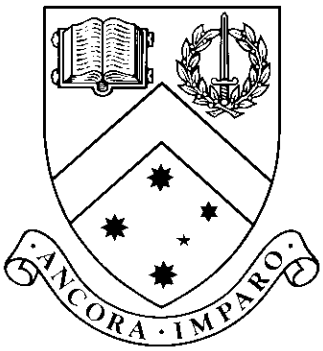
Quality of Service Support in IEEE 802.16 Broadband Wireless Access Networks

by

Ehsan Asadzadeh Aghdaee



Thesis

Submitted by Ehsan Asadzadeh Aghdaee

for fulfillment of the Requirements for the Degree of

Master of Engineering Science

Department of Electrical & Computer System Engineering
Monash University

October, 2006

Contents

List of Tables

List of Figures

Abstract

The IEEE 802.16-2004 standard for Broadband Wireless Access (BWA) defines the air interface specification in Wireless Metropolitan Area Network (MAN). The medium access control signalling has been well-defined in the IEEE 802.16 specification, but the admission control and scheduling mechanisms, which are important components for providing Quality of Service (QoS), are left unspecified in the standard and hence remain as open research problems.

This thesis presents a QoS architecture for the IEEE 802.16 standard and describes admission control and scheduling algorithms for the architecture. The scheduling mechanisms are important both at the base station (BS) and subscriber station (SS) to provide quality of service to different types of traffic such as constant-bit rate (CBR), real-time and non-real-time variable bit rate, and best effort (BE).

The initial part of the thesis reviews the litrature on IEEE 802.16 BWA network and various methods of admission control and scheduling. It proposes an admission control policy for multi-class admission control problem in IEEE 802.16 BWA. In this method, a certain amount of bandwidth is reserved for unsolicited grant of service (UGS) and real-time polling service (rtPS) traffic classes. The performance of the proposed algorithm is comparable to Complete Sharing (CS) and Complete Partitioning (CP) algorithms at lower loads, but outperforms both CS and CP algorithms at higher loads.

The main part of the thesis presents a set of bandwidth allocation and scheduling mechanisms for both BS and SS and their performance evaluation. Earliest Deadline First with Bandwidth Reservation (EDF-BR) and Multi-Class Allocation (MCA) algorithms are employed at the BS scheduler. At the subscriber station level, enhancements to Priority Queuing (PQ) algorithm are employed. The performance evaluation of the scheduling algorithms are carried out by simulation experiments to characterize the delay and packet loss rate of UGS, rtPS, nrtPS and BE class of services under various load conditions. The proposed enhancements are shown to have better performance.

Acknowledgment

For me, it has been a big journey from the start to finish. A journey that has been wrought with endless sleepless nights, disappointments, and a lot of struggle. Needless to say, that the only thing that kept me going was the support of a number of people. First and foremost my supervisor, Dr. Nallasamy Mani, "Without your unstinting support, faith in my work, there is just no way that I could have completed my thesis. You have been always there whenever I needed your help in any form and that is what saw me through times of confusion, and helplessness. I thank you for everything." I thank my associate supervisor, Prof. Bala Srinivasan for his ceaseless encouragement and invaluable advice.

On a more personal note, I would like to thank my parents for always encouraging me to do my best and strive for excellence.

Ehsan Asadzadeh Aghdaee

Declaration

I declare that, to the best of my knowledge, the research described herein is original except where the work of other is indicated and acknowledged, and that the thesis has not, in whole or in part, been submitted for any other degree at this or any other university.

Ehsan Asadzadeh Aghdaee
October 1, 2006

Chapter 1

Introduction

In recent years, there has been a considerable growth in demand for high-speed wireless Internet access and this has caused the emergence of new short-range wireless technologies (viz. IEEE 802.11) and also long-range wireless technologies (viz. IEEE 802.16).

Long-range wireless technologies, in particular IEEE 802.16, offer an alternative to the current wired access networks such as cable modem and digital subscriber line (DSL) links. The IEEE 802.16 has become an attractive alternative, as it can be deployed rapidly even in areas difficult for wired infrastructures to reach and also, it covers broad geographical area in more economical and time efficient manner than traditional wired systems.

In comparison to 802.11 standard, 802.16 can service a much greater number of simultaneous users and approximately 50 times greater (radial) coverage. Additionally, the set up and subscription costs are significantly less due to the lack of physical cabling. The IEEE 802.16a standard has a range of up to 30 miles with data transfer speeds of up 70 Mbps.

At the same time, the growth in adopting a broadband wireless access (BWA) network has significantly increased the customer demand for guaranteed quality of service (QoS). The provision of QoS has become a critical area of concern for BWA providers. The IEEE 802.16 standard appear to offer a solution for this problem, by establishing a number of unique and guaranteed QoS parameters in terms of delay, jitter and throughput. This enables service providers to offer flexible and enforceable QoS guarantees, a benefit that has never been available with other fixed broadband wireless standards [1].

The IEEE 802.16 has great potential in the broadband market. Recent investments in Australia [2] indicate that IEEE 802.16 may become prevalent in the coming years. The IEEE 802.16 standard is now supported by both the IEEE as well as the European Telecommunications Standards Institute (ETSI) HiperMAN standard. The changes adopted by either body are being reflected in the baseline technical requirements. It is also supported by parallel interoperability efforts through the industry forum known as Worldwide Interoperability for Microwave Access (WiMAX) [1] Forum™.

1.1 Quality of Service Fundamentals

Integrated telecommunication networks carry traffic of several different classes, including one or more real-time traffic classes, each with its own set of traffic characteristics and performance requirements. Two different approaches have been developed

[1]WiMAX is an acronym that stands for Worldwide Interoperability for Microwave Access, a certification mark for products that pass conformity and interoperability tests for the IEEE 802.16 standards.

to deal with this phenomenon. The first approach is circuit-switched, in which sufficient resources are allocated to each connection to handle its peak rate. This guarantees that the connection will obtain the quality of service (QoS) it requires, but at the cost of under-utilizing the network resources. The second popular approach is the packet-switched approach, in which traffic from all sources are broken into packets and statistical multiplexing techniques are used to combine all the network traffic through single switching node. This allows higher network utilization, but requires more sophisticated controls to ensure that the appropriate QoS is provided [3].

Packet-switched networks were originally designed to provide best effort services, but later the demand for differentiated service caused the packet-switched networks to be integrated with QoS architecture. QoS architecture introduces tools to treat packets in different way, for example, real-time packets will be given priority over non-real-time packets allowing them to traverse the network faster and arrive at the destination within their required delay bounds.

QoS in packet-switched networks can be characterized in terms of a specific set of parameters including delay, delay jitter, bandwidth and loss or error rate. Delay is the time that it takes for the packet to traverse from source to destination, it consists of transmission delay, propagation delay and queuing delay in intermediate routers. Delay jitter is the fluctuation or variation in end-to-end delay from one packet to the next within the same packet flow. Bandwidth is a measure of the amount of data that a network allows one flow to transmit over a period of time. Drop rate of one flow measures the number of packets which are dropped due to buffer overflow, transmission error or expiry due to over-staying than their maximum delay bound.

The ability to manage congestion and maintain QoS in a packet-switched network requires the collaboration of many components in the QoS architecture. The three main components are as follows:

- **Admission Control** determines whether a new request for resources can be granted or not based on the knowledge of total network capacity and the already accepted flows. It has a critical role of limiting the number of flows admitted into the network so that each individual flow obtains its required QoS.

- **Packet Scheduling** is a critical component in any QoS architecture, as the packets traverse different switches (or routers) along their way through the network. It determines the order in which packets belonging to different flows transmit on the output link, thus ensuring that packets from different applications meet their QoS constraints. The basic scheduler architecture is shown in figure 1.1, in which several inputs are buffered and there is a single scheduler (server). An optimal scheduling mechanism will provide the necessary QoS guarantees required by different classes of traffic while efficiently utilizing the network resources.

- **Buffer Management** has the responsibility of discarding one or more incoming packets before the output buffer overflows, in order to improve the performance of the network. One of the most common packet drop strategies is Random Early Detection (RED)[4].

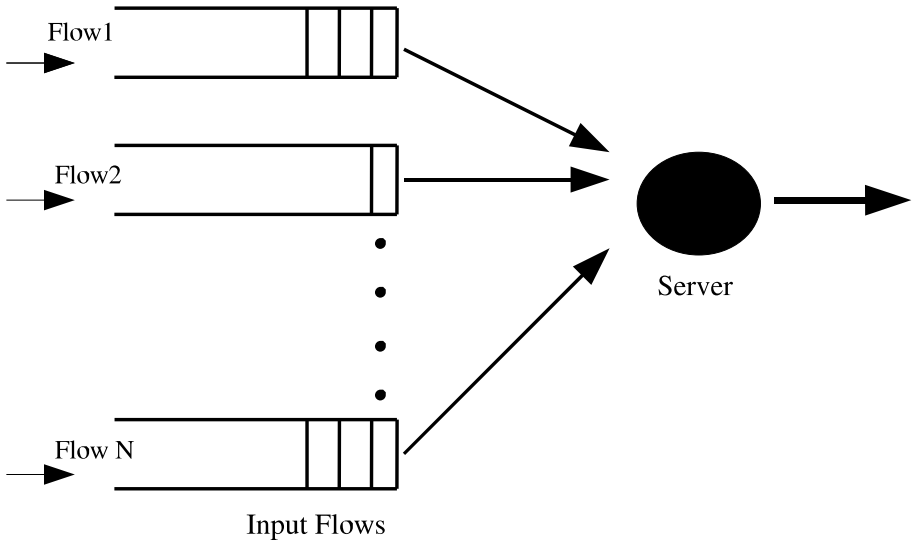


Figure 1.1: Simple scheduler architecture

1.2 QoS in IEEE 802.16

The IEEE 802.16 is designed with the aim of providing a guaranteed QoS. There are a number of features included in the current standard. However, the details of some of these features are still unspecified. In this section, we initially describe the current QoS architecture of the IEEE 802.16 and subsequently the unspecified features.

In IEEE 802.16, the QoS architecture is part of a MAC layer. For wireless networks it is natural to integrate the QoS architecture with the MAC protocol. The MAC protocol coordinates communication over the shared wireless medium. The 802.16 MAC provides QoS differentiation for different types of applications that might operate over 802.16 networks. IEEE 802.16 defines four classes of service to support the QoS [5].

- **Unsolicited Grant Service(UGS)** is designed to support real-time data streams consisting of fixed-size data packets issued at periodic intervals, such as T1/E1 and Voice over IP without silence suppression. UGS is prohibited from using any contention requests, and there is no explicit bandwidth request issued by subscriber station (SS). The base station (BS) provides fixed size access slots at periodic intervals to the UGS flows. However, the reserved bandwidth may be wasted when a corresponding UGS flow is inactive.

- **Real-Time Polling Service(rtPS)** is designed to support real-time data streams consisting of variable-sized data packets that are issued at periodic intervals, such as moving pictures experts group (MPEG) video. The mandatory QoS service flow parameters for this scheduling service are minimum reserved traffic rate, which is defined as the minimum amount of data transported on the connection over period of time and maximum latency, which is the upper bound on the waiting time of a packet in the network. The rtPS flows are polled by BS through a unicast request polling frequently enough to meet the delay requirements of the service flows.

- **Non Real-Time Polling Service(nrtPS)** is designed to support delay-tolerant data streams consisting of variable-sized data packets for which a minimum data rate is required. The nrtPS flows like an rtPS flow are polled through a unicast request polling but at time-scale of one second or less. The nrtPS flows can also receive a few request polling opportunities during network congestion, and are also allowed to use the contention requests.

- **Best Effort Service(BE)** is designed to support data streams for which no minimum service level is required and therefore may be handled on a bandwidth available basis.The BE flows are allowed to use contention request opportunity. The applications in this class receive any bandwidth remaining after it has been allocated to all other classes.

1.2.1 Admission Control

In IEEE 802.16 before a subscriber station (SS) can initiate a new connection, it must first make a request to the base station (BS) with the service contracts required. The BS may reject the request based on its ability to uphold the requirements. The admission control mechanism at the BS is not specified in the standard. It is possible that the BS admits a connection based on statistical QoS, where, on average, all connections would not have their QoS satisfied, or that the BS could have a more strict model where QoS is absolutely guaranteed, even in the worst case scenario. The admission control problem will be discussed in more detail in chapter 3.

1.2.2 Scheduling

IEEE 802.16 specifies only the outbound traffic scheduling goals, and not the methodology. Essentially, traffic is to be serviced such that service contracts have no or minimal disruption.

Bandwidth can be requested in the initialization of a connection, such as an UGS traffic class connection, or, due to the uplink burst profile being so dynamic in other classes of traffic, the SSs would indicate current uplink bandwidth requirements for each.

When a SS is granted bandwidth it may choose how this is allocated among its connections. This means that connections may "borrow" bandwidth from other connections within the same SS. The only exception is UGS traffic, which is granted a fixed amount of bandwidth per frame and may not use more than this. Of course it may use less, in which case the unused bandwidth passes to other connections to use.

1.3 Research Motivation and objectives

Our research is motivated by the fact that user expectations of wired and wireless networks have increased with regard to a large variety of services and applications with different QoS requirements. The implementation of QoS guarantees in such networks is a necessary prerequisite for the efficient support of services and applications over such networks. We believe that this topic requires extensive research on both the qualitative and the quantitative sides to achieve the most suitable QoS architecture.

We are also motivated by the fact that an ideal QoS architecture needs to seamlessly support the same QoS constraints across heterogeneous types of networks including those designed for the wired and wireless environments. Our emphasis would be directed towards broadband wireless access networks in which stringent requirements are imposed on the QoS architecture due to the numerous limitations of the wireless channel.

A further motivating factor is that IEEE 802.16 does not specify the methodology to use for scheduling and admission control. Research in this area has only recently

started to gain momentum, with many researchers proposing algorithms for efficient bandwidth allocation; however, limitations are still apparent in their work.

The objective of this thesis is to develops a new and efficient scheduling architecture to support bandwidth and delay QoS guarantees for the IEEE 802.16 broadband wireless access (BWA) standard. Our design objectives are simplicity and improved network performance. The architecture developed supports various types of traffic including constant bit-rate, variable rate bit-rate and best effort. Number of QoS aware scheduling algorithms is proposed to efficiently schedule data on both BS and SS.

A new priority admission control strategy is also introduced, which admits the connections into the system according to their QoS requirements. The new strategy can be used to minimize the connection blocking probability of higher priority connections which is acceptable from both users' and service providers' point of view.

1.4 Thesis Organization

This dissertation is organized as follows: Chapter 2 describes the MAC layer functionalities of the IEEE 802.16. It also presents a survey of literature in the area of traffic scheduling and admission control. Chapter 3 describes the proposed Admission Control framework for broadband wireless access (BWA). Chapter 4 describes the proposed QoS scheduling architecture for the IEEE 802.16 standard. It also explains the simulation models and the simulation methodology used in this study. The results and their significance are discussed in Chapter 5. Chapter 6 concludes

the thesis by summarizing the contributation of this research effort, and providing some directions for further research.

Chapter 2

Background and Related Work

This chapter describes the evolution of IEEE 802.16 BWA and also explain the other BWA technologies that are competing with IEEE 802.16 standard. It also provides an overview of IEEE 802.16 standard, and a review of literature on the previous contributions towards admission control and traffic scheduling in fixed and mobile wireless networks.

2.1 Evolution of IEEE 802.16

The IEEE began the development of the 802.16 WirelessMAN Air Interface specification in 1999 and was then published as IEEE 802.16-2001 on April 8, 2002. It was a global project involving hundreds of engineers. The initial version of the 802.16 standard operates in the 10-66GHz frequency band and requires line-of-sight between stations. A number of extensions to this standard have since been made.

An extension to the 802.16 standard, 802.16a, was accepted on January 29 and published on April 1st, 2003 [6]. It was designed to be a MAN/WAN wireless standard from the ground up. Products conforming to the 802.16a specification would take advantage of quality of service (QoS) implementations using Time Division Duplex (TDD) or Freqeuency Division Duplex (FDD) access methodologies instead of a wireline CSMA/CA approach.

The 802.16a standard has the following characteristics [6]:

- it is designed for the 2-11GHz band and supports both licensed and unlicensed bands

- it specifies the physical layer and medium access control layer of the air interface of interoperable fixed point-to-multipoint (PMP)

- it specifies an optional mesh topology

IEEE 802.16 addresses the "first-mile/last-mile" connection in Wireless metropolitan area networks. It focuses on the efficient use of bandwidth between 10 GHz and 66 GHz. It also defines a medium access control (MAC) layer that supports multiple physical layer (PHY) specifications customised for the frequency band of use. The 802.16 MAC allows multiple service flows with different quality of service (QoS) parameters on the same subscriber station.

Due to the use of lower frequency bands and orthogonal frequency division multiplexing (OFDM), 802.16a does not require optical line-of-sight (OLOS) for transmission. 802.16c, another extension, adds interoperability with other frequencies around 10-66 GHz [7].

In June 2004, the IEEE-SA (Standards Association) announced a new official standard IEEE 802.16-2004 [5], which consolidated 802.16, 802.16a and 802.16c. A key result of this was to clarify the standard within the industry.

The IEEE 802.16 topology consists of two types of fixed (not mobile) stations. The first one is the base station (BS), which is usually installed on top of buildings or towers to serve subscriber stations up to 50Km away from it. The second type of station is the subscriber station (SS), which is located in home and business premises. IEEE 802.16 may operate in either point to multi point(PMP) or mesh basis. In PMP mode, a central base staion(BS) assigns and regulates transmission periods for each subscrirber station(SS). In mesh mode, traffic can be routed explicitly between SSs or based on a hierarchical scheme. Figure 2.1 shows a possible PMP installation where subscriber stations provide network access for other PCs through base station.

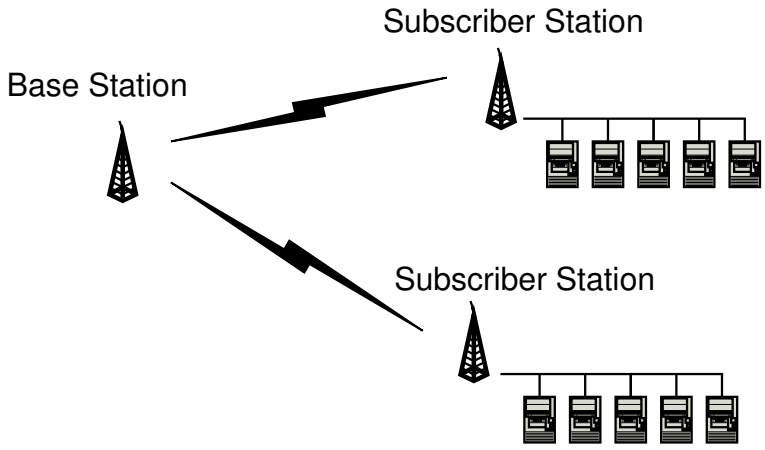


Figure 2.1: IEEE 802.16 possible installation

The IEEE 802.16 standard specifies the air interface for PMP broadband wireless access (BWA) systems providing multiple services. The specifications include the

MAC and PHY layers. The MAC layer is structured to support multiple PHY specifications, each suited to a particular operational environment [8].

2.1.1 Competing Technologies

Wideband Code Division Multiple Access (W-CDMA) is a wideband spread-spectrum 3G mobile telecommunication air interface utilizing the Code Division Multiple Access (CDMA) multiplexing scheme. It provides simultaneous support for a wide range of services with different characteristics on a common 5MHz carrier. Key features include [9]:

- Supports two basic modes of duplex: Frequency Division Duplexing (FDD) and Time Division Duplexing (TDD) modes
- Employs coherent detection on uplink and downlink based on the use of pilot symbols
- Inter-cell asynchronous operation
- Variable rate transmission
- Multi-code transmission
- Adaptive power control based on signal-to-interference ratio (SIR)
- Multiuser detection and smart antennas can be used to increase capacity and coverage

Universal Mobile Telecommunications System (UMTS) [10] using W-CDMA as the underlying standard is a third-generation (3G) mobile phone technology and was

standardised by the 3rd Generation Partnership Project (3GPP). UMTS supports up to 1920 kbit/s data transfer. UMTS is a direct competitor to WiMAX.

High-Speed Downlink Packet Access (HSDPA) [11] is a new mobile telephony protocol using W-CDMA. It is packet-based with data transmissions up to 8-10 Mbit/s or 20 Mbit for Multiple-Input Multiple-Output (MIMO). HSDPA can include Adaptive Modulation and Coding (AMC), MIMO, Hybrid Automatic Request (HARQ), fast scheduling, fast cell search, and advanced receiver design. Release 4 of this specification introduces IP support and release 5 provides data rates up to 10 Mbit/s specifically for multimedia services. Release 6 is aimed at further increasing the data rates to 20 Mbit/s.

CDMA2000 [12], being another 3G hybrid of the original CDMA technology, can provide access to voice and data services. The first version of this technology (CDMA2000 1xRTT) allows speeds of 144 kbit/s. Later versions, CDMA2000 1xEV can provide 3.1 Mbit/s on the downlink and 1.8 Mbit/s on the uplink.

2.1.2 Medium Access Control layer

The IEEE 802.16 MAC is comprised of three sublayers as shown in Fig 2.2. It includes service specific convergence sublayer (CS) that interface to higher layers, above the core MAC common part sublayer (CPS) that carries out the key MAC functions [5]. Below the common part sublayer is the privacy sublayer. In the following sections we will describe the details of each of these sublayers and also some key MAC operations will be reviewed.

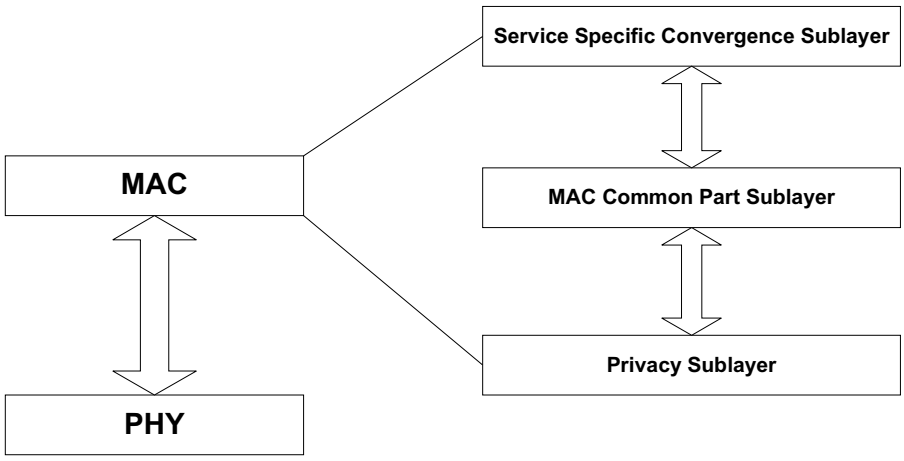


Figure 2.2: MAC layer of IEEE 802.16

Service Specific Convergence Sublayer (CS)

CS is responsible to map the transport-layer-specific traffic to a MAC layer that carry any traffic type. IEEE Standard 802.16 defines two general service specific CS sublayers for mapping services to and from 802.16 MAC connections. The ATM convergence sublayer is defined for ATM services, and the packet convergence sublayer is defined for mapping packet services such as IPv4, IPv6, Ethernet, and virtual local area network (VLAN) [13]. The main responsibility of this sublayer is to classify the external network service data units (SDUs) and associating them to the proper MAC service flow and connection identifier (CID). It may also include such functions as payload header suppression (PHS) and reconstruction. The internal format of CS payload is unique to the CS, and the MAC CPS is not required to understand the format syntax or parse any information semantics from the payload [5].

MAC Common Part Sublayer (CPS)

The MAC CPS provides the core MAC functionalities of system access, connection establishment, bandwidth allocation, and connection maintenance [14]. It receives data from the various CSs, through the MAC service access point (SAP), classified to particular MAC connections. Accordingly, certain QoS level is applied to the transmission and scheduling of data.

The IEEE 802.16 MAC protocol uses a connection-oriented approach. All services, including inherently connectionless services, are mapped to a connection. This provides a mechanism for requesting bandwidth, associating QoS and traffic parameters, transporting and routing data to the appropriate convergence sublayer. All data transmissions take place in the context of a connection. Upon entering the network, each SS creates one or more connections over which their data are transmitted to and from the BS. The MAC CPS schedules the usage of air link resources, and provides service differentiation.

Privacy Sublayer

Privacy Sublayer provides authentication, secure key exchange and encryption. IEEE 802.16's privacy protocol is based on the Privacy Key Management (PKM) protocol of the Data Over Cable Service Interface Specification (DOCSIS) Baseline Privacy Interface (BPI+) but has been enhanced to fit seamlessly into the IEEE 802.16 MAC protocol and to better accommodate stronger cryptographic methods, such as the recently approved Advanced Encryption Standard [13].

MAC Frame Format

Periods of communication are divided into frames, which, in time division duplexing (TDD), is then divided into uplink and downlink subframes as shown in figure 2.3. A frame is made up of physical slots (PSs) which are then made up of symbols representing bits of data. The unit of bandwidth allocation is the PS.

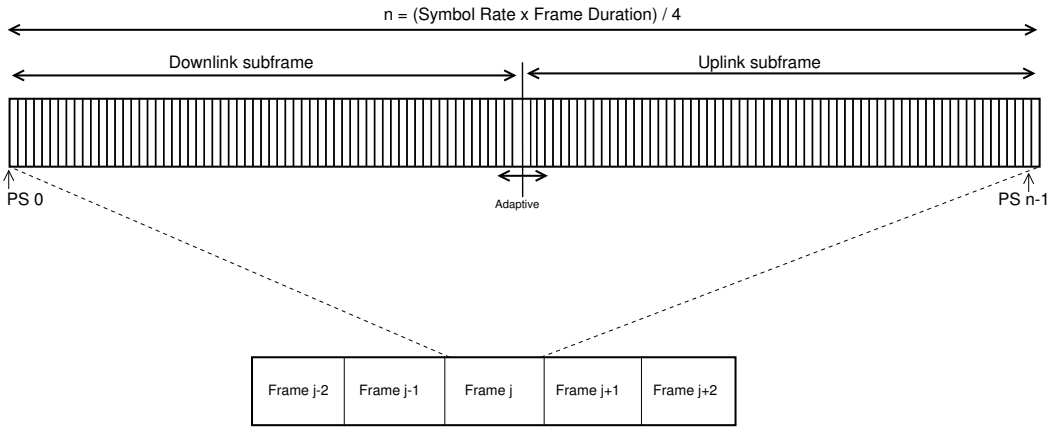


Figure 2.3: TDD frame structure

Whilst the frame size remains static within a network, the uplink and downlink subframes are dynamic in size determined by the BS. In the downlink direction (from BS to SS), the system operates in time division multiplexing (TDM) fashion. However, in the uplink direction, time division multiple access (TDMA) is used as the SSs share the uplink access to the BS on demand basis.

Network Entry and Initialization

In order to communicate on the network an SS needs to successfully complete the network entry process with the BS. Figure 2.4 shows the process a subscriber station follows to connect to a PMP network. After the SS decides on which channel

to attempt communication, the SS tries to find a downlink (DL) channel and to synchronize at the PHY level (it detects the periodic frame preamble), then the MAC layer looks for DL Channel Descriptor (DCD) and UL Channel Descriptor (UCD) to get information on modulation and other DL and UL parameters. When an SS has synchronized with the DL channel and received the DL and UL MAP for a frame, it begins the initial ranging process by sending a ranging request (RNG-REQ) MAC message. If it does not receive a response, the SS sends the ranging request again in a subsequent frame, using higher transmission power. Eventually the SS receives a ranging response, so that it can adjust local parameters e.g. its transmit power.

After successful completion of initial ranging, the SS sends a SS Basic Capability Request (SBC-REQ) message to the BS describing its capabilities. The BS accepts or denies the SS, based on its capabilities. The SS then sends a registration request (REG-REQ) message to the BS, and the BS sends a registration response (REQ-RSP) to the SS. From this point, the SS is able to use traditional 802 network devices to enable IP connectivity and assign attributes over the network. The BS and SS maintain the current date and time using the time of the day protocol. The SS then downloads operational parameters using trivial file transfer protocol (TFTP). Finally, the BS may establish one connection to the SS for each of the rtPS, nrtPS and BE classes of traffic for the purpose of sending control messages.

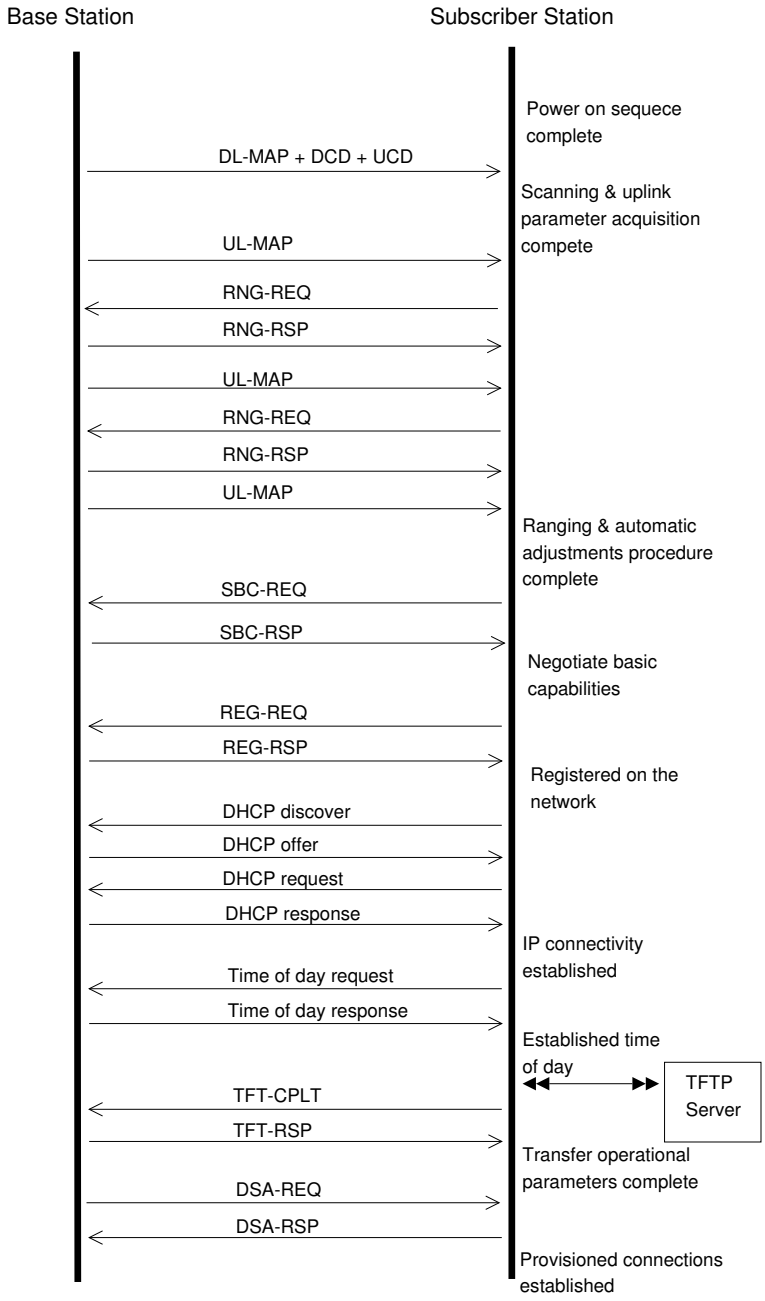


Figure 2.4: Typical PMP network entry and initialisation

Bandwidth Allocation and Request Mechanism

At the start of each frame, the BS broadcast the Downlink MAP (DL-MAP) and Uplink MAP (UP-MAP) which contains allocation transmissions and receptions for the participating SSs. The downlink frame is shown in figure 2.5. The frame starts with a frame control section that contains DL-MAP for the current downlink frame as well as the UL-MAP for a frame in future. The DL-MAP informs all SSs when to listen for transmissions destined for them in the current frame. The UL-MAP informs SSs of their transmission opportunities as a response to their dynamic bandwidth requests, or on the basis of prior service agreements.

Figure 2.5: Downlink Frame Structure

The downlink frame typically contains a TDM portion immediately following the uplink and downlink MAPs. This TDM portion carries downlink data for SS using a negotiated burst profile identified by the Downlink Interval Usage Code (DIUC). The TDM portion followed by a TDMA portion to allow better support for half-duplex SSs.

The uplink frame comprises transmissions from different SSs based on the discretion of the BS uplink scheduler as indicated in the UL-MAP. The scheduler may also earmark time for initial maintenance and bandwidth requests in any given frame. The uplink frame also contains guard times in the form of SS Transition Gaps. These gaps are used by the BS to re-synchronize to different SS transmissions. A typical uplink frame is shown in 2.6.

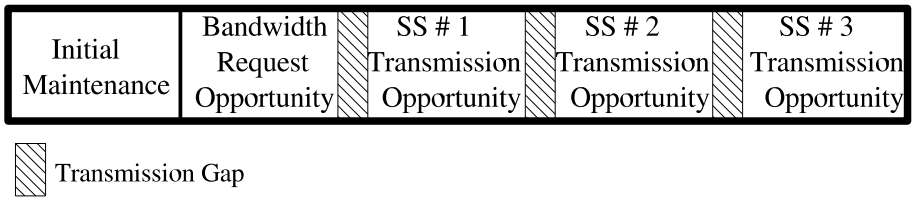


Figure 2.6: Uplink Frame Structure

There are three methods for a SS to be granted bandwidth.

- A request message can be sent to the BS as a stand-alone bandwidth request header or as a PiggyBack request. An requests for bandwidth may be incremental or aggregate. An incremental request increases the current allocation by the amount specified, and an aggregate request tries to replace the current allocation with the new value.

- A BS may simply grant bandwidth to a SS either in response to a request from a station, or because of an administrative policy providing some amount of bandwidth to a particular SS.

- Polling is used by the BS to grant small windows of bandwidth in order for SSs to request bandwidth. Polling is done on a per SS basis, bandwidth requested on a connection basis and bandwidth allocation is done on a SS basis.

2.2 Traffic Scheduling

Traffic scheduling is a process of determining the order in which packets get transmitted at the output of each router, thus ensuring that packets from different flows

meet their QoS constraints. Over the past few years, an enormous amount of research has been conducted in the area of traffic scheduling. With the advent of the wireline integrated service networks, many packet-level traffic scheduling policies have been proposed for switches and routers in order to meet the requirements of different service classes [15].

Since the main focus of this thesis is to discuss the performance of scheduling discipline in the context of IEEE 802.16, relevant work in the area of queuing algorithms are presented in the following sections.

FIFO Queuing

First-in First-out (FIFO) queuing considers as one of the simplest queuing policies, in which the order of arrival is the same as the order of service. When the queue become full due to traffic congestion, incoming packets are dropped. FIFO does not provide any protection against misbehaving flows and it is not very flexible in the sense that it does not provide performance guarantees for different flows. The main advantage of the FIFO queuing is its simple implementation.

Priority Queuing

Priority Queuing (PQ) [4] was one of the first queuing variation to provide a different level of treatment to different priority flows. All incoming packets are classified according to information in the packets as belonging to one of the priority classes and accordingly placed in the appropriate queue. The scheduler serves the queues in the priority fashion that a packet is scheduled from the head of service queue q as long as all queues of higher priority are empty. The main disadvantages of PQ are the lack of scalability and starvation of lower priority classes. Some suggest [4] to

use some sort of filter to police the amount of traffic that the higher priority classes are transmitting, in order to provide some level of service for lower priority classes.

Class-Based Queuing (CBQ)

The class based queuing (CBQ) is the variation of PQ [15]. It is designed to prevent complete resource denial to a particular class of service, thereby addressing the major weakness of strict priority queuing. With CBQ, each class is associated with a portion of link capacity (typically measured in byte), which specify the amount of traffic that should be transmitted from each queue in every scheduling cycle. A class represents a traffic stream or aggregate of traffic streams that are grouped according to traffic type, protocol or other criteria. The CBQ is more than a queuing scheme, it is also a QoS scheme that identifies different types of traffic and queues the traffic according to predefined parameters. In other word, certain portion of bandwidth is dedicated to a particular types of traffic.

FQ Queuing

Nagle [16] proposed a fair queuing (FQ) in order to overcome some of the drawbacks of FIFO queuing. This scheme is fair in that each busy flow get to send exactly one packet per scheduling cycle. However, in this scheme flows with shorter average packet size are penalized in compare to the flows with longer average packet size since the flows with longer average packet size receive more capacity. According to stalling [17], bit-round fair queuing (BRFQ) developed by to improve the performance of FQ. The BRFQ takes both packet length and flow identification into account to schedule packet. Although BRFQ is an improvement over FQ or FIFO in that it fairly allocates the available capacity among all active flows, it is not able to provide

differentiated service (QoS) by allocating different amount of capacity to different flows.

Parekh and Gallager [18] proposed a Generalized Processor Sharing (GPS) (also known as fluid fair queuing), which is an efficient and fair scheduling algorithm. With GPS, each flow is assigned a weight ϕ that determines the amount of service that this flow should receive. It is an idealized fluid-flow model that services all sessions simultaneously. The main use of GPS is as a reference model as we don't have fluid flow in reality. A Packet-by-packet GPS (PGPS) also refered to as WFQ [19], is an approximation of GPS behavior at packet level. It employs the same strategy as GPS in servicing the queues, that is after the current packet transmission finishes the next packet to be sent is the one with the smallest value of timestamp. One of the main drawbacks of PGPS is the complexity involved in computing the timestamp associated with each individual packet.

It is shown in [20] that when the number of connections increases WFQ does not provide a good approximation of GPS service discipline. The authors proposed Worst-case Fair Weighted Fair Queuing (WF^2Q), which provides a closer (more accurate) emulation of GPS than WFQ does. The main advantages of all these fair queuing algorithms are providing: bounded end-to-end delay on a per-flow basis and throughput guarantee to all traffic flows.

CBS Queuing

In another class of scheduling which is called cost based scheduling (CBS) [21], the time-varying cost functions are defined for each class of traffic. The aim is to schedule packets from different classes to minimize total cost. At any point of time each packet's cost can be calculated independently of any other queued packets.

However, it is not trivial to calculate the packets priorities except for simple cost functions and specifying service constraints using cost functions will be complicated.

EDF Queuing

Real-time scheduler such as Earliest Deadline First (EDF) [15] are designed for delay-sensitive applications and task sets with sophisticated characteristics. Each packet is assigned a deadline when it arrives. The deadline is defined in terms of the amount of time that the packet can stay in the queue before being transmitted. The scheduler selects the packet with the smallest deadline for transmission in every round. The dynamic nature of the priority in the EDF scheduler is evident from the fact that the priority of the packet increases with the amount of time it spends in the system. This ensures that packets with loose delay requirements obtain better service than they would in a static priority scheduler, without sacrificing the tight delay guarantees that may be provided to other flows. It is well known that for any packet arrival process where a deadline can be associated with each packet, the EDF policy is optimal in terms of minimizing the maximum lateness of packets. Here lateness is defined as the differences between the deadline of a packet and the time it actually transmitted on the link.

Wireless Specific Queuing

A detailed study conducted by [22] describes the issues and difficulties involved in using fair queuing in wireless networks. Wireless links generally posses characteristics that are quite different from those of wired links. Wireless link are subject to time and location dependent signal attenuation, fading, interference, and noise, which result in burst errors and time-varying channel capacities. Adapting fair

queueing to the wireless domain is very complicated, due to the above mentioned problems. The authors compared the performance of four wireless scheduling algorithms, which are Idealized Wireless Fair Queuing (IWFQ), the channel-condition Independent Fair Queuing (CIF-Q), the Server Based Fairness Approach (SBFA) and the wireless Fair Service algorithm (WFS). IWFQ [23, 24] is a realization of PGPS with a compensation mechanism for error-prone sessions . CIF-Q [25] and WFS [26] use similar virtual timestamping techniques to determine service order of arriving packets, but they differs in terms of how they compensate for errorneous transmissions.

The scheduling algorithms which are usually employed in the wireless code-division multiple access (CDMA) networks are Packet-by-Packet GPS, Dynamic Resource Scheduling (DRS), Wireless Multimedia Access Control Protocol with BER (WISPER), scheduled CDMA [27, 28, 29, 30]. The fundamental differences between scheduling in wired and wireless networks are the consideration of variable channel conditions, distributed channel access, and power consumption [24].

Scheduling in IEEE 802.16

Scheduling mechanisms are crucial for providing QoS guarantee in IEEE 802.16 networks. These mechanisms are important for both BS and SS to provide differentiated services among the different types of traffic corresponding to different connections. Some of the scheduling algorithms discussed above are suggested to be used in the IEEE 802.16 networks [31, 32, 33]. In [34], the authors proposed a QoS aware packet scheduling schemes for different types of traffic. In this scheme, PQ algorithm is suggested to be used as the BS scheduler, however the SS packet scheduler is not fully defined. A QoS architecture which uses a Wireless Fair Queuing (WFQ) as the

BS's scheduler and Multi-class Priority Fair Queuing (MPFQ) as the SS's scheduler is suggested in [35]. Cicconetti and Lenzini [36] have evaluated the performance of IEEE 802.16 MAC layer, using a Weighted Round Robin (WRR) as the uplink scheduler and Deficit Round Robin as the SS scheduler.

2.3 Call Admission Control

Call admission control (CAC) is a resource provisioning strategy to limit the number of call connections into the networks in order to reduce the network congestion and call dropping. There has been many research into admission control policies for circuit-switched networks with calls of multiple bit rates. Different admission control strategies allow different treatment of call requests which may or may not be admitted into service. In admission control calls maybe blocked or queued, or some combination of the two may be employed. Kraimeche [37] examines all three of these possibilities in detail, and investigate many heuristic policies, with an emphasis on the tradeoff between fairness and efficiency. Another work which has conducted by Kraimeche and Schwartz [38] suggest a class of restricted access policies in which incoming calls are divided into groups with same bit rates, and the total available bandwidth is partitioned among groups. These class of policies is shown to provide middle ground between the policies of complete sharing and complete partitioning.

In another work by Gopal and Stern [39], network is considered to have two classes of traffic and each one has different bandwidth requirements. The control problem is formulated as constrained maximization of expected throughput, and it is solved by the dynamic programming method of policy iteration. Ross and Tsang

[40] have shown that value iteration method would produce more efficient solution to this problem than dynamic programming.

Ferrari and Verma [41] propose joint scheduling and admission control algorithm for a system with two classes of traffic in packet switched network. They use a version of Earliset Due Date scheduling, along with a priority mechanism.

The admission control in wireless network is multi-faceted problem. Due to factors such as mobility and frequency reuse, the boundaries of the admission control problem have been expanded and are concerned with other functions as well. Whereas call admission control in wired networks assumes fixed start and end-point for the duration of the call, mobile wireless users move from cell to cell and change the network access point. Each time the user traverse a cell bounder, the handoff generates a bandwidth request to the system. Though this request is generated in mid-call, it is essentially a part of the admission control problem. As a result, the QoS requirements for hand off admission are stricter than for new user admission. Mobility analysis and channel holding time in a given cell proufoundly impact performance and have been studied extensively [42, 43, 44]

Liu and Lang [45] presented an admission control scheme for broadband multi-services wireless networks to limit the number of ongoing connections so that the QoS for each connection can be maintained at the desired level.

Hou and Fang [46] proposed a new mobility-based call admission schemes and new call bounding schemes for wireless mobile networks providing service to multiple classes of mobile users. The concept of influence curve is introduced to characterize the influence an active call exerts on adjacent cells, according to which the channel reservation can be adjusted dynamically and mobility-based CAC schemes can be designed. The authors also proposed a new call bounding scheme to put a direct

limitation on the number of new calls admitted to a cell, in order to overcome potential congestion.

In another work, which has done by Yoon and Lee [47], a distributed dynamic reservation scheme is proposed that can control the call admission of mobile multimedia traffic in a dynamic and distributed manner to support mobility in wireless multimedia communications. The proposed scheme approximate the channel occupancy distribution based on the observation of arrival rate, means and variance of total calls and handoff calls, then employs an elaborate two-regional approximation, in which a simple distribution model can be applied in each region. By using this method the number of reservation channel can be estimated very quickly, which reduces the computation dramatically.

A the multi-class admission control in wireless LANs where user are not mobile and thus no handoffs occurs is discussed in [48, 49]. In [50], admission control in ad-hoc networks is studied, This differs from other mobile networks since there are no base station.

In [51], there is a discussion of issues in network management and control in wireless multi-media networks. The multi-class admission control problem is discussed more recently in [52, 53, 54] among others.

Two different mulit-media reservation-based admission control algorithms were developed in [55] and [56]. They ensure the provision of adequate QoS to users in the system as defined by the probability of being dropped during service. The reservations are done in a manner, however, which does not take into account the differing QoS requirements of the individual traffic classes.

2.4 Summary

This chapter presented a detailed discussion of the IEEE 802.16 standard. The discussion focused mainly on the MAC layer aspects that were significant in this thesis. A review of related research in the area of traffic scheduling and admission control is also provided.

The next chapter discusses the admission control procedure in greater detail and presents a priority algorithm that attempts to minimize the blocking probability of real-time connections (application). Chapter 4 discusses the uplink bandwidth allocation and scheduling strategies proposed for providing QoS in IEEE 802.16 networks.

Chapter 3

Admission control in IEEE 802.16

Our approach towards supporting a QoS requirements of different classes in IEEE 802.16 BWA is in two directions. One is in the form of an admission control mechanism and the other in the form of packet scheduling. Scheduling mediates the low level contention for service between packet of different classes, while admission control determines the acceptance or blocking of a new connection. These two levels of control are related for example, if too many traffic is allowed to enter the network by an admission control policy, then no scheduler would be able to provide the requested QoS of all classes. A functioning admission control is thus a prerequisite for any guarantee of packet level QoS. Both are important in ensuring that a requested QoS can be provided for a certain number of connections, a number that is dictated by the amount of available bandwidth in the system.

This chapter describes admission control policy which describes the techniques used to determine the availability of uplink resources for providing the QoS requirements specified by the requesting connection. The bandwidth allocation techniques will be discussed in the next chapter.

3.1 objective

There is no question about the necessity of using admission control in any kind of networks that aims to provide QoS support for its connections. As the IEEE 802.16 is aiming to do so it is essential to have some sort of admission control being embedded in its QoS architecture, however, there is no admission control procedure is defined in the current standard. In this work we have suggested an efficient admission control module, targeting network utilization while giving a priority to the UGS and rtPS connection requests, since they carry a data of real-time applications.

3.2 Overview of Admission Control

We expressed the need for the admission control in order to control the usage and allocation of bandwidth resources for various traffic classes requiring certain QoS gurantee. Admission control is a key component in determinig whether a new request for a connection can be granted or not according to the current traffic load of the system. This assumes great significance when the BS needs to maintain a certain promised level of service for all the connections being admitted(served). If the admission control admits too few connections, it results in wastage of system resources. On the other hand, if too many connections are allowed to contend for resources, then the performance of the already admitted connections degrades rapidly in the presence of new connections. Therefore, judicious decision making mechanisms for allocating bandwidth to different classes of service are needed.

In IEEE 802.16, before an SS can initiate a new connection or changing or deleting an already admitted connection, it must first make a request to the BS.

As mentioned earlier in chapter 2, four types of MAC layer services exist in IEEE 802.16. These service flows can be created, changed, or deleted through the issue of dynamic service addition (DSA), dynamic service creation (DSC) and dynamic service deletion (DSD) messages. Each of these actions can be initiated by the SS or the BS and are carried out through a two or three-way-handshake.

3.3 Admission Control Policy

The task of admission controller is to accept or reject the arriving requests for a connection in order to maximize the channel utilization, by accepting as many connections as possible, while keeping the QoS level of all connections at the level specified in their traffic profile. In other words it ensures that already admitted connections QoS will not be affected by the decision made. Although it may seems to be very intuitive and simple procedure, it has great influence on QoS of the admitted connections.

This issue have been studied and researched extensively in the context of wired and wireless networking. Although the focus of this research was not admission control, whenever it comes to IEEE 802.16, putting a well rounded introduction seems to be indispensable, thus we have concisely introduced some of the fundamental works that need to be done in this area as there is no defined procedure in IEEE 802.16.

Admission control algorithms can be categorized into three classes of complete sharing, complete partitioning and hybrid policies which is a combination methods of the other two.

- The complete sharing (CS) allows all users equal access to the bandwidth available at all times. This strategy results in maximum utilization of the available bandwidth, specially in high traffic networks, which is what network providers aiming at. However, at the same time, it does not differentiate between connections of different priority that is a perverse outcome when connections of one class needs significantly less bandwidth than others. At this situations it might be desirable to reject calls of this type to increase the probability of future acceptance of a larger call. In other word it is not fair strategy to the wider bandwidth users as all request would be dealt with the same priority.

- The Complete partitioning (CP) policy, on the other hand, divides up the available bandwidth into non-overlapping pools of bandwidth in accordance with the type of user's connection. Therefore, number of existing users in each class would be prohibited to a maximum number M which admission decision will be made upon. This policy allows for more control of the relative blocking probabilitiy at the expense of overall usage of the network.

- The hybrid policies basically provide a compromise between the different policies by subdividing the available bandwidth into sections. Part of the bandwidth is completely shared and the other part is completely partitioned. Depending on the policy adopted, the partitioned division would be dedicated to some or all classes.This allows more live up to the QoS requirements of the different user types while maintaining higher network utilization.

In the above mentioned admission control policies, CP and CS, have no complexities to be elaborated and decision making would be a matter of checking a

single condition, though, the hybrid strategy is a more open ended problem. As an example, in the following we suggest an algorithm that could be considered as a hybrid method for admission control.

3.3.1 Admission control procedure

As in IEEE 802.16 UGS and rtPS connections have higher priority due to carrying data of real time applications, it would be desirable, from both user and network provider point of view, to prioritize the connection requests of this classes over an nrtPS and BE connections request in admission control policy. In order to meet the demands of these two classes we have taken advantage of hybrid models by setting aside a good share of total bandwidth for them, along the paradigm of CP. This amount can be determined arbitrary according to the policy of service provider as depending to the network in hand this amount has a significant impact on blocking probability of other classes.

The remainder of bandwidth will be shared between all classes for the CS operation, though the bandwidth that an nrtPS class request contends to get from the shared pool would not be allocated to it in a non-preemptive form. Thats where we try to accommodate connections with their required bandwidth, while seemingly there is not enough bandwidth to be granted. This is deployed by stealing some bandwidth from connections like nrtPS.

As nrtPS connections specify their bandwidth requirement by specifying a maximum sustained traffic rate ($r^{\max}$) and a minimum reserved traffic rate ($r^{\min}$), knowing that, admission controller can take a bandwidth up to sum of $r^{\max}_{\text{nrtps}} - r^{\min}_{\text{nrtps}}$ from nrtPS connections, the lower priority class.

The terminologies used in this chapter are as follow:

- B the total bandwidth
- R_{ugs} Reserved bandwidth for UGS
- R_{rtps} Reserved bandwidth for rtPS
- r_{ugs} traffic rate of ugs
- $B_{\text{allocated}}$ the share of the bandwidth which is allocated to the connections so far

In the following we describe the strategy we adopt to utilize the extra gained bandwidth:

- Whenever an ADC receives a UGS connection request the decision to accept or reject is made based on the following condition:

$$B_{\text{allocated}} \; + \; r_{\text{ugs}} \; \leq \; B - R_{\text{rtps}} \tag{3.1}$$

  It should be noticed that the required bandwidth would not be taken from UGS reserved bandwidth unless the shared bandwidth pool is drained. If the above condition was not hold the ADC would go to the next step which is downgrading the already admitted nrtPS connections. If there was no room for downgrading nrtPS connections then the request will be rejected.

  Downgrading is the act of cutting down the bandwidth of nrtPS connections in order to accumulate enough bandwidth for servicing the new request in the system. The downgrading procedure (shown in Algorithm 1) is carried

Algorithm 1 -Downgrade(r_{required})-

1: $B_{\text{aquired}} = (r^{\max}_{\text{nrtps}} - r^{\min}_{\text{nrtps}} - \text{totaldgl}) * n_{\text{nrtps}}$;
2: **if** $(r_{\text{required}} \leq B_{\text{aquired}})$ **then**
3: $dgl = \frac{r_{\text{required}}}{n_{\text{nrtps}}}$ *//Current required downgrading portion*
4: **for** $i \leftarrow 1, n_{\text{nrtps}}$ **do**
5: reduce(i,dgl)
6: **end for**
7: totaldgl+ = dgl; *//total downgraded portion so far*
8: *return* 1
9: **else**
10: *return* 0
11: **end if**

out in two stages. At the first stage the maximum amount of bandwidth (B_{aquired}) that can be obtained by cutting down all the nrtPS connections (n_{nrtps}) bandwidth reservation to $r^{\min}_{\text{nrtps}}$ is calculated. In the second stage the required bandwidth of the new connection (r_{required}) is checked against the B_{aquired}, if it is less than or equal to the B_{aquired} the new connection can be be admitted and all the admitted nrtPS connections bandwidth reservation would be reduced by the required downgraded level (dgl). Otherwise, the new connection would be rejected, since it can not be admitted into the system even by downgrading the nrtPS connections.

Downgrading procedure potentially can end up to constantly servicing nrtPS connections at the minimum level of possible if we do not upgrade the downgraded connections when there is enough bandwidth available to provide them with. This issue springs from our policy in downgrading all the nrtPS connections at the same level. In order to avoid this problem to occur we introduce an upgrading procedure which is a part of the algorithm for accepting/rejecting nrtPS request. The upgrading procedure (as shown in Algorithm 2) consists

of checking the currently available bandwidth ($B_{\text{available}}$) of the system and increase the reserved bandwidth of the existing nrtps connections if they have been downgraded before. The aim of upgrading procedure is treating nrtPS connections with more fairness.

Algorithm 2 -Upgrade()-

1: $B_{\text{available}} = (B - B_{\text{allocated}})$;
2: **if** $(totaldgl * n_{\text{nrtps}} \leq B_{\text{available}})$ **then**
3: **for** $i \leftarrow 1, n_{\text{nrtps}}$ **do**
4: increment(i, totaldgl)
5: **end for**
6: **end if**
7: $totaldgl = 0$

- Whenever the ADC receives new nrtps connection request, the connection admits to the system at the rate ($r_{\text{nrtps}}^{\max} - totaldgl$) if:

$$B_{\text{allocated}} + (r_{\text{nrtps}}^{max} - totaldgl) \leq B - R_{\text{ugs}} - R_{\text{rtps}} \tag{3.2}$$

It means that the ADC will reserve the same amount of bandwidth for the new nrtPS connection as the already admitted nrtps connections.

- Whenever an ADC receives an rtPS connection request the decision to accept or reject it is made by checking whether

$$B_{\text{allocated}} + r_{\text{rtps}} \leq B - R_{\text{ugs}} \tag{3.3}$$

It should be noticed that the required bandwidth would not be taken from rtPS reserved bandwidth unless the shared bandwidth pool is drained. If the

condition 3.3 was not hold the ADC would go to the next step which is downgrading the already admitted nrtPS connections in order to accomodate the new rtps connection. If there was no room for downgrading nrtps connections then the new request will be rejected.

The ADC always accept the BE connection request but no bandwidth is reserved for this class of connections, as it is not necessary to provide any guaranteed service for this class.

The explained policy leaves room for servicing the higher priority classes at situations in which system is highly loaded. The idea of taking bandwidth from nrtPS class at high traffic times is a sort of implicit or hidden second level bandwidth reservation for classes like UGS and rtPS and also maximizing utilization.

3.3.2 Result and Analysis

An event driven simulator was developed to evaluate the performance of different admission control policies discussed in this chapter. We consider a single BS within a wireless network with a fixed bandwidth of B. We assume there a three different types of traffic, each of which generates new connection establishment's request according to mutually independent Poisson processes with rates $\lambda_{\text{ugs}}, \lambda_{\text{rtps}}, \lambda_{\text{nrtps}}$ and have channel holding times of $\frac{1}{\mu_{\text{ugs}}}$ $\frac{1}{\mu_{\text{rtps}}}$ $\frac{1}{\mu_{\text{nrtps}}}$ which follow the ngative exponential distribution. For the hybrid methods we have assumed $R_{\text{ugs}} = 20\%$ and $R_{\text{rtps}} = 15\%$ of the total bandwidth (B). The only difference between hybrid I and II is that in hybrid II a downgrading procedure is employed but not in hybrid I. The assumed simulation parameters are shown in Table 3.1

Table 3.1: System Parameters

Parameter	Value
r_{ugs}	16 Kbps
r_{rtps}	32 Kbps
r_{nrtps}^{max}	48 Kbps
r_{nrtps}^{min}	16 Kbps
B	5 Mbps

The blocking probability of UGS class under different admission control policy is shown in figure 3.1. We note that the probability of blocking a new UGS connection is significantly lower under a hybrid schemes than the CS and CP policy. As it can be seen hybrid II (the reservation scheme with downgrading) policy performs better than the hybrid I. This is because under a hybrid II policy the system can accommodate more connections since more bandwidth is available. This is because of using a downgrading procedure which reduces the bandwidth reservation of nrtPS connections. CP policy performs slightly better than the CS when the system load is less than 0.5. As the offered load increases further, the CS policy outperforms the CP . This is because when the load increases the CS policy begins to heavily favor the narrow banded UGS connections over the others.

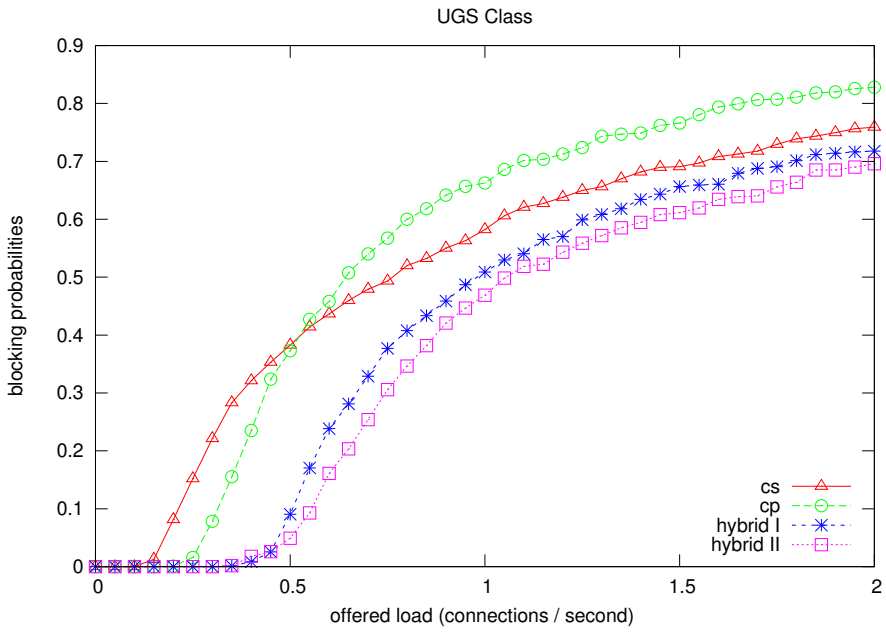


Figure 3.1: Performance of different admission control policies for UGS

As it can be seen in Figure 3.2 the hybrid policies are providing a much lower blocking probability than CS and CP policies for rtPS connections as well. The hybrid II policy performs better than the hybrid I policy except under a heavy load which both methods performs similarly and that is because the amount of bandwidth that can be freed ,by applying a downgrading procedure on a nrtPS connections, is negligible in comparison to the offered load. The provided blocking probability by CS is smaller than that of the CP, but when the load increases to more than 1.5 times they both converge to roughly the same value.

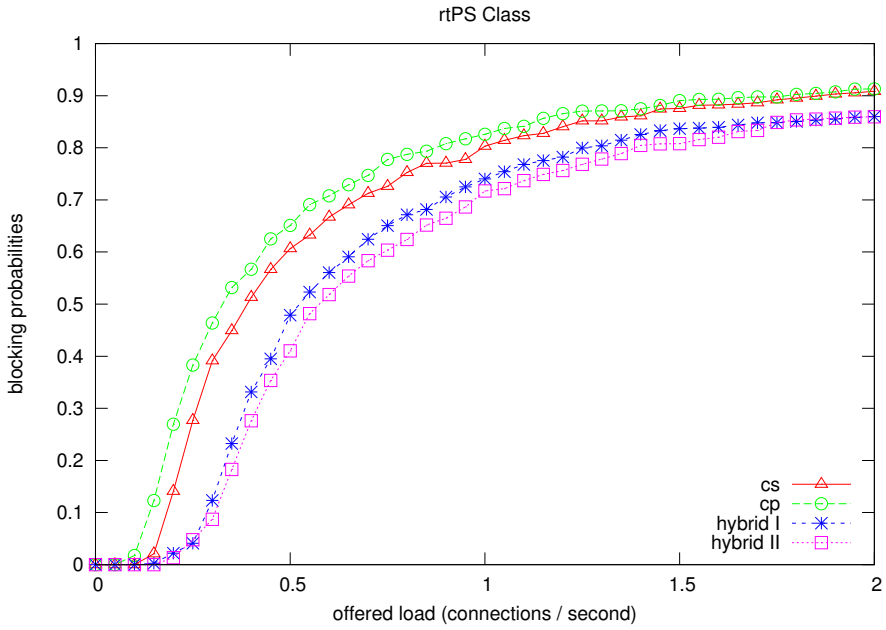


Figure 3.2: Performance of different admission control policies for rtPS

The hybrid policies provided a higher blocking rate for nrtPS connections unlike the UGS and rtPS as shown in figure 3.3. This is due to the fact that under a hybrid I and II policies we have reserved a certain amount of bandwidth for UGS and rtPS classes, thus less bandwidth remains available in the shared pool that nrtPS connections can contend for. The hybrid II performs better than the hybrid I this is because when we downgrade the nrtPS connections then more connections can be accomodated into system resulting in smaller blocking rate when the system load is less than o.5. Under a CP the nrtPS blocking rate is lower than that of provided by CS. This is because the nrtPS connection bandwidth request is larger than the UGS and rtPS requests as a result the CS favors them over the nrtPS.

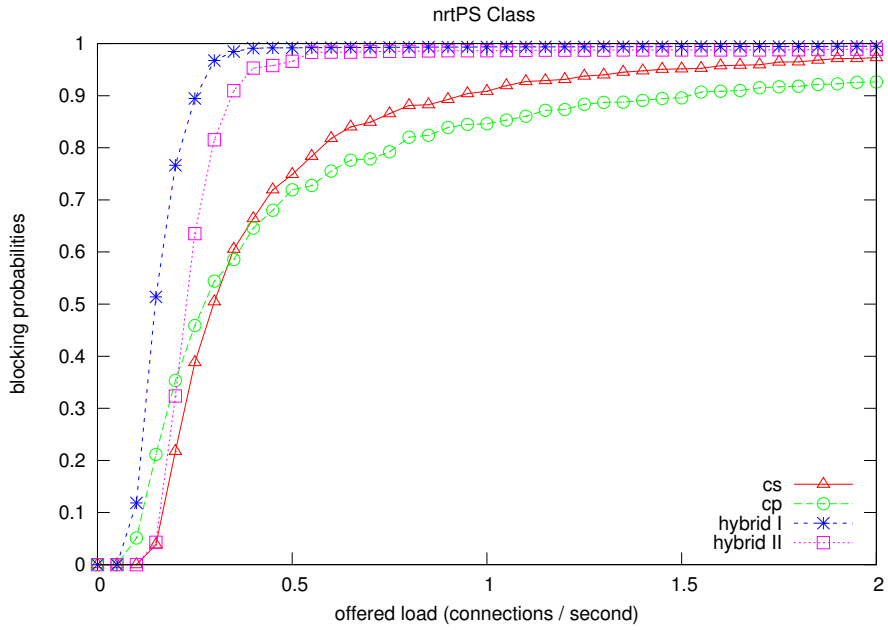


Figure 3.3: Performance of different admission control policies for nrtPS

3.4 Summary

As shown in this chapter the proposed hybrid policies are more efficient in providing priority admission control than CS and CP, which do not differentiate between connections of different classes. The proposed hybrid policy provides a significantly lower blocking probability for real-time connections than the non-real time connections. This is what the network provider is aiming to do. The traffic scheduling methods are discussed in the next chapter.

Chapter 4

Traffic Scheduling

This chapter describes the bandwidth allocation techniques developed in the course of this research. As stated earlier, our research focuses on developing and comparing a set of efficient, practically feasible and simple to implement algorithms, to provide QoS support in IEEE 802.16 BWA. This paves the ground for being able to set customized network policy by selecting well-matched scheduler algorithms to optimise performance while maintaining the required level of QoS.

In this chapter, we investigate a new set of bandwidth allocation techniques for the BS, which are based on soft reservation of bandwidth for each class of service and sharing of the bandwidth among connections with the same class of service according to their immediate bandwidth requirements.

As an uplink packet scheduling is performed at the SS in IEEE 802.16 BWA, it is necessary to study the impact of using different SS packet schedulers on the performance of the system in terms of the QoS parameters such as bandwidth, delay, and drop.

4.1 Scheduling in IEEE 802.16 Broadband Wireless Access

In the IEEE 802.16 standard, the BS and SS must reserve resources to meet their QoS requirements. The principal resource to be reserved is bandwidth. The BS controls and also allocates the required bandwidth to downlink and uplink connections according to their traffic profile. The IEEE 802.16 supports UGS, rtPS, nrtPS and BE class of service as discussed in chapter 1.2. Each connection in the uplink direction is mapped into one of four existing types of uplink scheduling service. Some of these scheduling services prioritize the uplink access to the medium for a given service, for example by using unsolicited bandwidth grants. Others use polling mechanisms for real time and non-real time services or even pure random access for non real time BE services[5].

Scheduling in IEEE 802.16 is divided into two related scheduling tasks (see Figure 4.1). The first task, performed at the BS is the scheduling of the airlink resources (uplink subfrarme) among the SSs according to the information provided with the bandwidth requests, which is received from the SSs. The second scheduler task is the scheduling of individual packets at SSs and BS. The SS scheduler is responsible for the selection of the appropriate packets from all its queues, and sends them through the transmission opportunities allocated to the SS within each subframe.

The scheduling algorithms in the BS and SS may be very different since the SS may use bandwidth in a way that is unforeseen by the BS. The BS sees bandwidth requests on a per connection basis and based on this, grants bandwidth to the SS while attempting to maintain QoS and fairness. However, according to the IEEE 802.16 standard, the SS may be granted an aggregated amount of bandwidth

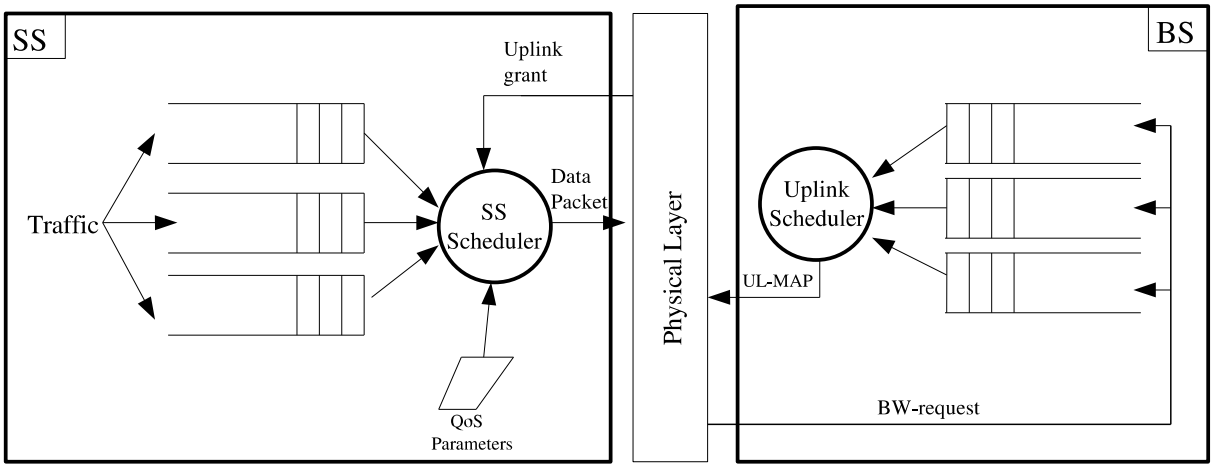


Figure 4.1: QoS function within the BS and SSs

(Grant Per Subscriber Station (GPSS)) rather than bandwidth on a connection basis (Grant Per Connection (GPC)). Thus, when an SS receives an uplink grant, it cannot deduce from the grant which of its connections it was intended for by the BS. Consequently, the SS scheduler is responsible for sharing the bandwidth among the different connections whilst maintaining QoS and fairness. Since different connections must be provided with individual QoS levels, the importance of building a QoS aware scheduler is undeniable.

As explained in chapter 2, many packet-level traffic scheduling techniques have been proposed for providing QoS in wired and wireless networks. However, the objective of most of these algorithms is to provide strict QoS guarantees (i.e., hard QoS) to traffic streams by requiring them to strictly conform to predetermined traffic profiles. Non-conforming traffic is not guaranteed, and so could suffer substantial performance degradation, even though over a period of time the traffic stream may have under-utilized its allocated bandwidth. The reality of unpredictable workload

is one of the biggest challenges for all the schedulers. In wireless multimedia networks it is difficult to predetermine profiles of real-time traffics; this can lead to degradation of the expected QoS level of the non-conforming traffic, which may be detrimental to the overall QoS experienced by the end user. Soft QoS provisioning can be defined as graceful acceptance of traffic profile violation when excess bandwidth is available, provided that the session does not exceed its overall reserved bandwidth in the long term. In wireless networks, soft QoS provisioning can be ideal for scalable multimedia applications which can tolerate occasional degradation in network performance due to channel errors[57]. This motivates us to develop algorithms that have soft QoS provisioning properties. Soft QoS provisioning differs from best-effort QoS in that the later does not consider the QoS requirements of the applications at all [58].

4.2 Base Station Bandwidth Allocation Architecture

As shown in figure 4.2, the BS bandwidth allocation architecture consists of connection classifier, scheduling database module, bandwidth scheduler and MAP generation module. The connection begins to send out the bandwidth requests after being admitted into the system by the admission control. All bandwidth requests from the connections are classified by the connection classifier according to their connection identifier (CID) and its service type, they are then forwarded to the appropriate queues in scheduling database module. The scheduler allocates bandwidth to the connections according to the bandwidth requests retrieved from the queues.

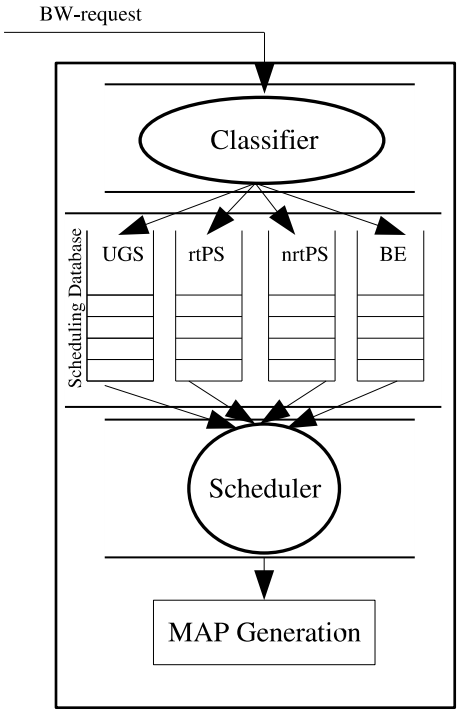


Figure 4.2: BS bandwidth allocation architecture

Based on the result of the bandwidth allocation the UL-MAP will be generated and broadcasted to the SSs.

The following information are kept in our scheduling database module for each active connection in the system: Connection Identifier (CID), traffic classe (UGS, rtPS, nrtPS, BE), queue length status and time of expiry (TOE) value. CID and traffic class are fixed values and are entered into the database when the connection is first admitted into the system. In our model we assume that every BW-request message transmitted by the SS to the BS, carries the reservation request information which consist of traffic queue-length status and also time of expiry (TOE) value of the first packet in the connection traffic queue. We take a TOE value of the first

packet of each queue into account, as a means of prioritizing a more urgent BW-requests. A TOE value of the first packet in the queue represents the amount of time that the scheduler has before allocating a bandwidth to this connection. If the scheduler does not allocate a bandwidth to this connection before the TOE deadline, at least the first packet of this connection and probably all packets belonging to the same traffic burst, will be expire. The BS updates the TOE and queue length information for each of the connection whenever it receives a new BW-request from each connections. In the current draft of the IEEE 802.16 standard, BW-request header quantifies the bandwidth request in the number of bytes that are required by the corresponding connection. This is a quantity that merely represents the queue length or changes in queue length. As the queue length is just an indirect measure of indicating the current traffic demand or load, the sole measure cannot be used to deal with the QoS requirements, especially for delay-sensitive or loss-sensitive applications. It needs to be used in combination with some other parameters. A new QoS management message was proposed in [59], that does not change any part of the existing IEEE 802.16 standard. A proposed message can carry various dynamic traffic parameters, thus providing an opportunity for the SS to transmit more information about the current status of each of its traffic queues. Adopting this new management message is necessary to improve the QoS support of the current standard.

4.2.1 Base Station Scheduler Framework

As it is stated before the goal of our BS scheduler is to provide a soft QoS support for the connections which are admitted into the system, while wisely maximizing the channel utilization.

The following notations have been used throughout this section:

B_{uplink}: total bandwidth (bps) allocated for uplink transmission.

$B_{\text{i}}^{\text{rtPS}}, B_{\text{i}}^{\text{nrtPS}}, B_{\text{i}}^{\text{be}}$: the amount of bandwidth which each connection i has requested from the BS.

$P(t)$: the remainder of bandwidth at time t.

$EB(t)$: the excess bandwidth at time t.

R_{total}: aggregated reserved bandwidth for all service classes.

$R_{\text{ugs}}, R_{\text{rtps}}, R_{\text{nrtps}}, R_{\text{be}}$: the amount of reserved bandwidth in bit for each of the UGS, rtPS, nrtPS and BE service classes, respectively.

$N_{\text{ugs}}, N_{\text{rtps}}, N_{\text{nrtps}}, N_{\text{be}}$: the number of admitted UGS, rtPS, nrtPS, BE connection in the system at time t, respectively.

$q_{\text{i}}(t)$: queue size (bit) of connection i at time t.

$l(t)$: the sum of leftover (bit) bandwidth at time t.

$S_{\text{rtps}}, S_{\text{nrtps}}, S_{\text{be}}$: The share of bandwidth that can be allocated to each connection in the current frame

The scheduling algorithms which are proposed here are based on an idea of reserving a minimum amount of bandwidth for each class of service during each frame-time, and then distributing the reserved bandwidth for each class of service among the corresponding connections of that class. The excess bandwidth is distributed among all the connections according to their instantaneous bandwidth requirements. The distribution of the excess bandwidth among different classes follows priority logic, from highest to lowest: UGS, rtPS, nrtPS and BE. In another word the procedure boils down into the following two stages:

- Reserved bandwidth of each priority class is distributed among the admitted connections of that class, according to the desired scheduling policy.
- The excessed Bandwidth would be allocated to those connections that have not been granted a part or total of their requested bandwidth.

  The excess bandwidth is defined as follows:

$$R_{\text{total}} = \sum_{c=1}^{N} R_c \quad , \quad c = \{ugs, rtps, nrtps, be\} \quad , \quad N = 4 \tag{4.1}$$

$$P(t) = B_{\text{uplink}} - R_{\text{total}} \tag{4.2}$$

$$EB(t) = P(t) + l(t) \tag{4.3}$$

The excess $EB(t)$ is comprised of a left over of the reserved bandwidth of each traffic class, if any, and also the remainder of bandwidth ($P(t)$). The main purpose of the second round of scheduling is to maximize the system utilization and to provide soft QoS, by allocating the left over bandwidth of one class of service to another class of service which is requiring more bandwidth than its reserved bandwidth, due

to the arrival of a traffic in burst. In figure 4.3, an instance of the two stages of the scheduler framework is illustrated.

As it is hinted before, the above explained stages can be defined as the skeleton of scheduling algorithms at an abstract level. Different strategies can be adopted, specially for the second round, to skin over this frame. This is done in the following sections.

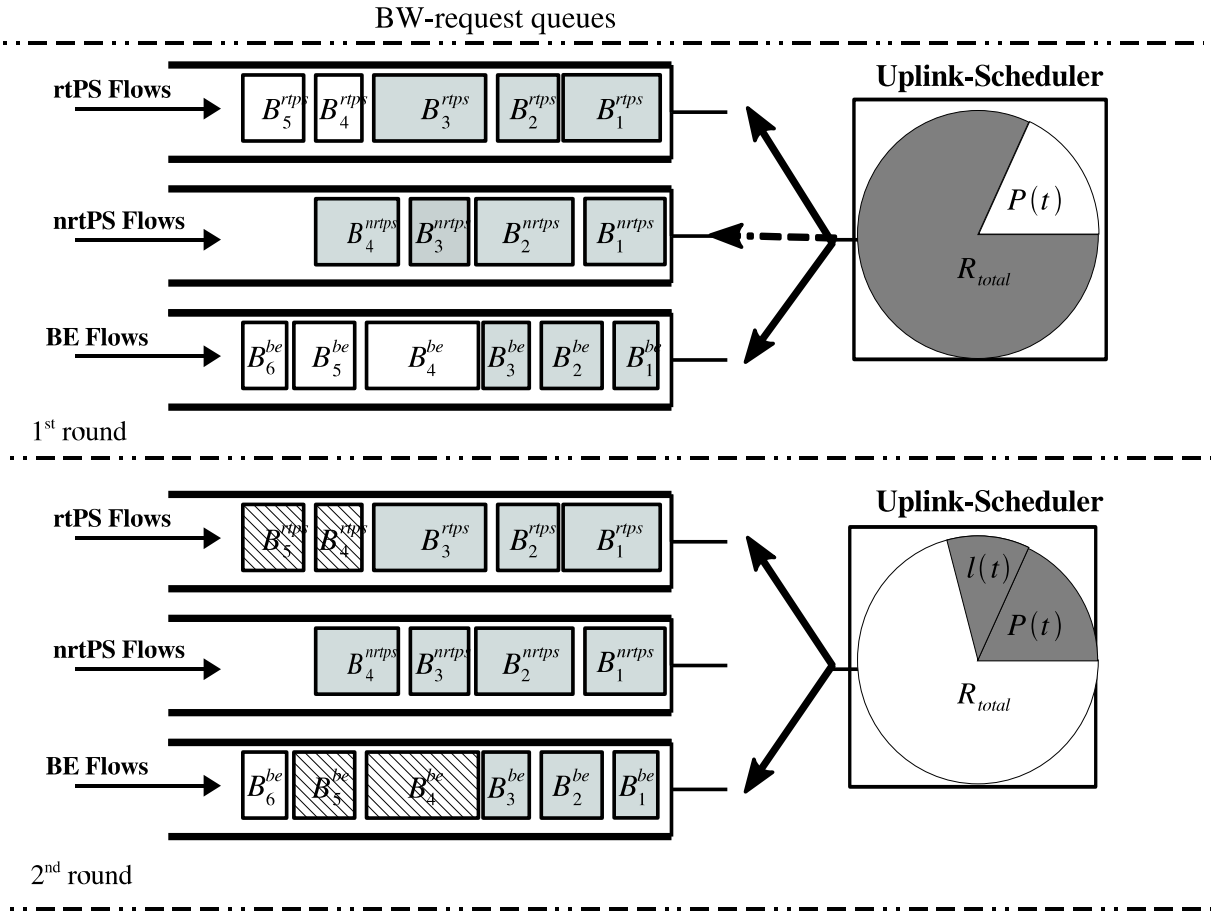


Figure 4.3: BS scheduler framework

While our proposed airlink scheduler follows the priority discipline to meet delay and loss requirements of different classes, it also tries to maintain its fairness by

reserving a minimum amount of bandwidth for each class of service during each frame time. The size of this amount of reserved bandwidth for each class of service reflects a degree of scheduler fairness towards that class. It can also be said that size of the reserved bandwidth is a trade-off between providing a better QoS support for higher priority classes and being fair to lower priority classes. The calculation of the amount of reserved bandwidth to be allocated to each class reflects the policy of the scheduler. Reserving a zero amount of bandwidth for each class of service would turn the scheduler into a pure strict priority scheduler and reserving a large amount of bandwidth for lower priority classes would have a negative impact on the QoS support of higher priority classes. The values of these reserved bandwidths can be adjusted dynamically by policy of the admission controller, which is described in chapter 3. The admission control policy determines the maximum and minimum values of these parameters according to the current traffic load of each class of service.

The allocation of bandwidth to the UGS connections in our scheduling schemes follows the IEEE 802.16 policy mechanism, which requires the allocation of a fixed size data grants at periodic intervals.

The scheduling disciplines which are considered here for allocation of bandwidth within the connections of each class of service are Earliest Deadline First With Bandwidth Reservation (EDF-BR) and Multi Class Allocation (MCA).

Earliest Deadline First with Bandwidth-Reservation (EDF-BR)

The EDF-BR algorithm is carried out in four stages. In the first stage, the scheduler visits each non-empty priority (rtPS, nrtPS, BE) queue and sorts the BW-requests in the ascending order of their TOE.

At the second stage (shown in Algorithm 3), the requested data size of the first BW-request packet in the rtPS queue is checked, if it is less than or equal to R_{rtps}, the reserved bandwidth is reduced by the size of the BW-request, and all the connection's requested bandwidth is allocated to it. Otherwise, the BW-request is reduced by the size of R_{rtps} and the connection's bandwidth requirement is partially allocated to it. This process is repeated until either the R_{rtps} is no more greater than zero or the request queue is empty. In the case that the request queue is empty but R_{rtps} is greater than zero, R_{rtps} will be added to $l(t)$. When any of the above conditions occurs, the scheduler moves on to service the nrtPS request queue. After the nrtPS connections being serviced the scheduler moves on to service the BE request queue.

At the third stage, the scheduler allocates the excess bandwidth $EB(t)$ to the connections which have not been serviced in the first round of bandwidth allocation, starting from the highest priority queue. The scheduler moves to service the next priority queue if $EB(t)$ is greater than zero and if the queue is not empty. The procedure shown in Algorithm 3 continues for the third stage of scheduling, but this time instead of distributing the reserved bandwidth, the $EB(t)$ would be distributed within the connections that still require bandwidth.

At the fourth stage, when the bandwidth size to be allocated to each connection is determined, the allocations in the frame are made so that each SS gets contiguous allocation time. This is because in IEEE 802.16 the bandwidth allocation is per SS not per connections.

Among the connections with the same TOE and within the same service class, bandwidth can be granted in one of the following strategies:

Algorithm 3 -EDF-BR()-

1: **if** $(N_{\text{rtps}} > 0)$ **then** ▷ Allocate Bandwidth to *rtps* connections
2: **while** $(i < N_{\text{rtps}}$ && $R_{\text{rtps}} > 0)$ **do**
3: **if** $(B_i^{\text{rtps}} < R_{\text{rtps}})$ **then**
4: AllocateBandwidth(B_i^{rtps}) ▷ allocate bandwidth to connection i
5: $R_{\text{rtps}} \leftarrow R_{\text{rtps}} - B_i^{\text{rtps}}$
6: **else**
7: AllocateBandwidth(R_{rtps})
8: $R_{\text{rtps}} \leftarrow 0$
9: **end if**
10: $i++$
11: **end while**
12: **end if**
13: $l(t) = l(t) + R_{\text{rtps}}$
14: **if** $(N_{\text{nrtps}} > 0)$ **then** ▷ Allocate Bandwidth to *nrtps* connections
15: **while** $(i < N_{\text{nrtps}}$ && $R_{\text{nrtps}} > 0)$ **do**
16: **if** $(B_i^{\text{nrtps}} < R_{\text{nrtps}})$ **then**
17: AllocateBandwidth(B_i^{nrtps})
18: $R_{\text{nrtps}} \leftarrow R_{\text{nrtps}} - B_i^{\text{nrtps}}$
19: **else**
20: AllocateBandwidth(R_{nrtps})
21: $R_{\text{nrtps}} \leftarrow 0$
22: **end if**
23: $i++$
24: **end while**
25: **end if**
26: $l(t) = l(t) + R_{\text{nrtps}}$
27: **if** $(N_{\text{be}} > 0)$ **then** ▷ Allocate Bandwidth to *be* connections
28: **while** $(i < N_{\text{be}}$ && $R_{\text{be}} > 0)$ **do**
29: **if** $(B_i^{\text{be}} < R_{\text{be}})$ **then**
30: AllocateBandwidth(B_i^{be})
31: $R_{\text{be}} \leftarrow R_{\text{be}} - B_i^{\text{be}}$
32: **else**
33: AllocateBandwidth(R_{be})
34: $R_{\text{be}} \leftarrow 0$
35: **end if**
36: $i++$
37: **end while**
38: **end if**
39: $l(t) = l(t) + R_{\text{be}}$
40: $EB(t) = l(t) + P(t)$

- Random: One of the connections is randomly selected to get the bandwidth, this can be an advantage in a sense that all connections have equal chance to utilize the available bandwidth. However, sometimes servicing the connection with higher queue length is more urgent.

- Queue length based: The connection with higher queue size is selected, this is because the connection with higher queue size is experiencing higher level of congestion, and it's buffer management module is more likely to drop its packets due to overflow or to avoid overflow. However, this method potentially can increase the average queue length of other connections. Therefore, we sometimes get better result in random selection.

The simulation result of both strategies are shown in chapter 5. In the course of this thesis we call the first strategy as EDF-RA and the second strategy as EDF-QL.

Multi-Class Allocation (MCA)

The MCA algorithm is carried out in five stages. At the first stage, the scheduler sorts the BW-requests in each of the non-empty priority queues in the ascending order of their TOE. The connections with the same TOE values are arranged in the descending order of their queue length value.

At the second stage, the scheduler first allocates all the bandwidth except the amount reserved for nrtPS and BE ($B_{\text{total}} - R_{\text{nrtps}} - R_{\text{be}}$), uniformly among all the rtPS BW-requests in the queue. If any of the connection BW-requests's size is less than the S_{rtps}, the remaining bandwidth will be added to to those rtPS connection's BW-requests which still require more bandwidth. Within those BW-requests the

priority is given to one with the highest queue length among the ones with the same TOE value.

At the third stage, the R_{nrtps} and the left over bandwidth (if any, excluding the reserved bandwidth for BE) are then allocated to the nrtPS connections by the scheduler, in a similar fashion to the second stage.

At the fourth stage, the R_{be} and the left over bandwidth if any are then allocated to the BE connections. The scheduler follows the same style as nrtPS.

At the fifth stage, when the bandwidth size to be allocated to each connection is determined, the allocations in the frame are made so that each SS gets contiguous allocation time. This is because in IEEE 802.16 the bandwidth allocation is per SS not per connections. The pseudo code is given in Algorithm 3.

In MCA scheduling scheme all the BW-requests in a traffic class receive a minimum share of bandwidth during each frame time. The pseudo code of this procedure is shown in Algorithm 4(PART I) and 5(PART II).

Algorithm 4 MCA(PART I)

Ensure: $N_{\text{rtps}} \neq 0, N_{\text{nrtps}} \neq 0, N_{\text{be}} \neq 0$

if $(N_{\text{rtps}} > 0)$ **then**
  $S_{\text{rtps}} = [(B_{\text{total}} - R_{\text{nrtps}} - R_{\text{be}}) \ / \ N_{\text{rtps}}]$
  for $i \leftarrow 1, N_{\text{rtps}}$ **do**
    if $(B_i^{\text{rtps}} < S_{\text{rtps}})$ **then**
      AllocateBandwidth(B_i^{rtps})
      $l(t) = l(t) + (S_{\text{rtps}} - B_i^{\text{rtps}})$
      $B_i^{\text{rtps}} = 0$
    else
      AllocateBandwidth(S_{rtps})
      $B_i^{\text{rtps}} = B_i^{\text{rtps}} - S_{\text{rtps}}$
    end if
  end for
  while $(l(t) > 0$ && $i < N_{\text{rtps}})$ **do**
    if $(B_i^{\text{rtps}} \neq 0)$ **then**
      if $(B_i^{\text{rtps}} < l(t))$ **then**
        AllocateBandwidth(B_i^{rtps})
        $l(t) = l(t) - B_i^{\text{rtps}}$
      else
        AllocateBandwidth($l(t)$)
        $l(t) = 0$
      end if
    end if
    $i++$
  end while
end if
if $(N_{\text{nrtps}} > 0)$ **then**
  $S_{\text{nrtps}} = [(l(t) + R_{\text{nrtps}})/N_{\text{nrtps}}]$
  for $i \leftarrow 1, N_{\text{nrtps}}$ **do**
    if $(B_i^{\text{nrtps}} < S_{\text{nrtps}})$ **then**
      AllocateBandwidth(B_i^{nrtps})
      $l(t) = l(t) + (S_{\text{nrtps}} - B_i^{\text{nrtps}})$
      $B_i^{\text{nrtps}} = 0$
    else
      AllocateBandwidth(S_{nrtps})
      $B_i^{\text{nrtps}} = B_i^{\text{nrtps}} - S_{\text{nrtps}}$
    end if
  end for

▷ The algorithm will continue on the next page

Algorithm 5 MCA(PART II)

```
39:     while (l(t) > 0 && i < N_nrtps) do
40:         if (B_i^nrtps ≠ 0) then
41:             if (B_i^nrtps < l(t)) then
42:                 AllocateBandwidth(B_i^nrtps)
43:                 l(t) = l(t) − B_i^nrtps
44:             else
45:                 AllocateBandwidth(l(t))
46:                 l(t) = 0
47:             end if
48:         end if
49:         i++
50:     end while
51: end if
52: if (N_be > 0) then
53:     S_be = [(l(t) + R_be)/N_be]
54:     for i ← 1, N_be do
55:         if (B_i^be < S_be) then
56:             AllocateBandwidth(B_i^be)
57:             l(t) = l(t) + (S_be − B_i^be)
58:             B_i^be = 0
59:         else
60:             AllocateBandwidth(S_be)
61:             B_i^be = B_i^be − S_be
62:         end if
63:     end for
64:     while (l(t) > 0 && i < N_be) do
65:         if (B_i^be ≠ 0) then
66:             if (B_i^be < l(t)) then
67:                 AllocateBandwidth(B_i^be)
68:                 l(t) = l(t) − B_i^be
69:             else
70:                 AllocateBandwidth(l(t))
71:                 l(t) = 0
72:             end if
73:         end if
74:         i++
75:     end while
76: end if
```

To compare the performance of our proposed algorithms, we have also developed a strict class based queuing (SCBQ), which does not reserve any bandwidth for different priority classes. It begins by allocating the bandwidth to the UGS connections based on their bandwidth requirements, if any bandwidth would be left then it will be allocated to rtPS, and then to nrtPS and BE.

4.3 Subscriber Station Scheduler

In the IEEE 802.16 BWA standard, an uplink packet scheduler is located at the SS, which schedules packets from the connection queues into transmission opportunities allocated to the SS within each frame. The packet scheduling at the SS occurs just after the BS allocates the bandwidth to the SS.

At the SS each class of service (UGS, rtPS, nrtPS, BE) is associated with a delay bound of (D_{ugs}, D_{rtps}, D_{nrtps}, D_{be}), whenever an incoming packet of class i arrives at its queue at time t, it is stamped with a TOE of $t + D_i$, and packets are getting served in an increasing order of their TOE.

The first algorithm, which was developed employs priority queuing (PQ) algorithm. The algorithm allocates all the allocated bandwidth ($B_{allocated}$) to the first UGS connection. The remaining bandwidth would then pass on the second, and so on. After UGS connections were looped through, the scheduler moves on to rtPS connections in a similar fashion. After rtPS, then nrtPS and finally BE connections. The main problem with this scheduler was that it could starve lower priority connections of bandwidth.

A policing mechanism was introduced in order to prevent higher priority connections to use all SS share of bandwidth, which limits the maxim amount of bandwidth

allocated for each connection to its service contract rate. The problem with this method was that a connection's queue could grow indefinitely because its queue was never guaranteed to be allocated more than it could produce during each frame. An improvement on this policing method was made so that the maximum allocation was lifted to twice the contract rate. This method also had issues as it was a fixed value and never concentrated on individual connection's service contracts being upheld. Therefore, more intuitive policing methods were introduced.

In this method the scheduler first calculate the packet's expiry level of each connection's queue and accordingly a severity multiplier (β) would be determined for each of them. Two different approaches are proposed for calculating the packet's expiry level of each queue, namely IPQ1 and IPQ2. Following are the two methods that can be applied to calculate the severity multiplier.

- **Improved Priority Queuing 1 (IPQ1):** This method checks each station's connection queues for whether any of the PDUs would be due to expire soon. A PDU has critical delay tolerance if it would expire within the next n uplink subframes. The value of n should be experimentally determined according to the network characteristics. The simulation results showed that the value of 5 provides an optimal results for our model. Number of PDUs due to expire determines the severity of the expiration and accordingly a severity multiplier is determined and assigned against the connection's service contract rate (r_i). Table 4.1 shows the contract rate severity multiplier (β), this values have been selected experimentally.

- **Improved Priority Queuing 2 (IPQ2):** In this method, the scheduler first compute the weighted average waiting time (T_k) of the existing packets in

Table 4.1: Severity of traffic expiration with contract rate multiplier (β)

Number of PDUs due to expire	Level	β
0	None	0.0
< 10	Low	0.6
< 100	Medium	0.8
> 100	High	1.0

each connection's queue from (4.4). The Weighted average time (T_k)takes into account the overall criticality of packet's delay bounds together with the length of queues by giving higher priority/weight to packets that are pending to expire and in the same situation the packets in the longer queue are prioritized.

Since we have different priority classes in our system, and each of them are associated with different delay tolerance, in order to compare a T_k of different classes, we normalize them into scale of 0 to 1 by dividing the average delay to the maximum delay tolerance of that class. The smaller the normalized weighted average waiting time (δ_k), the lower the severity of the packet expiry at that queue.

Let d_{i_k} denotes the amount of time that packet i waited in the queue k, w_{i_k} represents the weight of packet i in queue k, and n_k denotes the total number of packets waiting to be served at the queue k. D_k indicates the maximum delay tolerance of queue k. Therefore, we can write the following expressions.

$$w_{i_k} = \frac{n_k - i}{n_k} \tag{4.4}$$

$$T_k = \frac{\sum_{i=1}^{n_k} d_{i_k} w_{i_k}}{\sum_{i=1}^{n_k} w_{i_k}} \tag{4.5}$$

$$\delta_k = \frac{T_k}{D_k} \tag{4.6}$$

The normalized waiting time (δ_k) determines the severity of the packet expiration and accordingly a β will be assigned against the connection's service contract rate (r_i). Table 4.2 shows, the details of the contract rate severity multiplier against the δ_k. We have selected the severity rates experimentally by running the simulation for number of times, though they can be determined more deliberately depending on the traffic model of the network.

Table 4.2: Severity of traffic expiration with contract rate multiplier (β)

	UGS & rtPS				
δ_k	< 0.20	< 0.30	< 0.40	< 0.50	> 0.50
β	0.6	0.65	0.70	0.85	1
	BE & nrtPS				
δ_k	< 0.15	< 0.25	< 0.50	< 0.75	> 0.75
β	0.2	0.4	0.60	0.80	1

Once the multipliers is determined, the scheduler allocates bandwidth to each connections queue, first starting with UGS. The connection i would be allocated bandwidth (according to expression 4.7) up to its service contract rate (r_i) multiplied by the determined multiplier (β_i) .

$$\text{Bonus}_i \;=\; r_i * \beta i \tag{4.7}$$

If the size of the packet at the head of the highest priority queue is less than or equal to the Bonus_i , the value of Bonus_i is reduced by the number of bits in the packet and the packet is transmitted to the output port. The process will be repeated until either the Bonus_i is no more than zero or the queue is empty. When any of these conditions occur, the scheduler moves to serve the next priority queue which has a Bonus_i greater than zero.

After the first round of bandwidth allocation (bonus allocation), if the remaining bandwidth would be greater than zero the scheduler would allocate the remaining bandwidth to the queues in the order of their priority, similar to the classic PQ algorithm. This three stages are illustrated in figure 4.4.

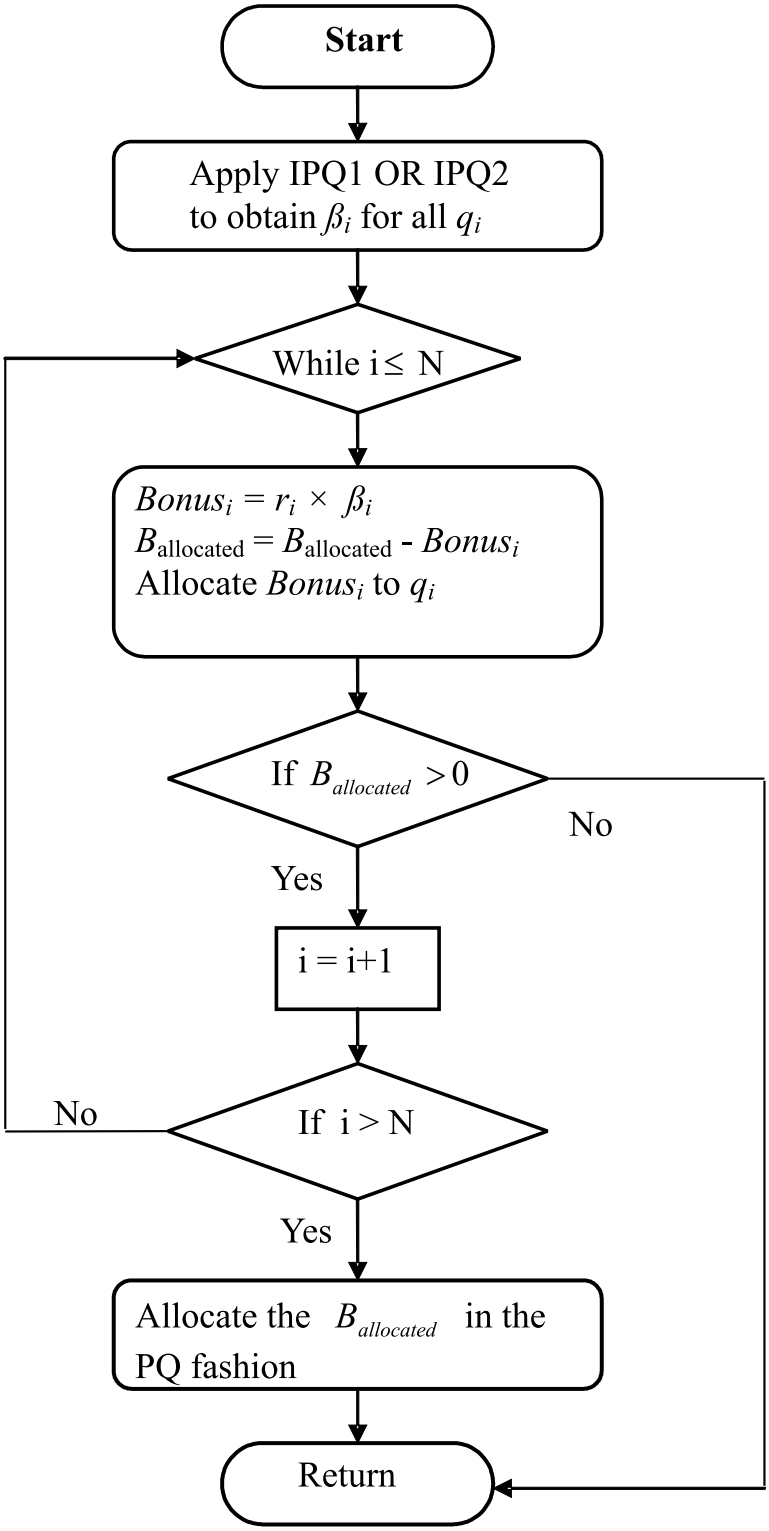


Figure 4.4: Flow chart of IPQ algorithm

4.4 Simulation Assumptions

Making several assumptions were necessary to limit the scope of the problem while developing the simulation model. The objective behind making these assumptions were to keep the simulation complexity manageable while still meeting the research goals. This section briefly discusses the assumptions made in the modeling process. They are listed here:

- Since the performance of the scheduling algorithms are of major concern, we disregard the contention resolution process among the reservation requests during the contention period and assume that the contending SSs can successfully transmit their reservation request during this period.
- Sufficient amount of buffer is assumed to be available at SSs for each of their queues so that no data packet is lost due to buffer overflow.
- No bandwidth is required for stations to relay any control message. This included bandwidth requests and initialization on the network.
- It is also assumed that fragmentation of PDUs does not increase the traffic overhead.
- Any dropping of PDUs would not cause retransmission from originator.
- The uplink and downlink subframe are assumed to be equal.
- Each connection has specific QoS parameters in terms of maximum delay and bandwidth requirement.
- The channel utilization is calculated without taking control overheads into account.

4.5 Discrete Simulation

Simulation looks to map certain features of an abstract system onto features of another, generally much simpler, system. Discrete event simulations, or time-step simulations, model processes as a sequence of events ordered by the time of each event occuring. Each event will have a beginning and an end, and events themselves may be "begin events" or "end events" for processes, for example, the simulation may have an event for the beginning of a moving object and another event for the end of that object's movement.

With the discrete points in time comes the ability to have state variables that measure the state of the process being simulated. As the simulation moves through a series of events, the process will be observed as a succession of state changes.

4.6 Simulation Model

A C-coded event-driven simulator is developed using discrete simulation method as described in section 4.5, in order to evaluate the performance of the proposed packet scheduling schemes in IEEE 802.16 WirelessMAN.

Our Simulator is comprised of two main components, the BS and the SS. The BS is made of scheduling database module and airlink scheduler module. The scheduling database module contains the detailed information about the status of all the active connections in the network, based on this information the airlink scheduler allocate a bandwidth to each SSs. Our SS module consists of packet scheduler module and fragmentation module. Packet scheduler module selects the appropriate packets from different queues and passes them to the fragmentation module.

Fragmentation module checks to see if the packet can be fit in the transmission opportunity that is available, if not then the packet is fragmented into smaller pieces, as many fragmented packets that can be sent in the current opportunity is sent and the remainders are kept to be transmitted in the next time that the transmission opportunity is allocated to this SS. The interaction between modules is described in Figure 4.5.

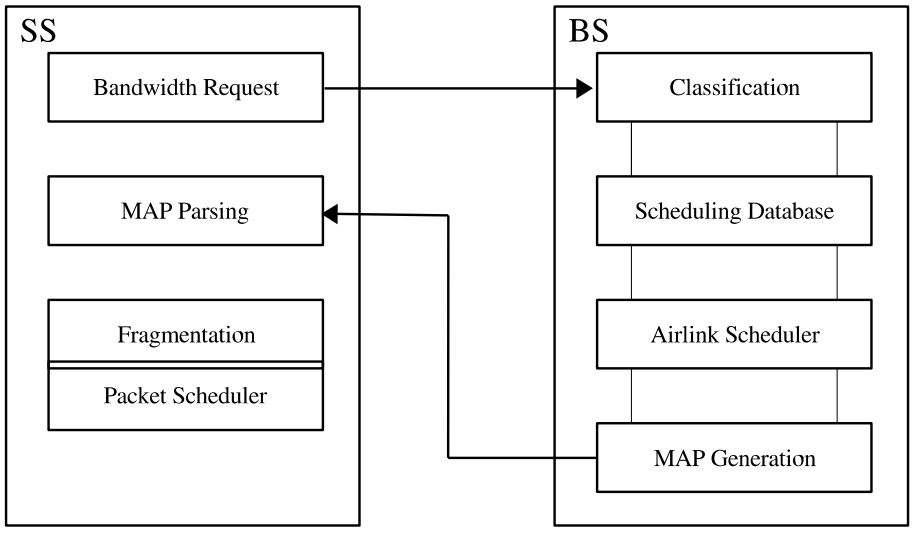


Figure 4.5: Module Interaction within 802.16

4.6.1 Events

In the simulator, the event set is comprised of the following events:

- Generate Traffic
- Airlik-Scheduling

- SS-Scheduling

Generate Traffic event has four different types: Generate-UGS-Traffic, Generate-nrtPS-Traffic, Generate-BE-Traffic, while executing the Generate-UGS-Traffic, Generate-nrtPS-Traffic and Generate-BE-Traffic events, packets are generated according to the specified arrival rate at different queues of SSs. When a Generate-UGS-Traffic event is happening, a new packet will be inserted into the UGS traffic queue of the specified SS. Also a new Generate-UGS-Traffic event is inserted into the event list for that SS. For the Generate-nrtPS-Traffic, Generate-BE-Traffic events the similar procedure will be followed. All of these events are happening for all SSs but independent from each other, as the traffic sources are independent from each other at each SS.

Airlink-Scheduling event occurs periodically with the period being equal to the length of a frame. When the Airlink-Scheduling event occurs, scheduling calculation for the next uplink subframe is performed and the output of the scheduling algorithm will be UL-MAP for the next frame, and also the corresponding data structures are updated. UL-MAP contains the order in which SSs must transmit their data and also specifies the beginning and end of each time period that each SS is allowed to transmit. Based on this information the SS-Scheduling event is inserted into the event list for all the SSs that are allowed to transmit during the next uplink subframe.

Table 4.3: System Parameters

Prameters	Value
Channel bandwidth	25 *MHZ*
Channe rate	80 *Mbps*
Frame duration	1 *ms*
Number of PS	2500
PS Size	4 *byte*
Propagation delay	500 *microsec*

4.6.2 System Parameters

A frame was fixed at 1 ms with 5000 physical slot (PS) in each. Each PS was able to hold 4 bytes of data. This gave the system a total bandwidth of 80Mbit/s. Transmission of data was simulated to have a fixed propagation delay of 500 *microseconds*.

A Maximum Transfer Unit (MTU) was introduced to stop large PDUs not being sent due to smaller allocation in bandwidth per frame. Therefore, fragmentation on PDUs has to occur which gave a smoothing of traffic flow. The MTU was set to 53 octets. The simulation input parameters are shown In table 4.3.

4.6.3 Traffic Classes

The rtPS traffic is modeled according to 200Mbps over 1 hour of MPEG-2 VBR stream of Jurassic Park. For the other classes packet arrival follows Poisson process distribution with rates $\lambda_{ugs}, \lambda_{nrtps}, \lambda_{be}$.

The UGS traffic has a fixed packet size, nrtPS and BE packet sizes are drawn from a negative exponential distribution. Each connection has specific QoS parameters in terms of maximum delay and bandwidth requirement. The delay requirements

Table 4.4: Traffic Sources Description

Service class	UGS	rtPS	nrtPS	BE
Maximum delay *(ms)*	30	40	1000	3000
Mean PDU size *(byte)*	52	NA	100	150
Average bandwidth *(KB)*	48.8	NA	39.06	36.62

and traffic descriptions of each class of services, used in the simulation are outlined in table 4.4.

PDUs were queued ready to send after creation but were dropped if they stayed in the queue longer than their timeout. BE traffic could be simulated as not having a timeout, however, (TCP) retransmissions would simply cause a recreation of the PDU. The simulation expired BE PDUs to stop memory hogging due to infinite queue lengths.

4.7 Summary

This chapter presented the simulation methodology used throughout this research. The simulation model and the BS and SS bandwidth allocation architectures were discussed in details. A detailed discussions of the BS and SS scheduling algorithms and simulation model were also presented. The analysis of the results and their performance evaluations are discussed in chapter 5.

Chapter 5

Performance Analysis

This chapter presents the simulation results and analysis. The aim of the different simulation experiments was to characterize the performance of the system under different pairs of BS and SS scheduler in terms of three performance requirements, namely the service level, the average packet delay and the average packet drop (loss) rate. The average packet delay of one class was measured by averaging the packet delay of all the connections within that class for varying incoming loads. The average packet drop rate for one class was measured by averaging the packet drop rate of all the connections within that class. The average packet drop rate of a connection was measured as the ratio of the of packets (in bit) dropped to the total number of packets successfully transmitted.

In the first part of this chapter, we demonstrate the obtained results from the simulation of each BS scheduling algorithm under a number of different SS scheduler over 100 seconds of simulation time. We then present a detailed comparison between all different combinations of BS and SS scheduling algorithms in the final section of this chapter.

5.1 Baseline Experiment

The baseline simulation model uses SCBQ algorithm at the BS and PQ at the SS. We consider two cases, which differ in terms of the number of SSs served by the BS (35, 40 SSs). The metrics observed were the service level, packet drop rate and delay.

5.1.1 Bandwidth Measurement

Figure 5.1 shows the service level (allocated bandwidth) of each class of traffic when 35 SSs are served by the BS. scheduler at the SS gives the highest priority to the UGS connections. It has been found, as rtPS service curve increases, it overrides nrtPS and BE service curves. When the bandwidth requirements of rtPS class reaches its maximum, during a 14-18 second, 24-26 second and 40-42 second period of time, the service level of BE suddenly drops to zero. This shows that the scheduler does not impose any restriction on the bandwidth usage of the higher priority classes. This is because the BS scheduler prioritize the rtPS BW-requests over nrtPS and BE BW-requests, and the same thing is happening in the SS scheduler. As it can be seen the nrtPS service level does not degrade as much as the BE does, this is because the BS scheduler gives nrtPS BW-requests higher priority over the BE BW-requests and at the end, any unallocated bandwidth is allocated to the BE connections.

As shown in Figure 5.2, with 40 SSs (the system is overloaded) the BE service level begins to deteriorate further. As the rtPS bandwidth requirement fluctuates drastically during a time period of 12-44 seconds, no BE traffic is able to leave the system. According to the graph, whenever the rtPS service level increases to its peak rate, the nrtPS service curve decreases to the minimum available bandwidth which

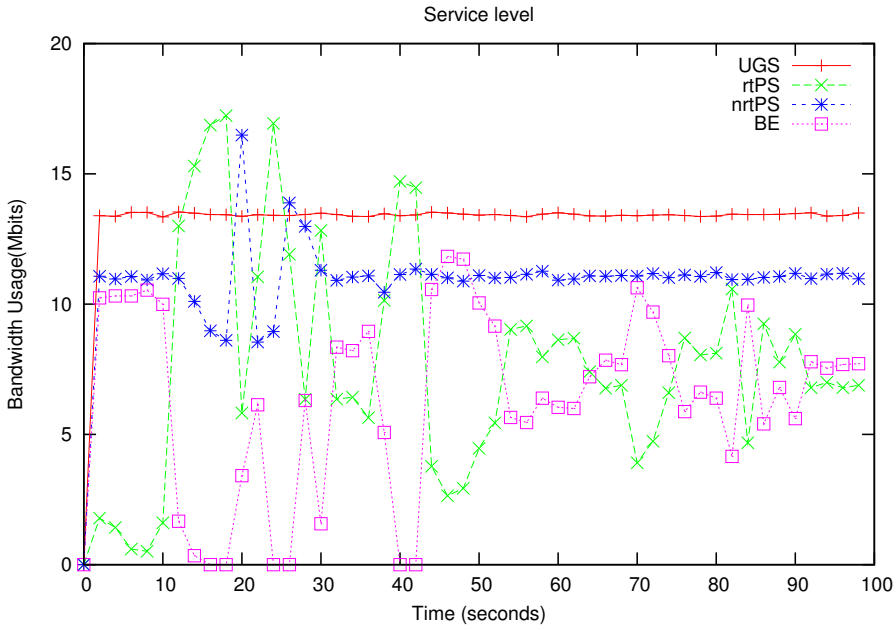


Figure 5.1: Service level in a baseline simulation with 35 stations

is less than its required bandwidth. This causes most of its traffics get queued. As soon as the rtPS bandwidth requirement reduces, the BS allocates more bandwidth to nrtPS connections. This is because the nrtPS connections have suffered during the previous time frames and are currently having large queue size and are requesting more bandwidth in order to transmit all their backlogged traffic.

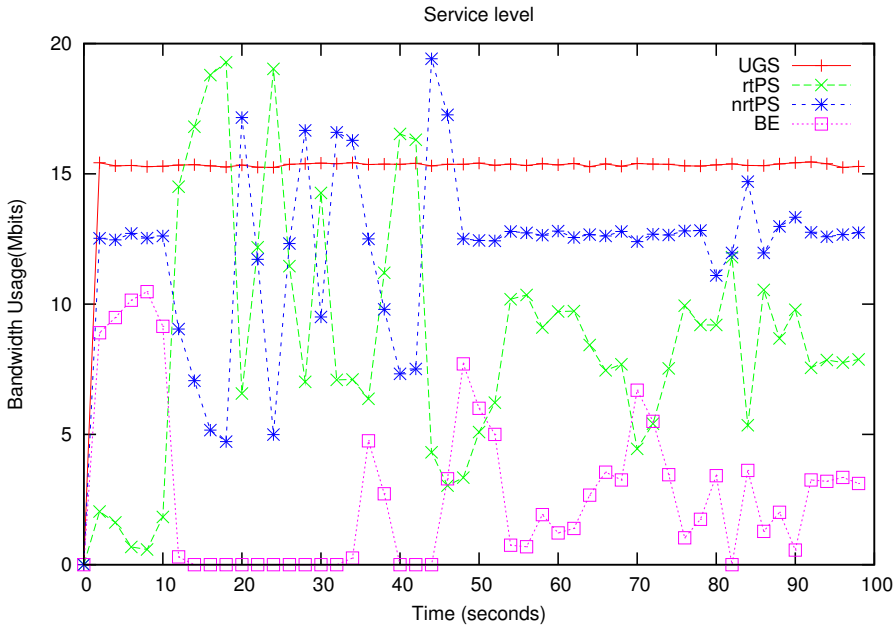


Figure 5.2: Service level in a baseline simulation with 40 stations

5.1.2 Average Packet Delay Measurement

The average delay for 35 and 40 subscriber stations are shown in Figure 5.3 and Figure 5.4 respectively. Delay is well behaved for UGS, rtPS and nrtPS. The shape of the rtPS and nrtPS delay curves are not very smooth, which can be justified by the bursty nature of the rtPS traffic sources and its negative effect on the nrtPS delay curve. The nrtPS delay curve fluctuates significantly between the period of 12-46 seconds, this is because according to the service level graph the nrtPS service level curve fluctuates considerably during this period too. The differences between Figure 5.3 and Figure 5.4 shows that when the traffic load increases the rtPS delay slightly rises, but nrtPS delay increases significantly. This is because the scheduler prioritize the rtPS traffic over the nrtPS traffic.

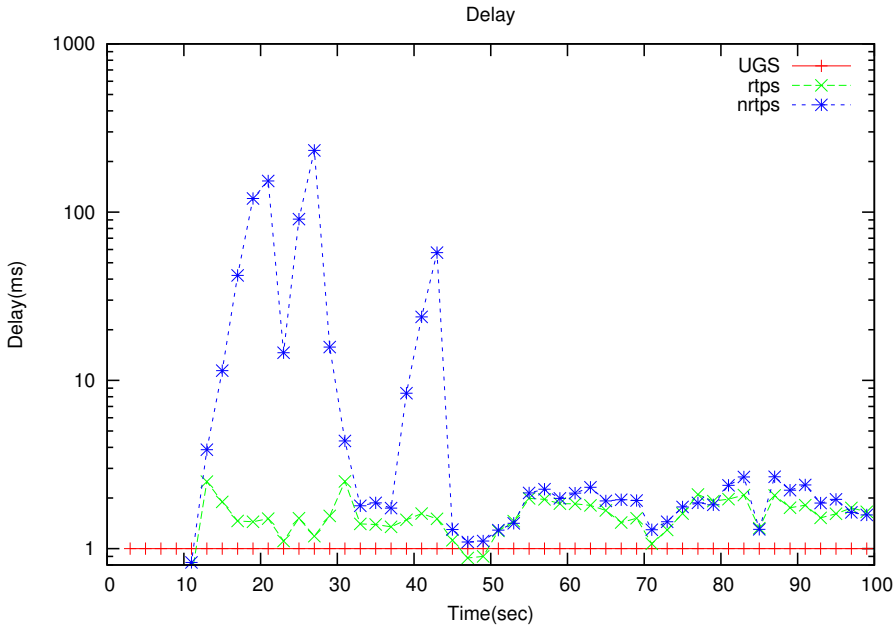


Figure 5.3: Delay in a baseline simulation with 35 stations

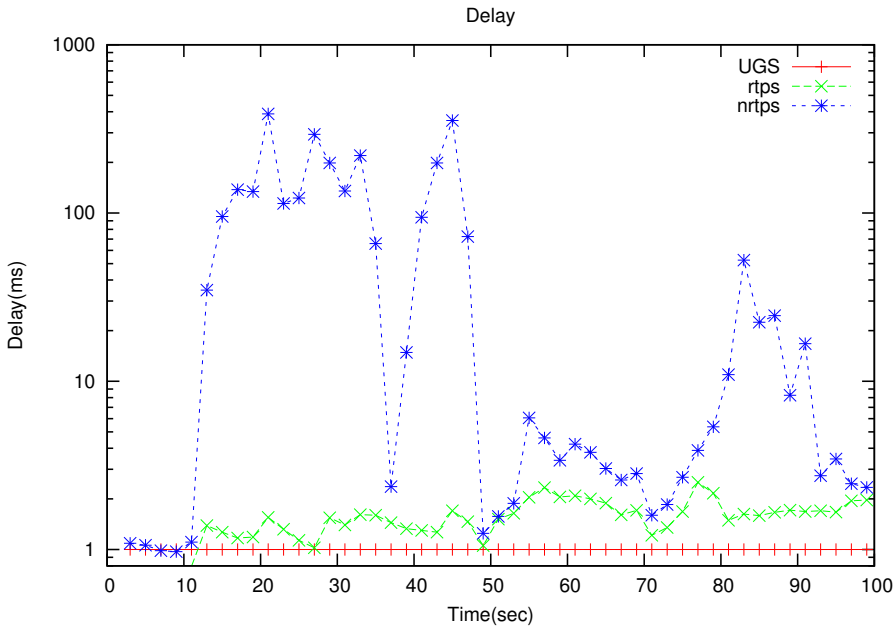


Figure 5.4: Delay in a baseline simulation with 40 stations

5.1.3 Average Packets Drop Measurement

The amount of BE traffic that was dropped due to being unserviceable are shown in figure 5.5 and 5.6. It can be clearly seen that the BE drop rate curve closely follows its service level curve, that high drop rate happens during the periods that not enough bandwidth is allocated to BE connections and as a result high volume of data get expired in the connection queues. When the number of SSs increases to 40 (shown in Figure 5.6) the drop rate of BE traffic also increases, because of not given enough bandwidth. The result show that the BS scheduler is not fair to lower priority classes an the lower priority classes starve.

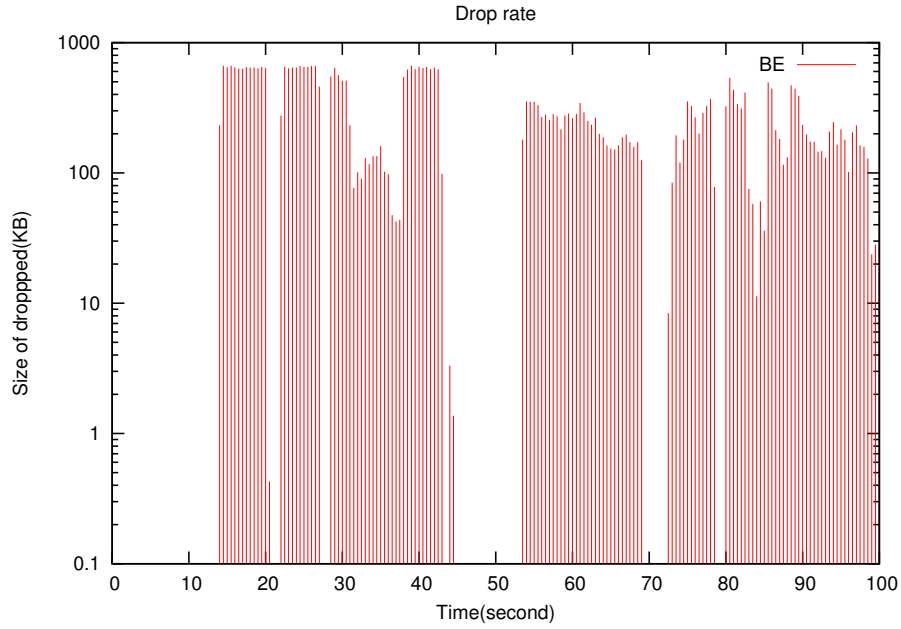


Figure 5.5: Packet loss in a baseline simulation with 35 stations

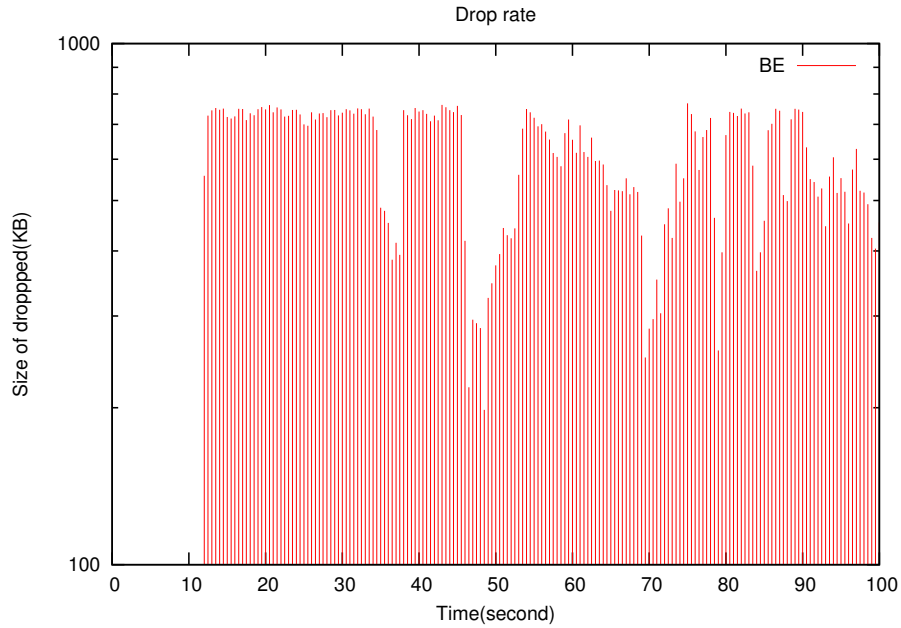


Figure 5.6: Packet loss in a baseline simulation with 40 station

5.2 EDF Experiment

The intent of this experiment was to study the performance of the EDF-RA scheduling discipline under two different SS scheduler. We have considered two scenarios as shown in Table 5.1 . EDF-RA was BS scheduler, but they differ in terms of the applied SS scheduler. In the first scenario PQ used as SS scheduler, and in the second one IPQ1 used as SS scheduler. We have evaluated the performance of theses two scenarios against the same input data set.

5.2.1 Bandwidth Measurement

Figure 5.7 shows the bandwidth allocation of each type of service class for the first scenario (EDF-RA-PQ) under a load of 35 SSs. During the first ten seconds of the

Table 5.1: Different Experiment Scenarios

	Number of SS	BS scheduler	SS scheduler
First scenario	35	EDF-RA	PQ
	40	EDF-RA	PQ
Second scenario	35	EDF-RA	IPQ1
	40	EDF-RA	IPQ1

simulation both nrtPS and BE classes receive a fairly steady amount of bandwidth, as the rtPS connections do not generate that much of traffic. According to the graph, the bandwidth requirement of rtPS class fluctuates dramatically after the tenth second with the peak of 17 Mbps at the 16,18 and 26 seconds. The increase in bandwidth requirement of rtPS class has a direct negative influence on the service level of nrtPS and BE classes because the PQ allocates considerable amount of the received bandwidth to the rtPS connections, however, it is most likely that only a portion of that bandwidth was actually intended by the BS to be allocated to rtPS connections. This has more serious effect on the service level of BE class than nrtPS class, as there are a period of time between 16-18 sec, 24-26 sec, 40-42 sec that no BE traffic is able to leave the system. The nrtPS service level fluctuates between a time period of 10-28 seconds since the allocated bandwidth to this class is less than its actual bandwidth requirement, after this period its service level becomes more stable as the rtPS bandwidth requirement falls.

According to the graph, after the 42th second the rtPS and BE service curves, go up and down as if they are inverse of each other. The reason behind this is because the BE service class always has enough traffic to transmit and hence it uses all the left-over bandwidth from the other classes. The above explanations

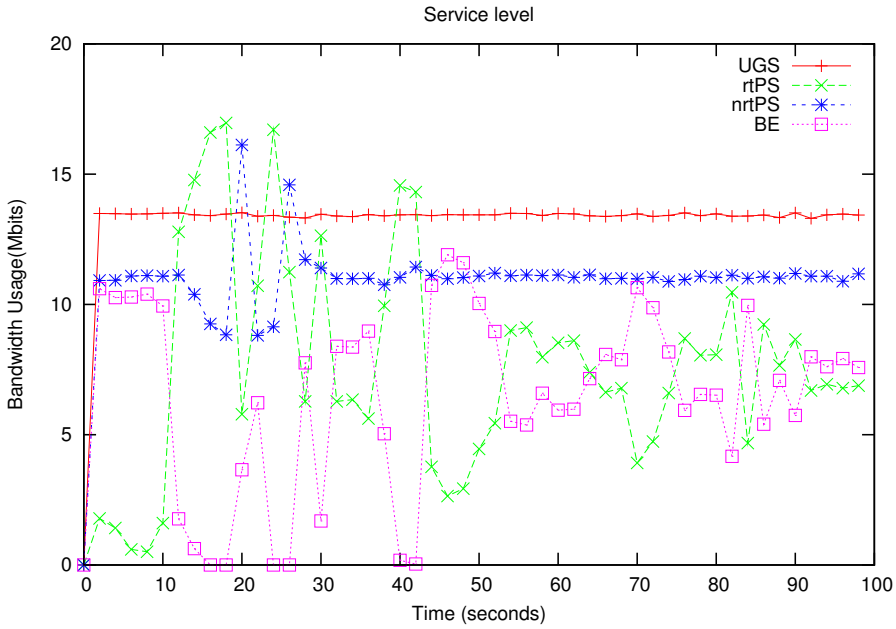


Figure 5.7: Service level in a first scenario with 35 stations

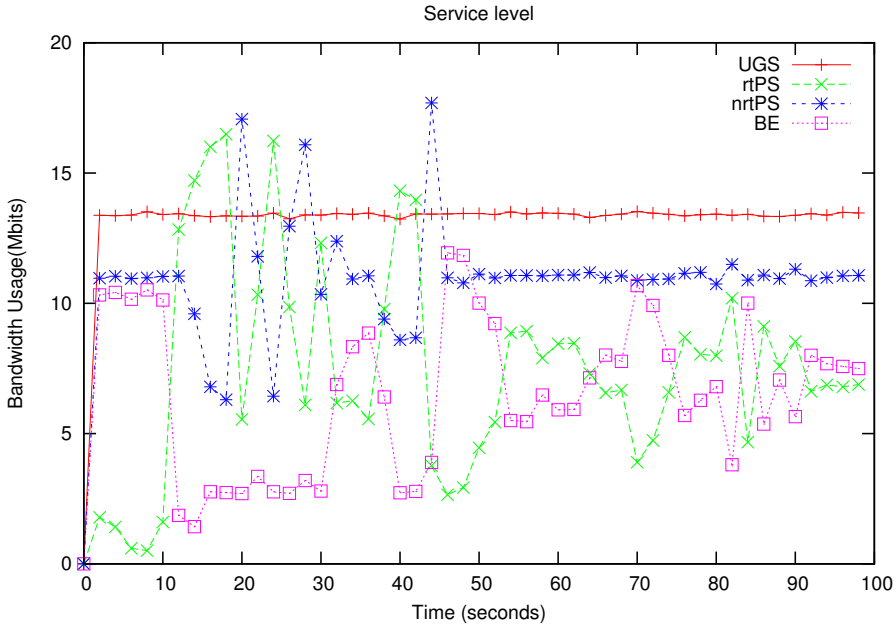


Figure 5.8: Service level in a second scenarion with 35 stations

are also valid for the Figure 5.9 in which the number of SSs increased to 40. The starvation of BE service class is more apparant in this figure as the BE does not receive any bandwidth for a longer period of time (between 12-34 sec and 40-44 sec) in comparison to when the traffic load was 35 SSs.

If we look at the Figure 5.8, which shows the service level of the second scenario under a load of 35 SS, it can be clearly seen that the service level of BE never drops to zero even during a 10-42 seconds that the bandwidth requirement of rtPS rises considerably. It can be justified by the fact that the IPQ1 scheduler takes the severity of the backlogged queues into account by selecting an appropriate *beta* (severity multiplier) for each connection and allocating some bandwidth according to this severity multiplier to each connection in order to prevent higher priority classes to starve the lower ones. Figure 5.10 shows the service level when the system is under heavy load. The graph shows that even in this case the BE service level never drops to zero and it consistently receives some minimum amount of bandwidth. It can also be seen that the peak service level of rtPS is bounded as the IPQ1 does not let the rtPS to consume all the BE and nrtPS bandwidth shares.

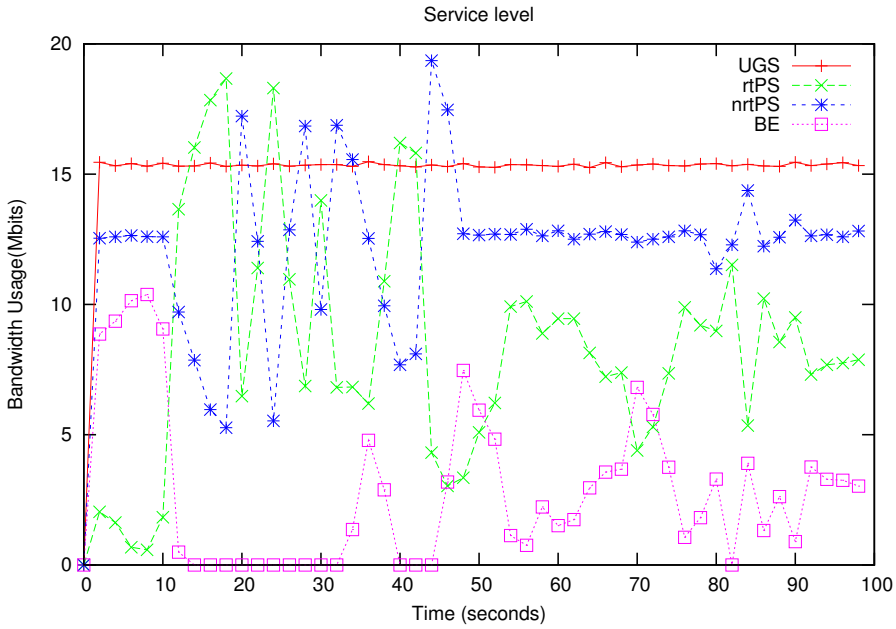


Figure 5.9: Service level in a first scenario with 40 stations

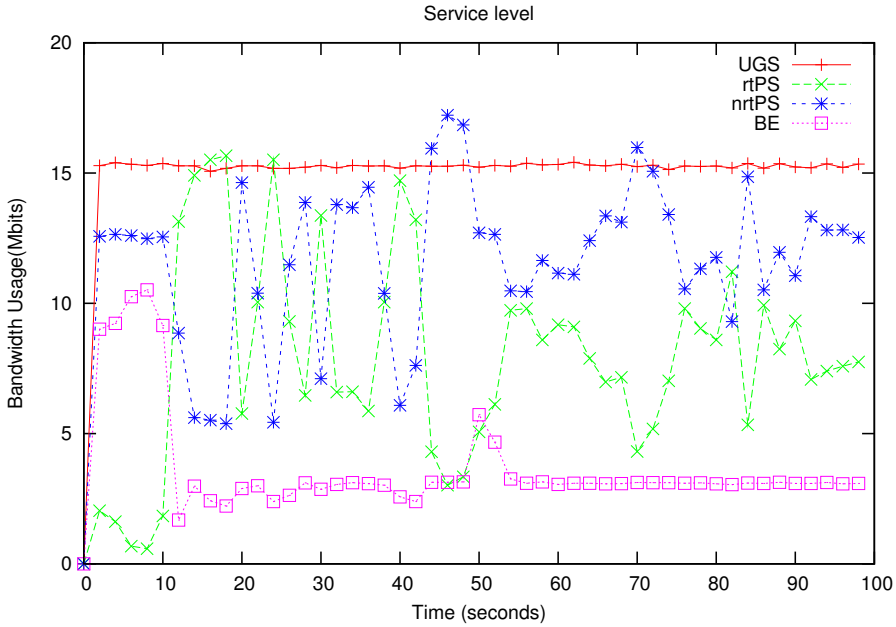


Figure 5.10: Service level in a second scenarion with 40 stations

5.2.2 Average Packet Delay Measurement

The improvement on the service level of lower priority classes under second scenario comes at the expense of increase in an average delay of UGS, rtPS and nrtPS classes. Figure 5.11, 5.13, 5.12 and 5.14 shows the average delay for the first and second scenarios under a load of 35 and 40 station. As it can be seen in figure 5.12 the average delay of the UGS and rtPS classes slightly increases in the second scenario in comparison to the first scenario (Shown in Figure 5.11). As it can be seen in Figure 5.11 and 5.12 the nrtPS delay curve fluctuates during the period of 10-46 second, which is due to variation in bandwidth requirements of rtPS connections because of using highly bursty rtPS traffic generator. The differences between the two scenarios is that the nrtPS peak delay is much higher under a IPQ1 than the PQ scheduler. This is because the nrtPS connections are receiving less share of bandwidth under a IPQ1 in comparison to PQ, since the scheduler also allocates a some bandwidth to the BE connections as well. It should be mention here that the provided delay for nrtPS class under a second scenario is still far below its maximum delay tolerance.

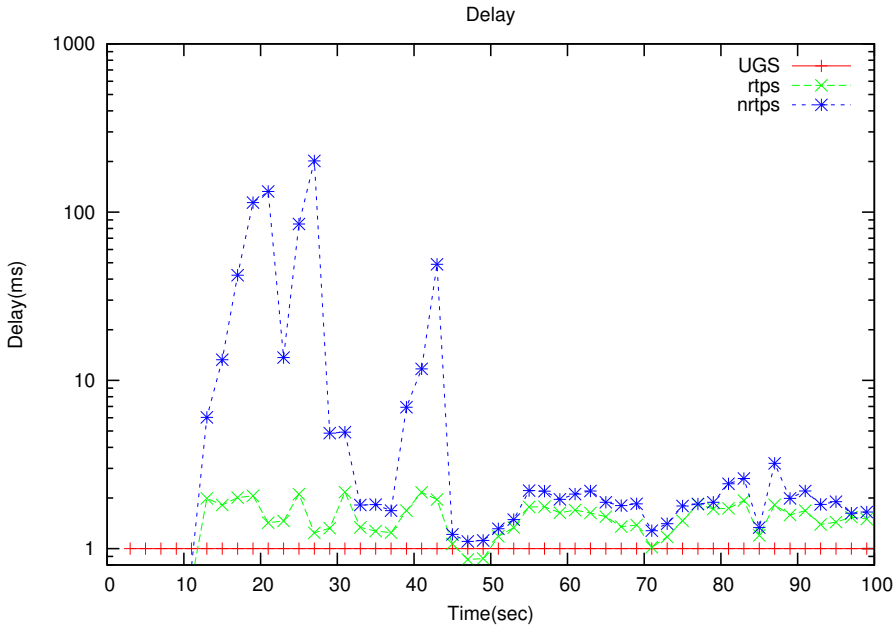


Figure 5.11: Delay in a first scenario with 35 stations

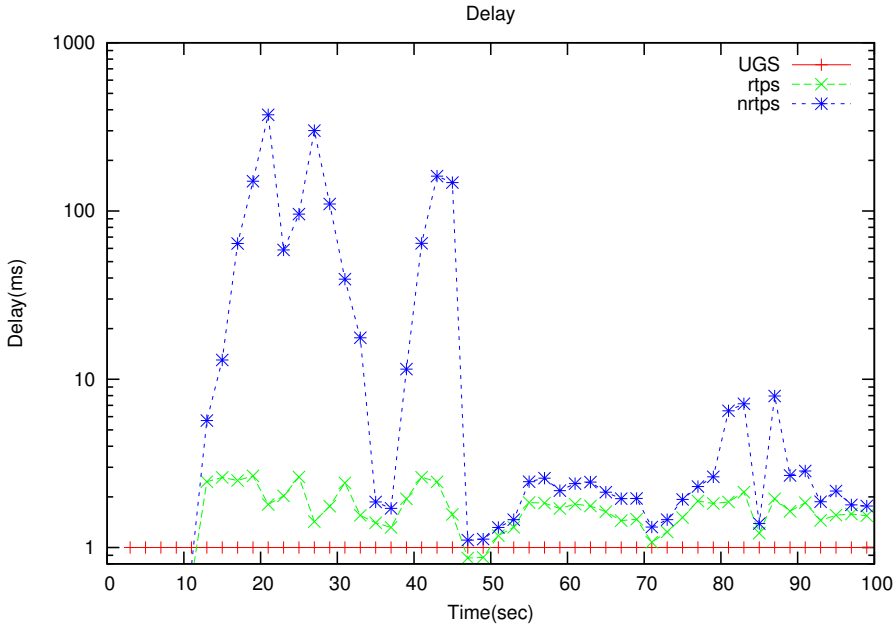


Figure 5.12: Delay in a second scenario with 35 stations

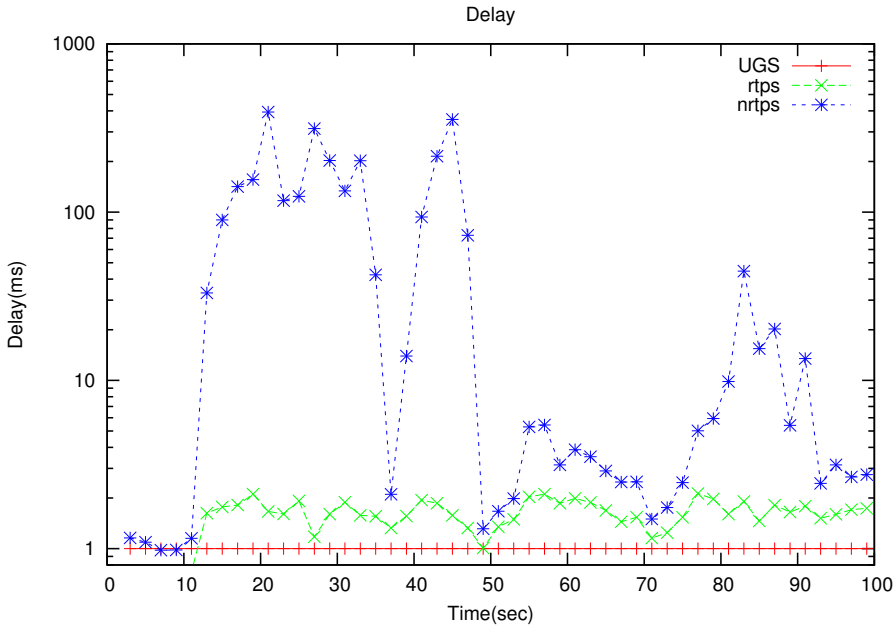


Figure 5.13: Delay in a first scenario with 40 stations

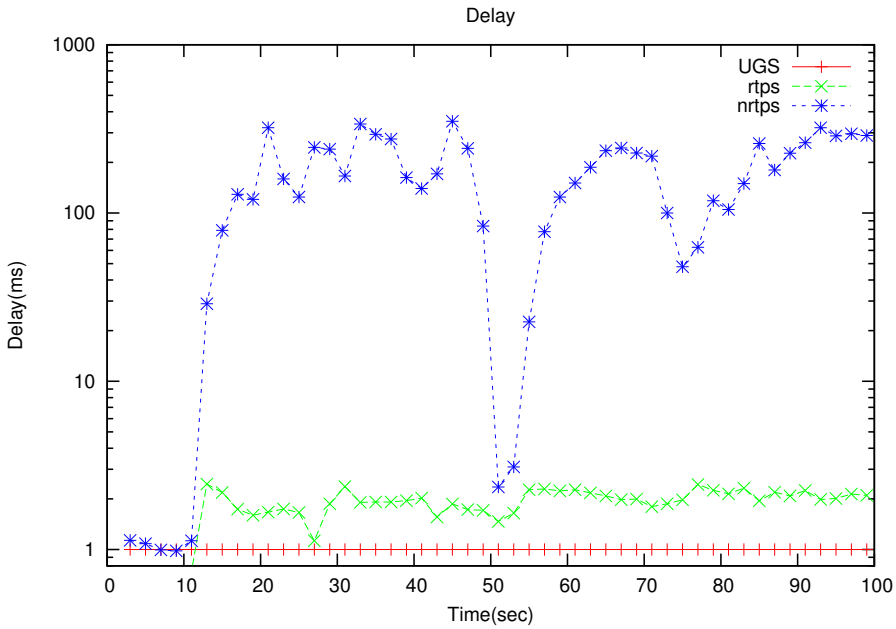


Figure 5.14: Delay in a second scenario with 40 stations

5.2.3 Average packet Drop Measurement

The figure 5.15 and 5.16, show the drop rate of the BE class under a traffic load of 35 and 40 SSs. The BE drop rate reduces considerably under a IPQ1 scheduling discipline in comparison with PQ for both 35 and 40 SSs. This is because the IPQ1 algorithm consistently allocates bandwidth to the BE connections even under a heavy traffic loads, as a result reduces the BE drop rate. The results show that the IPQ1 algorithm provides a better services for a lower priority classes in comparison to PQ service discipline which starve the BE connections.

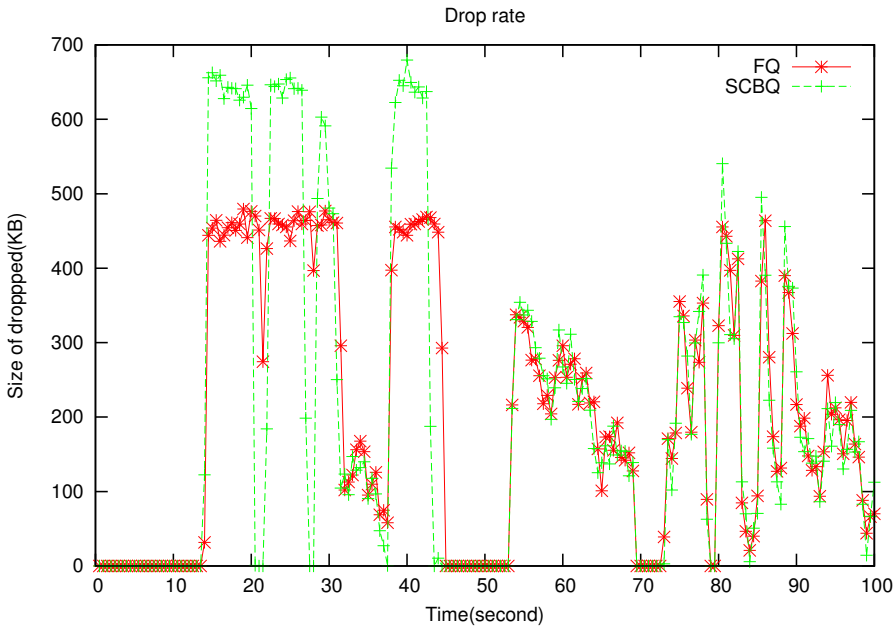


Figure 5.15: Drop rate under 35 stations for BE service class

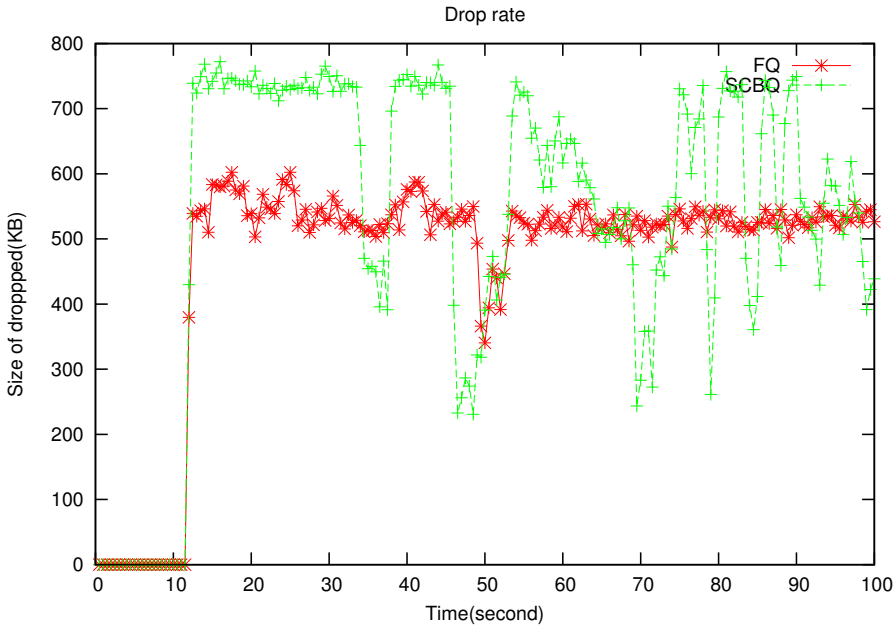


Figure 5.16: Drop rate under 40 stations for BE service class

5.3 MCA Scheduling Discipline

The intent of this experiment was to study the performance of the MCA scheduling discipline under two different SS scheduler. We have considered two scenarios as shown in Table 5.2. The MCA was the BS scheduler in borth cases, but two different schedulers are applied at the SS. In the first scenario PQ scheduling mechanism used as the SS scheduler, and in the second one IPQ1 used as a SS scheduler.

5.3.1 Bandwidth Measurement

Figures 5.17 and 5.19, show the service level of the first scenario with 35 SSs and 40 SSs respectively. The service level of UGS class is fairly constant in both graphs, but the rtPS service level curve fluctuates quite drasitically after the first ten seconds as

Table 5.2: Different Experiment Scenarios

	Number of SS	BS scheduler	SS scheduler
First scenario	35	MCA	PQ
	40	MCA	PQ
Second scenario	35	MCA	IPQ1
	40	MCA	IPQ1

its bandwidth requirement chnages over a different period of time (as we have used vbr traffic source). The negative impact of this fluctuation on the service level of the lower priority classes (nrtPS, BE) is quite clear as whenever the service level of the rtPS class reaches its peak rate the nrtPS service level curve decreases and at the same time BE service curve pushes to zero. When number of SSs increases to 40 SSs, for around 20 seconds (between the time period of 12 sec- 32 sec) the service level of BE class drops to zero as PQ allocates all the received bandwidth to rtPS and nrtPS connections and nothing would be left to be allocated to BE connections.

But if we look at the service level graphs of the second scenario Figure 5.18 and Figure 5.20 , it can be observed that the BE service level never drops to zero as the IPQ1 algorithm considers the expiry status of the packets of each queue. As it can be seen, with 40 SSs, BE service curves is fairly stable. It is because the scheduling algorithm detects the severity of packet expiry at the BE queue in the first round of bandwidth allocation and accordingly allocates some minimum amount of bandwidth to the BE connections.

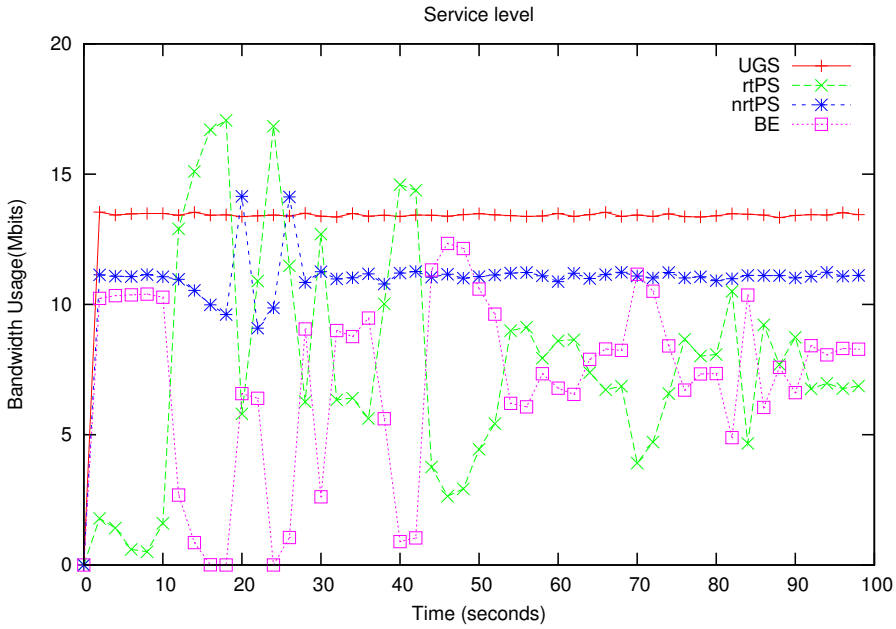


Figure 5.17: Service level in a first scenario with 35 stations

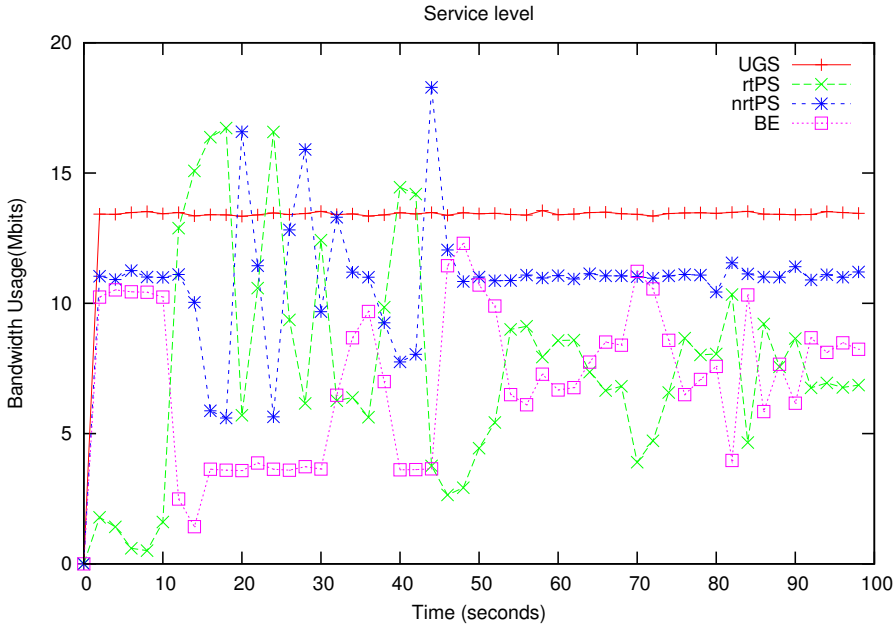


Figure 5.18: Service level in a second scenarion with 35 stations

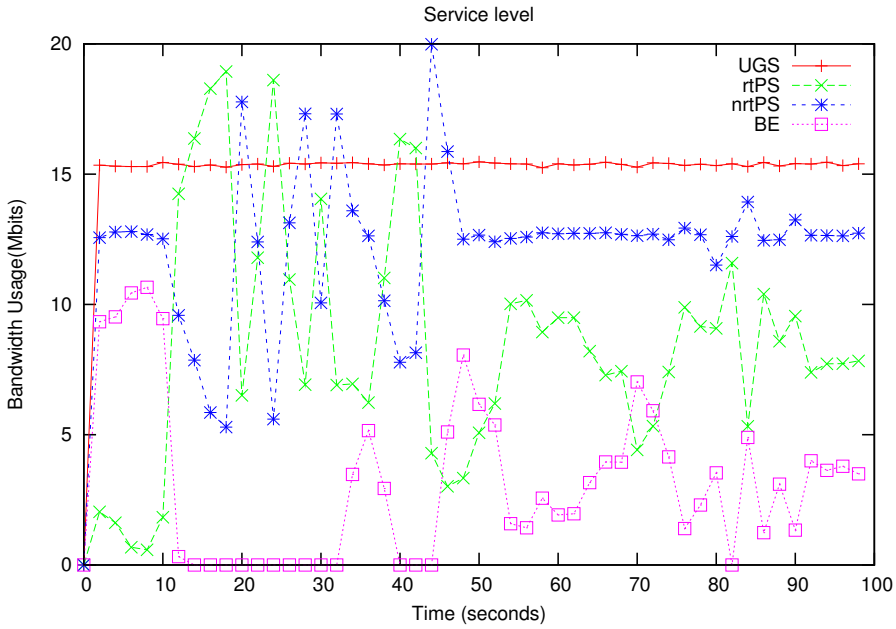


Figure 5.19: Service level in a first scenario with 40 stations

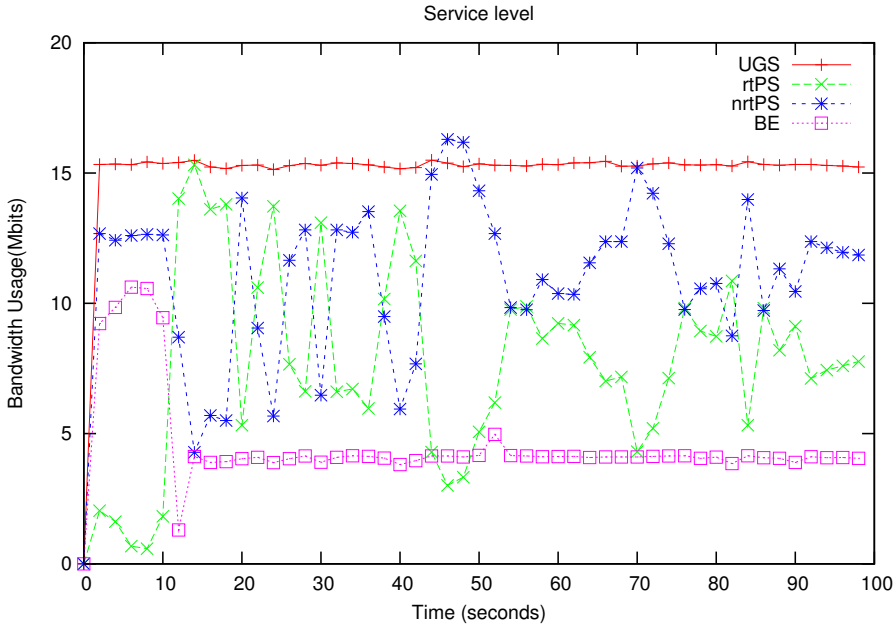


Figure 5.20: Service level in a second scenarion with 40 stations

5.3.2 Average Packet Delay Measeurement

The average delay of scenario one and two under a load of 35 and 40 SSs, are shown in figure 5.21, 5.23, 5.22 and 5.24. The average delay of UGS class is constant for both scenario one and two. The overall rtPS average delay increases slightly in the second scenario, but the overall average delay of the nrtPS class has increased significantly in the second scenario in comparison to the first scenario. The reason behind this rise is IPQ1 allocates less bandwidth to nrtPS connections to be able to allocates some bandwidth to BE connections as well. But it should be mentioned that the average delay of nrtPS class is still below its maximum delay bound, and the algorithm provided a reasonable delay for the nrtPS connections.

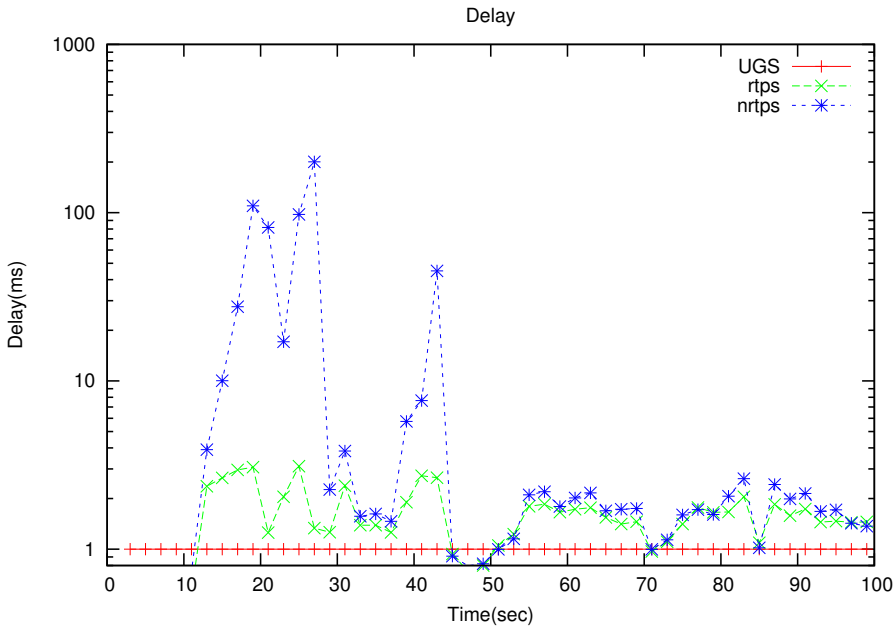


Figure 5.21: Delay in a first scenario with 35 stations

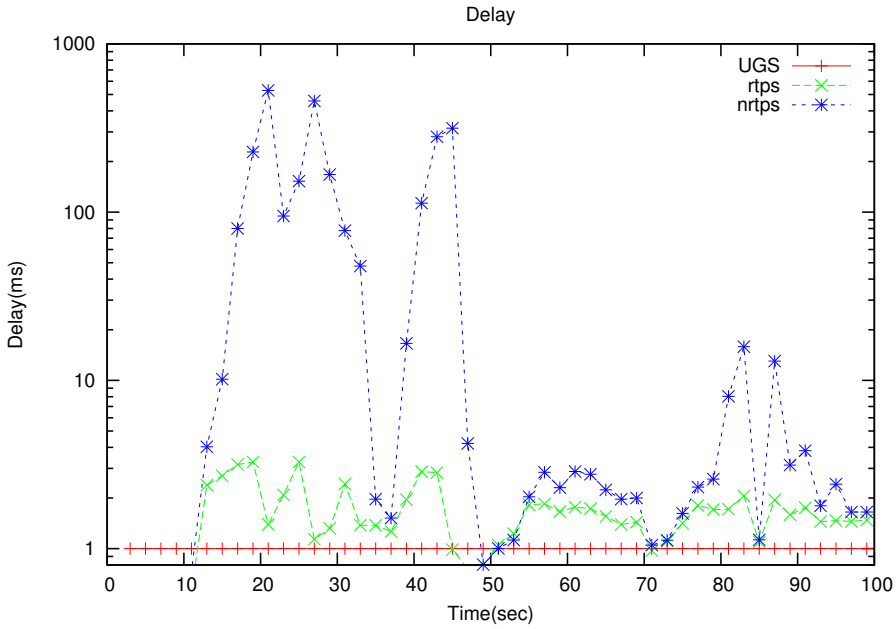


Figure 5.22: Delay in a second scenario with 35 stations

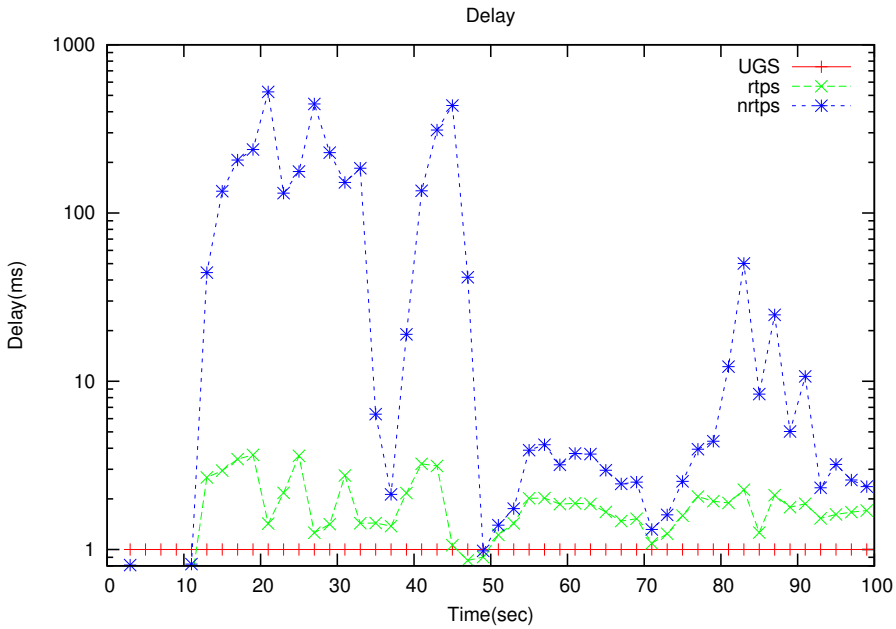


Figure 5.23: Delay in a first scenario with 40 stations

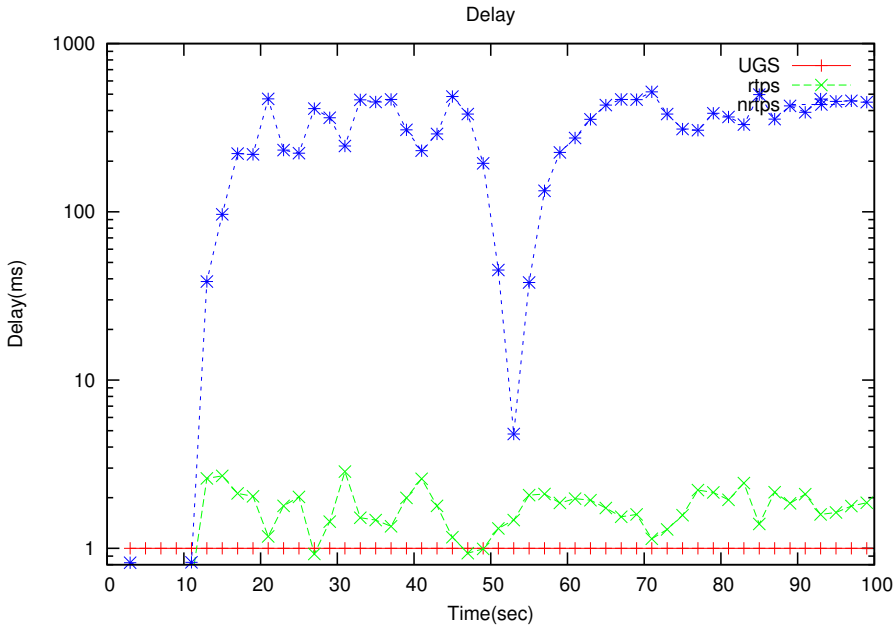


Figure 5.24: Delay in a second scenario with 40 stations

5.3.3 Average Packet Drop Measurement

The average drop of BE class under a IPQ1 is significantly lower than the PQ under both 35 SSs (Figure 5.25) and 40 SSs (Figure 5.26). This can be justified by the fact that the BE class receives more bandwidth under IPQ1 than the PQ. But this bandwidth is not enough to reduce its drop rate to zero as the system is overloaded under both 35 and 40 SSs. Under a load of 35 SSs there are just a few times that PQ drop rate is less than IPQ1. If you look at the service level graphs at those points, you can see that the PQ allocates higher amount of bandwidth to the BE connections than IPQ1 does at those points. The reason behind this is under a PQ algorithm, the BE connections receive the bandwidth whenever the other connections do not

need that bandwidth, however under a IPQ1 the BE connections consistently receive some share of the bandwidth.

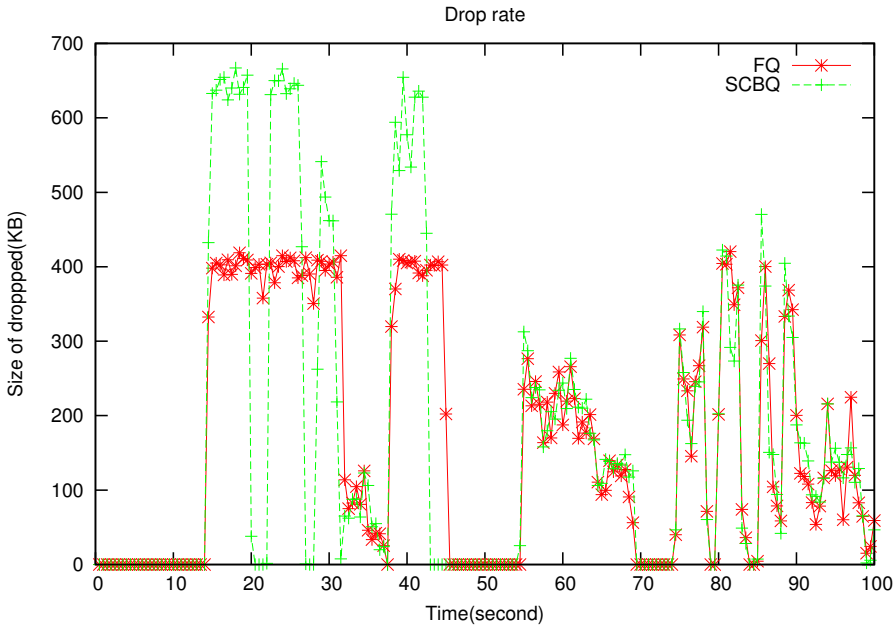


Figure 5.25: Drop rate under 35 stations for BE service class

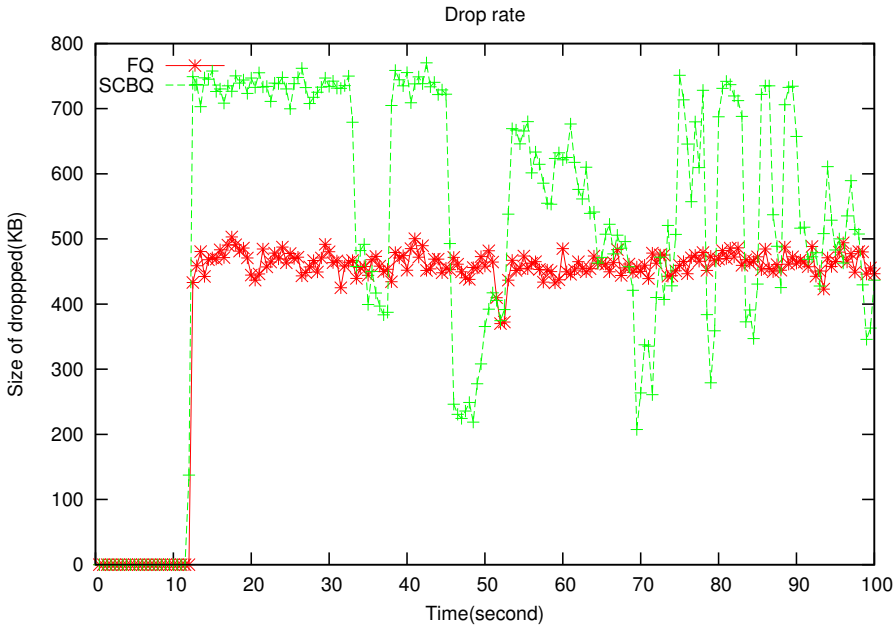


Figure 5.26: Drop rate under 40 stations for BE service class

5.4 Performance Comparison

In this section the performance of all the proposed algorithms are compared against each other. The parameters which have been considered here for comparison are delay, packet drop rate and channel utilization.

Figure 5.27 shows the average delay of UGS class under different traffic load for different combinations of scheduling schemes. As it can be seen, the average delay curves of all the algorithms increase smoothly when the system is underloaded (i.e., the number of SSs is smaller than or equal to 30). Under this condition, UGS queues are lightly occupied. Thus, when a packet is received by the SS, it is very likely to be serviced in the subsequent frames. However, when the system becomes overloaded (i.e., the number of SSs becomes greater than 30), the average packet delay has higher rate of increase.

The lowest delay for UGS class is provided by MCA-PQ (MCA used at the BS and PQ at the SS) and followed by MCA-IPQ1, MCA-IPQ2, SCBQ-PQ, EDF-QL-PQ, EDF-RA-PQ, EDF-QL-IPQ1, EDF-RA-IPQ1, EDF-QL-IPQ2, EDF-RA-IPQ2. As it can be seen in the graph, the scenarios in which IPQ2 is used as the SS scheduler are overally having higher delay comparing to all other combinations. This behavior can be explained by the way that the IPQ2 calculates the severity multiplier for each class of service. In severity multiplier calculation the UGS connections are getting lower bonus rate than other classes as their normalized average delay is fairly low. Although, IPQ2 gives a lower delay when it is paired with MCA as BS scheduler. According to the graph the MCA-PQ and MCA-IPQ1 provide lower average delay than the other schemes for UGS connections. This is due to the way in which capacity has been provisioned to the SSs by the MCA BS scheduler.

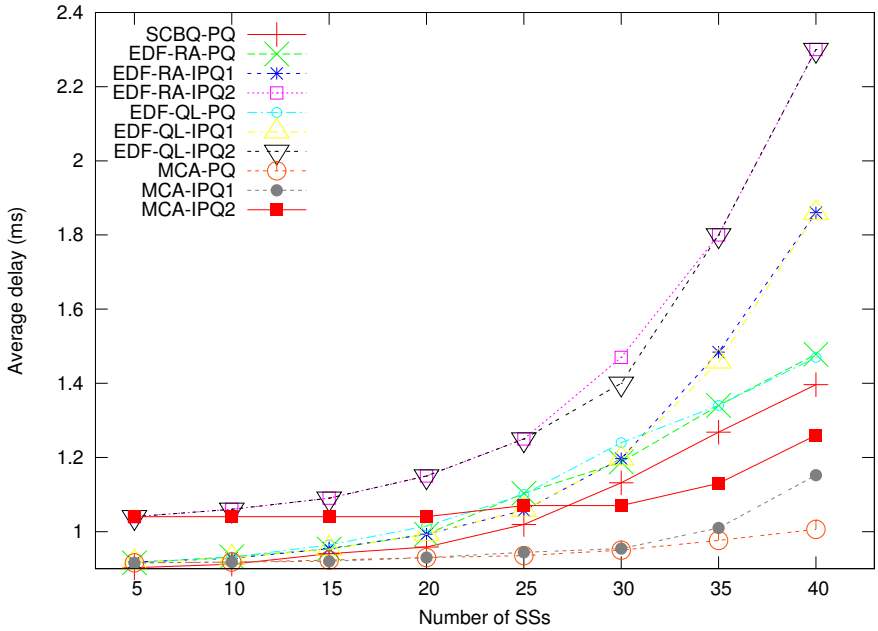


Figure 5.27: UGS average delay vs. number of SSs

Figure 5.28 shows the rtPS average delay (the second priority class), when the number of SSs increases from 5 to 40. The average delay increases smoothly from 2 ms to 9.8 ms, 10.5 ms and 11.2 ms for the MCA-PQ, MCA-IPQ1 and MCA-IPQ2 scheduling scenarios respectively. The EDF-QL-PQ, EDF-RA-PQ and SCBQ-PQ mean delay curves begin from 2 ms and they all converge to approximately 10 and 12 ms under traffic load of 35 and 40 SSs. The EDF-QL-IPQ2 and EDF-RA-IPQ2 delay curves approximately begin from 2 ms and rise to 12 ms and 12.9 ms respectively. The rtPS delay curves of EDF-RA-IPQ1 and EDF-QL-IPQ1 increases from 2 ms to 16 ms and 17 ms respectively. When the system becomes overloaded (i.e., the number of SSs becomes greater than 30) the EDF-RA-IPQ1 and EDF-QL-IPQ1 delay, increase sharply. That is because under a IPQ1 scheduling algorithm, the

rtPS connections are receiving less bandwidth in compare with the scenarios that PQ or IPQ2 is used as SS scheduler.

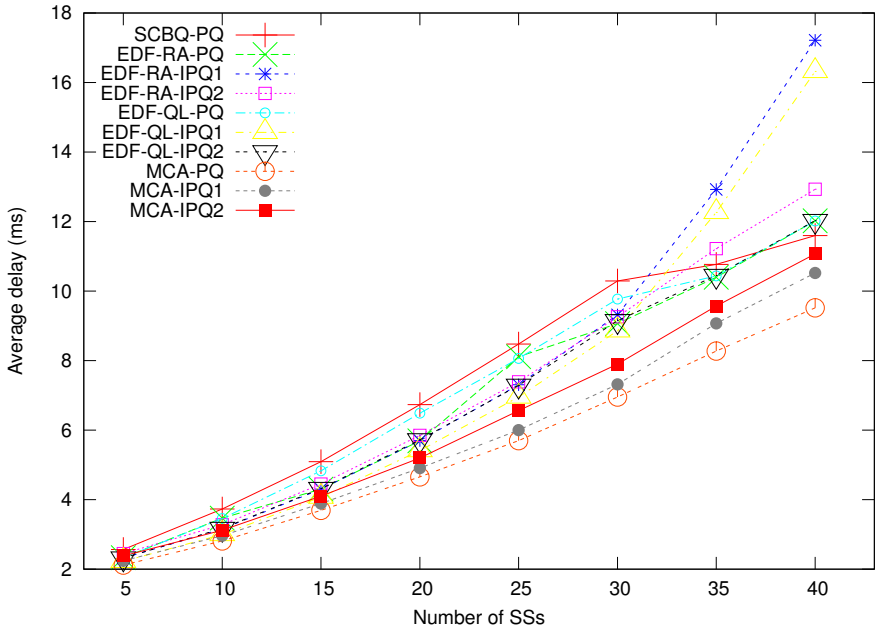


Figure 5.28: rtPS average delay vs. number of SSs

Under a PQ algorithm there is no limit on the bandwidth usage of rtPS connections, so when the traffic load increases the rtPS connections starve the lower priority connections and receive all their required bandwidth by the PQ scheduler without considering the other classes. As a result we got lower delay for rtPS class with PQ than the IPQ1. The IPQ2 calculates the severity multiplier by considering all the packets in connection's queue and not only the packets which are due to expire within the next couple of frames like an IPQ1. Therefore, when the traffic load increases and rtPS connections are having a higher queue length (as the overall bandwidth should be distributed among more connections), the calculated severirty multiplier by IPQ2 algorithm is larger than the ones calculated by IPQ1. As a result

the rtPS connections are experiencing a higher delay under a IPQ1 scheme when the system becomes overloaded. The only exception here is MCA-IPQ1, which is providing a lower delay for rtPS connections than other IPQ1 schems. This is the effect of using MCA in the BS, that distribute the bandwidth equally among all the connections of one class (according to their SLA). Using a MCA in conjunction with IPQ1 provides a very good delay for the rtPS connections.

Figure 5.29 shows the average delay of nrtPS traffic, for a better illustration the results are also shown in Table 5.3. According to the graph, the mean nrtPS delay curves increase linearly when the system is not overloaded (i.e, the number of SSs is smaller than or equal to 30). However, when the system becomes overloaded (e.g., the number of SSs becomes greater than 30), the mean delay curves increase sharply. When the number of SSs is equal to 30, the lowest nrtPS delay is provided by MCA-PQ = 7.57 and MCA-IPQ1 = 7.73, followed by MCA-IPQ2 = 8.2, EDF-QL-IPQ2 = 9.15, EDF-QL-PQ = 9.28, EDF-RA-IPQ2 = 9.31, EDF-RA-PQ = 9.5, EDF-QL-IPQ1 = 9.73, EDF-RA-IPQ1 = 9.87, SCBQ-PQ = 10.67. However, when the system load increases to 35 SSs (system become overloaded), the lowest mean delay is provided by MCA-PQ = 49.12, followed by EDF-QL-PQ = 79.87, EDF-RA-PQ = 81.48, SCBQ-PQ = 86.82, MCA-IPQ2 = 112.27, EDF-QL-IPQ2 = 117.21, EDF-RA-IPQ2 = 118.78, EDF-QL-IPQ1 = 194.09, EDF-RA-IPQ1 =197.83 and MCA-IPQ1 = 215.54 which shows the significant rise in delay. As can be seen, the mean delay of the schemes which use a IPQ1 as a SS scheduler are significantly higher (when the system becomes overloaded), than the ones which use a PQ as their SS scheduler. In the case of using a IPQ1 as a SS scheduler instead of PQ, the nrtPS connection queues grow as the scheduler allocates only a portion of the remaining bandwidth to the nrtPS connections, as it also considers the BE bandwidth requirements. The

same justification as the one given for rtPS is valid about the IPQ2. However, it should be noted here that the provided delay for nrtPS connections by the IPQ1 scheduling discipline is higher than the PQ and IPQ2 scheduling scenarios but it is still far below the maximum delay bound of nrtPS class.

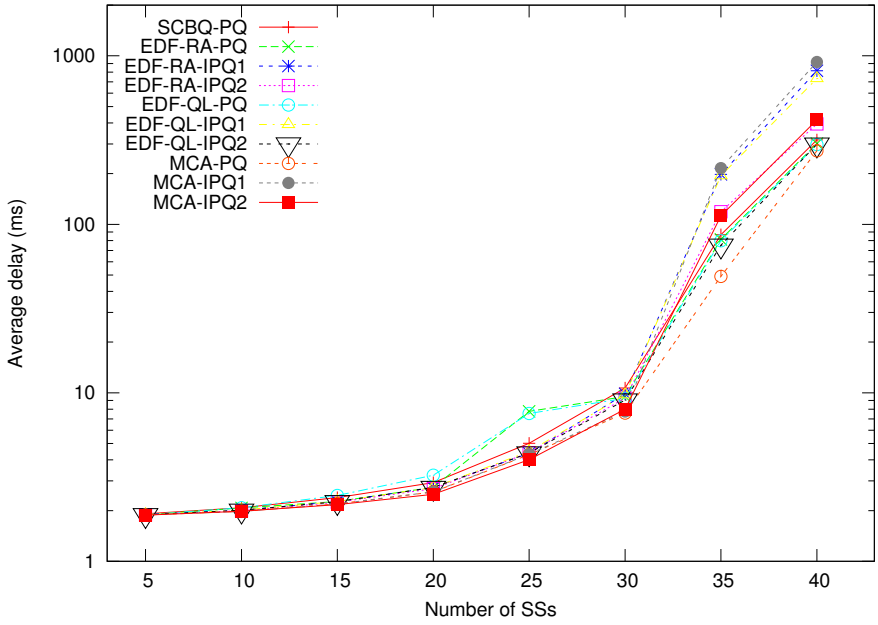


Figure 5.29: nrtPS average delay vs. number of SSs

Table 5.3: nrtPS average delay

	Num SSs							
$nrtPS_{(ms)}$	5	10	15	20	25	30	35	40
SCBQ-PQ	1.91	2.08	2.39	2.93	5.01	10.67	86.82	317.42
EDF-RA-PQ	1.88	2.07	2.26	2.75	7.76	9.5	81.48	301.28
EDF-RA-IPQ1	1.89	2.03	2.26	2.74	4.38	9.87	197.83	816.33
EDF-RA-IPQ2	1.89	2	2.24	2.71	4.42	9.31	118.78	394.07
EDF-QL-PQ	1.89	2.07	2.45	3.24	7.53	9.28	79.87	297.52
EDF-QL-IPQ1	1.89	2.04	2.26	2.73	4.48	9.73	194.09	733.85
EDF-QL-IPQ2	1.89	2.01	2.26	2.74	4.39	9.15	117.21	401.77
MCA-PQ	1.88	1.97	2.19	2.58	4.3	7.57	49.12	273.12
MCA-IPQ1	1.88	1.97	2.18	2.58	4.32	7.73	215.45	873.41
MCA-IPQ2	1.88	1.98	2.17	2.5	4.01	8.2	112.27	417.52

The delay of BE connections is shown in figure 5.30, for better illustration the results are also shown in Table 5.4. When the number of SSs are less than 20, the BE delay increases smoothly. The mean delay considerably rises when the number of SSs increases from 20 to 25 SSs but it is still less than 40 ms. When the number of SSs increases to more than 25, the delay rises sharply. When the number of SSs increases from 35 to 40 there is not much increase in the average delay as most of the packets would be expired before being served. It can be clearly seen that the schemes that use a PQ as the SS scheduler are providing a much higher delay than the ones which are using a IPQ1 and IPQ2. This is because the PQ scheme starve the BE connections as they have the lowest priority. Figure 5.30 shows that that the schemes which are paired with IPQ2 provide a much lower delay comparing to all other combinations.

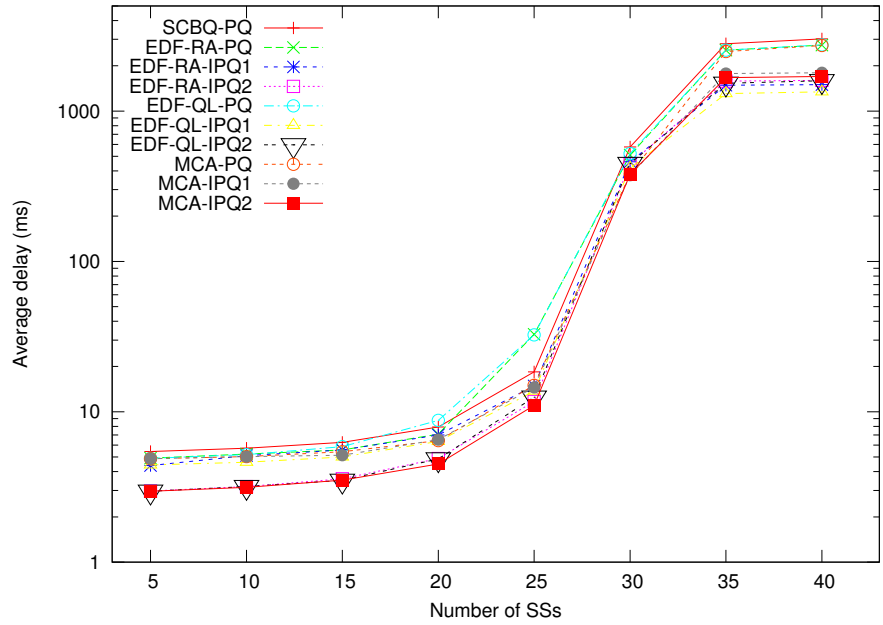


Figure 5.30: BE average delay vs. number of SSs

Table 5.4: BE average delay

	Num SSs							
$BE_{(ms)}$	5	10	15	20	25	30	35	40
SCBQ-PQ	4.94	5.19	5.68	7.2	16.77	525.28	2547.9	2747
EDF-RA-PQ	4.91	5.22	5.56	7.09	32.74	511.05	2543	2733
EDF-RA-IPQ1	4.39	5.11	5.56	7	14.86	459.45	1483.7	1513
EDF-RA-IPQ2	2.97	3.18	3.58	4.88	11.71	445.66	1582.7	1607.49
EDF-QL-PQ	4.9	5.19	5.86	8.76	32.5	520.78	2542	2721
EDF-QL-IPQ1	4.92	5.12	5.56	7.06	15.22	463	1450	1490.53
EDF-QL-IPQ2	2.9	3.18	3.59	4.86	12.54	447.94	1581.4	1590.43
MCA-PQ	4.87	5.03	5.42	6.42	15	383	2489	2728
MCA-IPQ1	4.87	5.03	5.14	6.52	14.5	374.07	1776	1823
MCA-IPQ2	2.96	3.14	3.51	4.49	11.03	397.77	1667.6	1693.7

Typical results on variations in average drop rates ($drop_{ugs}$, $drop_{rtps}$, $drop_{nrtps}$, $drop_{be}$) and channel utilization (U) for different scheduling scenarios are shown in Table 5.5 and Table 5.6. The channel utilization was measured as the ratio between the amount of a bandiwdth used for packets transmission to the total available bandwidth of the system. In all the scheduling scenarios and under all different traffic loads the drop rate of UGS class is zero. This is because the UGS has the highest priority in the system and it receives the best service under different scheduling scenarios.

When the traffic load is less than 25 SSs, we got the drop rate of zero percent for rtPS. When the load increases to 25 SSs we got a negligible drop for rtPS class which is due to the use of VBR (variable bit rate) traffic sources, that sometimes the total generated traffic by rtPS sources exceed the available system capacity over a 30 ms period of time which is the delay bound of rtPS connections. Under a load of 30 SSs the lowest drop rate is provided by MCA-IPQ2 = 0.005, MCA-IPQ1 = 0.006, MCA-PQ = 0.006 and SCBQ-PQ = 0.006 followed by EDF-OL-IPQ2 = 0.008, EDF-RA-IPQ2 = 0.009, EDf-RA-PQ = 0.009, EDF-RA-IPQ1 = 0.01, EDF-QL-IPQ1 = 0.01 and EDF-QL-PQ = 0.01. As the traffic load increases further, the rtPS packet drop rate increases too. When the system becomes overloaded the scenarios that use a PQ as a SS scheduler tend to provide a slightly lower drop rate for rtPS class in compare to others. This is because the PQ algorithm starve the lower priority classes in order to provide the best service for the rtPS class.

Under a traffic load of 30 SSs or less the nrtPS drop rate is zero. There are two reasons for why we get a lower drop rate for nrtPS class than rtPS class. The first reason is that we are using different types of traffic sources for each of them. As it is mentioned above the rtPS traffic sources are very bursty and are generating variable

bit rates data, however the nature of nrtPS traffic sources are not as bursty as VBR sources. The second reason behind this is that the rtPS packets can only tolerate a delay of 40 ms as they are carrying a real time application data, but the nrtPS packts can tolerate a delay of up to 1 minute before become expire. Therefore, the scheduler has a more time to service the nrtPS packets than the rtPS ones. When the number of SSs increases to more than 30 (the system becomes overloaded), the scenarios which use a IPQ1 as their SS scheduler are having a higher nrtPS drop rate than the others. This is because the IPQ1 takes into account a severity of packet expiry only within the next few frames in comparison to IPQ2, which consider the overall criticality of packet expiry at each connection. The IPQ1 provides a higher drop rate for nrtPS class than PQ, since PQ does not consider the severity of packet expiry at all and starve the BE connections.

Under a load of less than 30 SSs, the BE drop rate is zero for all the scheduling scenarios. When the traffic load increases furthur the lowest drop rate for BE is provided by MCA-IPQ1 followed by EDF-QL-IPQ1, EDF-RA-IPQ1, MCA-IPQ2, MCA-PQ, EDF-QL-IPQ2, EDF-RA-IPQ2, EDF-QL-PQ and SCBQ-PQ. As can be seen, the scenarios that use a IPQ1 as their SS scheduler are having a significantly lower drop in comparison with the scenarios that use a PQ.

As it is shown in Table 5.5 and Table 5.6 the channel utilization (U) considerably rises when traffic load increases from 5 SSs to 30 SSs. But increasing the load to more than 30 SSs (overloading the system) causes only a slight increase in U. A very similar results on channel utilization were obtained for all scheduling scenarios, when the system load is less than or equal to 30 SSs. Although, the channel utilization slightly increases by overloading the system but at the same time the overall drop rate goes up significantly.

Table 5.5: Simulation result

Num SSs	5	10	15	20	25	30	35	40
	SCBQ-PQ							
$drop_{\text{ugs}}$	0	0	0	0	0	0	0	0
$drop_{\text{rtps}}$	0	0	0	0	0.004	0.006	0.01	0.04
$drop_{\text{nrtps}}$	0	0	0	0	0	0	0	0.03
$drop_{\text{be}}$	0	0	0	0	0	0.02	0.34	0.73
U	0.15	0.30	0.45	0.61	0.76	0.91	0.96	0.97
	EDF-RA-PQ							
$drop_{\text{ugs}}$	0	0	0	0	0	0	0	0
$drop_{\text{rtps}}$	0	0	0	0	0.006	0.009	0.02	0.05
$drop_{\text{nrtps}}$	0	0	0	0	0	0	0	0.02
$drop_{\text{be}}$	0	0	0	0	0	0.05	0.33	0.72
U	0.15	0.31	0.45	0.61	0.78	0.91	0.97	0.97
	EDF-RA-IPQ1							
$drop_{\text{ugs}}$	0	0	0	0	0	0	0	0
$drop_{\text{rtps}}$	0	0	0	0	0.002	0.01	0.04	0.09
$drop_{\text{nrtps}}$	0	0	0	0	0	0	0.002	0.05
$drop_{\text{be}}$	0	0	0	0	0	0.02	0.32	0.63
U	0.15	0.30	0.45	0.61	0.76	0.91	0.96	0.96
	EDF-RA-IPQ2							
$drop_{\text{ugs}}0$	0	0	0	0	0	0	0	0
$drop_{\text{rtps}}$ 0	0	0	0	0	0.004	0.009	0.04	0.08
$drop_{\text{nrtps}}$ 0	0	0	0	0	0	0	0	0.03
$drop_{\text{be}}$ 0	0	0	0	0	0	0.05	0.32	0.70
U	0.15	0.38	0.45	0.61	0.74	0.90	0.94	0.95
	EDF-QL-PQ							
$drop_{\text{ugs}}$	0	0	0	0	0	0	0	0
$drop_{\text{rtps}}$	0	0	0	0	0.007	0.01	0.02	0.05
$drop_{\text{nrtps}}$	0	0	0	0	0	0	0	0.02
$drop_{\text{be}}$	0	0	0	0	0	0.07	0.33	0.72
U	0.15	0.31	0.46	0.62	0.78	0.91	0.96	0.97

Table 5.6: Simulation result

Num SSs	5	10	15	20	25	30	35	40
	EDF-QL-IPQ1							
$drop_{\text{ugs}}$	0	0	0	0	0	0	0	0
$drop_{\text{rtps}}$	0	0	0	0	0.002	0.01	0.04	0.09
$drop_{\text{nrtps}}$	0	0	0	0	0	0	0.001	0.05
$drop_{\text{be}}$	0	0	0	0	0	0.01	0.32	0.63
U	0.15	0.30	0.45	0.61	0.76	0.91	0.96	0.96
	EDF-QL-IPQ2							
$drop_{\text{ugs}}0$	0	0	0	0	0	0	0	0
$drop_{\text{rtps}}$ 0	0	0	0	0	0.002	0.008	0.04	0.08
$drop_{\text{nrtps}}$ 0	0	0	0	0	0	0	0	0.02
$drop_{\text{be}}$ 0	0	0	0	0	0	0.02	0.31	0.70
U	0.15	0.36	0.45	0.61	0.76	0.91	0.96	0.97
	MCA-PQ							
$drop_{\text{ugs}}$	0	0	0	0	0	0	0	0
$drop_{\text{rtps}}$	0	0	0	0	0.001	0.006	0.01	0.04
$drop_{\text{nrtps}}$	0	0	0	0	0	0	0	0.01
$drop_{\text{be}}$	0	0	0	0	0	0.004	0.27	0.69
U	0.15	0.30	0.45	0.61	0.76	0.91	0.98	0.98
	MCA-IPQ1							
$drop_{\text{ugs}}$	0	0	0	0	0	0	0	0
$drop_{\text{rtps}}$	0	0	0	0	0.0009	0.006	0.03	0.08
$drop_{\text{nrtps}}$	0	0	0	0	0	0	0.005	0.08
$drop_{\text{be}}$	0	0	0	0	0	0.001	0.24	0.55
U	0.15	0.30	0.45	0.61	0.76	0.91	0.98	0.97
	MCA-IPQ2							
$drop_{\text{ugs}}$ 0	0	0	0	0	0	0	0	0
$drop_{\text{rtps}}$ 0	0	0	0	0	0.001	0.005	0.02	0.06
$drop_{\text{nrtps}}0$	0	0	0	0	0	0	0	0.01
$drop_{\text{be}}$ 0	0	0	0	0	0	0.004	0.25	0.65
U	0.15	0.30	0.45	0.61	0.74	0.91	0.98	0.98

5.5 Summary

A comprehensive analysis of the system performance under a range of traffic loads for number of scheduling combinations was illustrated in this chapter. As shown in the previous sections, different scheduling combinations have different performance in terms of mean delay, drop rate and utilization for different classes. As a result it is up to the service provider to select a pair of shcedulers that meets its network objectives.

From the results presented in this chapter we can conclude that when the system is not overloaded, IPQ1 and IPQ2 algorithms either reduce the drop rates of rtPS, nrtPS and BE service classes or keep it at the same rate as the PQ algorithms does. This is because the IPQ1 and IPQ2 algorithms accurately measure the packet expiry level of each connection's queue and accordingly distribute the bandwidth among the connections. However, when the system becomes overloaded the IPQ1 and IPQ2 algorithms provide a slightly higher drop for rtPS connections than PQ, but at the same time they reduce the BE drop rate significantly. This is because when the system becomes overloaded all connections queue grow at higher rates. Therefore, IPQ1 and IPQ2 allocate a fair share of bandwidth to each of the connections according to the criticality of the packet expiry conditions at each queue. As a result in the overloaded conditions they prevent the higher priority classes to starve the lower ones. It can also be concluded from the results that the overall provided QoS support by IPQ2 outperforms that of IPQ1; considering all classes of service together.

It is also evident from the results that the MCA algorithm provides a better QoS support in terms of packet delay and drop rate for all types of traffic than that of provided by EDF-RA and EDF-QL algorithms.

Chapter 6

Conclusion

In this chapter we conclude this thesis with a summary of the research and the significant results. We also suggest directions for future research related to the work reported here.

6.1 Significant Result and Conclusion

The key contributions of this research were the development of an efficient bandwidth allocation architecture and admission control procedure in order to provide a QoS differentiation for different classes of service for the IEEE 802.16 BWA standard. The performance of different scheduling algorithms was investigated using simulation. The simulation models could be used for future research projects in this area.

A set of QoS oriented bandwidth allocation algorithms, namely Earliest Deadline First with Bandwidth Reservation (EDF-BR) and Multi-Class Allocation (MCA) are proposed for employment at the BS. At the subscriber station (SS) level, we introduced two different enhancement to the original priority queuing (PQ) and

evaluated their performance in comparison with traditional PQ. The comparisons were made under different load condition. We have also evaluated the performance of different combinations of BS and SS schedulers.

Simulation results show that MCA and EDF-BR schemes provide better performance than the strict class based queuing (SCBQ) in different traffic scenarios. It was determined from the simulations that MCA provides a better QoS support in terms of packet delay and drop rate for all types traffic than EDF-BR. It was also established via the simulation that the improved priority queuing 1 (IPQ1) and IPQ2 schemes outperform the PQ by preventing the starvation of a lower priority classes. The simulation results show that the IPQ1 and IPQ2 algorithms provide a lower drop rate for BE service class than PQ. The IPQ1 and IPQ2 measure the packet expiry level of each connection's queue and accordingly distribute the bandwidth among the connections. It was also shown that our proposed algorithms provide a service differentiation in terms of delay and loss rate for real-time (served via UGS and rtPS) and non real-time (served via nrtPS and BE) traffic.

We have also developed a simple but practical priority admission control policy called hybridI. The performance of hybridI and hybridII were compared to that of complete sharing(CS) and complete partioning(CP). The comparisons were made under different incoming traffic load for different priority flows. It was established from simulations that the hybrid methods minimize the blocking probability of UGS and rtPS classes which maximize the revenue for service provider.

6.2 Suggested Future Research

There are much scope for future work with regard to the performance of EDF-BR and MCA. The understanding of how the queuing algorithms at the BS behave when paired with different queuing algorithm at the SS, are very important because this could be a significant step in presenting an analytical explanation. This would help to estimate an upper bound delay for different priority classes in order to minimize the packet loss. This is a key step to possible future work for finding a closed form expression for the mean delay of different queuing strategy.

Since many different scheduling schemes can be applied in the IEEE 802.16, it would be interesting to study the performance of these schemes under a error-prone wireless channel since the wireless links are not always ideal. An analytical model for error-prone wireless channel based on Game theory approach and the performance evaluation would be an interesting research problem.

Appendix A

Publications

- E.A Aghdaee, N. Mani, and B. Srinivasan, An Enhanced Bandwidth Allocation Algorithms for QoS provision in IEEE 802.16 BWA, *submitted in The International Conference on Information Networking (ICOIN)*, January 2007.

References

[1] M. Pidutti. 802.16 Tackles Broadband Wireless QoS Issues. *CommsDesign*, December 2004.

[2] T. Gaden. Intel pumps $37 million into Unwired. *Whirlpool Broadband Multimedia*, August 2005.

[3] W. Stalling. *Data and Computer Communications*. Prentice Hall, NewJersey, 7 edition, 2004.

[4] G. Huston. *Internet Performance Survival Guide: QoS Strategies for Multiservice Networks*. John Wiley & Sons, 1 edition, 2000.

[5] IEEE 802.16 Working Group on Broadband Wireless Access. *IEEE Standard for Local and Metropolitan Area Networks*. Part 16: Air Interface for Fixed Broadband Wireless Access Systems. IEEE, 2004.

[6] IEEE 802.16 Working Group on Broadband Wireless Access. *IEEE Standard for Local and Metropolitan Area Networks*. Part 16: Air Interface for Fixed Broadband Wireless Access Systems - Amendment 2: Medium Access Control Modifications and Additional Physical Layer Specifications for 2-11 GHz. IEEE, 2003.

[7] IEEE 802.16 Working Group on Broadband Wireless Access. *IEEE Standard for Local and Metropolitan Area Networks*. Part 16: Air Interface for Fixed Broadband Wireless Access Systems - Amendment 1: Detailed System Profiles for 10-66 GHz. IEEE, 2002.

[8] 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*, 08:213–2128, August 2004.

[9] W-CDMA. *Wikipedia - The Free Encyclopedia*, October 2005.

[10] UMTS. *Wikipedia - The Free Encyclopedia*, October 2005. `http://en.wikipedia.org/wiki/UMTS`.

[11] High-Speed Downlink Packet Access. *Wikipedia - The Free Encyclopedia*, October 2005.

[12] CDMA2000. *Wikipedia - The Free Encyclopedia*, October 2005. `http://en.wikipedia.org/wiki/CDMA2000`.

[13] C. Eklund, R. B. Marks, K. L. Stanwood, and S. Wang. IEEE standard 802.16: A Technical Overview of the WirelessMAN^{TM}Air Interface for Broadband Wireless Access. *IEEE Communications Magazine*, 40(6):98–107, June 2002.

[14] A. Ghosh, D. R. Wolter, and J. G. Andrews. Broadband Wireless Access with WiMax/8O2.16: Current Performance Benchmarks and Future Potential. *IEEE Communications Magazine*, 43(2):129–136, February 2005.

[15] R. Guerin and V. Peris. Quality-of-service in packet networks: Basic Mechanisms and directions. *Cmputer Networks, special Issue on Internet Telephony*, 31(3):169–189, Feb 1999.

[16] J Nagle. On Packet Switches with Infinit Early Detection. *IEEE Trans. on Communications*, 35:435–438, 1987.

[17] W. Stalling. *High-Speed Networks and Internets: Performance and Quality of Service*. Prentice Hall, NewJersey, 2 edition, 2002.

[18] A. K. Parekh and R. G. Gallager. A Generalized Processor Sharing Approach to flow control- The single Node Case. *IEEE INFOCOM*, 2:915–24, May 1992.

[19] A. Parekh. A Generalized Processor Sharing Approach to flow control in Integrated Service Networks. *Phd dissertation, Massachusetts Institute Of Technology*, 35, Feb 1992.

[20] J. C. R. Bennet and H. Zhang. WF^2Q: Worst-case Fair Weighted Fair Queuing. *IEEE INFOCOM*, pages 120–128, Mar 1996.

[21] J.M Peha and F.A Tobagi. A cost-based scheduling algorithm to support integrated services. *Proceedings - IEEE INFOCOM*, 2:741 – 753, 1991.

[22] V. Bharghavan, S. Lu, and T. Nandagopal. Fair Queuing in Wireless Networks: Issues and Approaches. *IEEE Personal Commun*, pages 44–53, Feb 1999.

[23] V. Bharghavan S. Lu and R. Sirkant. Fair scheduling in wireless packet networks. *IEEE/ACM Trans. Networking*, 7(4):473 – 489, August 1999.

[24] Y. Cao and V.O.K Li. Scheduling Algorithms in Broad-Band Wireless Networks. *Proc. IEEE*, 89(1):76 – 78, January 2001.

[25] I. Stoica E. Ng and H. Zhang. Packet fair queuing algorithms for wireless networks with location-dependent errors. *IEEE INFOCOM*, pages 1103–1111, Mar 1998.

[26] S. Lu, T. Nandagopal, and V. Bharghavan. Design and analysis of an algorithm for fair service in error-prone wireless channels. *Wireless Networks*, 6(4):323 – 343, 2000.

[27] M. A. Arad and A. Leon-Garcia. A Generalized Processor Sharing Approach to Time Scheduling in Hybrid CDMA/TDMA. *IEEE INFOCOM*, 3:1164–1171, Mar 1998.

[28] M. A. Arad and A. Leon-Garcia. Scheduled CDMA: A Hybrid Multiple Access for Wireless ATM Network. *Indoor and Mobile Radio Commun. Conf.*, 3:913–917, Mar 1996.

[29] O. Gurbuz and H. Owen. Dynamic Resource Scheduling Strategies for QoS in W-CDMA. *IEEE Global Telecommun. Conf.*, 1:183–87, Dec 1999.

[30] D.A Levine I.F Akyildiz and I. Joe. A Slotted CDMA Protocol with BER Scheduling for Wireless Multimedia Networks. *IEEE/ACM Trans. Networking*, 7(2):146–158, Apr 1999.

[31] D. Cho, J. Song, M. Kim, and K. Han. Performance Analysis of the IEEE 802.16 Wireless Metropolitan Area Network. *First International Conference on Distributed Frameworks for Multimedia Applications*, pages 130–136, February 2005.

[32] Y. Shang and S. Cheng. An Enhanced Packet Scheduling Algorithm for QoS Support in IEEE 802.16 Wireless Network. *LNCS*, 3619:152–661, September 2005.

[33] H. Lee, T. Kwon, and D. Cho. An efficient uplink scheduling algorithm for VoIP services in IEEE 802.16 BWA systems. *IEEE 60th Vehicular Technology Conference*, 5:3070–3074, September 2004.

[34] K. Wongthavarawat and A. Ganz. IEEE 802.16 Based Last Mile Broadband Wireless Military Networks with Quality of Service Support. *IEEE Military Communications Conference*, 2:779–784, October 2003.

[35] D. Wang G. Chu and S. Mei. A QoS architecture for the MAC protocol of IEEE 802.16 BWA System Communications. *Circuit and systems and West Sino Expositions, IEEE International Conference*, 1:435–39, July 2002.

[36] C. Cicconetti, L. Lenzini, and E. Mingozzi. Quality Of Service Support in IEEE 802.16 Networks. *IEEE Network*, pages 50 – 55, April 2006.

[37] B. Kraimeche and M. Schwartz. Analysis of traffic access control strategies in integrated service networks. *IEEE Transactions on Communications*, CM-33(10):1085–1093, 1985.

[38] B. Kraimeche and M. Schwartz. Circuit access control strategies in integrated digital networks. *Proceedings of the IEEE INFOCOM 84*, pages 230–235, 1984.

[39] I. S. Gopal and T. E. Stern. Optimal call blocking policies in an integrated services environment. *Proceeding of the Sventh Conference on Information Science and Systems*, 1984.

[40] K.W. Ross and D.H.K. Tsang. Optimal circuit access policies in an isdn environment: a markov decision approach. *IEEE Transactions on Communications*, 37(9):934–939, September 1989.

[41] D. Ferrari and D. C. Verma. Scheme for real-time channel establishment in wide-area networks. *IEEE Journal on Selected Areas in Communications*, 8(3):368 – 379, 1990.

[42] K.K Leung, W.A Massey, and W. Whitt. Traffic models for wireless communication networks. *IEEE Journal on Selected Areas in Communications*, 12(8):1353 – 1364, October 1994.

[43] P.V Orlik and S.S Rappaport. A model for teletraffic performance and channel holding time characterization in wireless cellular communication with general session and dwell time distributions. *IEEE Journal on Selected Areas in Communications*, 16(5):788 – 803, June 1998.

[44] Y. Fang and I. Chlamtac. A new mobility model and its application in the channel holding time characterization in pcs networks. *IEEE INFOCOM '99.*, 1:20 –27, 1999.

[45] W. Wu Y. Ruan G. Liu, W. Lang. QoS guaranteed call admission scheme for Broadband Multi-services Mobile Wireless Networks. *Proc. IEEE Int Conf. Comm*, 1:454–459, July 2004.

[46] J. Hou and Y. Fang. Mobility-based call admission control schemes for wireless mobile networks. *Wireless Communications and Mobile Computing*, 1(3):269 – 282, July 2001.

[47] I.-S. Yoon and B.G. Lee. A distributed dynamic reservation scheme that supports mobility in wireless multimedia communications. *IEEE Journal on Selected Areas in Communications*, 19(11):2243 – 2253, 2001.

[48] S.K. Biswas and D. Reininger. Bandwidth allocation for vbr video in wireless atm links. *Mobile Multimedia Communications*, pages 199 – 213, 1997.

[49] A.S. Acampora and Z. Zhang. A throughput/delay comparison: narrowband versus broadband wireless lan's. *IEEE Transactions on Vehicular Technology*, 42(3):266 – 273, August 1993.

[50] C.M. Barnhart, J.E. Wieselthier, and A. Ephremides. Admission-control policies for multihop wireless networks. *Wireless Networks*, 1(4):373 – 387, 1995.

[51] M. Schwartz. Network management and control issues in multimedia wireless networks. *IEEE Personal Communications*, 2(3):8 – 16, June 1995.

[52] Jeu-Yih Jeng, Yi-Bing Lin, and Herman Chung-Hwa Rao. Flexible resource allocation scheme for gsm data services. *IEICE Transactions on Communications*, E81-B(10):1797 – 1802, 1998.

[53] O. Baldo, Lee Kok Thong, and A.H. Aghvami. Performance of distributed call admission control for multimedia high speed wireless/mobile atm networks. *1999 IEEE International Conference on Communications*, vol.3:1982–6, 1999.

[54] D.S. Eom, M. Sugano, M. Murata, and H. Miyahara. Call admission control for qos provisioning in multimedia wireless atm networks. *IEICE Transactions on Communications*, E82-B(1):14 – 23, January 1999.

[55] J. Misic, S.T. Chanson, and F.S. Lai. Admission control for wireless networks with heterogeneous traffic using event based resource estimation. *Proceedings. Sixth International Conference on Computer Communications and Networks*, pages 262–269, 1997.

[56] C. Oliveira, J. B. Kim, and T. Suda. Quality-of-service guarantee in high-speed multimedia wireless networks. *IEEE International Conference on Communications*, 2:728–734, 1996.

[57] W.K Wong, H Zhu, and V.C.M Leung. Soft QoS Provisioning Using The Token Bank Fair Queuing Scheduling Algorithm. *IEEE Wireless Communications*, 10, June 2003.

[58] J. Kurose. Open issues and challenges in providing quality of service guarantees in high-speed networks. *Computer Communication Review*, 23(1):6–15, January 1993.

[59] k.Y Tcha, Y.B Kim, S.C Lee, and C.G Kang K.J.H. Lee. QoS Managment for Facilitation of Uplink Scheduling. *IEEE P802.16e/D4-2004*, August 2004.