UPLINK SCHEDULING FOR SUPPORTING PACKET VOICE TRAFFIC IN IEEE 802.16 BACKHAUL NETWORKS

UPLINK SCHEDULING FOR SUPPORTING PACKET VOICE TRAFFIC IN IEEE 802.16 BACKHAUL NETWORKS

By

LIZHONG DAI, B. Eng.
Beijing University of Posts and Telecomms, P.R.China

A Thesis
Submitted to the School of Graduate Studies
in Partial Fulfilment of the Requirements
for the Degree
Master of Applied Science

McMaster University
September 2007

MASTER OF APPLIED SCIENCE (2007) MCMASTER UNIVERSITY

(Electrical and Computer Engineering) Hamilton, Ontario

TITLE: **Uplink Scheduling for Supporting Packet Voice Traffic in IEEE 802.16 Backhaul Networks**

AUTHOR: Lizhong Dai
B. Eng.
Beijing University of Posts and Telecomms, P.R.China

SUPERVISOR: Dr. Dongmei Zhao

NUMBER OF PAGES: xiii, 60

Abstract

Wireless metropolitan area networking based on IEEE 802.16 is expected to be widely used for creating wide-area wireless backhaul networks, where each subscriber station (SS) is responsible for forwarding traffic for a number of connections. Quality of Service (QoS) provisioning is an important aspect in such networks. The IEEE 802.16 standard specifies that the bandwidth requests sent by the SS are for individual connections and pass only the number of bytes requested from each connection. This is inefficient for backhaul networks where each SS may be responsible for forwarding packets for a relatively large number of connections and the bandwidth request messages consume much bandwidth unnecessarily. Furthermore, the standard does not include latency information, which makes it difficult for the base station (BS) to schedule real-time traffic.

In this thesis we study real-time voice traffic support in IEEE 802.16-based backhaul networks. We propose a simple enhancement to the bandwidth request mechanism in 802.16 for supporting packet voice traffic. First, the SS combines the bandwidth requests of multiple voice connections, which are associated to it and have the same traffic parameters, and aggregates the bandwidth requests to the BS. This makes the bandwidth request process more efficient by saving transmission time of both the BS and the SSs. Second, in order to facilitate the BS to make resource allocation decisions, the aggregate bandwidth requests include information about the latency requirements of buffered real-time packets at the SSs. We propose three different bandwidth request and packet scheduling schemes, each of which requires a different amount of information in the bandwidth requests. Our results show that the proposed bandwidth request and scheduling schemes achieve significantly lower packet loss probability than standard 802.16 bandwidth requests and weighted round robin. The results further show that there is an optimum point about how much delay information the SS should report to the BS in order to best utilize the uplink resources while providing satisfactory real-time performance for the voice traffic.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, Dr. Dongmei Zhao, for her consistent guidance and support during the process of this thesis work. Without her invaluable insight and patient encouragement, it is not possible for me to complete this thesis. Secondly, I would like to acknowledge Dr. Terry Todd and Dr. Polychronis Koutsakis for their time in reviewing this thesis and valuable comments.

It is my real pleasure to have the opportunity staying with members of wireless networking laboratory in the past two years. I am very grateful for their help and friendship.

Last but not the least, I would also like to thank all of the professors and staff members in the department, who always seemed to be there when I need help, for their kindly opinion and consistent support during my graduate study.

Contents

List of Figures

List of Tables

List of Abbreviations

ARQ	Automatic Repeat Request
ATM	Asynchronous Transfer Mode
BE	Best Effort
BER	Bit Error Rate
BR	Bandwidth Request
BS	Base Station
BW	Bandwidth
BWA	Broadband Access network
CBFQ	Credit Based Fair Queueing
CID	Connection Identifier
CPS	Common Part Sublayer
CRC	Cyclic Redundancy Check
CS	Service-Specific Convergence Sublayer
DL	Downlink
DOCSIS	Data Over Cable Service Interface Specification
DRR	Deficit Round Robin
DS	Dynamic Service
DSL	Digital Subscriber Line
EDD	Earliest Due Date

EDF	Earliest Deadline First
ertPS	Extended Real-time Polling Service
FBWA	Fixed Broadband wireless Access
FCFS	First Come First Serve
FDBI	Full Delay Budget Information
FDD	Frequency Division Duplexing
FDM	Frequency Division Multiplexing
FDMA	Frequency Division Multiple Access
FIFO	First In First Out
FQ	Fair Queueing
GPS	General Process Sharing
HDBI	Head-of-line-packet Delay Budget Information
HO	Handover
hrtPS	Hybrid Real-time Polling Service
MAC	Medium Access Control
NLOS	Non Line of Sight
nrtPS	Non-real-time Polling Service
OFDMA	Orthogonal Frequency Division Multiple Access
OSI	Open Systems Interconnection
PDBI	Partial Delay Budget Information
PDU	Protocol Data Units
PHY	Physical Layer
PMP	Point to Multi-point
PEDD	Proactive Earliest Due Date
QAM	Quadrature Amplitude Modulation
QoS	Quality of Service

QPSK	Quadrature Phase Shift Keying
RF	Radio Frequency
rtPS	Real-time Polling Service
SDU	Service Data Units
SS	Subscriber Station
TDD	Time Division Duplexing
TDM	Time Division Multiplexing
TDMA	Time Division Multiple Access
UGS	Unsolicited Grant Service
UL	Uplink
VoIP	Voice over IP
WC-EDD	Work Conserving-Early Due Date
WFQ	Weighted Fair Queue
WIFI	Wireless Fidelity
WIMAX	Worldwide Interoperability for Microwave Access
WLAN	Wireless Local-area Network
WMAN	Wireless Metropolitan-area Network
WPAN	Wireless Personal Area Network
WRR	Weighted Round Robin

Chapter 1

Introduction

1.1 Overview

Wireless metropolitan area networks (MANs) based on IEEE 802.16, also known as WiMax[1], are expected to have wide deployment in the near future. Beyond just providing a single last hop access to wireline backbone networks, such as the Internet, the IEEE 802.16-based technology can be used for creating wide-area wireless backhaul networks, such as backhaul for connecting radio network controllers with base stations in cellular networks and for connecting Wi-Fi-based routers, for coverage extension with rapid and low-cost deployment. The IEEE 802.16 standard is designed for point-to-multipoint configurations, where several subscriber stations (SSs) are associated with a central base station (BS). Optional mesh deployment is also available, where SSs can communicate with each other. When used for backhaul transmissions in either a point-to-multipoint or mesh mode, each SS is usually responsible for forwarding traffic for more than one connection. Quality of service (QoS) provisioning, such as packet transmission scheduling, is one of the important topics for supporting multimedia services in such networks.

[1] WiMax Forum, a nonprofit organization, was established in 2001 with an aim to support wireless metropolitan area networking products on IEEE 802.16 basis.

There are mainly two types of packet scheduling schemes. The first type is based on the general process sharing (GPS), such as weighted fair queueing (WFQ) or weighted round robin (WRR), and aims at achieving fair throughput for best effort traffic. Since no latency information is involved in making scheduling decisions, they generally result in poor delay (or packet loss due to intolerable delay) performance. The other type of scheduling is for real-time traffic and scheduling decisions are made based on latency requirements of packets. Among all the real-time scheduling schemes earliest deadline first (EDF) scheduling scheme [5] is known to provide the optimal delay performance in the deterministic environment. It is also shown [6] [8] that the advantages of EDF over GPS-based scheduling are carried over to the statistical setting. One of the features of uplink traffic scheduling in all wireless networks is that stations distributed in different places of the network should send bandwidth requests to the central station, which otherwise does not have information of the current backlogged packets. Extensive work has been done in this area. In most of the scheduling schemes, the bandwidth requests pass only the number of packets (or bytes) requested from a particular connection. The priority of packet scheduling, such as the ones in [9] [10], is based on QoS requirements of each connection, instead of individual packets. Statistical delay performance, e.g., mean packet transmission delay and probability of packet losses due to longer transmission delay then tolerable, can be achieved if the long-term statistical properties of the packet arrival process of the real-time connections are known at the BS. However, this information is difficult to obtain in a backhaul network, where packets may traverse other networks, e.g., IEEE 802.11-based wireless local area network (WLAN), via one or more hops and experience random and variable delay before arriving at the buffer of an SS in an 802.16 network. The tolerable delay for packets from different connections may be significantly different after they arrive at the 802.16 SSs, even if they may have exactly the same end-to-end performance requirement. To make the problem even more challenging, packets from the same connection may experience completely different latency before arriving at the 802.16 SS and have different delay budgets in their transmissions to the 802.16 BS. Therefore, the bandwidth request messages in an 802.16-based backhaul network should include information of latency requirement of the backlogged packets. The IEEE 802.16 standard leaves details of scheduling and reservation management for the vendors to differentiate their equipment.

1.2 Motivation and Overview of the Proposed Work

In the IEEE 802.16 standard each bandwidth request message from an SS is for a single connection and specifies the ID number of the connection and the number of bytes that the connection is requesting to transmit. With only the number of bytes requested from each connection, information provided in the bandwidth request messages is insufficient for the BS to make accurate scheduling decisions for real-time services. When there are multiple connections associated to an SS, the SS should send a different bandwidth request message to the BS for each individual connection. This bandwidth request mechanism is neither efficient nor effective, especially when there are multiple connections associated to one SS. In this thesis, we propose a more efficient bandwidth request mechanism which aggregates the bandwidth requests of multiple connections with the same traffic parameters. In order to effectively support real-time traffic, the bandwidth request messages incorporate a certain amount of latency-related information about the buffered real-time packets. The 802.16 standard specifies that the bandwidth grant from the BS is aggregate to the SS but not explicitly to individual connections. In the case when the SS is granted a less amount of resources than requested, the SS decides how the available resources are allocated among its associated connections.

Besides the mismatch between individual bandwidth requests and aggregate bandwidth grants, the 802.16-based networks also have some unique features that should be taken into consideration when designing a QoS provisioning scheme. First, the uplink operation in

802.16 is TDMA-based, and resource allocation decisions for the uplink transmissions are done before every uplink subframe. Therefore, the frequency at which resource allocations are updated in an 802.16-based network is limited by the MAC frame duration, and a longer MAC frame results in slower resource allocation updates, which may negatively affect the QoS provisioning, especially for real-time traffic. Second, each SS can only be granted with one transmission burst in every uplink subframe. Since the BS cannot switch back and forth between different SSs during an uplink subframe, every time when an SS is permitted to transmit, it should transmit as many packets as it is allowed for the whole MAC frame.

In this thesis, we propose a simple enhancement to the bandwidth request mechanism in the 802.16 and design resource allocation and scheduling schemes for supporting real-time traffic. Both the bandwidth requests and grants are aggregate for multiple connections associated to the same SS. The bandwidth request process is a simple extension to the real-time polling service (rtPS) or extended real-time polling service (ertPS) defined in the IEEE 802.16 protocol. Each bandwidth request message includes a certain amount of information regarding the number of buffered real-time packets and their latency information. A scheduling scheme is performed at the BS based on the information received from the SSs to coordinate the resource allocations among different SSs. Three schemes are proposed for the SSs to make bandwidth requests and for the BS to make scheduling decisions, each requiring a different amount of information in the bandwidth request messages. In the scheme with full delay budget information (FDBI), each SS passes the delay budgets of all real-time packets in its buffer; in the scheme with delay budget information of the head-of-line packet (HDBI), the SS passes the least delay budget of all its buffered packets; and in the scheme with partial delay budget information (PDBI), the SS informs the BS of the number of packets with their delay budget falling in certain intervals.

1.3 Thesis Organization

The remainder of this thesis is organized as follows. The IEEE 802.16 protocol is introduced in Chapter 2, where QoS provisioning, scheduling, and bandwidth requests in supporting real-time services in 802.16 are briefly described. Chapter 3 presents the enhanced bandwidth request messages and three uplink scheduling schemes. An analytical model is developed in Chapter 4 for analyzing the packet loss rate performance of the PDBI scheme. Simulation results are demonstrated in Chapter 5, where we compare the performance of the three real-time scheduling schemes and compare the PDBI scheme with the WRR scheme in supporting real-time voice traffic. Chapter 6 concludes the thesis.

Chapter 2

Introduction to the IEEE 802.16 Standard

This chapter begins with a brief overview of the IEEE 802.16 protocol, followed by a description of the 802.16 MAC frame structure. The QoS provisioning and scheduling services in the 802.16 standard are introduced, followed by a description of the bandwidth request and grant mechanism in 802.16. Finally some research work related to scheduling and resource management for supporting real-time voice application in 802.16 networks is presented.

2.1 Overview of the Standard

With the increasing demand for fixed broadband wireless access (FBWA) systems, the IEEE 802.16 Working Group was formed in 1998 and the first version of the IEEE 802.16 standard was completed in October 2001. Driven by the need for non-line-of-sight (NLOS) operation, the IEEE 802.16a, an amendment of the first version, extends the air interface support to 2-11GHz band, including both licensed and license-exempt spectra. Approved in 2004, the IEEE 802.16d incorporates 802.16a with a target to provide last mile broadband access alternatives to DSL and cable modems, thereby it is also known as the fixed WiMax standard [1]. In order to add mobility features based on IEEE 802.16d and offer an option to complement existing mobile networks, the 802.16e standard was approved in Dec. 2005 by IEEE as the mobile WiMax standard [2].

The IEEE 802.16 physical layer (PHY) operates either at 10-66 GHz or 2-11 GHz band. It adopts multiple modulations and coding schemes to enhance the data transmission performance. In the 10-66 GHz band, line-of-sight (LOS) propagation and single carrier modulation is used, and the air interface is referred as WirelessMAN-SC. In the 2-11 GHz band, three different air interfaces for non-line-of-sight (NLOS) communication are defined: (i) WirelessMAN-SCa for single-carrier modulation, (ii) WirelessMAN-OFDM for OFDM-based transmission with time-division multiple access (TDMA), which uses OFDM with 256 carriers, and (iii) WirelessMAN-OFDMA for orthogonal frequency-division multiple access (OFDMA), which uses OFDMA with 2048 and 4096 carriers. Multiple modulations such as QPSK, 16-QAM and 64-QAM can be selected based on specific channel conditions.

The IEEE 802.16 MAC is a connection-oriented protocol that defines a variety of mechanisms for the SSs to exchange data from the central BS. All SSs synchronize with the BS clock. The IEEE 802.16 MAC is mainly designed for point-to-multipoint (PMP) configurations, where several SSs are associated with a BS. In the PMP topology all SSs communicate only with the BS. The BS occupies the full downlink channel and broadcasts to all associated SSs. Optional mesh deployment is also available, where SSs can communicate between each other, and thus extend the coverage range.

In the backhaul network, each SS is associated with multiple connections. Resource allocation, admission, and scheduling are controlled by the BS. Queue states and QoS requirements for the uplink traffic are obtained through the bandwidth requests process.

2.2 MAC Layer Frame Structure

The 802.16 MAC supports both Time Division Duplexing (TDD) and Frequency Division Duplexing (FDD) configurations. In the TDD case, the MAC uses one single radio carrier, which is shared by both the uplink and downlink. There is guard time between the uplink and downlink subframes. Each MAC frame includes a downlink subframe followed by an uplink subframe. The BS determines the lengths of the downlink and uplink subframes on a per frame basis, and broadcasts the frame structure and other management information to the SSs through downlink and uplink map messages (UL-MAP and DL-MAP) at the beginning of each frame. The downlink transmission is based on time-division multiplexing (TDM). Each downlink subframe may consist of multiple downlink bursts. Each downlink burst is a sequence of data that use identical physical mode (such as modulation and encoding schemes) and are destined to one or multiple SSs. The first downlink burst contains the channel descriptors and UL-MAP and DL-MAP messages. The channel descriptor is used to provide characteristics of the physical channel, and UL-MAP and DL-MAP to specify bandwidth allocations in the uplink and downlink, respectively, such as timeline allocations of the bursts, burst-to-connection mapping list, and burst-based physical modes. Each individual SS receives all the downlink stream from the BS but is only able to extract the data addressed to itself. The uplink transmission is based upon time division multiple access (TDMA) mechanism. Each uplink subframe consists of one or multiple uplink bursts for carrying uplink user data. Each uplink burst corresponds to one individual SS and is designed to carry variable-length MAC protocol data units (PDUs). The uplink subframe also includes contention periods for bandwidth requests and initial ranging. All SSs follow the instructions of the management messages from the BS and make the appropriate actions. Each transmission burst is separated from the others by a preamble field and contains several MAC PDUs. The synchronization between the SS and BS is done by using predefined time slot.

In the FDD case, the 802.16 MAC uses a fixed duration frame in both the uplink and downlink transmissions, and both full-duplex and half-duplex are supported. The MAC can transmit simultaneously in both the uplink and downlink directions over separate carriers at different frequencies, whereas the uplink subframe should be slightly delayed with respect to the downlink subframe so that the SSs can receive necessary information about the uplink channel access from the downlink. In an FDD frame structure, the downlink and uplink subframes are allocated in a different frequency band without necessity of guard time.

2.3 802.16 QoS Provisioning and Uplink Scheduling

A robust QoS strategy is crucial for accomplishing real-time services in a wireless network. The IEEE 802.16 standard defines a wide variety of mechanisms in both the PHY and MAC layers in order to provide end-to-end QoS provisioning.

The PHY adopts adaptive burst profiling mechanism and can dynamically assign burst profiles to each uplink or downlink burst in both the TDD and FDD configurations depending on link conditions. Multiple modulation and coding schemes are employed to adjust various transmission parameters for each individual SS on a per frame basis.

The 802.16 MAC is connection-oriented. All services are mapped to connections and the BS assigns each connection with a unique connection ID (CID), and this is true even for inherently connectionless services. Each transport connection is associated with one particular QoS level, which corresponds to a set of QoS parameters such as latency, jitter, and throughput assurances. Through this the MAC realizes the transmission scheduling on the air interface, thus providing end-to-end QoS to manage the performance of the whole transmission. The MAC defines a service flow by mapping it to a MAC connection (one connection per service flow). Service flows provide a mechanism for uplink and downlink QoS management, which applies to all processes including making bandwidth requests, associating QoS parameters, and delivering data.

To complement the QoS implementation, the MAC defines five scheduling services for the uplink operations: unsolicited grant service (UGS), real-time polling service (rtPS), extended real-time polling service (ertPS), non-real-time polling service (nrtPS), and best effort (BE) service. The first three scheduling services are designed for real-time traffic, and the rest two are for traffic without strict delay performance requirements.

UGS

The UGS scheduling service is designed to support real-time service flows that generate fixed size data packets periodically and expected to suit real-time traffic with the most stringent delay requirement, such as T1/E1 and Voice over IP without silence suppression. In this service the BS offers a fixed size burst in time slots to an SS periodically, and it is not necessary for the SS to make any explicit bandwidth requests. The bandwidth grant is negotiated in the initialization process of the communication session. Thus, this scheduling service can minimize MAC overhead and uplink access delay caused by the bandwidth request process of the SSs and achieve the best delay performance. However, the BS must provide fixed size data grants at periodic intervals to the UGS flows, and the reserved bandwidth may be wasted when a corresponding UGS flow is inactive. Assigning fixed size bandwidth grants to a voice connection can waste the uplink resources when the connection is in silent periods.

rtPS

The rtPS scheduling is designed for real-time uplink service with variable packet generation rates such as MPEG video and offers real-time, periodic, and unicast request opportunities. In this service, the rtPS flows are polled by the BS regardless of the network load. The polling rate should meet the QoS requirements, such as delay and packet loss rate, of the flows. The BS allows the subordinate SSs to make bandwidth requests at specified uplink time slots designated by the BS via the polling process. In receipt of the requests, the BS makes decisions on bandwidth allocations and broadcasts the bandwidth grants to the subordinate SSs. For voice applications, the BS assigns uplink time periods that are sufficient for unicast bandwidth requests to the voice connections. These periods are negotiated in the initialization process of the voice sessions. The rtPS service can optimize the data transport efficiency due to its capability of supporting variable grant sizes, however, it has larger MAC overhead and longer access delay than the UGS service.

ertPS

The ertPS is a new addition in IEEE 802.16e. The ertPS scheduling service is designed to support real-time service flows that generate variable rate data packets on a periodic basis, such as VoIP services with silence suppression. It is intended to combine the low latency performance of the UGS service and flexibility of the rtPS service for supporting real-time services. The BS keeps offering the same amount of bandwidth to the SS and does not have to poll the SS unless explicitly requested by the SS. If the SS generates packets at a constant rate, the ertPS service works in the same way as the UGS service and thus saves the latency of making bandwidth requests. On the other hand, when the packet generation rate from the source is changed, the SS can update the bandwidth request change, like in the rtPS service, and therefore prevent uplink resource waste as in the UGS algorithm.

nrtPS

The nrtPS service is designed to support time-insensitive data streams with variable rate data packets. This service has certain guaranteed minimum throughput by means of the Minimum Reserved Traffic Rate parameter, suitable for Internet applications such as FTP and HTTP. The BS determines how to perform the non-real-time polling for different connections, and what mechanism is used to serve delay-tolerable traffic.

BE

The BE service is designed to support the best effort uplink traffic without any QoS guarantees on delay or throughput, such as emails. The bandwidth allocation to BE applications is subject to the bandwidth distribution policy for the other scheduling service classes. In particular, the BE traffic receives the residual bandwidth after bandwidth assignment to the other service classes.

2.4 Bandwidth Request Mechanisms

Bandwidth request and grants process is founded on the concept of connections. An SS sends bandwidth request messages via the header of its uplink burst to the BS on a per connection basis. After receiving the bandwidth request messages, the BS then makes decisions and broadcasts the bandwidth grant information to the SSs via UL-MAP transmitted in the beginning of each downlink subframe. The SS then transmits user data in the specified time slots. The bandwidth assignment decision is based on the resource availability at the BS and bandwidth request information from all SSs. Only in UGS service flows, the SS does not need to send bandwidth requests because the grant is unsolicited. Each SS has a Basic CID, and each bandwidth grant is addressed to the SS's Basic CID, not to individual CIDs. That means, the BS allocates bandwidth to an SS as an aggregate in response to per-connection request from each SS. Since it is non-deterministic which request is being honored, the BS may grant an amount of bandwidth less than the total required from the SS, and the SS may redistribute bandwidth among the associated connections and maintain the QoS policy at local site.

The 802.16 MAC supports several bandwidth requests. In the uplink direction, there are two modes of bandwidth requests: contention mode and contention-free (polling) mode. In the contention mode, the SSs send bandwidth request messages during the contention period, and contentions are resolved using back-off resolution. In the contention-free mode, the BS polls each SS by offering an amount of bandwidth that is sufficient for the SS to send a bandwidth request message. This polling information is indicated in the UL-MAP field through the downlink transmission. The SS responds to the polling by sending a bandwidth request message in the specified uplink time slots as stand-alone packets or piggybacked with other packets. The BS may poll the SS individually or in a group. There are three types of polling: unicast polling, multicast and broadcast polling, and station initiated polling. Unicast polls are used to check for inactive stations. Multicast and broadcast polls are used to poll a group or all SSs if required. Station initiated polls are used at request from the SS.

In the five scheduling services, except UGS, all other four scheduling services need to make bandwidth requests, and have the queues to collect packets waiting for transmission. The rtPS may use the real-time polling mechanism, which is flexible and time delay guaranteed but requires some overhead to offer periodic dedicated request opportunities. The ertPS service uses the same bandwidth request mechanism as the rtPS when the SS needs to update the bandwidth requests. Both nrtPS and BE typically employ contention-based mechanisms to make bandwidth requests in response to broadcast/multicast polls advertised by the BS.

2.5 Related Work on Resource Allocations in 802.16 Networks

Resource management is a key part for supporting multimedia applications, but details of resource allocations are not specified in the 802.16 standard. There have been some efforts to investigate, evaluate, and revise the IEEE 802.16 MAC protocols in order to support multimedia applications especially real-time services.

In [31], the authors present three polling methods, i.e., the SSs are polled at the end of every uplink subframe, or polled at the start of the uplink subframe, or piggybacked polls. Meanwhile analytic models are also developed to evaluate the delay performance in accommodating multimedia traffic, where the impact of various system parameters like the MAC subframe lengths is also considered. In [32] the authors compared two bandwidth request mechanisms specified in the standard, contention based random access vs. polling, under both error-free and error-prone channel conditions, and investigated the influence of channel noise on the bandwidth request mechanisms. The performance of a polling-based bandwidth request mechanism is evaluated in [13], where an analytical model is proposed to investigate the resource utilization and multicast and broadcast polling mechanisms are considered.

A queue-aware bandwidth allocation and rate control mechanism is proposed in [19] for polling services. In [21] QoS support in 802.16 MAC is studied, and an admission control scheme, which uses bandwidth cross-allocation between different QoS levels, is proposed based on the scheduling services defined in the 802.16 specifications. Instead of focusing on central resource management in the BS, in [34] the authors investigate the reservation scheme in the mesh mode and propose an analytical model for a distributed scheduling algorithm. In [35] a bandwidth allocation and admission control algorithm is presented for an WiMAX network with special consideration on tele-medicine traffic. In the QoS performance study of [4], the authors propose a scheduling scheme to support multimedia applications in 802.16 PMP networks, where DRR is used as the downlink scheduler and WRR as the uplink scheduler. Since this scheme uses piggybacked bandwidth request mechanism to request uplink bandwidth for real time traffic, the overall transmission performance of real-time traffic is highly dependent on the delay introduced by the bandwidth request mechanism. A distributed scheduling scheme is proposed in [14], where schedulers are implemented at both the BS side and the SS side, and two-level queues are defined to differentiate the traffic priority. The higher priority queue is designed to accommodate the time-delay-critical traffic by adopting the UGS service and the First-in-first-out (FIFO) discipline, and the lower priority queue is designed to dispatch less-time-sensitive traffic with a guarantee of the minimum reserved bandwidth for each service flow. A scheduling scheme based on fair queueing (FQ) is introduced in [17] for both IEEE 802.16 and DOCSIS scenarios.

Real-time traffic is going to occupy a bigger portion of the total load of the existing packet radio networks, and hence providing services with satisfactory QoS requirements becomes a critical task to researchers and developers. Some research work on real-time scheduling in 802.16-based networks has been done in the literature. In [27] a scheduling scheme is proposed for voice over IP (VoIP) connections with alternate ON and OFF activities, and one reserved bit is used in the MAC header for the SS to inform the BS of the status transitions of its voice connection. In the quality of service performance study of [4], deficit round robin and weighted round robin, respectively, are used for downlink and uplink scheduling. The minimum reserved rate of a VoIP connection is computed as the sum of the VoIP sources peak rates. Reference [28] proposes a hybrid real-time polling service (hrtPS) for packet voice traffic with alternative active and silent periods in IEEE 802.16-based backhaul networks. Ref. [20] develops a bandwidth report algorithm with a model to pre-estimate the packets arrivals for supporting rtPS service. In [15] a hybrid scheduling scheme is proposed to support real time traffic and non-real-time traffic using earliest due date (EDD) and WFQ algorithm, respectively. The unsolicited grant service (UGS) is employed to support real-time packet voice and the packet-based end-to-end delay performance is evaluated. Several scheduling schemes based on FIFO EDD, Preemptive EDD and Round Robin are compared in [18] for the 802.16 networks. The performance of EDF is compared with WFQ in [11] when used for IEEE 802.16-based networks. A bandwidth allocation and admission control scheme is proposed in [12] for real-time and non-real-time polling services in 802.16 based networks, where the amount of bandwidth offered to a new connection is determined by the Nash equilibrium and game theory techniques are applied.

Currently, the available real-time scheduling schemes do not consider efficient resource utilization and strict latency guarantee. In the next chapter, we propose an enhancement to the 802.16 bandwidth request message, and based on which design three packet scheduling schemes for support real-time voice traffic in 802.16-based backhaul networks.

Chapter 3

Proposed Real-Time Scheduling Schemes

In this chapter, we propose and study three uplink scheduling schemes for supporting real-time voice in an 802.16-based backhaul network. We first describe the system that this work is based on, and then propose an enhanced bandwidth request format to support real-time voice connections. Thereafter follow the uplink scheduling schemes.

3.1 System Description

We consider an IEEE 802.16-based backhaul network as shown in Figure 3.1, where there is one BS associated with S SSs, denoted as $i = 1, 2, \ldots, S$. Each SS in the backhaul network may be connected to a WLAN access point or a cellular radio network controller, and be responsible for forwarding packets for a number of connections. We emphasize voice traffic in this work. The system may also have other types of traffic, such as real-time video traffic and best effort traffic. Having real-time video traffic in the same network affects the amount of available resources for voice traffic. However, we assume that a certain amount of the uplink resources is reserved for the voice traffic. Having best effort traffic

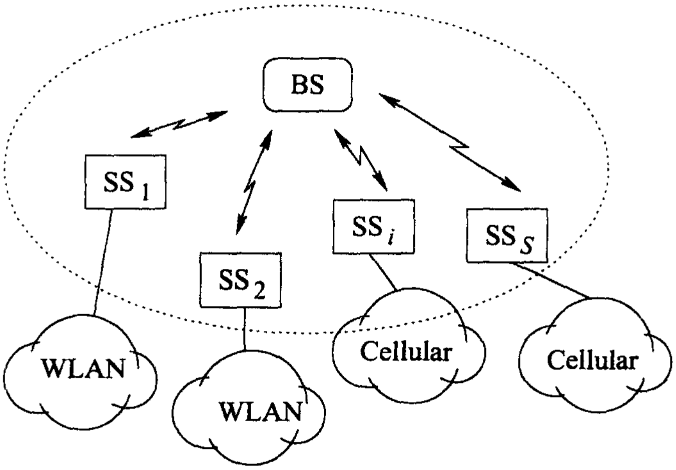


Figure 3.1: IEEE 802.16-based backhaul network

in the network should not affect the voice traffic performance, since voice traffic is usually given a higher priority than best effort traffic. Denote N_i the number of voice connections associated to SS i. For each voice connection of the SS, we refer to the basic two-state model, i.e. each voice connection has alternative active and silent periods. We consider that both the active and silent periods follow an exponential distribution with mean T_{ON} and T_{OFF}, respectively. During the silent periods, there is no packet generated. During the active periods, there is a constant packet generation rate, R_{voice}. We use the parameter β = $1/T_{OFF}$ to represent the rate of transition out the silent state, and α = $1/T_{ON}$ to denote the rate of transition out of the active state. Denote P_{ON} as the probability that the ON-OFF voice connection is ON, then we have $P_{ON} = T_{ON}/(T_{ON} + T_{OFF})$. Each SS keeps a data buffer for temporarily storing the incoming packets from the voice connections for uplink transmissions. All voice packets have the same length, which is known by the BS when the connections are set up. The bit sequence from an active voice source in every T_{voice} seconds is packed into a voice packet. Therefore, the transmission time required for each packet is $T_p = \frac{R_{voice} \times T_{voice} + L_{head}}{R_b}$, where R_b is the physical layer transmission throughput in bps, and L_{head} is the total header size at the physical, MAC and higher layers. Let $D_{i,k}$ represent the maximum tolerable delay for the kth packet in the buffer of SS i, and $d_{i,k}$ the remaining delay budget of the same packet, where $i = 1, 2, \ldots, S$; $k = 1, 2, \ldots, K_i$, and K_i is the buffer occupancy of SS i. That is, $(D_{i,k} - d_{i,k})$ is the amount of delay that the kth packet in the buffer of SS i has experienced when arriving at the SS. In order to guarantee that packets with less delay budget to be transmitted earlier, packets with a smaller $d_{i,k}$ value is placed closer to the head of the buffer. That is, $d_{i,k} \leq d_{i,k+1}$, where $i = 1, 2, \ldots, S$ and $k = 1, 2, \ldots, K_i - 1$. We consider that connection admission control is performed at the BS before each new connection is accepted in the system, so that the BS has the knowledge of the number of connections currently in the system.

We consider to use the bandwidth request mechanism in the rtPS and ertPS services defined in the 802.16 standard, since they are much more efficient than the UGS service when serving packet voice traffic. For the rtPS service, the BS periodically polls an SS in every downlink subframe and specifies the time slots when the SS can make a bandwidth request. Upon receiving the polling information, a bandwidth request message is sent by an SS through the next uplink subframe. The BS, after receiving the bandwidth requests and delay information of buffered packets from all the associated SSs, makes a scheduling decision about how many packets each of the SSs can transmit in the next uplink subframe. The decision is then broadcast in the following downlink subframe, as shown in Figure 3.2.

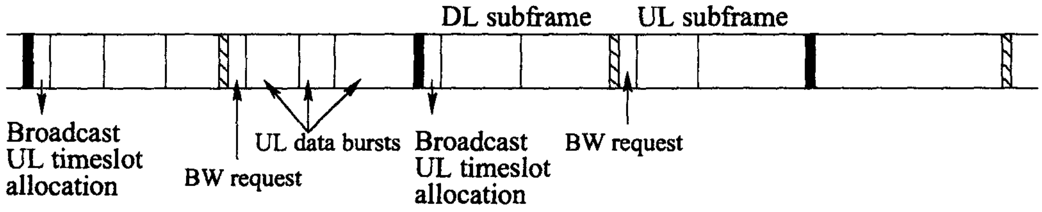

DL subframe
UL subframe
Broadcast UL timeslot allocation
BW request
UL data bursts
Broadcast UL timeslot allocation
BW request

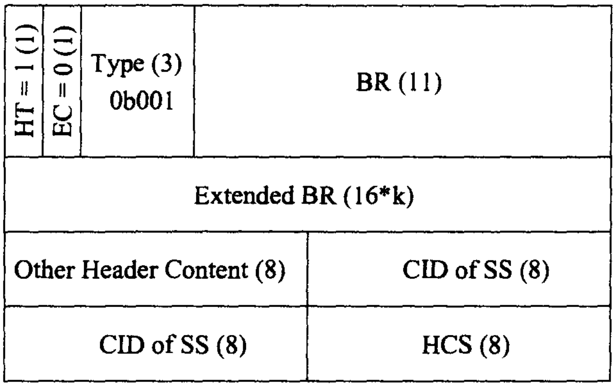
Figure 3.2: Bandwidth requests and grants

In the ertPS service, the SS keeps using its current transmission rate until the aggregate packet generation rate that it requires changes. When this occurs, the SS updates its bandwidth request. According to the standard, the SS can use allocated bandwidth, piggyback its new bandwidth request, or use contention-based transmission opportunities if the current available transmission rate is zero. As discussed in the previous chapter, since contention-based transmission does not guarantee a strict delay requirement, and piggyback request is optional for the SS and only for incremental requests, we consider to use the bandwidth request mechanism that the SS sends a standalone bandwidth request message in the specified uplink time slots to update its bandwidth request.

Each bandwidth request message aggregates the requests of multiple connections and reports information of delay budgets and number of buffered packets to the BS. With an equal packet size for all packets, the BS can find the number of required bytes from each SS. For connections with different traffic parameters, separated bandwidth requests should be used either following the standard IEEE 802.16 bandwidth request format or the one proposed in this thesis, and this is beyond the scope of this work, which focuses on the effect of aggregating bandwidth requests of multiple voice connections with the same traffic parameters and passing a certain amount of delay budget information to the BS.

3.2 Enhanced Bandwidth Requests

Fig. 3.3 shows the format of the enhanced bandwidth request, where the number between each pair of the brackets represents the number of bits for the corresponding field. Compared to the original bandwidth request message in the 802.16 standard, there are several differences. First, the CID field of individual connections is replaced with the CID of the SS in order to aggregate the bandwidth requests of multiple connections associated to the same SS. Second, the bandwidth request (BR) field, which originally uses 11 bits for specifying the number of requested bytes from each connection, is now used to represent the length of the bandwidth request message and the total number of bytes to be transmitted. In addition, the extended BR field is used for passing information about the number of buffered packets and their latency information. The size of the extended BR field is a multiple of 16 bits and depends on specific schemes used.



<table>
<tr><td>HT = 1 (1)</td><td>EC = 0 (1)</td><td>Type (3)
0b001</td><td colspan="3">BR (11)</td></tr>
<tr><td colspan="6">Extended BR (16*k)</td></tr>
<tr><td colspan="4">Other Header Content (8)</td><td colspan="2">CID of SS (8)</td></tr>
<tr><td colspan="4">CID of SS (8)</td><td colspan="2">HCS (8)</td></tr>
</table>

Figure 3.3: Bandwidth request message format

We propose three bandwidth request and scheduling schemes. In each of the schemes, the SS passes a different amount of information to the BS. Because of this, a different amount of resources is required for the bandwidth requests, and the resultant real-time performance is different. In the scheduling with Full Delay Budget Information (FDBI), the SS passes to the BS the delay budgets of all packets in its buffer. In the second scheme, instead of transmitting the delay budget values of all buffered packets, the SS only transmits the delay budget value of the head-of-line packet to the BS. This is referred to as scheduling with the Head-of-line Delay Budget Information, or HDBI. In this case, the SS reports the $d_{i,1}$ to the BS together with the total number of packets waiting for transmission in its buffer. Both the FDBI and HDBI require periodical bandwidth requests in every uplink subframe. In the scheduling with Partial Delay Budget Information (PDBI), the maximum tolerable delay, $D_{\max}$, is divided into L consecutive intervals, where $D_{\max} = \max_{1 \le i \le S, 1 \le k \le K_i} D_{i,k}$. The lth interval is $[(l-1)\tau, l\tau]$ for $l = 1, 2, \ldots, L-1$, and $[(l-1)\tau, D_{\max}]$ for $l = L$. That is, $L = \lceil D_{\max}/\tau \rceil$. Instead of reporting the delay budget of every packet in its buffer, the SS now reports the number of packets with delay budgets falling in each of the intervals. Let $u_{i,l}$, $i = 1, 2, \ldots, S$ and $l = 1, 2, \ldots, L$, represent the number of packets in the buffer of SS i and with their delay budgets falling in the lth interval. The values of $u_{i,l}$, $l = 1, 2, \ldots, L$, will be reported by SS i to the BS when making a bandwidth request, and used by the BS to make scheduling decisions for the next uplink subframe.

With a different amount of information included in the bandwidth request messages, the number of bits required for the extended BR field and the time for making a bandwidth request is different for the three schemes. Let $T_{BW\_req,i}$ represent the time for transmitting a bandwidth request, and $L_{BW\_req,i}$ the number of bits used for the extended BR field. Then,

$$T_{BW\_req,i} = \frac{48 + \lceil \frac{L_{BW\_req,i}}{16} \rceil \times 16}{R_b}, \tag{3.1}$$

where R_b is the physical layer transmission throughput in bps. Since the resource for bandwidth requests is preallocated when the BS polls an SS, $L_{BW\_req,i}$ should be calculated conservatively based on the maximum number of bits that are possibly required.

For the FDBI scheme, $L_{BW\_req,i}$ can be found as

$$L_{BW\_req,i} = \lceil \log_2 D_{\max} \rceil \times \left\lceil \frac{T_{MAC}}{T_{voice}} \right\rceil \times N_{\max}, \tag{3.2}$$

where $\lceil \log_2 D_{\max} \rceil$ is the number of bits required for reporting the delay budget of each buffered packet, T_{MAC} is the duration of one MAC frame, $\lceil\frac{T_{MAC}}{T_{voice}}\rceil$ is the maximum number of packets that can possibly be generated during one MAC frame from each voice connection, and $N_{\max}$ is the maximum number of voice connections that each SS can associate with and its value can be known from the admission control performed at the BS.

Using the HDBI scheme, the BR field passes the value of total number of buffer packets, and the extended BR field passes the delay budget of the head-of-line packet. Since only one delay budget value between 0 and $D_{\max}$ should be transmitted, we have

$$L_{BW\_req,i} = \lceil\log_2 D_{\max}\rceil. \tag{3.3}$$

Using the PDBI scheme, an expression for $L_{BW\_req,i}$ can be found as

$$L_{BW\_req,i} = \left\lceil\frac{D_{\max}}{\tau}\right\rceil \times \log_2\left(\left\lceil\frac{\tau}{T_{voice}}\right\rceil \times N_{\max}\right), \tag{3.4}$$

where $\lceil\frac{D_{\max}}{\tau}\rceil$ is the maximum number of intervals of length τ, and $\lceil\frac{\tau}{T_{voice}}\rceil \times N_{max}$ is the maximum number of packets that can possibly be generated by connections associated to one SS during an interval of length τ.

For the uplink transmission capacity, the maximum number of packets that the BS can receive in one uplink frame can be found as

$$M_{capa} = \left\lfloor\frac{T_{MAC\_UP} - T_{overhead} - \sum_{i=1}^{S} T_{BW\_req,i}}{T_p}\right\rfloor, \tag{3.5}$$

where T_{MAC_UP} is the duration of one uplink subframe, and $T_{overhead}$ is the time for contention-based transmission requests and ranging in the uplink subframe.

3.3 Packet Transmission Scheduling

We propose the corresponding scheduling schemes based on information provided in the aggregate bandwidth request messages.

3.3.1 Scheduling with FDBI

With the delay budget information of all buffered packets available at the BS, the packet scheduling can strictly follow the EDF rule. If the total number of packets requested for transmission from all the SSs is less than M_{capa}, then all packets can be transmitted, and the SSs are granted the bandwidth that they requested. Otherwise, the first M_{capa} packets with the smallest delay budgets are scheduled to transmit.

The following calculations are used by the BS to decide how many packets each SS can transmit in the next uplink subframe. Define a set $X_{i,k}$ as

$$X_{i,k} = \{(i^{'}, k^{'}) | d_{i^{'},k^{'}} \leq d_{i,k}, i^{'} = 1, \ldots, S, k^{'} = 1, \ldots, K_{i^{'}}\}, \tag{3.6}$$

for $i = 1, 2, \ldots, S$ and $k = 1, 2, \ldots, K_i$. That is, $X_{i,k}$ is a set of the packets with their delay budgets smaller than the kth packet from SS i's buffer. Define

$$Y_i = \{(i, k) |\ |X_{i,k}| \leq M_{capa}, k = 1, 2, \ldots, K_i\} \tag{3.7}$$

for $i = 1, 2, \ldots, S$, where $| \cdot |$ denotes the number of elements in a set. Then Y_i is a set of packets that can be transmitted from the SS i's buffer in the next uplink subframe.

3.3.2 Scheduling with HDBI

The BS finds the SS with the minimum value of $d_{i,1}$ and transmits as many packets from the SS as possible. Scheduling decisions are made at the BS following Algorithm 1, where Y represents the total number of packets that can be scheduled to transmit in the next uplink subframe, and Y_i the number of packets that can be scheduled to transmit in the next uplink subframe from SS i's buffer.

Algorithm 1: Scheduling with HDBI

```
Let $Y_i = 0$ for $i = 1, 2, \ldots, S$, $Y = 0$, and $\tilde{S} = \{i | i = 1, 2, \ldots, S\}$.
while $Y < M_{capa}$ and $\tilde{S} \neq \emptyset$ do
    $j = \arg\min_{i \in \tilde{S}} d_{i,1}$
    $Y_j = \min\{M_{capa} - Y, K_j\}$, $Y = Y + Y_j$, and $\tilde{S} = \tilde{S} - \{j\}$.
end while
```

3.3.3 Scheduling with PDBI

When making scheduling decisions, the BS first satisfies the bandwidth requirement of the buffered packets with their delay budgets falling into earlier delay budget intervals. The algorithm performed at the BS for making resource allocation decisions is as follows.

Algorithm 2: Scheduling with PDBI

```
Let $l = 0$, $Y = 0$, and $Y_i = 0$ for $i = 1, 2, \ldots, S$.
while $l \leq L$ do
    if $Y + \sum_{i=1}^{S} u_{i,l} \leq M_{capa}$ then
        $Y_i = Y_i + u_{i,l}$ for $i = 1, 2, \ldots, S$
        $Y = Y + \sum_{i=1}^{S} u_{i,l}$
        $u_{i,l} = 0$
```

```
7:    else
8:      Let $\tilde{S} = \{1, 2, \ldots, S\}$
9:      while $\tilde{S} \neq \emptyset$ do
10:       Find $j = \arg\max_{i \in \tilde{S}} u_{i,l}$
11:       if $Y + u_{j,l} \leq M_{capa}$ then
12:         $Y_j = Y_j + u_{j,l}$
13:         $Y = Y + u_{j,l}$
14:         $u_{j,l} = 0$
15:         $\tilde{S} = \tilde{S} - \{j\}$
16:       else
17:         $Y_j = Y_j + (M_{capa} - Y)$
18:         $Y = M_{capa}$
19:         $u_{j,l} = u_{j,l} - (M_{capa} - Y)$
20:         $\tilde{S} = \emptyset$
21:         $l = L$
22:       end if
23:     end while
24:   end if
25:   $l = l + 1$
26: end while
```

Starting from the first delay budget interval, i.e., $l = 1$, the BS first checks if its available capacity is sufficient for transmitting all the packets with delay budgets falling in the interval. If there is sufficient resource available, the number of packets that each SS is allowed to transmit in the next uplink subframe is increased by $u_{i,l}$ (line 4), and the BS keeps checking the next delay budget interval; otherwise, the BS decides how many packets in the lth interval can be transmitted. It does so by first selecting the SS with the largest number of buffered packets (line 10) and transmitting as many packets from the SS as possible.

If there is still bandwidth available after serving all packets in that interval from the SS, packets from the SS with the next largest number of buffered packets are transmitted until all the BS resource is used up. The performance of this scheduling scheme depends on the values of τ and L. When τ is smaller, L is larger, then the BS can get more accurate delay budget information about the buffered packets. On the other hand, a smaller value of τ (a larger value of L) requires more information to be transmitted from the SSs to the BS, and may consume more resources for bandwidth requests and leave less resources for data packet transmissions.

When performing the above algorithm, the BS updates values of $u_{i,l}$'s as shown in lines 6, 14 and 19. If $\sum_{i=1}^{S}\sum_{l=1}^{L} u_{i,l} \geq 2M_{capa}$ before the algorithm starts, then the updated value $\sum_{i=1}^{S}\sum_{l=1}^{L} u_{i,l}$ after performing the algorithm is larger than M_{capa}. That is, the SSs do not need to request bandwidth again in the following MAC frame. The SSs are informed of this status in the downlink bandwidth grant messages. In the PDBI scheme, the SSs do not make an explicit bandwidth request until it is notified that $\sum_{i=1}^{S}\sum_{l=1}^{L} u_{i,l} \leq M_{capa}$. This type of bandwidth request process is similar to that specified in ertPS, and can save more bandwidth for data packet transmissions.

Chapter 4

Analysis of Packet Loss Performance of PDBI

We formulate an analytical model to investigate the packet loss performance for the PDBI scheme. Analyzing the exact performance of the proposed scheduling schemes is difficult. The PDBI service system does not fit into any classical queueing model, but a G/G/1 model. Although formulae for loose delay bounds in a G/G/1 queue are available, they require the second moments of the arrival and service processes, and the latter is very difficult to obtain in the PDBI service system. In this thesis, we choose to borrow some results from the concept of effective bandwidth. The effective bandwidth approach can be used to find the packet loss probability when a number of ON-OFF sources are multiplexed and share a first-in first-out (FIFO) buffer. The basic definition of the effective bandwidth is the constant service rate required to serve a certain source in order to guarantee a small performance violation probability, where the performance can be that the maximum packet transmission delay is below a certain bound. The effective capacity is the dual of the effective bandwidth and defined as the maximum constant arrival rate that can result in a certain small performance violation probability for a given service process. Both deal with relationship among the arrival process, service process, and performance violation probability, from which the packet loss probability can be found.

We develop an analytical model for deriving the packet loss performance of the PDBI scheme and examine the effect of τ on the packet loss performance. For simplicity and mathematical tractability, we consider that all SSs have the same number of voice connections associated, i.e., $K_i = K$ for all i, and all packets have the same maximum tolerable delay, i.e, $D_{i,k} = D_{max}$ for all i and k values. We first assume that all packets are served according to their delay budgets, the packet with a smaller delay budget served first. This is equivalent to first-come-first-serve (FCFS) with the above assumption on $D_{i,k}$'s, as shown in Figure 4.4.

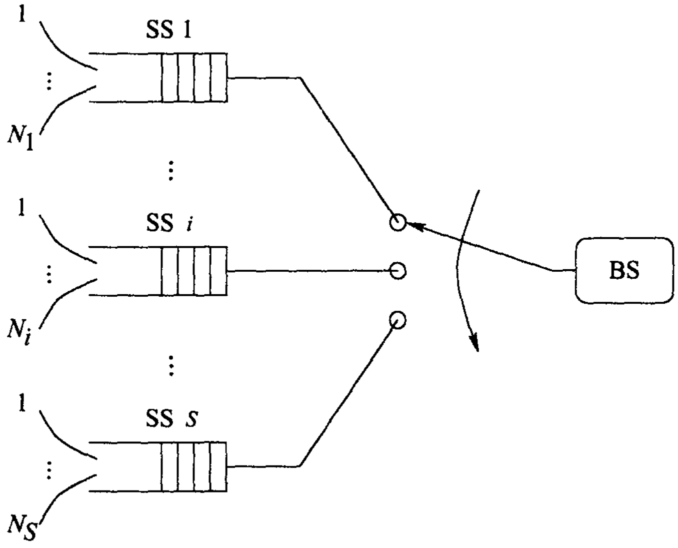


Figure 4.4: Illustration of packet buffers

We then analyze the performance of this service system using the fluid flow model, where we assume the number of packets generated during active state is so large that it appears like a continuous data flow. The buffer occupancy thus becomes a continuous random variable. Due to the slotted delay budget reports, not all packets are served according to their arrival times in the PDBI scheme. Consider a reference packet, k^*, which is associated to a reference SS i^*. Let l^* represent the delay budget interval that packet i^* falls in. According to the PDBI scheme, a packet k associated to SS i, $i \neq i^*$, will cause extra delay to the transmission of packet k^* if packet k (i) has its delay budget falling in the same delay budget interval l^*, (ii) arrives later than packet k^*, and (iii) the number of packets falling in delay budget interval l^* in the buffer of SS i is larger than that in the buffer of SS i^*, i.e., $u_{i,l^*} > u_{i^*l^*}$. Let $T_{ext} = nT_{Packet}$ represent the total extra delay caused by slotted delay budget reports. Then the overall packet loss probability of PDBI can be expressed by

$$P_{\text{PDBI}} = \sum_{n=0}^{K(S-1)} \Pr\{N_{\text{e}} = n\} P_{\text{FCFS}} \left(D_{\text{max}} - nT_{\text{Packet}} \right), \tag{4.8}$$

where $K(S-1)$ is the total number of voice connections which is also the maximum number of packets that can cause extra delay to the reference packet, N_e is a random variable representing the number of packets that can cause extra delay to the reference packet, $P_{\text{FCFS}}(d)$ is the packet loss probability with FCFS and maximum delay budget of d for all packets. In the remaining part of this section, we will derive $\Pr\{N_e = n\}$ and $P_{\text{FCFS}}(d)$.

First, we consider that all packets are served in an FCFS manner. The peak traffic arrival rate from each ON-OFF voice connection is $R_p = \frac{1}{T_{\text{voice}}}$ packets/second, and the service rate is $C = \frac{M_{capa}}{T_{\text{MAC}}}$ packets/second, where M_{capa} is given in 3.5. It is shown in [36] that the packet loss probability can be approximately found as

$$P_{\text{FCFS}}(d) = e^{\alpha r d C / R_p}, \tag{4.9}$$

where

$$r = \frac{(1-\rho)(1+\beta/\alpha)}{\frac{C}{KSR_p} - 1}. \tag{4.10}$$

and $\rho = KSP_{ON}R_p/C$.

It is indicated in [36] that (4.9) tends to overestimate the packet loss probability due to that the derivation does not take into consideration of the effect of statistical multiplexing for multiple connections to share the buffer space.

We then consider the effect of slotted delay budget reports on the packet transmission delay. We consider a small time interval of length τ, which is much less than T_{voice}. Then the probability that there is one voice packet generated in an interval of length τ can be approximately found as

$$p = P_{\text{ON}} \frac{\tau}{T_{\text{voice}}}. \tag{4.11}$$

For a given SS i, the probability that there are $U_i = u$ packets arriving in an interval of length τ is given by

$$\Pr\{U_i = u\} = \binom{K}{u} p^u (1-p)^{K-u}, \tag{4.12}$$

for $u = 0, 1, \ldots, K$. Let $N_{e,i}$ represent the number of packets that are in the buffer of SS i and cause extra delay to the reference packet. Then we have

$$\Pr\{N_{e,i} = n\} = \sum_{u=1}^{K-1} \sum_{v=\max\{u+1,n\}}^{K} \Pr\{U_{i^*} = u\}\Pr\{U_i = v\}Q_{n,v}, \tag{4.13}$$

where

$$Q_{n,v} = \int_0^{\tau} \frac{1}{\tau} \binom{v}{n} q(t)^n [1 - q(t)]^{v-n} dt \tag{4.14}$$

is the probability that there are n packets in SS i causing extra delay to the reference packet (that is generated at time t of an interval of length τ), and

$$q(t) = P_{\text{ON}} \frac{\tau - t}{T_{\text{voice}}} \tag{4.15}$$

is the probability that there is one packet generated from a voice connection in the interval of length τ after t. Similar to (4.11), (4.15) is more accurate for small values of τ. Let $\mathbf{P}_{N_{e,i}}$ be an $1 \times (K+1)$ vector whose $(n+1)$st element is $\Pr\{N_{e,i} = n\}$, where $i = 1, 2, \ldots, S$ and $\mathbf{P}_{N_e}$ be an $1 \times [(S-1)K + 1]$ vector whose $(n+1)$st element is $\Pr\{N_e = n\}$. Then

$$\mathbf{P}_{N_e} = \mathbf{P}_{N_{e,1}} \circledast \mathbf{P}_{N_{e,2}} \circledast \ldots \circledast \mathbf{P}_{N_{e,i^*-1}} \circledast \mathbf{P}_{N_{e,i^*+1}} \circledast \ldots \circledast \mathbf{P}_{N_{e,S}}, \tag{4.16}$$

where $\circledast$ represents a convoluting operation.

Chapter 5

Numerical Results

This chapter presents the simulation results and performance analysis of the proposed schemes. The discussion begins with a description of the simulation model and the parameters setting for the simulation, and then the simulation results of the three proposed scheduling schemes are presented. After that, we compare the performance of the proposed schemes with the WRR scheme. Analytical results based on the model in Chapter 4 will be shown and compared with the simulation results.

5.1 Simulation Model and Parameters Setting

We consider an IEEE 802.16 backhaul network where 6 SSs are connected to the BS. Both the uplink and downlink share the same frequency channel. In each MAC frame, the uplink and downlink subframes are of equal size. We examine the voice packet transmission performance using the proposed schemes in terms of voice packet loss rate, mean transmission delay and delay jitter, and compare their performance. An equal number of voice connections are associated to each SS. Each voice connection has exponentially distributed active and silent periods with means T_{ON} and T_{OFF}, respectively. A connection generates a constant rate bit sequence at R_{voice} bps when the source is in the active states. The bit sequence

Table 5.1: Default Simulation Parameters

Parameter	Value
MAC frame duration T_{MAC}	10 ms
Initial ranging period duration $T_{overhead}$	312μs
Uplink burst preamble T_{pre}	11.11μs
Physical transmission throughput R_b	6.91Mbps
Voice mean ON time T_{ON}	240 ms
Voice mean OFF time T_{OFF}	400 ms
Voice packet generation rate R_{voice}	64 kbps
Voice packetization time T_{voice}	20 ms
Max. number of voice connections per SS $N_{\max}$	18
Voice packet header size L_{head}	40 bytes

from an active voice source in every T_{voice} seconds is packed into a voice packet. Therefore, the transmission time required for each voice packet is $T_p = (R_{voice}T_{voice}+L_{head})/R_b$, where L_{head} is the total header size at the physical, MAC and higher layers. Except for the initial ranging period, the remaining time in the uplink subframe can be used for voice traffic, including both bandwidth requests and packet transmissions. Default parameters are listed in Table 5.1. In our simulation, we consider two different cases: (i) All voice packets have the same latency requirement, i.e., $D_{i,k} = D_{\max} = 60$ms; and (ii) $D_{i,k}$ is uniformly distributed between 30ms and 60ms. In order to isolate the effect of the scheduling services on transmission latency performance, we assume that all transmissions are error-free.

5.2 Performance of PDBI, HDBI, and FDBI

We assume that all voice packets have the same latency requirement, i.e., $D_{i,k} = D_{\max}$ for all i and k. Figs. 5.5-5.7 compare the performance of the proposed schemes, in terms of average packet loss rate, delay and delay jitter, respectively. Fig. 5.5 shows that HDBI has the worst packet loss performance among all the three proposed schemes. This is due to the fact that the BS has the least amount of delay budget information available for making scheduling decisions. When $\tau = D_{\max}$, the PDBI scheme reports the same amount of delay

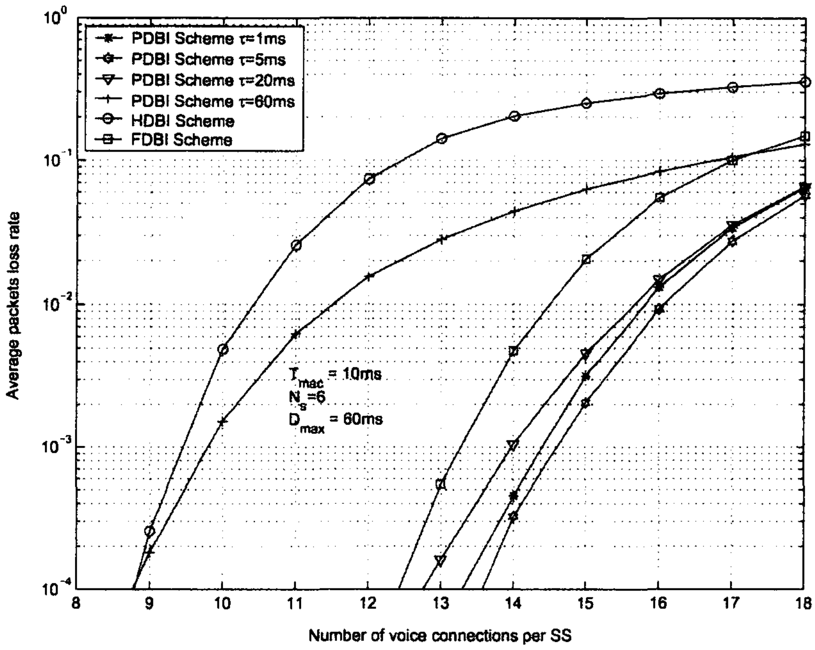


Figure 5.5: Comparison of average loss rate

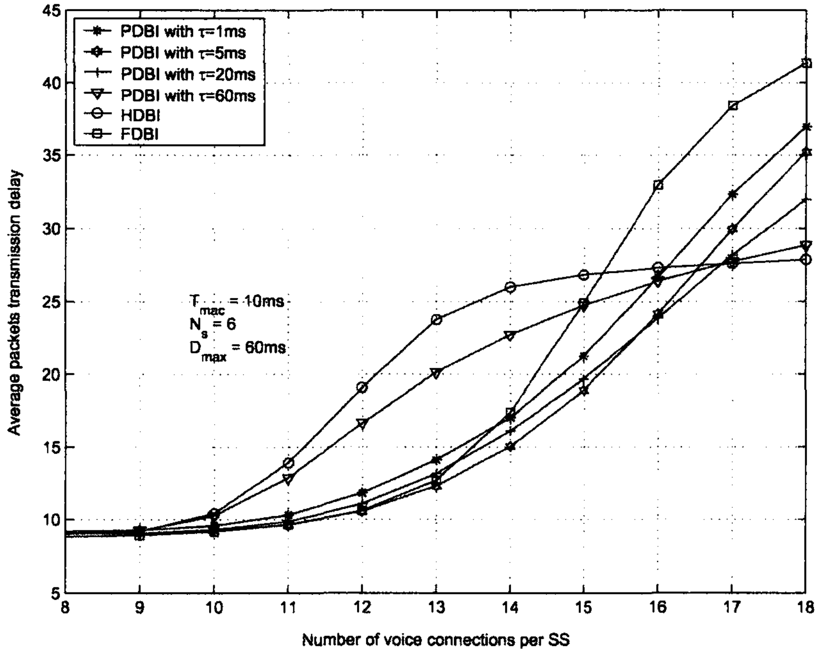


Figure 5.6: Comparison of average transmission delay

budget information to the BS as the HDBI does, but achieves lower packet loss probability, since in PDBI the SS saves resources in updating bandwidth. Comparing the PDBI and FDBI we can find that using PDBI achieves better packet loss performance than using FDBI when τ is relatively small, e.g., $\tau < 20$ms in the simulated system. As a special case, the PDBI with $\tau = 1$ reports the same amount of delay budget information to the BS as the FDBI does, but achieves lower packet loss performance due to that the PDBI does not update bandwidth requests in every uplink subframe, which allows more bandwidth for data packet transmissions.

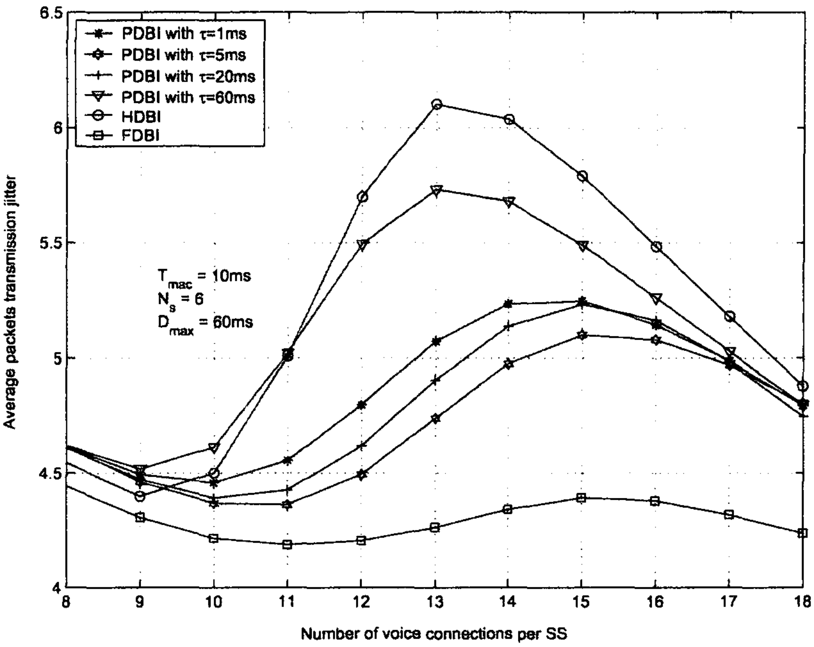


Figure 5.7: Comparison of average delay jitter

Fig. 5.6 shows that HDBI has the highest packet transmission delay among all three schemes at a low or moderate traffic load, and the FDBI achieves the lowest packet transmission delay at a low traffic load. The reason is that among the three schemes the FDBI reports the most complete delay budget information to the BS for bandwidth grant decisions while

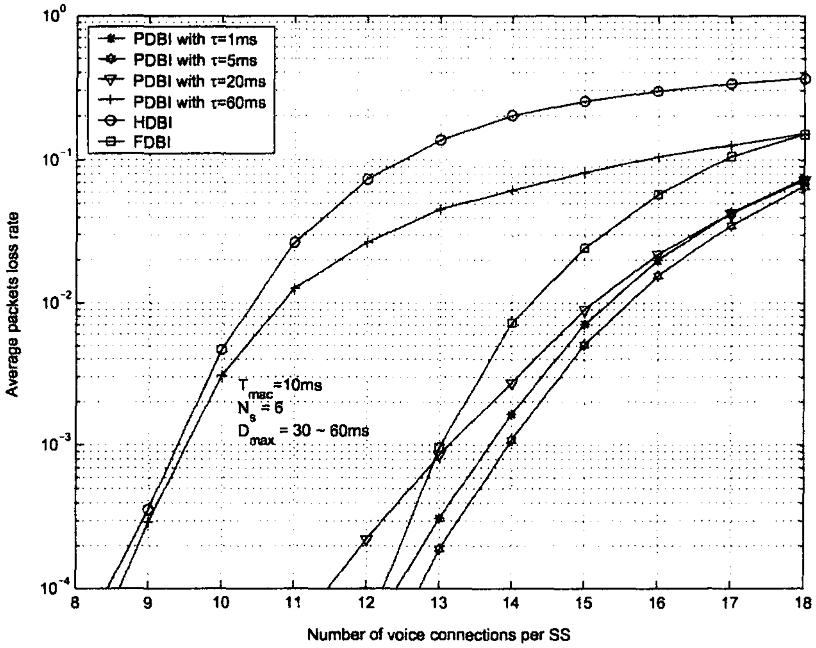


Figure 5.8: Comparison of average loss rate

HDBI reports the least. It is also observed that all the three schemes achieve the required latency performance, since we set the delay upper-bound as D_{max}=60ms in our simulation system. In Fig. 5.7 we can see that FDBI achieves the best jitter performance, which is below 4.5ms. The packet delay jitter performance in FDBI benefits from the full delay budget information. HDBI results in the worst delay jitter performance among the three schemes due to lack of delay budget information in the BS. The delay jitter performance using PDBI is in between, and dependent on the value of τ.

Figs. 5.8-5.10 show the performance results when $D_{i,k}$ is uniformly distributed between 30ms and 60ms. Fig. 5.8 shows that the PDBI scheme has the best packet loss performance and the HDBI is the worst one among all the three proposed schemes. Fig. 5.9 shows that the HDBI has the highest transmission delay among all three schemes at a low or moderate traffic load, and PDBI with small values of τ has very close average delay performance as

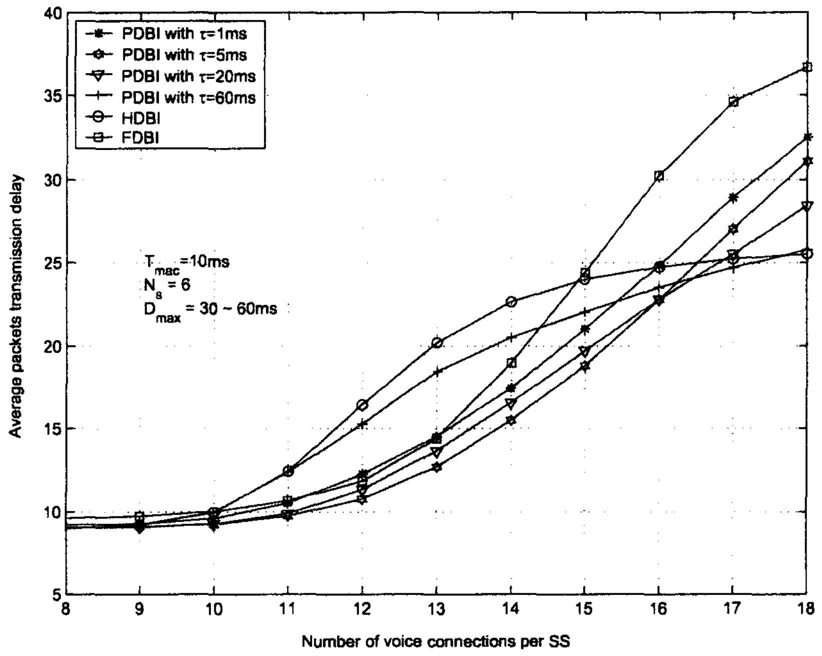


Figure 5.9: Comparison of average transmission delay

FDBI. These observations are consistent with those in Fig. 5.5 and Fig. 5.6 respectively. Fig. 5.10 shows that FDBI does not necessarily achieve the best delay jitter performance, unlike in the Fig. 5.7. All schemes achieve similar delay jitter performance.

In summary, we can have the following conclusions. Without sufficient delay budget information available, the HDBI is unable to give a higher priority to packets with smaller delay budget. Therefore, although HDBI is easy to implement, its performance is the worst among all three schemes. On the other hand, the FDBI scheme does not always achieve the best performance, since it consumes more resources for bandwidth updates than the PDBI, and leaves less resources for packet transmissions. The PDBI can tradeoff between the resources for bandwidth requests and for data transmissions by selecting different values of τ. The effect of selecting values of τ in PDBI on the packet transmission performance is studied next.

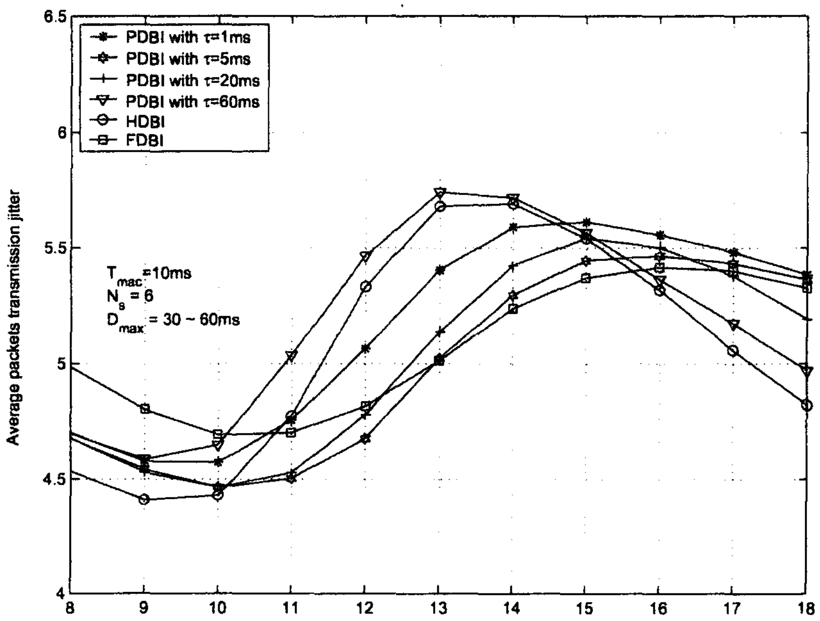


Figure 5.10: Comparison of average delay jitter

5.3 Optimal Point of τ

The value of τ plays an important role in PDBI and can affect the performance greatly. Figs. 5.11 and 5.12 show the performance of the PDBI scheme vs. values of τ. As τ increases, less detailed delay budget information is available at the BS, while a less amount of resource is required for the SSs to report the delay budget information to the BS. Therefore, overall there is an optimum value of τ when the packet loss rate is minimized. It is shown in Fig. 5.11 that the optimum value of τ is around 5ms in the simulated system when T_{MAC}=10ms, and Fig. 5.12 shows that the optimal value of τ is around 7ms when T_{MAC}=5ms. Figs. 5.13 and 5.14 show the effect of different values of τ on the packet loss probability based on the PDBI model developed in Section 4, where the range of τ is from 2 to 20ms. Although the packet loss probability is in general higher than that shown

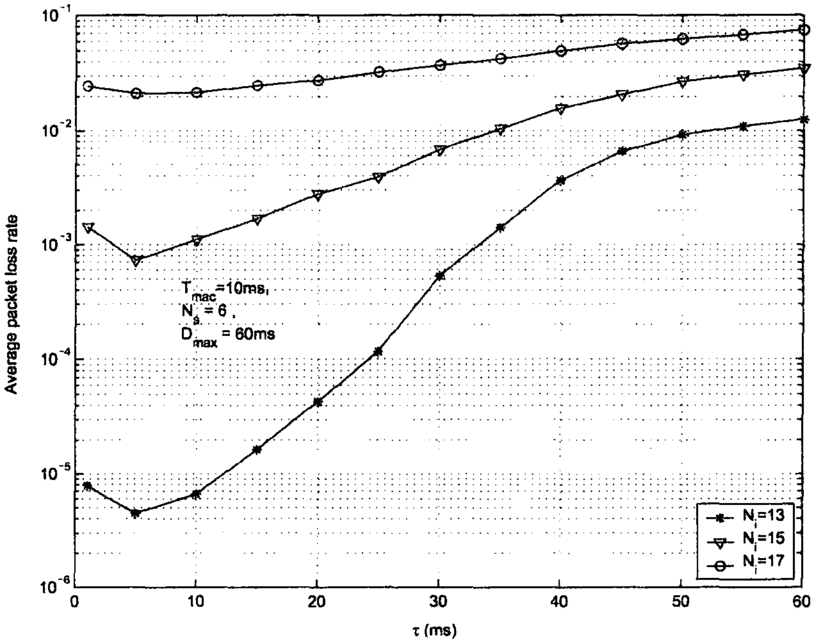


Figure 5.11: PDBI: average packet loss rate vs. τ, T_{MAC} = 10ms

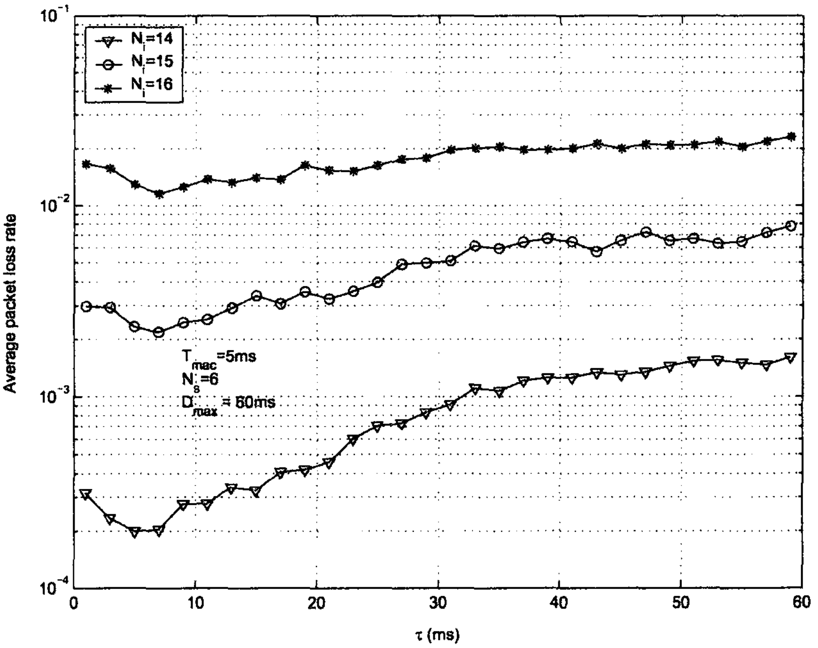


Figure 5.12: PDBI: average packet loss rate vs. τ, T_{MAC} = 5ms

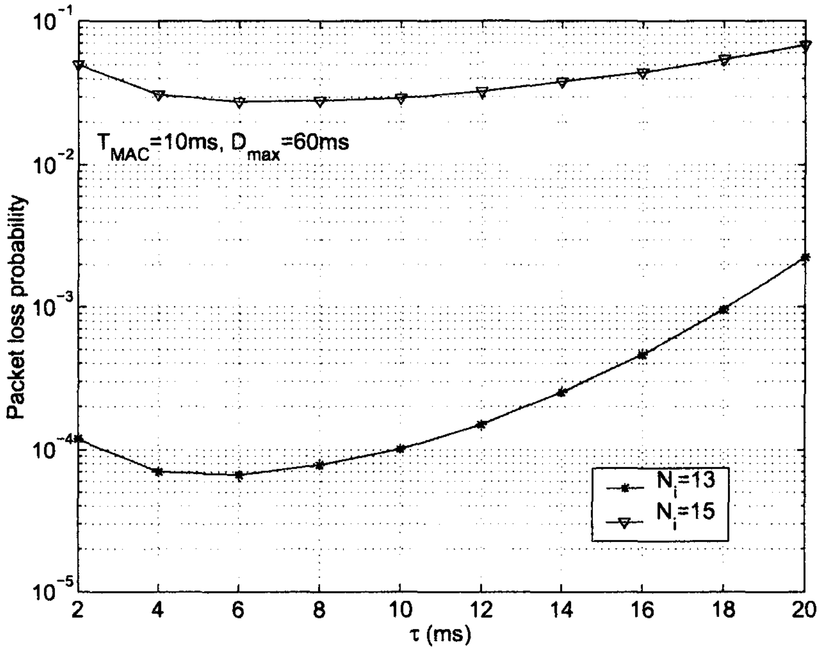


Figure 5.13: PDBI: Analytical packet loss probability, $T_{\text{MAC}} = 10\text{ms}$

in Fig. 5.11 as (4.9) tends to overestimate the loss probability, the curves in Fig. 5.13 indicate that there is an optimum value of τ to minimize the packet loss probability, and the optimum value is around 5ms. This observation is consistent with that in Fig. 5.11. Meanwhile, Fig. 5.14 shows that the optimum value of τ is around 7ms, which leads to the minimum packet loss probability when T_{MAC}=5ms, slightly larger than the value when T_{MAC}=10ms. This is due to that the scheduling decisions can be more frequently updated with a shorter MAC frame, which reduces the pressure on accurate delay budget reports in order to achieve the same scheduling performance. This optimum value of τ shown in Fig. 5.14 is also very close to that in Fig. 5.12. When $D_{i,k}$ varies between 30 and 60ms, Fig. 5.15 show that the optimal value of τ is around 5ms.

Figs. 5.16-5.18 also indicate that the MAC frame duration affects the packet transmission performance. This will be demonstrated in the next section.

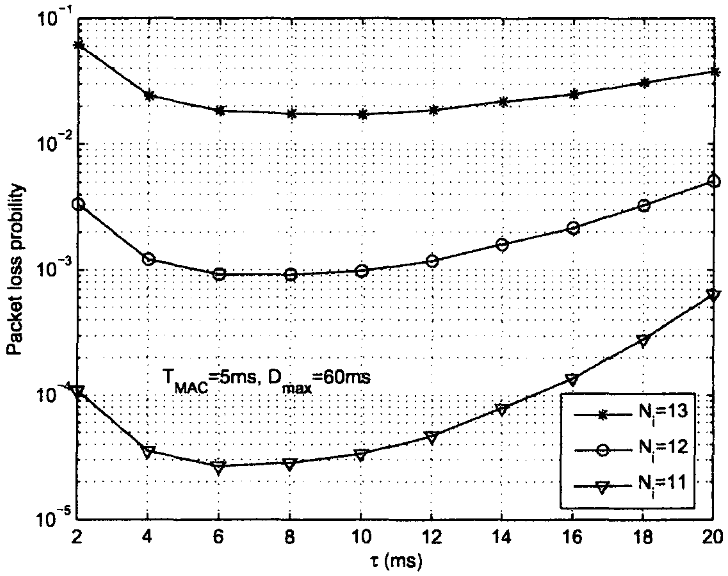


Figure 5.14: PDBI: Analytical packet loss probability, $T_{\text{MAC}} = 5\text{ms}$

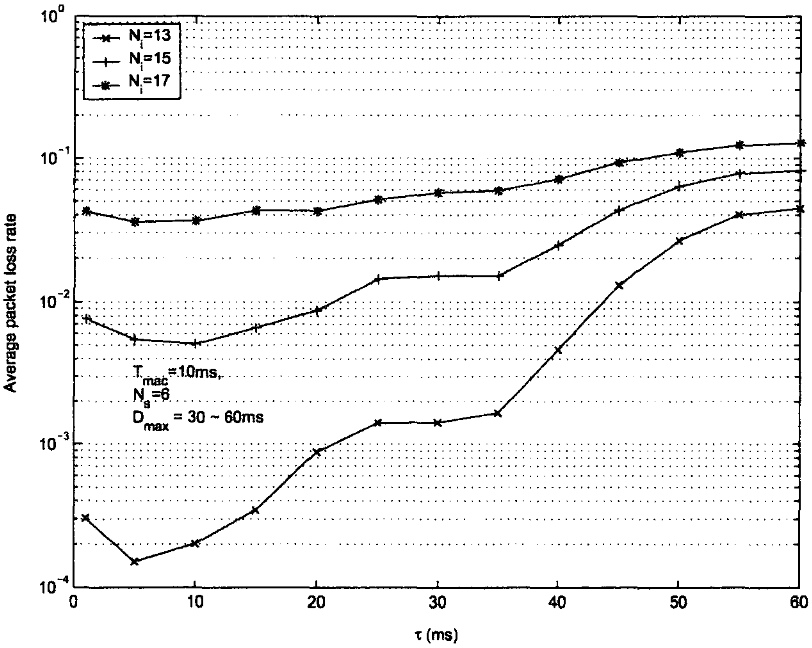


Figure 5.15: PDBI: average packet loss rate vs. τ

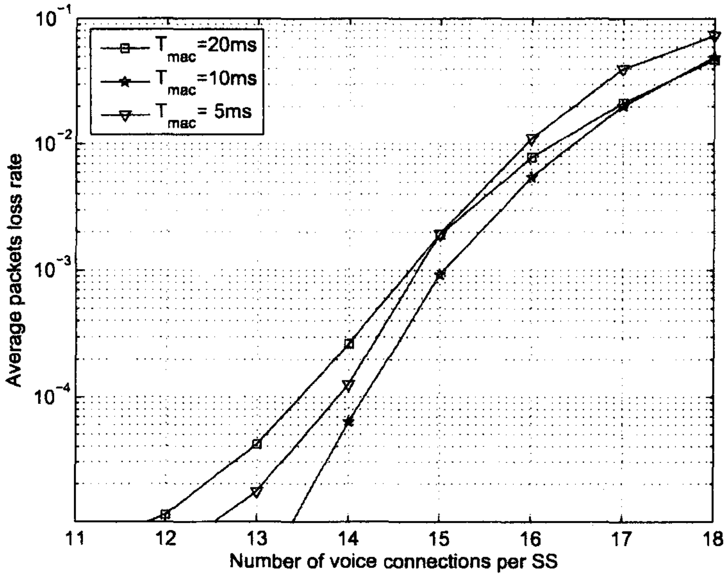


Figure 5.16: PDBI: average packet loss rate vs. number of voice connections

5.4 Impact of MAC Frame Size

We look at the performance of the PDBI for different MAC frame sizes, where τ = 5ms and $D_{i,k}$ = D_{max} = 60ms. Fig. 5.16 shows that when T_{MAC} = 10ms, the PDBI scheme achieves the best packet loss performance, compared with the case when T_{MAC} = 20ms and 5ms. This indicates that selecting the MAC frame size can affect the voice transmission performance. The reason for this is that, when the MAC frame is short, the percentage of MAC layer overhead together with the time for bandwidth requests is relatively higher, leaving less resources available for voice packet transmissions; On the other hand, if the MAC frame is too long, bandwidth cannot be updated promptly, and packet transmission performance will be degraded.

Fig. 5.17 shows that at a low traffic load, choosing T_{MAC} = 10ms achieves almost as good delay performance as T_{MAC} = 5ms and much lower delay than T_{MAC} = 20ms; at a high

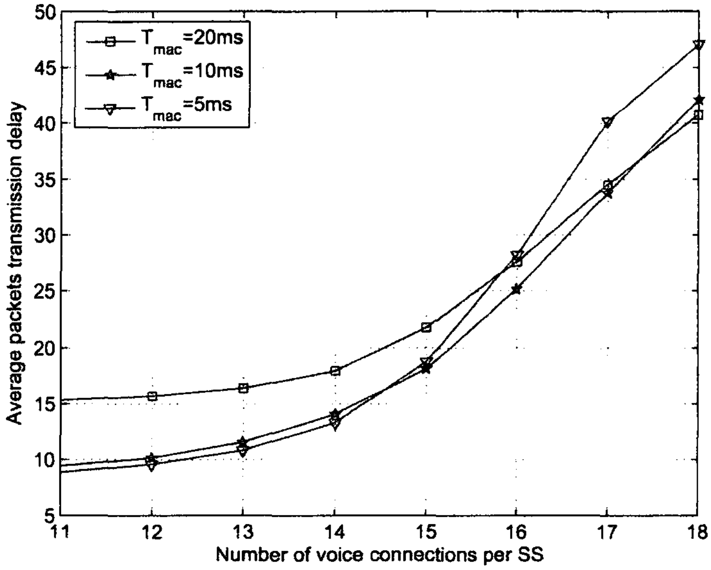


Figure 5.17: PDBI: average packet transmission delay vs. number of voice connections

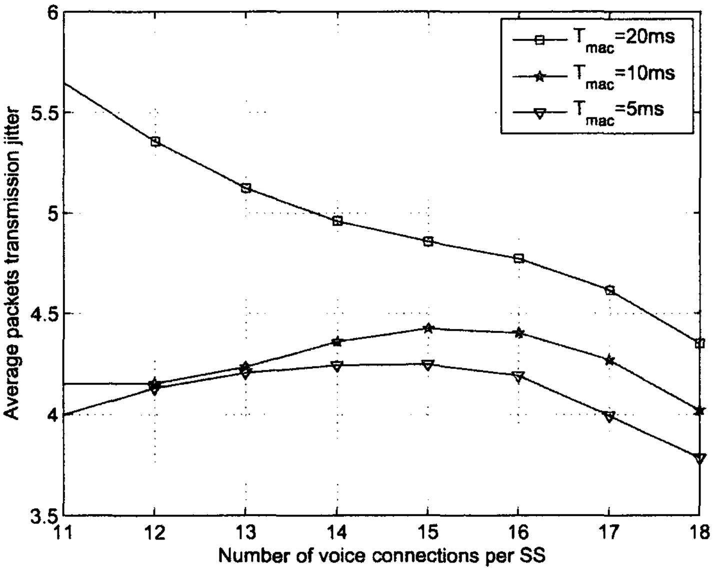


Figure 5.18: PDBI: average packet delay jitter vs. number of voice connections

traffic load, choosing T_{MAC} = 10ms achieves the better delay performance than T_{MAC} = 5ms and 20ms. Fig. 5.18 also shows that choosing T_{MAC} = 10ms achieves similar delay jitter performance as T_{MAC} = 5ms and much better delay jitter performance than T_{MAC} = 20ms.

In Figs. 5.19-5.21 we compare the performance of the PDBI at different MAC frame sizes, where τ = 5ms, and $D_{i,k}$ is uniformly distributed between 30 and 60ms. We find that the loss rate of T_{mac} = 20ms is affected significantly by this new simulation condition, and choosing T_{MAC} = 10ms and T_{MAC} = 5ms achieves almost same loss rate, both are much better than T_{mac} = 20ms. We also find that the delay and delay jitter performance of T_{mac} = 20ms is the worst among the three MAC sizes. This is due to the same reason that the bandwidth request cannot be updated timely in the case of a long MAC frame.

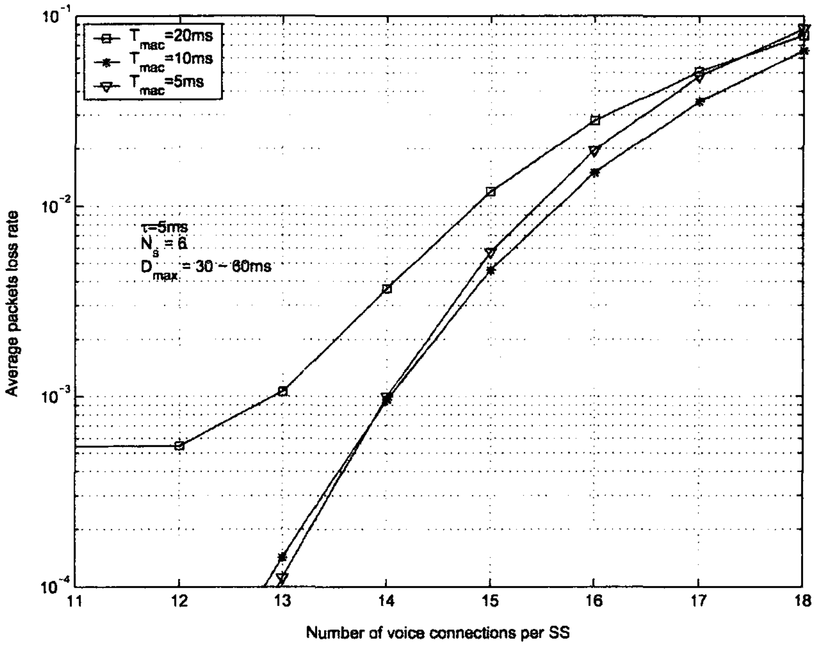


Figure 5.19: PDBI: average packet loss rate vs. number of voice connections

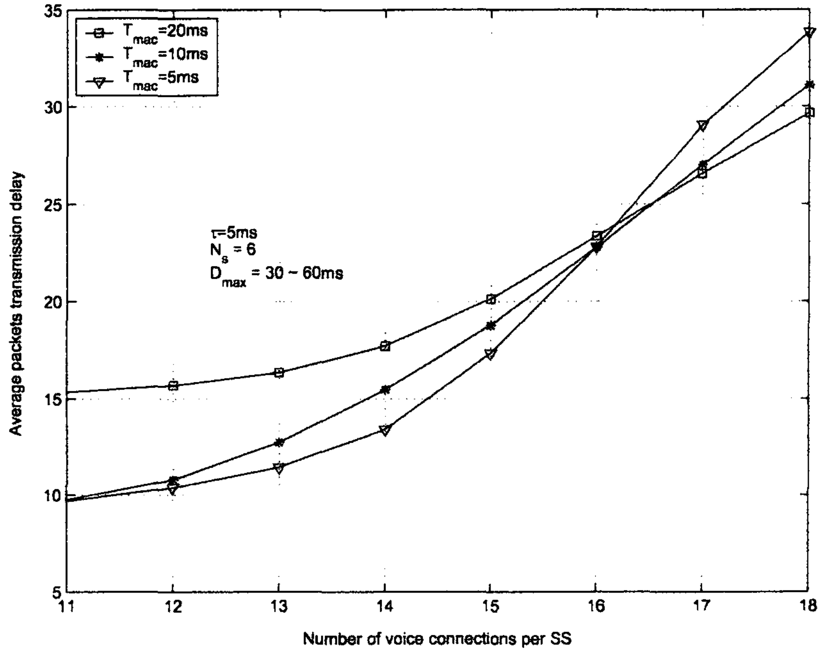


Figure 5.20: PDBI: average transmission delay vs. number of voice connections

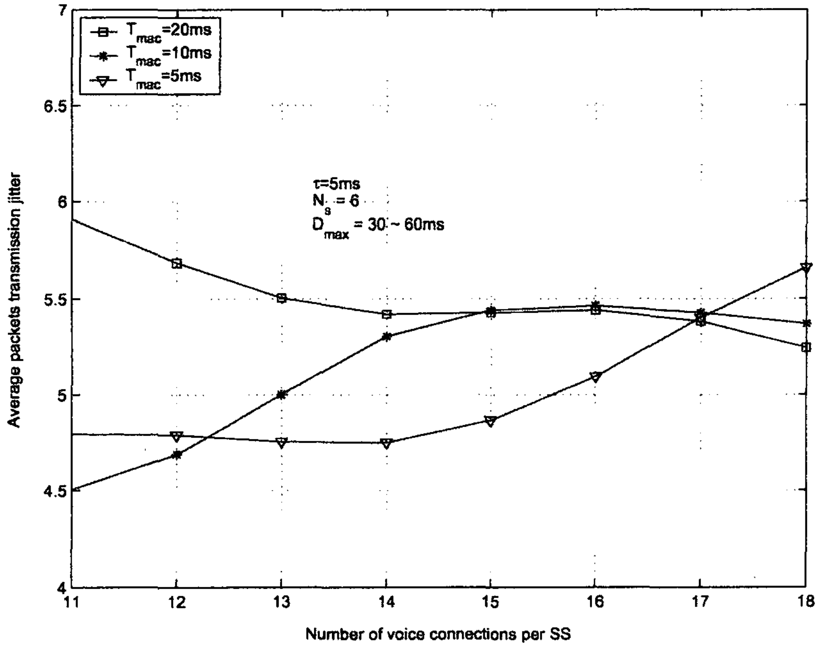


Figure 5.21: PDBI: average delay jitter vs. number of voice connections

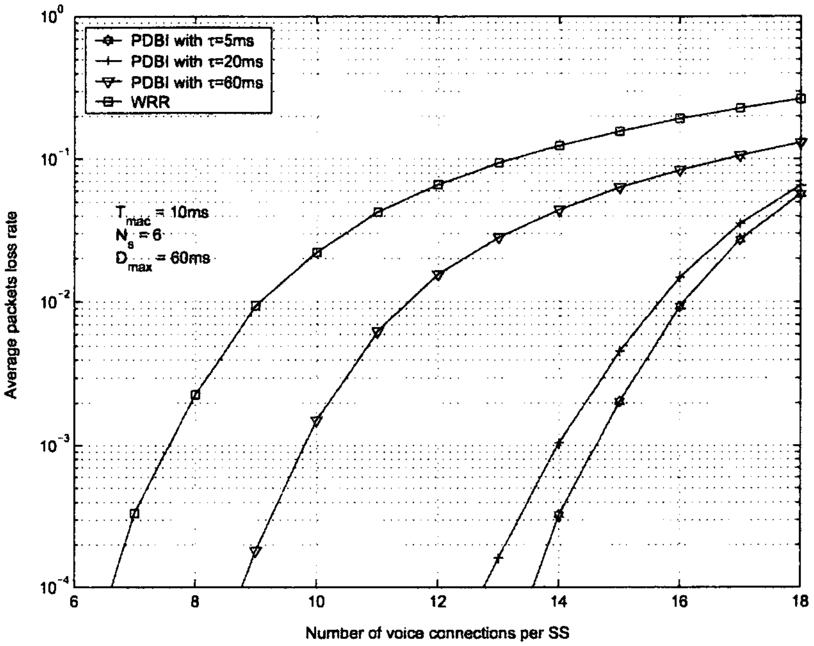


Figure 5.22: Comparison of packet loss performance between PDBI and WRR

5.5 Comparison with WRR

In this section we compare the packet loss performance of the proposed PDBI scheme to the weighted round robin (WRR) algorithm with the bandwidth message defined in the 802.16 standard. The original bandwidth request messages in 802.16 do not include delay budget information. Therefore, the BS cannot use any real-time scheduling scheme to make resource allocation decisions. WRR is a typical scheduling scheme that the BS can use in this situation. When performing the weighted round robin scheduling, the BS receives the number of buffered bytes from each active connection. That is, at the beginning of each uplink subframe, the SS sends a standard 802.16 bandwidth request for each active connection. Since the BS allocates resources on a per SS basis, the weights in the WRR are for individual SSs, not connections. The weight for an SS is equal to the number of voice connections. In each uplink subframe, all packets allowed to transmit from the

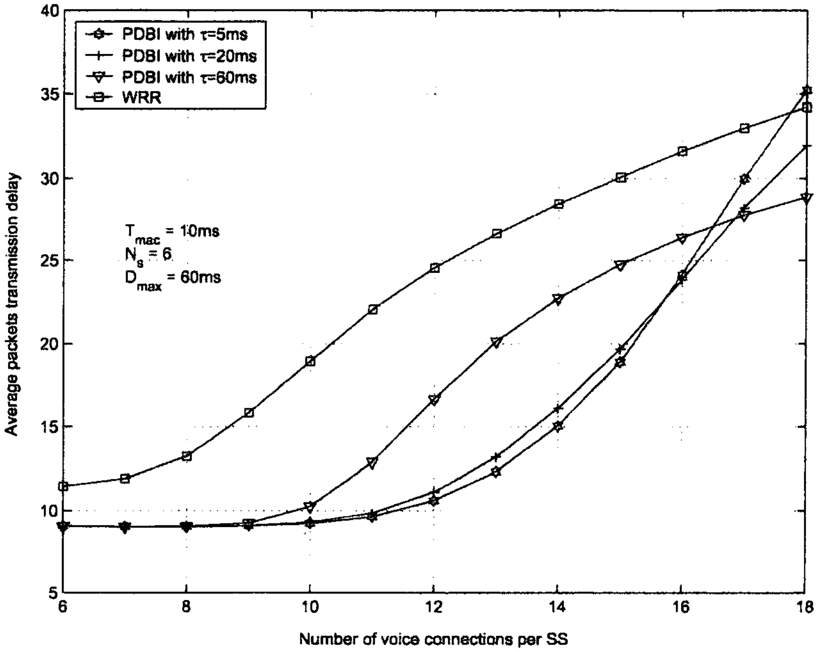


Figure 5.23: Comparison of transmission delay performance between PDBI and WRR

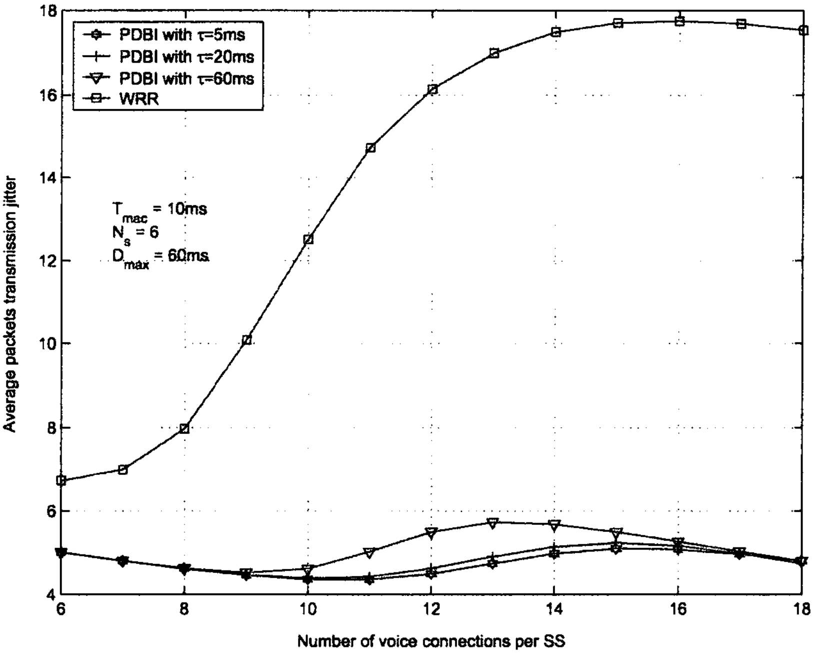


Figure 5.24: Comparison of delay jitter performance between PDBI and WRR

same SS are grouped into a burst so that the BS only needs to switch to the SS once in each uplink subframe. Fig. 5.22 shows the packet loss performance of WRR compared to PDBI, where using WRR results in much higher packet loss rate. Fig. 5.23 and Fig. 5.24 show that PDBI achieves much better results than WRR does with regard to the delay and delay jitter, respectively. The reason is that standard bandwidth request message in WRR scheme does not include any latency information. Without any latency related information available at the BS, the BS using WRR cannot efficiently allocate the bandwidth resource, thus resulting in higher loss rate, longer delay and larger delay jitter. Figs. 5.25-5.27 show

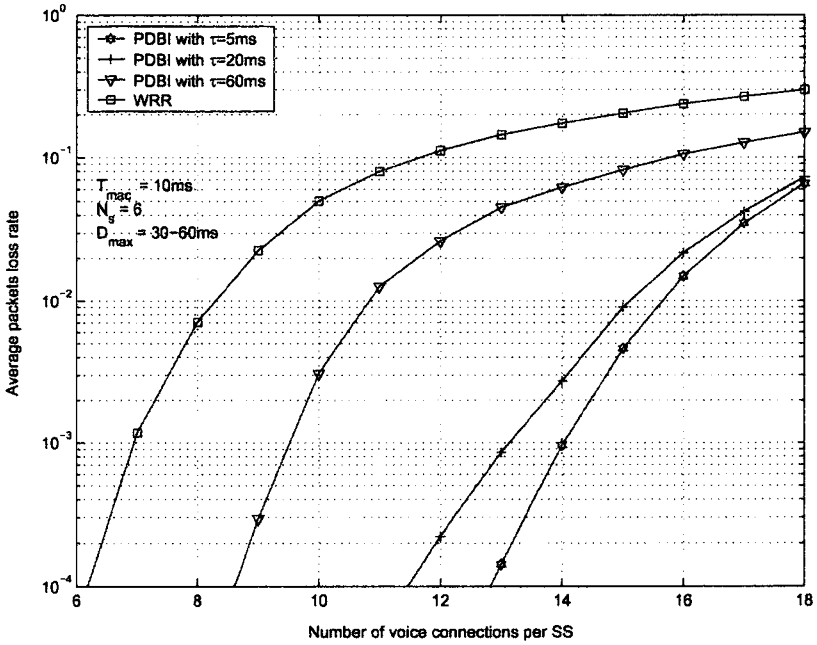


Figure 5.25: Comparison of packet loss performance between PDBI and WRR

the performance of WRR in comparison with PDBI when $D_{i,k}$ is uniformly distributed between 30ms and 60ms. The performance results are similar to Figs. 5.22-5.24. Without latency information in WRR bandwidth request messages, the WRR scheme has overall worse performance than the PDBI, in all terms of loss rate, delay and delay jitter.

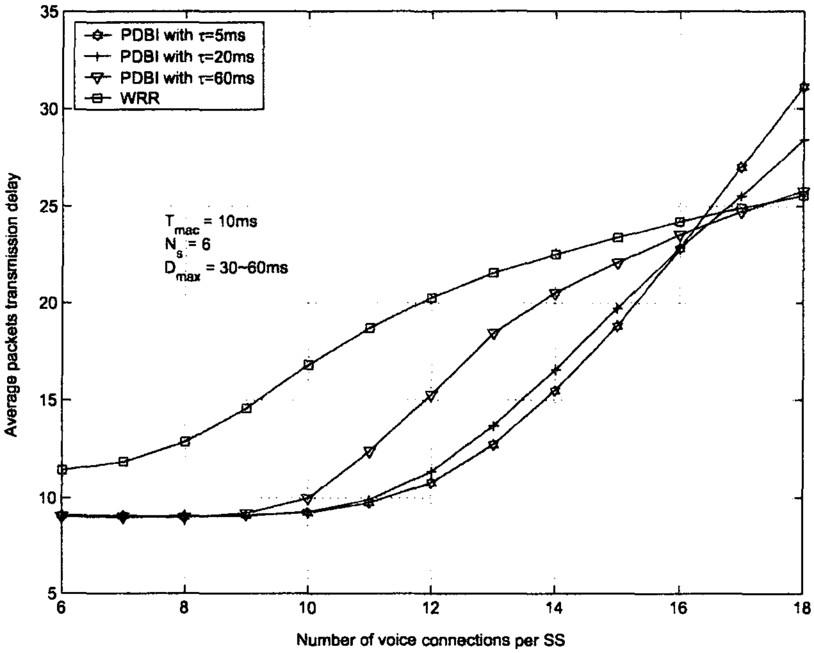


Figure 5.26: Comparison of transmission delay performance between PDBI and WRR

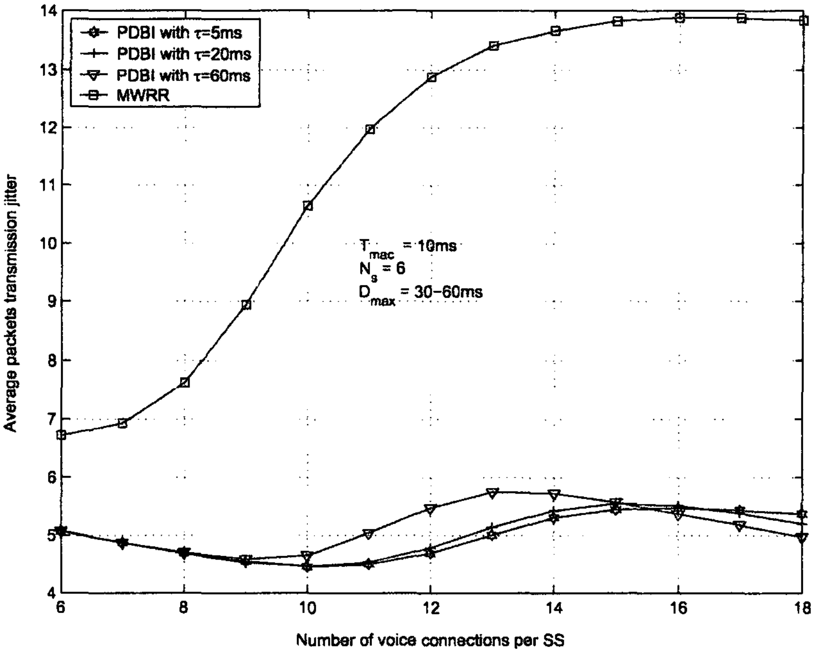


Figure 5.27: Comparison of delay jitter performance between PDBI and WRR

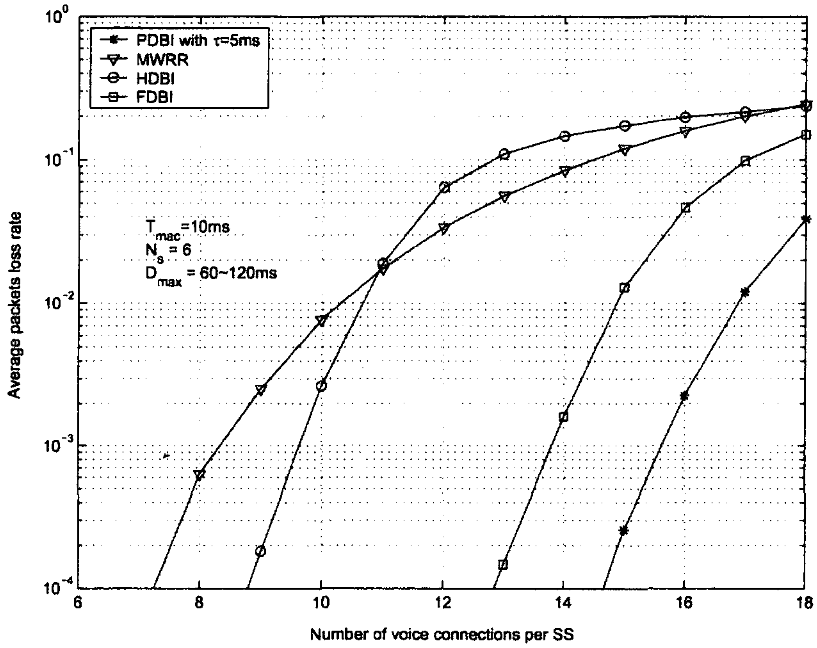


Figure 5.28: Comparison of packet loss performance D_{max}=60∼120ms

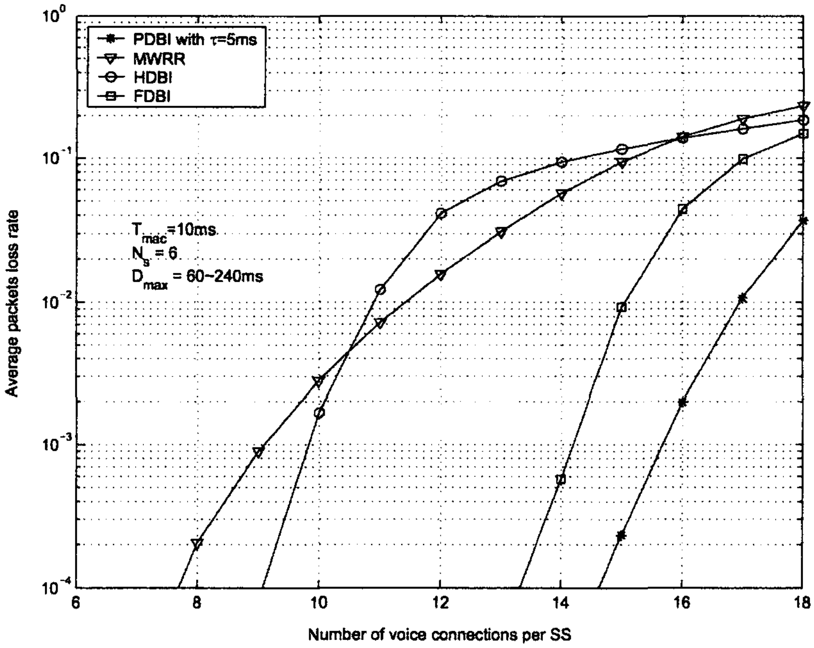


Figure 5.29: Comparison of packet loss performance D_{max}=60∼240ms

Figs. 5.28-5.29 further show that the PDBI scheme has overall better loss rate performance than all other schemes, namely the FDBI, PDBI, and WRR schemes.

5.6 Summary

From the numerical results we can find that:

- Having latency information in the bandwidth request messages helps the BS make better scheduling decisions for real-time voice traffic;
- Aggregate bandwidth requests of multiple connections associated with the same SS can save network resources;
- There is a tradeoff between the amount of latency information transmitted in the bandwidth request messages and the resultant packet transmission performance;
- With a properly chosen τ, PDBI achieves significantly better performance than other schemes; and
- The MAC frame size can affect the real-time packet transmission performance.

Chapter 6

Conclusions and Future Work

In this thesis we have proposed an enhancement to the bandwidth request mechanism in IEEE 802.16 and designed scheduling schemes for supporting real-time voice traffic in 802.16 backhaul networks. Our results show that by passing an appropriate amount of latency-related information in the bandwidth requests and aggregating the bandwidth requests of multiple connections from the same SS to the BS, the PDBI scheme can achieve significantly better real-time packet transmission performance and higher resource utilization efficiency.

By comparing different scheduling schemes, we have found that the original 802.16 bandwidth request format can be improved by incorporating latency information in order to better support real-time traffic in backhaul networks, and there is an optimum amount of latency information that the SS should report to the BS in order to achieve the best real-time transmission performance and maximizing the system capacity.

Quality of service provisioning and radio resource management have been important issues in IEEE 802.16-based networks. Due to the high transmission rate and wide coverage areas, the IEEE 802.16-based networks are expected to support various services, such as variable rate video traffic and data traffic. Packet transmission scheduling and resource management for supporting heterogeneous traffic will be studied in the future.

Furthermore, wireless channel propagation can negatively impact the packet transmission performance. Resource allocation and quality of service provisioning by incorporating the physical channel fading and co-channel interference is another future research topic.

In addition, we are planning to study the scheduling issue in a mesh topology based on the IEEE 802.16 MAC protocol.

Bibliography

[1] IEEE Std 802.16^{TM}-2004, "IEEE standard for local and metropolitan area networks - part 16: air interface for fixed broadband wireless access systems", *IEEE Standards*, Oct. 01, 2004

[2] IEEE Std 802.16e^{TM}-2005 "IEEE standard for local and metropolitan area networks - part 16: air interface for fixed and mobile broadband wireless access systems", *IEEE Standards*, Feb. 28, 2006.

[3] IEEE Standards, Wireless LAN Medium Access Control (MAC) And Physical Layer Specifications — Medium Access Control (MAC) Quality of Service (QoS) Enhancements, IEEE Press, 2005.

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

[5] D. Ferrari and D. Verma, "A Scheme for Real-time Channel Establishment in Wide-area Networks", *IEEE Journal on Selected Areas in Communications*, vol.8, no.39, April 1990, pp.368-37.

[6] V. Sivaraman and F. Chuissi, "Providing End-to-End Statistical Delay Guarantees with Earliest Deadline First Scheduling and Per-Hop Traffic Shaping", *Proc. of IEEE INFOCOM 2000, Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 2, pp.631-640, Israel, March 26-30, 2000.

[7] L. Georgiadis, R. Guerin, and A. Parekh, "Optimal Multiplexing on a Single Link: Delay and Buffer Requirements", *IEEE Transactions on Information Theory*, vol.43, no.5, Sept. 1997, pp.1518-1535.

[8] J. Liebehem, D. Wrege, and D. Ferrari, "Exact Admission Control for Networks with a Bounded Delay Service", *IEEE/ACM Transactions on Networking*, vol.4, no.6, Dec. 1996, pp.885-901.

[9] S. Choi and K.G. Shin, "An Uplink CDMA System Architecture with Diverse QoS Guarantees for Heterogeneous Traffic", *IEEE/ACM Transactions on Networking*, vol.7, no.5, Oct. 1999, pp.616-628.

[10] S. Choi and K.G. Shin, "A Unified Wireless LAN Architecture for Real-time and Non-real-time Communication Services", *IEEE/ACM Transactions on Networking*, vol.8, no.1, Feb. 2000, pp.44-59.

[11] N. Ruangchaijatupon, L. Wang, and Y. Ji, "A Study on the Performance of Scheduling Schemes for Broadband Wireless Access Networks", *2006 International Symposium on Communications and Information Technologies, ISCIT '06.*, pp.1008-1012, Thailand, Oct. 19-21, 2006.

[12] D. Niyato and E. Hossain, "Radio resource management games in wireless networks: an approach to bandwidth allocation and admission control for polling service in IEEE 802.1", *IEEE Wireless Communications*, vol.14, no.1, Feb. 2007, pp.27-35.

[13] L. Lin, W. Jia, and W. Lu, "Performance Analysis of IEEE 802.16 Multicast and Broadcast Polling based Bandwidth Request", *IEEE Wireless Communications and Networking Conference, WCNC 2007*, pp.1854-1859, Hongkong, March 11-15, 2007.

[14] J. Sun, Y. Yao, and H. Zhu, "Quality of Service Scheduling for 802.16 Broadband Wireless Access Systems", *2006 IEEE 63rd Vehicular Technology Conference, VTC 2006-Spring*, vol.3, pp.1221-1225, Melbourne, Australia, May 7-10, 2006.

[15] K. Vinay, N. Sreenivasulu, D. Jayaram, and D. Das, "Performance Evaluation of End-to-end Delay by Hybrid Scheduling Algorithm for QoS in IEEE 802.16 Network", *2006 IFIP International Conference on Wireless and Optical Communications Networks*, pp.5-9, Bangalore, India, April 11-13, 2006.

[16] J. Song, H. Choi, H. Kim, S. Kwon, D. Cho, H. Chang, G. Lim, and J. Kim, "Performance comparison of 802.16d OFDMA, TD-CDMA, cdma2000 1xEV-D0 and 802.11a WLAN on voice over IP service", *IEEE 61st Vehicular Technology Conference*, vol.3, pp.1965-1969, Stockholm, Sweden, May 30-June 1, 2005.

[17] M. Hawa and D.W. Petr, "Quality of Service Scheduling in Cable and Broadband Wireless Access Systems", *The Tenth IEEE International Workshop on Quality of Service, IWQoS 2002*, pp.247-255, Miami Beach, USA, May 15-17, 2002.

[18] R. Jayaparvathy, G. Sureshkumar; and P. Kanakasabapathy, "Performance Evaluation of Scheduling Schemes for Fixed Broadband Wireless Access Systems", Jointly held with *2005 IEEE 7th Malaysia International Conference on Communication and 2005 13th IEEE International Conference on Networks*, vol.2, pp.6-11, Kuala Lumpur, Malaysia, Nov. 16-18, 2005.

[19] D. Niyato and E. Hossain, " Queue-aware Uplink Bandwidth Allocation and Rate Control for Polling Service in IEEE 802.16 Broadband Wireless Networks ", *IEEE Transactions on Mobile Computing*,vol.5, no.6, June 2006, pp.668-679.

[20] R. Mukul, P. Singh, D. Jayaram, D. Das, N. Sreenivasulu, K. Vinay, and A. Ramamoorthy, "An Adaptive Bandwidth Request Mechanism for QoS Enhancement in

WiMax Real Time Communication", *2006 IFIP International Conference on Wireless and Optical Communications Networks*, pp.5-9, Bangalore, India, April 11-13, 2006.

[21] H. Wang, W. Li, and D.P. Agrawal, "Dynamic Admission Control and QoS for 802.16 Wireless MAN", *Fourth Annual Wireless Telecommunications Symposium, WTS2005*, pp.60-66, California, USA, April 28-30, 2005.

[22] S.A. Xergias, N. Passas, and L. Merakos, "Flexible Resource Allocation in IEEE 802.16 Wireless Metropolitan Area Networks", *The 14th IEEE Workshop on Local and Metropolitan Area Networks*, pp.6-11, Chania Crete, Greece, Sep 18-21, 2005.

[23] Q. Liu, X. Wang, and G. B. Giannakis, "A Cross-layer Scheduler Design with QoS Support for Wireless Access Networks", *Proc. of the 2nd IEEE International Conference on Quality of Service in Heterogeneous Wired/Wireless Networks (QShine'05)*, pp. 21-27, FL, USA, Aug. 22-24, 2005.

[24] C. Hoymann, "Analysis and Performance Evaluation of the OFDM-based Metropolitan Area Network IEEE 802.16", *Computer Networks*, vol.49, no.3, June 2005, pp.341-363.

[25] D. Niyato and E. Hossain, "Joint Bandwidth Allocation and Connections Admission Control for Polling Services in IEEE 802.16 Broadband Wireless Networks", *Proc. of IEEE International Conference on Communications*, vol.12, pp.5540-5545, Istanbul, Turkey, June 2006.

[26] D. Niyato and E. Hossain, "A Radio Resource Management Framework for the IEEE 802.16-based OFDM/TDD Wireless Mesh Networks", *Proc. of IEEE International Conference on Communications*, vol.9, pp.3911-3916, Istanbul, Turkey, June 2006.

[27] 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*, vol.9, no.8, Aug. 2005, pp.691-693.

[28] D. Zhao and X. Shen, " Packet Voice Transmissions in IEEE 802.16 Backhaul Networks", *IEEE Communication Magazine*, vol.14, no.1, Feb. 2007, pp.44-51.

[29] M. Ergen, S. Coleri, and P. Varaiya, "QoS Aware Adaptive Resource Allocation Techniques for Fair Scheduling in OFDMA Based Broadband Wireless Access Systems", *IEEE Transaction on Broadcasting*, vol.49, no.4, Dec. 2003, pp.362-370.

[30] Bo Rong, Yi Qian, and Hsiao-Hwa Chen, "Adaptive Power Allocation and Call Admission Control in Multiservice WiMax Access Networks", *IEEE Wireless Communications*, Vol.14, no.1, Feb. 2007, pp.14-19.

[31] R. Iyengar, P.Iyer, and B. Sikdar, "Analysis of 802.16 Based Last Mile Wireless Networks", *IEEE Global Telecommunications Conference, GLOBECOM '05*, vol.5, pp.3123-3127. Misso. USA, Nov.28-Dec.2, 2005.

[32] Q. Ni, A. Vinel, Y. Xiao, A. Turlikov, and T. Jiang, "Investigation of Bandwidth Request Mechanisms under Point-to-Multipoint Mode of WiMax Networks", *IEEE Communications Magazine*, vol.45, no.5, May 2007, pp.132-138.

[33] L. Badia, A. Baiocchi, A. Todini, S. Merlin, S. Pupolin, A. Zanella, and M. Zorzi, "On the Impact of Physical Layer Awareness on Scheduling and Resource Allocation in Broadband Multicellular IEEE 802.16 Systems", *IEEE Wireless Communications*, vol. 14, no.1, Feb. 2007, pp.36-43.

[34] M. Cao, W. Ma, Q. Zhang, and X. Wang, "Investigation of Bandwidth Request Mechanisms under Point-to-Multipoint Mode of WiMAX Network", *IEEE Communications Magazine*, vol.45, no.4, April 2007, pp.1455-1464.

[35] D. Niyato, E.Hossain, and J. Diamond, "IEEE 802.16/WiMAX-based Broadband Wireless Access and Its Application for Telemedicine/e-health Services", *IEEE Wireless Communications*, vol.14, no.1,Feb. 2007, pp.72-83.

[36] D. Anick, D. Mitra, and M. Sondhi, "Stochastic Theory of a Data-Handling System with Multiple Sources", *Bell system Tech. Journal*, vol.61, no.8 , Oct. 1982, pp.1871-1894.

[37] L. Dai and D. Zhao, "Uplink Scheduling for Supporting Real-time Voice Traffic in IEEE 802.16 Networks", *Proc. of International Conference on Heterogeneous Networking for Quality, Reliabiility and Rebustness*, Vancouver, BC, Canana, Aug. 14-17, 2007.