Processing ICIP*, volume 2, pages 173–6, Sept. 2005.

[59] G. Van der Auwera, P. David, and M. Reisslein. Traffic and quality characterization of single-layer video streams encoded with the H.264/MPEG-4 advanced video coding standard and scalable video coding extension. *IEEE Transactions on Broadcasting*, 54(3):698–718, Sept. 2008.

[60] P. Viswanath, D. Tse, and R. Larioa. Opportunistic beamforming using dumb antennas. *IEEE Transactions on Information Theory*, 48:1277–1294, June 2002.

[61] J. Wang, M. Venkatachalam, and Y. Fang. System architecture and cross-layer optimization of video broadcast over WiMAX. *IEEE Journal on Selected Areas in Communications*, 25(4):712–721, May 2007.

[62] WiMAX Forum. WiMAX system evaluation methodology V2.1, 2008.

[63] D. Wu, Y. Hou, and Y.-Q. Zhang. Scalable video coding and transport over broadband wireless networks. *Proceedings of IEEE*, 89(1):6–20, Jan. 2001.

[64] D. Wu and R. Negi. Effective capacity: a wireless link model for support of quality of service. *IEEE Transactions on Wireless Communications*, 2(4):630–643, Jul. 2003.

[65] G. Yanfeng and H. Aiqun. Bandwidth allocation algorithm of VoIP based on the adaptive linear prediction in the IEEE 802.16 system. In *Proceedings of 6th International Conference on ITS Telecommunications*, pages 16–19, Jun. 2006.

[66] Z.-L. Zhang. *End-to-end support for statistical quality-of-service guarantees in multimedia networks.* PhD thesis, University of Massachusetts, Amherst, Jan. 1997.

[67] D. Zhao and X. Shen. Performance of packet voice transmission using IEEE 802.16 protocol. *IEEE Wireless Communications*, 14:44–51, 2007.

[68] X. Zhu and B. Girod. Video streaming over wireless networks. In *Proc. European Signal Processing Conference (EUSIPCO)*, pages 1462–1466, 2007.