A FRAMEWORK FOR EFFICIENT BANDWIDTH MANAGEMENT IN BROADBAND WIRELESS ACCESS SYSTEMS

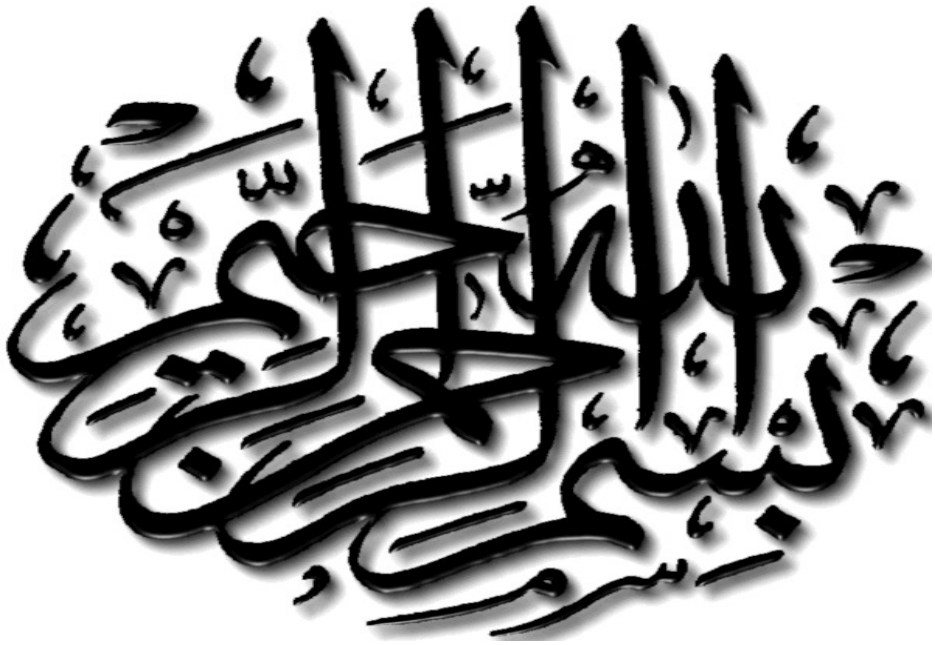
BY

BADER S. AL-MANTHARI

A thesis submitted to the School of Computing
in conformity with the requirements for
the degree of the Doctor of Philosophy

Queen's University
Kingston, Ontario, Canada
March 2009

ISBN:978-0-494-48208-7



IN THE NAME OF ALLAH,
THE MOST GRACIOUS, THE MOST MERCIFUL

"O my Lord, increase me in knowledge"[1]

[1] The Holy Qur'an, surat "Taha", verse: 114.

To My Parents,
My Family,
And My Friends

Abstract

Broadband Wireless Access Systems (BWASs) such as High Speed Downlink Packet Access (HSDPA) and the Worldwide Interoperability for Microwave Access (WiMAX), pose a myriad of new opportunities for leveraging the support of a wide range of "content-rich" mobile multimedia services with diverse Quality of Service (QoS) requirements. This is due to the remarkably high bandwidth that is supported by these systems, which was previously only available to wireline connections. Despite the support for such high bandwidth, satisfying the diverse QoS of users while maximizing the revenues of network operators is still one of the major issues in these systems. Bandwidth management, therefore, will play a decisive role in the success of such wireless access systems. Without efficient bandwidth management, network operators may not be able to meet the growing demand of users for multimedia services, and may consequently suffer great revenue loss. Bandwidth management in BWASs is, however, a challenging problem due to many issues that need to be taken into consideration. Examples of such issues include the diverse QoS requirements of the services that BWASs support, the varying channel

quality conditions of mobile users, and hence the varying amount of resources that are needed to guarantee certain QoS levels during the lifetime of user connections, the utilization of shared channels for data delivery instead of dedicated ones and network congestion.

In this thesis, we address the problem of bandwidth management in BWASs and propose efficient economic-based solutions in order to deal with these issues at different bandwidth management levels, and hence enhance the QoS support in these systems. Specifically, we propose a bandwidth management framework for BWASs. The framework is designed to support multiple classes of traffic with different users having different QoS requirements, maximize the throughput of BWASs, support inter- and intra-class fairness, prevent network congestion and maximize the network operator's revenues. The framework consists of three related components, namely packet scheduling, bandwidth provisioning and Call Admission Control-based dynamic pricing. By efficiently managing the wireless bandwidth prior to users' admission (i.e., pre-admission bandwidth management) and during the users' connections (i.e., post-admission bandwidth management), these schemes are shown to achieve the design goals of our framework.

Keywords: BWASs, bandwidth management, packet scheduling, bandwidth provisioning, Call Admission Control, dynamic pricing, revenues and fairness.

Co-Authors

Chapter 2

- B. Al-Manthari, N. Nasser and H. Hassanein, "Dynamic Pricing in Wireless Cellular Networks", submitted to *IEEE Communications Surveys and Tutorials*, June 2008.

- B. Al-Manthari, N. Nasser and H. Hassanein, "Packet Scheduling in 3.5G High Speed Downlink Packet Access Networks: Breadth and Depth", *IEEE Networks Magazine*, vol. 21, no. 1, pp. 41-46, February 2007.

Chapter 3

- B. Al-Manthari, H. Hassanein, N. A. Ali and N. Nasser, "Fair Class-based Downlink Scheduling in Next Generation Broadband Wireless Access Systems", accepted at *IEEE Transactions on Mobile Computing*, to appear.

- B. Al-Manthari, N. A. Ali, N. Nasser and H. Hassanein, "QoS-based Resource Management Scheme for Multimedia Traffic in High Speed Wireless Networks", *Proceedings of the Annual IEEE Global Telecommunications Conference (Globecom)*, Washington, DC., U.S.A, pp. 5236- 5241, November 2007.

- B. Al-Manthari, N. A. Ali, N. Nasser and H. Hassanein, "A Generic Centralized Downlink Scheduler for Next Generation Wireless Cellular Networks", *Proceedings of the IEEE International Conference on Communications (ICC)*, Glasgow, Scotland, pp. 4566-4572, June 2007.

Chapter 4

- B. Al-Manthari, N. A. Ali, N. Nasser and H. Hassanein, "Dynamic Multiple-Frame Bandwidth Provisioning with Fairness and Revenue Considerations for Broadband Wireless Access Systems", Journal paper submitted for publication.

- B. Al-Manthari, N. A. Ali, N. Nasser and H. Hassanein, "Dynamic Bandwidth Provisioning with Fairness and Revenue Considerations for Broadband Wireless Communication", *Proceedings of the IEEE International Conference on Communications (ICC)*, Beijing, China, pp. 4028- 4032, May 2008.

- B. Al-Manthari, N. A. Ali, N. Nasser and H. Hassanein, "Frame-Level Dynamic Bandwidth Provisioning for QoS-Enabled Broadband Wireless Networks", *Proceedings of the ACS/IEEE International Conference on Computer Systems and Applications (AICCSA)*, Doha, Qatar, pp. 999- 1004, April 2008.

Chapter 5

- B. Al-Manthari, N. Nasser, N. A. Ali and H. Hassanein, "Efficient Admission-level Bandwidth Management in Broadband Wireless Networks: An Economic Approach with Monetary Incentives", Journal paper submitted for publication.

- B. Al-Manthari, N. Nasser, N. A. Ali and H. Hassanein, "Congestion Prevention in Wireless Networks: An Economic Approach", Conference paper submitted for publication.

- B. Al-Manthari, N. Nasser, N. A. Ali and H. Hassanein, "Efficient Bandwidth Management in Broadband Wireless Access Systems Using CAC-based Dynamic Pricing", accepted at the *IEEE conference on Local Computer Networks (LCN)*, Montreal, Canada, to appear.

Acknowledgements

First of all, I thank ALLAH (God) for His mercy, help and guidance, without which this document would not be made possible.

I am grateful to my supervisors professor Hossam Hassanein and Dr. Nidal Nasser for giving me the opportunity to be a member of the Telecommunications Research group and to pursue my Ph.D. degree under their supervision. Their expert knowledge, excellent supervision, wonderful personalities and most of all motivation and support were a great source of help throughout the course of my research.

I am also grateful to Dr. Najah Abu Ali from UAE University whose collaboration provided motivation and support that were a great source of inspiration during the course of my research. Her expertise, knowledge and enthusiasm have made a significant contribution to my graduate experience and have certainly helped improve the quality of my thesis.

I would like to express my deepest gratitude to my parents, siblings and my friends. Mother and father thank you very much for your sincere love. Without your continuous support and prayers this work would not have been accomplished. Brothers and sisters, you cannot imagine how I am grateful to you for your unbelievable support, invaluable advice and standing by me at all times.

I owe Dr. Hanady Abdulsalam special thanks for her humongous unconditional help, caring and support. She has generously devoted to me much of her knowledge and time and I deeply acknowledge her for that. I am also very grateful to my friend Mohammed Hussain who has encouraged me in many occasions and was a great source of help to me.

I would like to thank all members of the Telecommunications Research Lab and the School of Computing at Queen's University for their support and friendship.

Special thanks to Queen's University and the Ministry of Higher Education in the Sultanate of Oman for their financial support.

Statement of Originality

I hereby certify that this Ph.D. thesis is original and that all ideas and inventions attributed to others have been properly referenced.

Table of Contents

List of Figures

List of Tables

List of Acronyms

1G	First Generation
2G	Second Generation
3G	Third Generation
3GPP	3rd Generation Partnership Project
AMR	Adaptive Multi-Rate
BWAS	Broadband Wireless Access System
CAC	Call Admission Control
CAC-bDP	Call Admission Control-based Dynamic Pricing
CCAC	Conventional Call Admission Control
CDMA	Code Division Multiple Access
EDN	Early Delay Notification
FCBPS	Fair Class-Based Packet Scheduling
FFT	Fast Fair Throughput
FM-LWDF	Fair Modified Largest Weighted Delay First

HSDPA	High Speed Downlink Packet Access
ILP	Integer Linear Programming
JFI	Jain Fairness Index
LP	Linear Programming
Max CIR	Maximum Carrier to Interference Ratio
M-LWDF	Modified Largest Weighted Delay First
PDUs	Protocol Data Units
PF	Proportional Fairness
QoS	Quality of Service
SB	Score-Based
UMTS	Universal Mobile Telecommunications System
VoIP	Voice over IP
WiMAX	Worldwide Interoperability for Microwave Access
WTP	Willingness to Pay

List of Definitions

Charge	The amount that is billed for a service.
Congestion externality	The degradation of quality of service that occurs to other users when a certain user transmits when the network is congested.
Handoff connection	An active connection that moved from one cell to another and is requesting the service of its base station.
Handoff connection dropping probability	The probability of dropping a handoff connection.
Interference-limited networks	Networks that are affected by the amount of transmitted power by users.
Modulation	Superimposing the information bits on the carrier frequency.

Multi-user diversity	Exploiting the variations of the channel conditions of the users by serving those with more favorable channel conditions for the benefit of user and/or system throughput.
New connection	A new connection that is requesting to access the network.
New connection admission probability	The probability of admitting a new connection.
New connection blocking probability	The probability of rejecting a new connection.
Price	The amount of money associated with a unit of service.
Price elasticity of demand	The change in demand for a certain product or service due to a change in its price.
Social fairness	The state of economy, where the majority of people are able to buy certain products regardless of their incomes. In the context of this thesis, it refers to the ability to buy or use network services.
Social welfare	Aggregate utility of people.
System capacity	The maximum data rate the system can support.
User's Willingness to Pay	The amount of money the users are willing to pay for a certain product or service.

Chapter 1

Introduction

The evolution of today's wireless communication technology began in the early 1980's with the introduction of first generation (1G) wireless cellular systems. These systems utilized analog interface technology and supported voice-only capabilities, which were typically low in quality. Nevertheless, these systems marked the beginning of a new era in personal communications and were the first step towards achieving the prominent communication concept: “any time, anywhere and in any form”. With the high demand for cellular services and the increased need for enhanced quality and more features, the second generation (2G) of wireless cellular systems was introduced. 2G is characterized by its digital air interface. It is primarily a voice-centric technology, but it does provide higher bandwidth, better voice quality and limited data services. 2G wireless cellular systems have gained tremendous popularity, where they attracted many users and were successfully deployed in many parts of the world. The splendid success of these systems,

however, coupled with the enormous incessant growth of the Internet have led to an increase of demand for mobile wireless data services. This necessitated the need for a higher capacity[2], better Quality of Service (QoS) support and more efficient systems beyond the capabilities of 2G wireless systems. The evolution towards 3G wireless cellular systems was, therefore, inevitable. One of the most famous 3G systems is Universal Mobile Telecommunications System (UMTS) that was developed by the 3rd Generation Partnership Project (3GPP) [1]. UMTS promises a transmission rate of up to 2 Mbps, which makes it possible to support new data services and enhance the ones that are supported by current 2G systems.

Forecasts for emerging mobile wireless markets, however, anticipate that bandwidth will be squeezed by services like multimedia on demand. This will spur the need for data rates beyond what is offered by current 3G wireless systems. To boost the support for such high data rates, Broadband Wireless Access Systems (BWASs) have been developed. For example, 3GPP has standardized a 3.5G BWAS called High Speed Downlink Packet Access (HSDPA) [2] as an extension to the existing 3G UMTS. HSDPA can theoretically support up to 14.4 Mbps, 7 times larger than the data rate offered by the UMTS. Another BWAS is the Worldwide Interoperability for Microwave Access (WiMAX), which has been standardized by the IEEE 802.16 group [3] and [4]. WiMAX is a BWAS that could support up to 70 Mbps. The high data rates offered by these systems allow them to deliver a competitive advantage for mobile data network

[2] The system capacity is defined as the maximum data rate it can support.

operators by boosting network performance to improve the user[3] experience of new converged services such as audio and video streaming, mobile Internet browsing and Voice over IP (VoIP), to name but a few. Such services necessitate the support of different classes of traffic with widely different QoS requirements, which need to be guaranteed by the wireless networks.

Although two different technologies, HSDPA and WiMAX have many common features, among which is the use of a shared channel for data transmission. The shared channel is divided into transmission time frames, where the base station decides which users are allowed to transmit during each time frame. Using shared channels for data transmissions instead of dedicated ones facilitates the support for higher number of mobile users, hence improving the system utilization and reducing the per-bit transmission costs. Sharing such channels among mobile users is, however, a challenging problem due to their varying channel quality conditions, their diverse QoS requirements and anticipated high traffic demands. BWASs, therefore, require more careful and efficient bandwidth management and resource sharing techniques in order to satisfy the QoS requirements of existing and new multimedia services and to maximize the system throughput at the same time. Bandwidth management in BWASs can be done at three levels, namely admission-level, class-level and packet-level as shown in Figure 1.1. Admission-level bandwidth management is typically realized by employing a Call Admission Control (CAC) scheme, which is a pre-admission provisioning strategy that is responsible for accepting or rejecting new connections. CAC aims at satisfying the long-

[3] The terms "user" and "mobile user" are used interchangeably throughout this thesis.

term QoS of users by maximizing the number of admitted user connections while maintaining the QoS of ongoing ones. Class-level and packet-level bandwidth management are post-admission bandwidth management strategies that deal with already admitted users' connections. Specifically, class-level bandwidth management considers the aggregate demand of admitted user connections at the class level. It determines the number of transmission time frames that each traffic class needs in order to maintain the QoS of its admitted users at acceptable levels throughout the lifetime of their connections. After the time frames are distributed amongst the different classes, packet-level bandwidth management is utilized in order to determine which of the users' packets are scheduled for transmission in a single time frame.

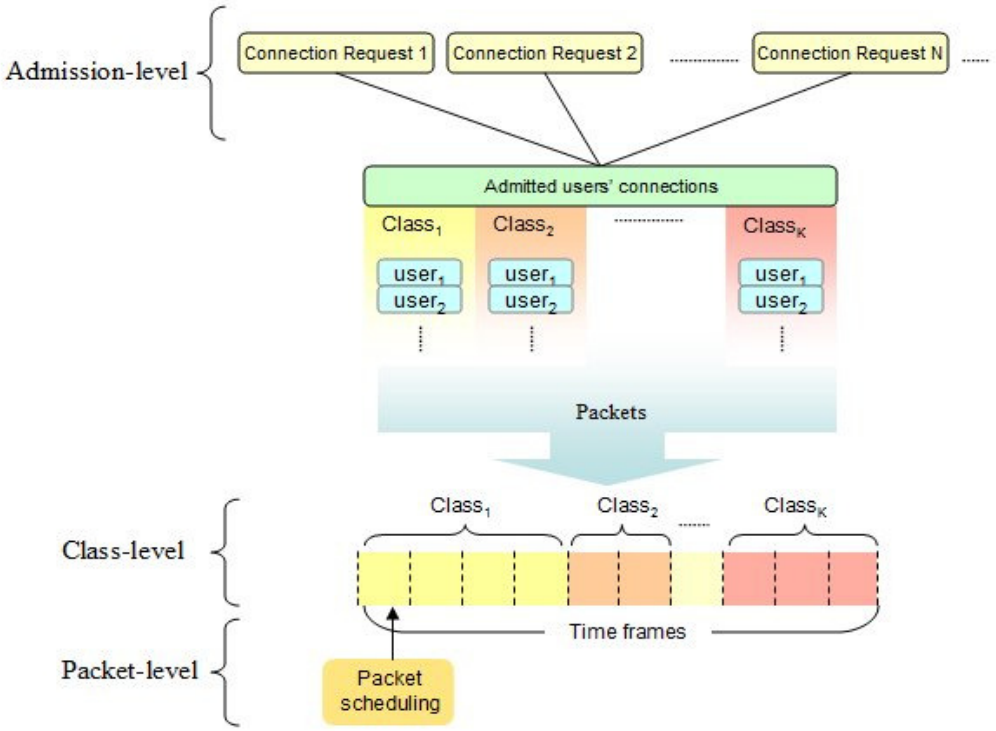


Figure 1.1: Levels of bandwidth management

In this thesis, we address the problem of bandwidth management in BWASs and propose effective economic-based solutions to enhance the QoS support in these systems. Our solutions are designed to balance between the conflicting requirements of users (e.g., guaranteed QoS) and network operators (e.g., high revenues). Particularly, we propose a bandwidth management framework, which consists of three related components corresponding to the three aforementioned levels of bandwidth management. The components are a packet scheduling scheme, a bandwidth provisioning scheme and a CAC-based dynamic pricing scheme.

1.1 Motivations and Objectives

As mentioned above, bandwidth management is crucial for the success of BWASs. Without efficient bandwidth management, network operators may find themselves incapable of meeting the escalating demand of users for multimedia services, and hence they may suffer immense revenue loss. Bandwidth management in BWASs is, however, a challenging problem due to many issues that need to be carefully taken into consideration. One of the major issues is the consideration of the mobile users' channel quality conditions. Mobile users experience varying channel conditions that affect their supportable data rates (i.e., maximum attainable data rates) from time to time due to their mobility, interference from other users, obstacles, etc, [5]. Ideally, bandwidth management schemes should exploit the information about the instantaneous channel quality conditions of the users by allocating bandwidth to those who are experiencing good channel quality conditions in order to maximize the system throughput. However,

favoring the users based on their channel quality conditions raises the issue of fairness as the users with bad channel quality conditions will not get served, and may consequently suffer from starvation. Therefore, fairness is another important issue that has to be taken into consideration while designing bandwidth management schemes. Achieving fairness, however, is not a trivial task because one needs to consider intra-class fairness (i.e., fairness between users within the same class) as well as inter-class fairness (fairness between different classes of traffic) at the same time. Another issue is the QoS requirements of different users. Since BWASs are envisaged to support a wide range of multimedia services with diverse QoS requirements, these requirements should be carefully taken into consideration to meet the user satisfactions.

Another important issue that must be taken into consideration is network congestion. Even though BWASs can support high data rates, it is expected that these systems will suffer from congestion due mainly to the wide support of bandwidth-intensive multimedia services. If a user transmits when the network is congested, the QoS of other users in the network such as packet delay and packet loss may become severely degraded. This phenomenon, which is known in economics as congestion *externality* [6], can indisputably result in user dissatisfaction, and hence potential revenue loss. Congestion in BWASs is typically dealt with by employing CAC to limit the number of admitted users in the system depending on the amount of available resources. CAC by itself, however, cannot guarantee a congestion-free system because it does not provide incentives to the users to regulate their usage of the wireless resources. Therefore, other strategies besides CAC must be employed to avert congestion.

The last issue of bandwidth management in BWASs is revenue loss due to serving low-revenue-generating users. Network operators experience different revenue losses from serving users depending on their channel quality conditions, the amount of the buffered data they have at the base station and the amount of money they are willing to pay for different services. A good design of a bandwidth management scheme should consider such revenue losses and aim at minimizing them.

Existing bandwidth management schemes deal with only some of these issues, and hence cannot optimize the performance of BWASs nor can they maximize the satisfaction of users. We, therefore, aim at considering all of the aforementioned issues in designing our framework. Since some of these issues conflict with one another (e.g. exploiting the variations of the channel quality conditions of users while achieving fairness), striking a proper balance between them is, therefore, a main focus of this work.

1.2 Thesis Contributions

BWASs such as HSDPA and WiMAX are envisaged to support a higher number of mobile users and variety of bandwidth-intensive "content-rich" wireless multimedia services. In this thesis, we propose a bandwidth management framework for BWASs consisting of three novel approaches to simultaneously achieve the following objectives:

1) Supporting different classes of traffic with users having different QoS requirements and bandwidth demands;
2) Maximizing the throughput of the wireless system;

3) Ensuring a fair distribution of wireless resources by supporting inter- and intra-class fairness;
4) Maximizing the network operator's revenues by limiting the revenue loss resulting from serving low-revenue generating users; and
5) Providing monetary incentives to the users to use the wireless resources efficiently and rationally in order to the prevent network congestion.

The thesis focus is on downlink (i.e., from the base station to mobile users) bandwidth management in BWASs. The main contributions of this thesis include the following:

1) Packet Scheduling Scheme

In Chapter 3, a novel packet scheduling scheme for BWASs is proposed in order to provide efficient bandwidth management at the packet level. The scheduling scheme optimally determines which of the users' packets are transmitted in any given time frame depending on many factors including the channel quality conditions of the users, their QoS requirements, fairness, the revenue earned from serving them and their priorities. We show that our packet scheduling scheme fulfills its design objectives of maximizing the social welfare of the system, supporting different types of traffic with different bandwidth and QoS requirements, improving inter- and intra-class fairness as well as increasing the network operator's revenues.

2) Bandwidth Provisioning Scheme

In Chapter 4, we propose a bandwidth provisioning scheme in order to achieve bandwidth management at the class level. To the best of our knowledge, our scheme is the first multiple frame bandwidth provisioning in BWASs to allocate an optimized number of time frames for each class based on the bandwidth requirements of its admitted users. After the optimal number of time frames for each class is determined, our proposed packet scheduling scheme can be used to distribute the frames among users of each class. To maximize inter-class fairness, a dynamic weight update scheme is proposed to dynamically adjust the class weights according to the classes' performance history. A distinctive feature of the weight update scheme is that it allows the weights of lower priority classes to be temporarily higher than those of higher priority classes while still maintaining service differentiations between them according to the requirements of network operators. We show that the overall system performance is improved when our bandwidth provisioning and packet scheduling schemes are employed together. In addition, we show that inter-class fairness can be better achieved using our proposed weight update scheme.

3) Call Admission Control-based Dynamic Pricing Scheme

In Chapter 5, we propose a Call Admission Control (CAC) scheme in order to provide effective bandwidth management for BWASs at the admission level. The proposed CAC scheme is designed to support users having different bandwidth requirements and belonging to different classes of traffic. To optimize bandwidth management at the

admission level, we integrate our proposed CAC scheme with a dynamic pricing component. Dynamic pricing is utilized to provide monetary incentives to the users to regulate their usage of the wireless system's resources to achieve the best possible system performance. Specifically, the pricing component dynamically computes the prices of wireless services depending on the load of the system using an ex-ante analysis of user demand behaviors towards price changes. A distinctive feature of our CAC-based dynamic pricing scheme is that the CAC and pricing functions are executed independently, hence simplifying their implementation and providing network operators the flexibility to use different CAC and user demand functions without affecting the computation of prices. We demonstrate that our CAC-based dynamic pricing scheme can guarantee a congestion-free system if the utilized user demand model is accurate in predicting their reaction to price changes. We also demonstrate that our proposed scheme can significantly increase the utilization of the wireless network, hence increasing the revenues of network operators.

1.3 Thesis Organization

This thesis is organized as follows. Chapter 2 presents some background material and previous work that are necessary for understanding the discussions to follow. Chapter 3 introduces our packet scheduling scheme that is designed to provide efficient bandwidth management at the packet level. Chapter 4 presents our bandwidth provisioning scheme, which aims at managing the bandwidth at the class level. Our dynamic weight update scheme is also presented in this chapter. Chapter 5 discusses our CAC-based dynamic

pricing scheme, which aims at preventing congestion as well as maximizing the system utilization. Chapter 6 presents the conclusions drawn from the thesis and discusses possible future research directions.

Chapter 2

Background and Related Work

This chapter provides the background material to help the reader follow the remainder of this thesis. Section 2.1 presents an overview of two of the most well-known BWASs, High Speed Downlink Packet Access (HSDPA) and Worldwide Interoperability for Microwave Access (WiMAX). Common features of HSDPA and WiMAX, which are utilized in this thesis, are discussed in this section. Sections 2.2, 2.3 and 2.4, respectively, provide detailed descriptions of packet-level, class-level and admission-level bandwidth management. A comprehensive literature review and comparisons are also provided in these sections.

2.1 Broadband Wireless Access Systems

High Speed Downlink Packet Access (HSDPA), labeled as 3.5G BWAS, has been introduced as an extension of UMTS [2], to optimize its support for data services. UMTS already offers fast data services, such as high-quality video transmissions at 384 Kbps. HSDPA, however, brings further enhancements to the provisioning of packet-data services, both in terms of system and end-user performance. This is because HSDPA is designed to achieve higher performance with a peak downlink data rate that is about 14.4 Mbps. As a result, network operators can offer their customers even more sophisticated multimedia services while on the move. HSDPA is particularly suited to extremely asymmetrical data services, which require significantly higher data rates for the downlink transmission than they do for the uplink.

The Worldwide Interoperability for Microwave Access (WiMAX) is another BWAS that has been standardized by the IEEE 802.16 group. WiMAX comes in two versions, fixed WiMAX based on the 802.16-2004 standard [3] and mobile WiMAX based on the 802.16e amendment [4] to the 802.16 standard. Both versions of WiMAX can theoretically support up to 70 Mbps and its base station can reach up to 50 km, hence enabling high-speed wireless access over large metropolitan areas. WiMAX promises compelling economics and a simplified IP-based architecture that reduces complexity and cost. The robust QoS support in WiMAX will enable it to efficiently handle real-time multimedia services such as video, high quality online gaming and streaming music in addition to providing "last mile" broadband connections, hotspots and high-speed connectivity for business customers.

HSDPA and WiMAX are two different technologies that differ in a number of aspects including air interface, coverage, system architecture, etc. These two technologies, however, rely on many similar features that allow them to achieve the high data rates they support. The two most important common features are utilizing adaptive modulation[4] and Hybrid Automatic Repeat Request (HARQ), which are tightly coupled and rely on rapid adaptation of the transmission parameters to the instantaneous radio conditions. Adaptive modulation techniques enable the use of spectrally efficient higher order modulation when channel conditions permit, and revert to more robust lower order modulation for less favorable channel conditions. This implies that users with good channel conditions will potentially enjoy higher supportable data rates by using higher order modulation, whereas users with bad channel conditions will experience lower data rates. HARQ rapidly requests the retransmission of missing data entities and combines the soft information from the original transmission and any subsequent retransmissions before any attempts are made to decode a message. The main advantage of HARQ is reducing the number of data retransmissions in BWASs, hence improving the delay latency of these systems.

Another key common feature between HSDPA and WiMAX is the use of a shared channel for data transmissions. The rationale behind using a shared channel for data transmissions instead of dedicated ones is twofold. First, it improves the utilization of the wireless resources of BWASs, and hence it enables these systems to accommodate more user connections. Second, it reduces the cost of per-bit transmission, hence lowering the

[4] Modulation refers to superimposing the information bits on the carrier frequency.

cost of providing wireless services. Although HSDPA and WiMAX use different channel structures, their shared channels are divided into transmission time frames consisting of a number of slots, where each slot is of some fixed size. In HSDPA, the frame size is fixed at 2 ms whereas in WiMAX it can be variable and can range from 2.5 ms to 20 ms. Data transmission in these systems is done at their base stations at the beginning of each time frame.

Using shared channels for data transmission, however, complicates the task of resource sharing and bandwidth management. More intelligent and sophisticated bandwidth management schemes are needed to distribute the wireless resources among mobile users who have diverse QoS requirements and different channel quality conditions. To maximize the efficiency of BWASs and improve user satisfactions, bandwidth management is done at three levels, namely packet-level, class-level and admission-level. Due to the importance of each of these levels, they are discussed in depth in the following sections. It is imperative to point out that bandwidth management in the downlink direction can be different from that of the uplink due to different channel characteristics, amount of assigned resources, etc. The thesis focus is on the downlink, and therefore, our discussion hereafter covers only bandwidth management in downlink communication.

2.2 Packet-Level Bandwidth Management

A key component of BWASs is packet-level bandwidth management, which is realized through the use of packet scheduling. Packet scheduling will play an increasingly

prominent role in BWASs, since these systems are characterized by using downlink shared channels to support the increasing number of mobile data users. A centralized downlink packet scheduler is implemented at the base stations of these systems to provide fast scheduling decisions for controlling the allocation of the downlink shared channels to the mobile users by deciding which of their packets should be transmitted during a given time interval. Thus, to a large extent, the scheduler determines the overall behavior of these systems. One important factor that has been added to the scheduling problem in BWASs is the channel quality conditions of the mobile users. Mobile users experience varying channel conditions due to mobility, interference caused by other users in the system, distance from the base station, etc. The packet scheduler in BWASs should track the instantaneous channel conditions of the users and select for transmission those users who are experiencing good channel conditions to maximize the system throughput [5]. However, exploiting user channel quality conditions in the scheduling decisions raises the issue of fairness, as those users with bad channel conditions may not get served, and thus they may suffer from starvation. Therefore, packet schedulers should be carefully designed to maximize the efficiency of BWASs, while ensuring fairness among mobile users.

Packet scheduling in BWASs works as follows. Each user regularly informs the base station of his channel quality condition by sending a report in the uplink direction to the base station. The report contains information about the instantaneous channel quality condition of the user. The base station, in turn, uses this information to select the appropriate user(s) according to the adopted scheduling scheme as shown in Figure 2.1.

For example, in HSDPA, users are able to measure their current channel quality conditions by measuring the power of the received signal from the base station and then, using a set of models described in [7], to determine their current supportable data rates.

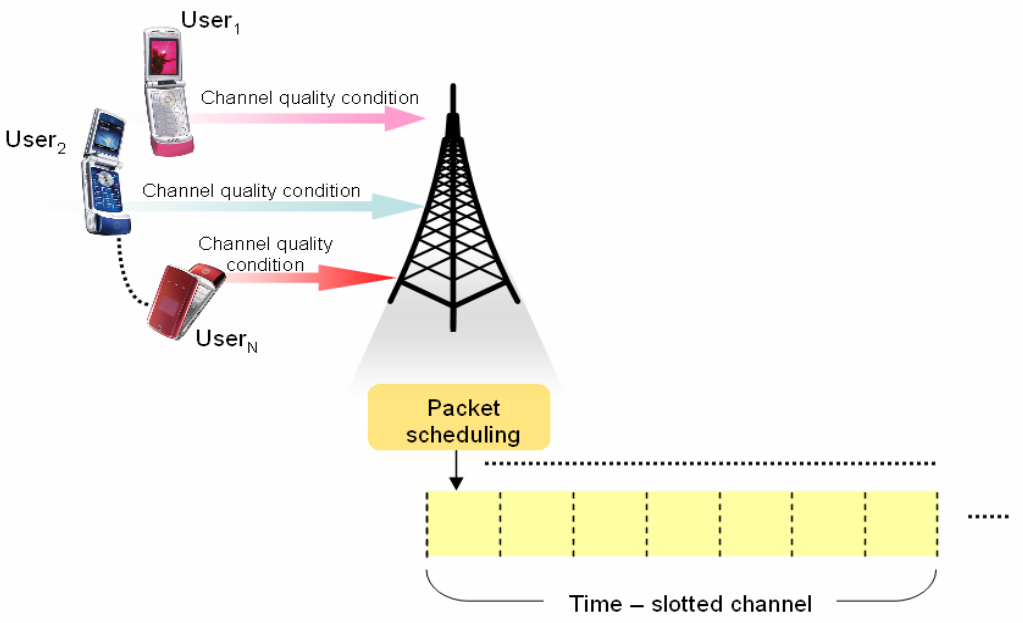


Figure 2.1: Packet scheduling

Several packet scheduling schemes have been proposed for BWASs. An overview of QoS provisioning techniques including state-of-the art packet scheduling in Code Division Multiple Access (CDMA) networks is presented in [8]. In our recent study [5], we surveyed HSDPA scheduling schemes, which we classified into two groups: real-time and non-real-time scheduling schemes. Non-real-time scheduling schemes are designed for non-real-time and best-effort traffic, where the user's average throughput is the main QoS metric. Real-time scheduling schemes are designed for multimedia traffic with QoS requirements such as minimum data rate or maximum delay requirements. The most

well-known non-real-time packet scheduling schemes in BWASs are the Maximum Carrier to Interface Ratio (Max CIR) [9] and Proportional Fairness (PF) [10]. Max CIR serves the users with the best channel quality conditions. Hence, maximizing the throughput of the wireless network at the expense of fairness. PF tries to balance the throughput-fairness trade-off by serving the users with the best relative channel quality condition, where the relative channel quality condition is the user's channel quality condition divided by his average throughput. Therefore, the PF scheme gives more priority to users as their average throughputs decrease in order to prevent users with good channel quality conditions from monopolizing the wireless resources as is the case with Max CIR.

It has been shown, however, that the PF scheme is fair only in ideal cases, where users experience similar channel conditions. The PF scheme, therefore, becomes unfair and unable to exploit multi-user diversity[5] in more realistic situations, where users usually experience different channel conditions [11] and [12]. To solve this problem, a Score-Based (SB) scheduling scheme is proposed in [12]. Unlike the PF scheme, the SB scheme selects the user whose current channel quality condition is high relative to his own rate statistics instead of selecting the one whose channel quality condition is high relative to his average throughput. Another proposal is Fast Fair Throughput (FFT) [13]. FFT modifies the PF scheme by multiplying the relative channel quality conditions of the users by an equalizer term to ensure a fair long-run throughput distribution among them.

[5] Exploiting the variations of the channel conditions of the users by serving those with more favorable channel conditions for the benefit of user and/or system capacity.

In [14] and [15] a packet scheduling scheme known as the Modified Largest Weighted Delay First (M-LWDF) is proposed to accommodate real-time traffic. M-LWDF uses the relative channel quality condition to compute the user's priority in the same manner as PF. To accommodate real-time traffic with delay requirements, M-LWDF multiplies the user's relative channel quality condition by a term representing the user's packet delay. This term ranges from 0 to 1, where it approaches 1 as the user's head of queue packet delay approaches its delay threshold. It is shown in [16] that M-LWDF may result in unfair distribution of wireless resources since if two users have the same head of queue packet delay, they will be assigned different priorities if their supportable data rates are different. Therefore, an enhancement of M-LWDF, referred as the Fair Modified Largest Weighted Delay First (FM-LWDF) is proposed in [16] to improve the fairness of M-LWDF. FM-LWDF borrows the equalizer term from the FFT scheme and adds it to M-LWDF in order to improve fairness among users.

In [17], another packet scheduling scheme known as the Max CIR with Early Delay Notification (EDN) is proposed. EDN tries to maximize the system throughput by scheduling the users using the Max CIR scheme as long as their packets' delays are below a certain threshold. If the packets delays of one or more users exceed a certain threshold, then the packets that have been queued the longest time are served first.

Another proposal for a packet scheduling scheme is proposed in [18]. The scheme represents the satisfaction of each user by a utility function and aims at maximizing the users' utilities. Two utility functions are proposed, one for delay-constrained traffic based on its delay and the other for best-effort traffic based on its average throughput. The

scheme, however, ignores users with data rate requirements. In addition, even though the scheme supports fairness among best-effort users, it ignores fairness among delay-sensitive users. Moreover, the scheme does not provide inter- and intra- class prioritization, which may limit its practicality.

Like the scheme in [18], the scheme in [19] uses different utility functions depending on the data rate requirements of users (e.g., stringent, flexible, etc). The scheme, however, ignores delay-sensitive users. In addition, the scheme does not take into account the instantaneous channel quality conditions of mobile users in the scheduling decisions, which is one of the most important features of packet scheduling in BWASs.

Therefore, packet-level bandwidth management in BWASs is still an open issue because of the need for a packet scheduling scheme that is capable of simultaneously supporting various QoS requirements in addition to providing effective inter- and intra-class prioritization and fairness. We further remark that none of the schemes discussed in this section considers the revenues of network operators, which may limit their viability.

2.3 Class-Level Bandwidth Management

As aforementioned, packet scheduling will play an imperative role in BWASs because of its key functionality in controlling the distribution of their shared wireless channels among users. Packet scheduling by itself, however, cannot achieve optimized bandwidth management. This is because it only considers the current time frame to make its decision. To augment the scheduling performance and maintain acceptable levels of QoS

throughout the lifetime of user connections, packet scheduling must be coupled with a longer term class-level bandwidth management scheme to span multiple time frames and decide how they are shared among the different classes of traffic, and hence their corresponding users. Class-level bandwidth management, which we refer to as "bandwidth provisioning", can be thought as a longer-term post admission bandwidth management that aims at satisfying the long-term bandwidth requirements of users for the lifetimes of their connections, as opposed to packet scheduling, which only allocates bandwidth over single time frames. Bandwidth provisioning works as follows. It first gathers the bandwidth requirements of each class based on the bandwidth requirements of its admitted users. It then determines how many frames are needed to satisfy each class. Once the number of time frames is determined for each class, packet scheduling can then be used to distribute them among the class's users on a frame by frame basis as shown in Figure 2.2[6].

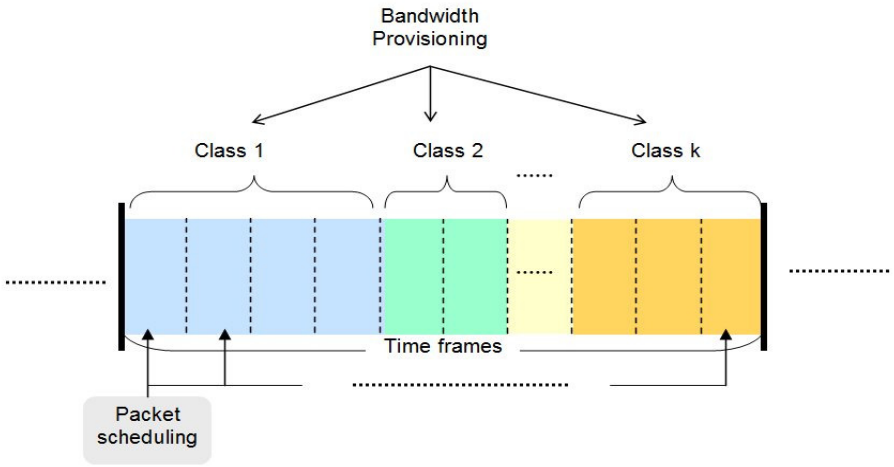


Figure 2.2: Class-level bandwidth management

[6] Note that the frames allocated per class need not be consecutive and they are only depicted this way for illustration purposes.

Most of the work on bandwidth provisioning has been done at the admission level [20], [21], [22], [23] and [24]. These schemes implement the CAC function and aim at maximizing the number of admitted users while satisfying the bandwidth requirements of different classes of traffic. Bandwidth management at admission level is very important in improving the performance of BWASs as discussed in the next section. There is a need, however, for bandwidth provisioning at the frame level (i.e., during the lifetimes of user connections). This is due to the varying bandwidth requirements of mobile users during the lifetimes of their connections as a result of their traffic burstiness and also due to their varying channel quality conditions, which affect the capacity of the base station, and hence the amount of bandwidth that it can sustain to each one of them. Little research work, has considered the problem of bandwidth provisioning across the user connections [25], [26], [27] and [28]. The scheme in [25] aims at minimizing the expected number of packets awaiting transmission for each user in order to reduce the overall system delay. It supports prioritization between users belonging to different classes of traffic. However, it does not support users with different bandwidth requirements. Therefore, users with higher number of packets in their corresponding queues can get more bandwidth regardless of the bandwidth required by other users in the system. In addition, to increase the efficiency of the system, the scheme assigns more priorities to users having higher "probability of connectivity" between them and the base station, where the probability of connectivity is used as a measure of the channel quality conditions of users. This measure, however, does not reflect the actual instantaneous data rates that the users can send or

receive at, which depend on their instantaneous channel quality conditions. Using this measure, the scheme may consequently assign more/less bandwidth than what is actually needed by users.

The scheme in [26] divides the number of slots in each time frame between different classes of traffic so that the frame-level connection blocking probability of each class (i.e., the probability that connections within each class are blocked and not assigned time slots in the current frame) is minimized. Unlike the scheme in [25], the scheme in [26] considers the instantaneous channel quality conditions of users as well as their minimum bandwidth requirements in the slot allocation process.

The bandwidth provisioning schemes in [27] and [28] do not consider the varying channel quality conditions of mobile users. Hence, they cannot achieve optimized bandwidth provisioning. In addition, these schemes provide very limited QoS support, and hence they are incapable of supporting many multimedia services in BWASs.

We remark that the schemes in [25], [26], [27] and [28] are designed to allocate slots within one time frame. However, as mentioned previously, to maintain the QoS of ongoing users at acceptable levels throughout the lifetime of their connections, there is a need for bandwidth provisioning over multiple time frames. In addition, these schemes lack support for fairness between different classes. Hence, they may result in unfair allocation of bandwidth, where users with good channel quality conditions and/or high bandwidth requirements may monopolize the whole bandwidth. Furthermore, none of these schemes considers the revenues of network operators when allocating the time slots. As a result, these schemes may not be desired by network operators, who are certainly

concerned about maximizing their revenues. Therefore, there is a need for a bandwidth provisioning scheme that is able to allocate multiple time frames and provide fairness between classes of traffic, while considering the revenues of network operators.

2.4 Admission-Level Bandwidth Management

Packet-level and class-level bandwidth management solutions improve the performance of BWASs but they cannot guarantee QoS to mobile users especially during congestion periods when the demand for bandwidth exceeds the system capacity. This necessitates the need for admission-level bandwidth management. Network operators typically employ CAC in order to manage the bandwidth of their wireless systems at the admission-level (i.e., prior to admission). By limiting the number of admitted user connections in the system, CAC can guarantee that the packet-level QoS (e.g., packet delay, average throughput, etc) of ongoing connections will not be adversely affected as a result of new incoming ones. There are two types of connections at admission level, new and handoff connections. A new connection occurs when a user initiates a new connection request, while a handoff connection occurs when an active user moves from one cell to another. Besides maintaining the packet-level QoS of ongoing connections at acceptable levels, CAC aims at enhancing admission-level QoS. The main QoS metrics at admission level are the new connection blocking and handoff connection dropping probabilities. The new connection blocking probability is the probability that a new connection is rejected and the handoff connection dropping probability is the probability that a handoff connection is dropped.

CAC has been extensively studied in the literature [20], [21], [22], [23], [24], [29], [30], [31], [32], [33], [34], [35], [36] and [37]. According to [38], existing CAC schemes can be classified as being either measurement-based, interactive, non-interactive, distributed, non-distributed, predictive and/or non-predictive. Measurement-based CAC schemes make their admission decisions based on measurement of actual current network traffic load [20], [21], [22], [23], [24], [29], [30], [31], [32], [33], [34], [35], [36] and [37]. Such measurements include the interference caused by the users in the network and the base station power. Non-interactive schemes instantaneously make their decisions on whether or not to accept a connection request to the system based on previously measured interference values or received power values. Interactive schemes allow users to interact with the system before making any admission decisions and monitor/predict their affect on the network if their connections are accepted [33]. Such interaction allows the system to gradually increase the power of new users until they are admitted instead of blocking them when there is insufficient power to support their connection requests at the time they are made. Distributed schemes [29], [30], [35], [36] and [37] consider status information for other base stations than the one that the connection requests are made to as opposed to non-distributed schemes [20], [21], [22], [23], [24], [31], [32], [33] and [34], which only interact with single base stations. Predictive schemes make predictions on future traffic conditions, which are then used to base their decisions on whether to accept new user connection requests or not [29], [30], [35], [36] and [37].

Existing CAC schemes have been shown to be very efficient in improving the packet-level QoS of ongoing connections amid congestion periods. However, they are not

as efficient in improving the admission-level QoS. This is because these schemes by themselves cannot avoid congestion because they do not provide incentives to users to share wireless system resources rationally and efficiently. Therefore, the connection blocking and dropping probabilities can reach high levels during congested periods. Recently, there has been some research on integrating admission-level dynamic pricing with CAC in order to control connection request arrivals to the system through monetary incentives, hence maintaining the admission-level QoS at the desired thresholds. Because of the relevance of this area of research to the work in this thesis, it is discussed in detail in the following subsection.

2.4.1 Admission-Level Dynamic Pricing

In admission-level dynamic pricing, the price for a unit of time or bandwidth is determined when the user initiates a connection request and before he is admitted to the system. The price in this case is fixed for the connection duration. This price is dynamically determined according to the network load based on an ex-ante analysis of user demand behaviors towards price changes. Dynamic pricing solutions in general assume that users are price-sensitive, which is normally the case with most users.

Dynamic pricing can competently promote rational and efficient usage of the shared wireless resources by influencing user behaviors. Dynamic pricing is, therefore, a promising solution to traffic control problems, which can help alleviate the problem of congestion and provide efficient bandwidth management. In addition, dynamic pricing can enhance economic efficiency, since it ensures that the wireless resources are given to

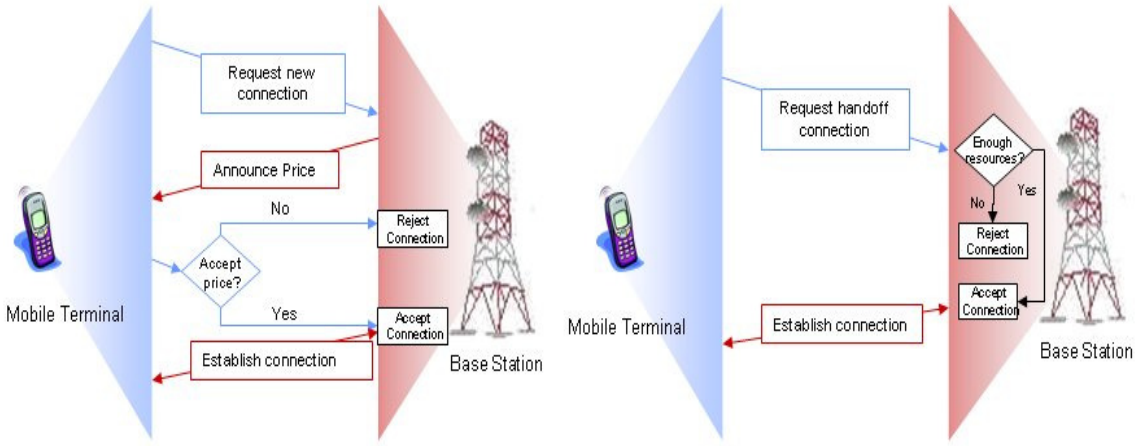
those who value them the most. Furthermore, dynamic pricing is cost-effective and can generate higher revenues. It should be noted that, since users are charged at admission level, handoff connections are not affected by dynamic prices since they were charged at the cell where the connections were first initiated. In general, the design of any dynamic pricing scheme depends primarily on two fundamental components:

1) User behavior: any dynamic pricing scheme must take into account the demand behaviors of users. Different users react differently to prices because some of them are more sensitive to prices than others. This is known in economics as the *price elasticity of demand*, which measures the responsiveness of a change in demand for a good or service to a change in price [39]. Different pricing schemes use different demand models. For example, some use exponential functions to represent the user demand for wireless services, whereas others use utility functions to represent the users' preferences and/or their *Willingness to Pay*[7] (WTP) for a certain service.
2) Price function: the price function determines how the price of a certain service is computed for a unit of time, bandwidth and/or power. Different schemes use different price functions depending on many factors including objectives of dynamic pricing, characterizations of resource usage, causes of congestion, assumptions about user behaviors, etc.

The general procedure for admission-level dynamic pricing is as follows. When a user makes a new connection request, the base station or any other centralized entity in

[7] The monetary value users are willing to pay for a certain service.

BWASs, computes the price for a unit of time or bandwidth according to the CAC scheme and announces this price to the user as shown in Figure 2.3 (a). If the user accepts the price, he can then establish the connection. Otherwise, he can retry later when the price is affordable. If the request is a handoff connection, then the base station only checks if there are enough resources for such a request; and consequently makes the decision to accept the connection or reject it based on this information without computing a new price as shown in Figure 2.3 (b). It should be noted that the prices need not be announced to users after they make connection requests. For example, the prices can be broadcasted periodically to users whether they make connection requests or not.



(a) (b)

Figure 2.3: Admission-level dynamic pricing procedure for (a) new connection and (b) handoff connection

Several CAC schemes with dynamic pricing have been proposed in the literature [40], [41] and [42]. The scheme in [40] dynamically computes the optimal price so that the price-affected connection arrival rates maximize the social welfare of the system (i.e., the summation of users' utilities).The user utility is assumed to be a function of the connection blocking probabilities, which are, in turn, a function of the arrival rates. However, the scheme is designed only to avert network congestion, where a flat rate pricing is assumed when the network is underutilized. Therefore, users are not given any incentives to increase their usage of the network when it is underutilized, which results in resource wastage, and hence potential revenue loss. In addition, the scheme lacks support for QoS, since it assumes that all connections require the same amount of resources. This makes it unsuitable for BWASs.

In [41], a CAC-based dynamic pricing scheme is proposed. In this scheme users are divided into two types, *priority users* and *conventional users*. When the network is underutilized, all users in this scheme are considered conventional users and are placed in the conventional queue awaiting admission, where they are charged a flat rate. During congested periods, a dynamic price is computed and the users are given the option to choose between being priority users, where they are charged a higher dynamic price and are placed in the priority queue to be served faster; or being conventional users, where they are charged a flat rate and are served more slowly. The dynamic price is determined so that the maximum number of users that the network can accommodate, and yet conform to the delay the users can spend in the admission queue is achieved.

However, even though the scheme considers the delay users experience in admission queues, it does not take into account the new connection blocking and handoff connection dropping probabilities. This may not be practical since wireless network operators have a limit on the number of connections they can block, which is usually determined by regulations. In addition, like the scheme in [40], this scheme assumes that connections require the same amount of resources, rendering it unsuitable for BWASs. Moreover, the scheme is only designed to prevent congestion. Therefore, it does not provide incentives to users to increase their demand for the network services when the network is underutilized.

The CAC-based dynamic pricing scheme in [42] aims at reducing congestion and maximizing revenues in wireless cellular networks. The scheme considers the effects of prices on connection arrivals, retrials (i.e., requesting the same service again after being blocked) and substitutions among services (i.e., substituting a service for another after being blocked). Using some assumptions about the new and handoff connection arrival rates, the scheme dynamically determines the prices of network services so as to encourage or discourage the arrival rates to the system to reserve some bandwidth for arriving handoff or higher-revenue-generating users.

Even though the scheme considers different classes of traffic, it assumes that users within each class request the same amount of bandwidth. This is still impractical in BWASs since, in these systems, each class can include various services each requesting different amount of bandwidth (e.g., audio streaming and video streaming in the

streaming class). In addition, the scheme is complex and requires many calculations to determine prices.

Note that the schemes in [40], [41] and [42] are based on assumptions about user demand models and cannot, therefore, be generalized to work with different demand models without affecting the way prices are computed. This limits their scalability, since different network operators might have different demand models depending on their subscribers. The schemes in [43], [44], [45], [46] and [47] apply dynamic pricing at admission level without using CAC. These schemes, therefore, cannot achieve optimized admission-level bandwidth management. In addition, the schemes in [48], [49], [50], [51], [52], and [53] apply dynamic pricing during the user connection (i.e., after being admitted). The users in these schemes are charged according to the amount of power they consume over the lifetimes of their connections. These schemes aim at mitigating interference in interference-limited[8] networks such as CDMA networks. However, these schemes may not be accepted by users because it is difficult to anticipate the total charge[9] for each connection, since dynamic prices vary amid the user's connection. More elaborate discussions about these schemes can be found in our recent survey paper on dynamic pricing in wireless networks [54]. A general overview on the role of pricing including dynamic pricing in radio resource planning and management is discussed in [55].

---

[8] In interference-limited networks, amount of power transmitted by each user causes interference to other users in the network.

[9] Charge is the amount that is billed for a service, whereas price is the amount of money associated with a unit of service. That is, price is used to compute the charge [6].

Therefore, there is a need for a CAC-based dynamic pricing scheme that is able to support different classes of traffic with different users having different bandwidth requirements, work with various demand models and compute the dynamic prices in a simple way.

2.5 Summary

This chapter presented existing research efforts in packet-level, class-level and admission-level bandwidth management in BWASs. Current state-of-the-art bandwidth management schemes at each one of these levels were surveyed. Advantages and limitations of the surveyed schemes were outlined. Table 2.1 provides comparisons between the schemes discussed in this chapter, which are most relevant to our work, using the following comparison criterion:

- Bandwidth management level: the level at which bandwidth management is performed.
- Channel quality consideration: whether the scheme considers the channel quality conditions of users or not.
- Supported QoS: the type of QoS the scheme supports. For example, some schemes support different classes of traffic, whereas others provide very limited QoS support.
- Fairness support: whether the scheme supports some form of fairness between users and/or classes or not.

- Revenue consideration: whether the scheme considers the revenues of network operators or not.

As observed from Table 2.1 and despite extensive research efforts, comprehensive bandwidth management in BWASs is still an open research problem. This is because existing bandwidth management schemes deal only partially with the issues of QoS support, considerations of users' channel quality conditions, fairness support and revenue considerations. Hence, such schemes cannot optimize the performance of BWASs nor can they maximize the satisfactions of users.

In this thesis, we aim at considering all the aforementioned issues as well as providing network operators the flexibility to determine the appropriate trade-offs between conflicting issues (e.g., maximizing throughput vs. achieving fairness). This is achieved by designing a bandwidth management framework consisting of different components that operate at different bandwidth management levels. Such a framework is of practical importance to network operators due to the expected increase of demand for multimedia services in BWASs and its consequences in terms of the need for better bandwidth management to ensure user satisfaction and increased revenues. Specifically, the framework consists of three components, a packet scheduling scheme, a bandwidth management scheme and a CAC-based dynamic pricing scheme.

Table 2.1: Comparison between different bandwidth management schemes

Criteria / Reference	Management Level	Channel consideration	QoS Support	Fairness Support	Revenue consideration
[9]	Packet-level	Yes	No	No	No
[10]	Packet-level	Yes	No	Yes	No
[11]	Packet-level	Yes	No	Yes	No
[13]	Packet-level	Yes	No	Yes	No
[14], [15]	Packet-level	Yes	Packet delay only	Yes	No
[16]	Packet-level	Yes	Packet delay only	Yes	No
[17]	Packet-level	Yes	Packet delay only	No	No
[18]	Packet-level	Yes	Packet delay only	Partial	No
[19]	Packet-level	No	Data Rate only	Yes	No
[25]	Class-level	Partial	Inter-class prioritization only	No	No
[26]	Class-level	Yes	Different classes of traffic with different bandwidth requirements	No	No
[27]	Class-level	No	No	No	No
[28]	Class-level	No	No	No	No
[40]	Admission-level	No	No	No	No
[41]	Admission-level	No	No	No	No
[42]	Admission-level	No	Different classes of traffic / same requirements within each class	No	Yes

Chapter 3

Optimal Packet Scheduling Scheme

Packet scheduling will have a great impact on the performance of BWASs because of its decisive role in distributing the wireless resources of these systems among mobile users. A distinctive feature of packet scheduling in BWASs is the adoption of the users' channel quality conditions in the scheduling decisions in order to maximize the capacities of these systems. This, however, adds a new dimension to the scheduling problem and complicates the task of scheduling as it raises the issue of fairness. Therefore, a good design of a packet scheduling scheme should properly balance the throughput-fairness trade-off. In addition, packet scheduling should be competent in supporting different QoS as BWASs support different types of multimedia services, which in essence, have diverse QoS requirements. Finally, any packet scheduling scheme should consider the revenues of network operators in the scheduling decisions. This will indubitably ensure the viability of the scheme.

In this chapter, we introduce our packet scheduling scheme, which aims at providing efficient bandwidth management at the packet level. The scheme is based on practical economic concepts to maximize the satisfactions of users as well as network operators. We also introduce the concept of opportunity cost and show how it can be used to limit the revenue loss resulting from scheduling low revenue generating users.

The rest of this chapter is organized as follows. Section 3.1 outlines our proposed packet scheduling scheme and discusses its objectives. Section 3.2 describes the system model. Section 3.3 presents the general formulation of our proposed packet scheduling scheme, which includes utility and opportunity cost functions, followed by specific definitions for the utility function's parameters to support different types of traffic with different QoS requirements. Section 3.4 illustrates the effectiveness and strengths of our proposed packet scheduling scheme through a comprehensive performance evaluation. Section 3.5 summarizes the chapter.

3.1 Scheme Outline and Objectives

We propose a novel packet scheduling scheme for BWASs. The proposed scheme is to be implemented at the base stations of these systems, where packet scheduling is performed as discussed in Chapter 2. Our proposed scheme is designed to simultaneously achieve the following objectives:

1) Supporting multiple classes of traffic with users having different QoS and traffic demands;

2) Satisfying the conflicting requirements of the users and network operators (i.e., guaranteed QoS vs. revenues);
3) Maximizing the throughput of the wireless system; and
4) Ensuring a fair distribution of wireless resources (e.g., bandwidth).

Unlike most existing schemes, where different users within each class are assumed to have the same QoS requirements, we consider a more generalized problem, supporting multiple users with different QoS requirements within each class. This is more practical since each traffic class in BWASs can include various services with different QoS requirements (e.g., video and audio streaming in the streaming class). Another problem that is dealt with in our scheme is satisfying the conflicting requirements of the network operator (i.e., high revenues) and the users (i.e., guaranteed QoS). In practice, different users may have different preferences depending on many factors including the types of wireless services they request, age, budgets, etc. These preferences are accounted for in our scheme by employing a utility function with certain practical properties. To this end, we provide specific definitions for the utility function to support three different types of traffic, namely best-effort traffic, traffic with minimum data rate requirements and traffic with maximum packet delay requirements. In addition, we show that the two well-known scheduling schemes, Maximum Carrier to Interface Ratio (Max CIR) [9] and Proportional Fairness (PF) [10] are special cases of our proposed scheme. This gives the network operator more flexibility in choosing between different scheduling schemes. The preferences of the network operator are represented in our scheme by an opportunity cost

function to bound revenue loss resulting from serving low revenue generating users. To maximize the system throughput, the proposed packet scheduling scheme utilizes the information of the channel quality conditions of the users in its scheduling decisions. Furthermore, we provide unique fairness parameters for the traffic cases that are considered in this chapter to ensure a fair distribution of the wireless resources (e.g., bandwidth).

3.2 System Model

We consider a BWAS comprising a downlink time-slotted shared channel. Data transmission is done in time frames of fixed or variable size duration, where each frame consists of a number of fixed size time slots. We consider that the base station serves N user connections. We also consider that there are K classes of traffic, where class i has higher priority than class i+1. Let N_i denote the number of class i user connections, and $N = \sum_{i=1}^{K} N_i$. We allow users within the same class to have different QoS requirements depending on the types of services they request. Also, and without loss of generality, we assume that each user has one connection request. Thus, the base station maintains one queue for every user.

Upon call arrival, the BWAS receives traffic in the form of IP packets from higher layers, which are segmented into fixed size Protocol Data Units (PDUs). These PDUs are stored in the transmission queue of the corresponding connection. Subsequently, the

PDUs are transmitted to the appropriate connection(s) according to the adopted scheduling scheme as shown in Figure 3.1.

As explained in Chapter 2, each user regularly informs the base station of his channel quality condition by sending a report in the uplink to the base station. The report contains information about the instantaneous channel quality condition of the user. The scheduling scheme would then use this information to select the appropriate connection(s) for transmission.

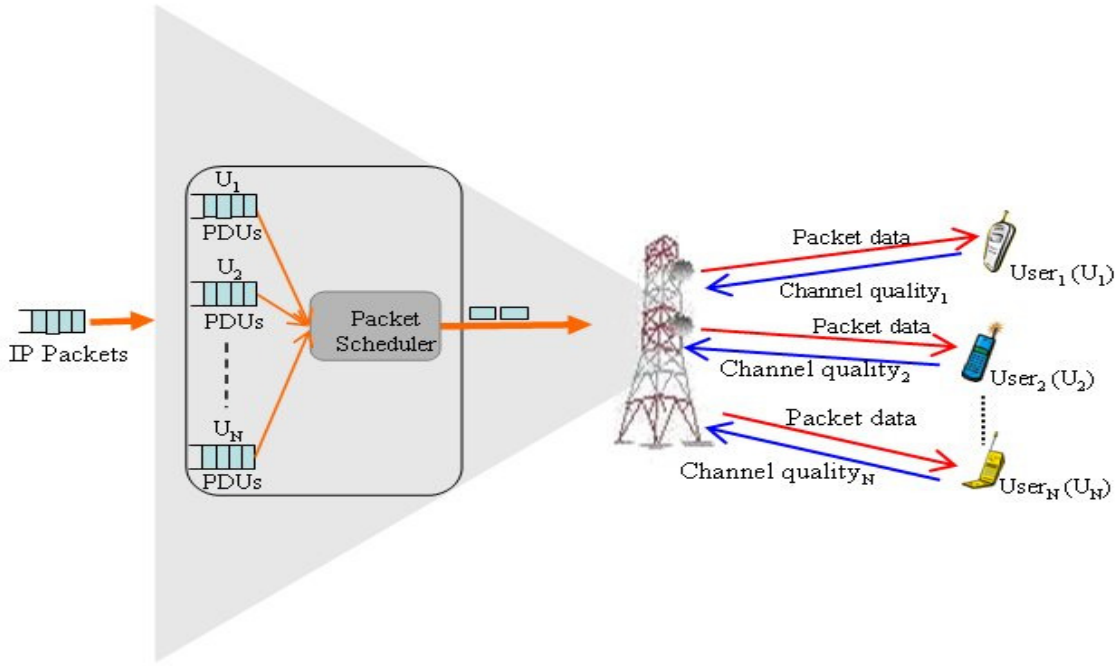


Figure 3.1: System model

3.3 Fair Class-based Packet Scheduling Scheme

In this section, we present our proposed packet scheduling scheme, which we refer to as Fair Class-Based Packet Scheduling (FCBPS). We first begin by outlining the general formulation of the scheduling problem, which includes a general utility function to represent the satisfactions of mobile users and an opportunity cost function to represent the cost of serving them (in terms of revenue loss). Next we state the conditions that the utility function should satisfy and we propose a plausible utility function that meets the stated conditions. After that, we provide specific definitions for the parameters of the proposed utility function to support three different types of traffic with different QoS requirements, namely best-effort traffic, traffic with minimum data rate requirements and traffic with maximum delay requirements.

The satisfaction of user *j* of class *i* at time *t* as perceived by the network operator can be expressed by a utility function of the form $U_{ij}\left(\{\mathbf{X}_{ij}^{z}(t)\}_{z=1}^{m_{ij}}\right)$, where $\{\mathbf{X}_{ij}^{z}(t)\}_{z=1}^{m_{ij}} = \{X_{ij}^{1}(t), X_{ij}^{2}(t), ..., X_{ij}^{m_{ij}}(t)\}$ [10], $X_{ij}^{1}(t), ..., X_{ij}^{m_{ij}-1}(t)$ are chosen QoS quantitative measures of the user's satisfactions with the wireless system such as the average throughput, current data rate, average delay, etc, $X_{ij}^{m_{ij}}(t)$ is a fairness measure that represents how fair the scheduling scheme is to the user, $z = 1, 2.., m_{ij}$ is an index that refers to any of the QoS measures and m_{ij} is the maximum number of chosen quantitative

[10] In this thesis, the notation $\{\ \}_{z=1}^{m_{ij}}$ is used to represent a set of elements indexed from 1 to m_{ij}, i.e., $z = 1, 2, ..., m_{ij}$.

measures for user *j* of class *i*. The main objective of our packet scheduling scheme is to find a subset of users ($\mathbf{N}^*$) to transmit their packets to in order to maximize social welfare, which is the summation of user utilities [39]. Thus, the scheduling scheme can be formulated as the following optimization problem

$$\begin{aligned}
&\text{Objective: } \max_{(i,j)\in\mathbf{N}^*,\mathbf{N}^*\subseteq\mathbf{N}} \sum_{i=1}^{K}\sum_{j=1}^{N_i} U_{ij}\left(\{\mathbf{X}_{ij}^{z}(t)\}_{z=1}^{m_{ij}}\right)\\
&\text{Subject to: } \nu_{ij}^{z,\min} \le X_{ij}^{z}(t) \le \nu_{ij}^{z,\max}, \forall j \in \mathbf{N},\ \forall z, 1 \le z \le m_{ij}\\
&\left(\sum_{(i,j)\in\mathbf{N}^*} R_{ij}(t)\right) \le C\\
&OC_{\mathbf{N}^*}(t) \le H
\end{aligned} \tag{3.1}$$

where $\mathbf{N}^* \subseteq \mathbf{N}$ is the set of users (represented by the tuple (i, j), where *i* the class index and *j* is the user's index within the class) that are selected to transmit to, **N** is the set of the total number of users in the system, the first constraint is used to ensure lower and upper bounds on QoS provided to users (e.g., minimum and maximum data rate), $\nu_{ij}^{z,\min} \in \{\mathbf{v}_{ij}^{z,\min}\}_{z=1}^{m_{ij}}$ and $\nu_{ij}^{z,\max} \in \{\mathbf{v}_{ij}^{z,\max}\}_{z=1}^{m_{ij}}$ are predefined values for the lower and upper bounds corresponding to the *z*th QoS measure for user *j* (i.e., $X_{ij}^{z}(t)$), respectively, $R_{ij}(t)$ is the current supportable data rate of user *j* at time *t*, which depends on his channel quality condition[11], *C* is the

[11] Note that $R_{ij}(t)$ is computed based on the channel quality condition of the user as explained in Section 2.2. However, if the user requires less than $R_{ij}(t)$ to empty his buffer, then we set $R_{ij}(t)$ to the data rate that is just enough to empty the user's buffer in order to avoid giving more slots than the user needs.

system capacity, $OC_{\mathbf{N}^*}(t)$ is a cost function representing the cost of serving the selected users at time t (i.e., the users in set $\mathbf{N}^*$) and H is a predefined value. We consider the opportunity cost[12] as our cost function. The concept of opportunity cost can be used to manage the trade-off between fairness and revenue. This is because fairness may force the scheduler to serve low-revenue-generating users resulting in revenue loss to the network operator. Therefore, $OC_{\mathbf{N}^*}(t)$ is used to bound this revenue loss. We define $OC_{\mathbf{N}^*}(t)$ as follows. Let:

- p_{ij}: price per bit for user j of class i.
- $\{\mathbf{Rv}^g\}_{g=1}^{N} = \{Rv_{ij}^1,\ Rv_{ij}^2, ..., Rv_{ij}^N \mid Rv_{ij}^g \geq Rv_{ij}^{g+1}\}$, where $Rv_{ij}^g = p_{ij} \cdot R_{ij}(t)$ is the revenue that the network operator will earn from user j given that this user is served in the current time frame. That is, the set $\{\mathbf{Rv}^g\}_{g=1}^{N}$ contains all users in descending order of the revenue that the network operator will earn from each one of them provided that they are served in the current time frame.
- $\mathrm{Re}\, v_{Max} = \sum_{g \in \{\mathbf{Rv}^g\}_{g=1}^{N}} Rv^g$, given that $\left(\sum_{(i,j) \in \{\mathbf{Rv}^g\}_{g=1}^{N}} R_{ij}(t) \right) \leq C$. $\mathrm{Re}\, v_{Max}$ is the maximum obtainable revenue in the current time frame (i.e., the maximum revenue the network operator can generate in the current time frame). $\mathrm{Re}\, v_{Max}$ is obtained by calculating the revenues of all users that could send in the current time frame (i.e., without

[12] The opportunity cost for a good is defined as the value of any other goods or services that a person must give up in order to produce or get that good [39].

exceeding the system capacity) and that if served, they will generate the maximum revenue to the network operator.

Therefore, $OC_{\mathbf{N}^*}(t)$ is defined as follows:

$$OC_{\mathbf{N}^*}(t) = \mathrm{Re}\, v_{Max} - \sum_{(i,j)\in \mathbf{N}^*} p_{ij} \cdot R_{ij}(t) \tag{3.2}$$

That is, the opportunity cost is a measure of how much revenue the network operator would forego if the users in set $\mathbf{N}^*$ are selected for transmission given that there are higher-revenue-generating users (i.e., the users that generate $\mathrm{Re}\, v_{Max}$). The network operator can determine the appropriate level of opportunity cost of fairness by choosing the value of *H*, and hence the appropriate level of fairness-revenue. For example, the network operator could restrict the revenue loss to be no more than 20% of the maximum obtainable revenue (i.e., $H = \zeta \cdot \mathrm{Re}\, v_{Max}$, where $\zeta = 0.2$). Note that if $H = 0$, then this implies that the network operator cannot tolerate any revenue loss, and therefore, only the highest-revenue-generating users are scheduled to transmit. On the other hand, if $H = \mathrm{Re}\, v_{Max}$ then the opportunity cost is ignored. In this case, all users are considered for transmission.

3.3.1 The Utility Function

To ensure the practicality of the scheduling scheme, we require $U_{ij}\left(\{\mathbf{X}_{ij}^{z}(t)\}_{z=1}^{m_{ij}}\right)$ to meet the following conditions:

1) $\frac{\partial U_{ij}(t)}{\partial X_{ij}^{z}(t)} \geq 0, \forall z, z \in \{1,2,...,m_{ij}\}; X_{ij}^{z}(t) \in \{X_{ij}^{1}(t), X_{ij}^{2}(t),..., X_{ij}^{m_{ij}}(t)\}$, the utility should be a non-decreasing function of $X_{ij}^{z}(t)$ to ensure that the user is satisfied with more allocated network resources (i.e., more $X_{ij}^{z}(t)$).

2) $U_{ij}\left(\{\mathbf{X}_{ij}^{z}(t)\}_{z=1}^{m_{ij}}\right) = U_{\min}$, if $\mathbf{X}_{ij}^{z}(t) = \mathbf{X}_{ij}^{z,\min}(t), \forall z, 1 \leq z \leq m_{ij}$,

$U_{ij}\left(\{\mathbf{X}_{ij}^{z}(t)\}_{z=1}^{m_{ij}}\right) \geq U_{\min}$, otherwise.

where $X_{ij}^{z,\min}(t)$ is the minimum value of the zth QoS measure. That is, if all QoS measures are at their minimum values, then the user's utility is at its minimum value (i.e., $U_{\min}$) reflecting that the user is dissatisfied with receiving low QoS. If only some QoS measures are at their minimum values, then the user's utility is larger than or equal to the minimum value.

3) $U_{ij}\left(\{\mathbf{X}_{ij}^{z}(t)\}_{z=1}^{m_{ij}}\right) = U_{\max}$, if $\mathbf{X}_{ij}^{z}(t) = \mathbf{X}_{ij}^{z,\max}(t), \forall z, 1 \leq z \leq m_{ij}$,

$U_{ij}\left(\{\mathbf{X}_{ij}^{z}(t)\}_{z=1}^{m_{ij}}\right) \leq U_{\max}$, otherwise.

where $X_{ij}^{z,\max}(t)$ is the maximum value of the zth QoS measure. That is, if all QoS measures are at their maximum values, then the user's utility is at its maximum

value (i.e., $U_{\max}$). If only some QoS measures are at their maximum values, then the user's utility is less than or equal to the maximum utility.

4) $$\lim_{\{\mathbf{X}_{ij}^{z}(t)\}_{z=1}^{m_{ij}} \to \{\mathbf{X}_{ij}^{z,\max}(t)\}_{z=1}^{m_{ij}}} U_{ij}\left(\{\mathbf{X}_{ij}^{z}(t)\}_{z=1}^{m_{ij}}\right) = U_{\max}$$

the higher the network resources the user receives, the higher the user's utility up to a certain maximum value $U_{\max}$, then the utility stays at that level reflecting that any additional allocated network resources will not increase the user's utility.

In addition to the above conditions, we require the utility function to support inter-class prioritization. Solving for the above conditions will not produce a unique solution. We, hence, introduce a plausible utility function in Eq. (3.3), with constants $a_i > 0$, $\forall i, 1 \le i \le K$, to capture the feasible area of the solution

$$U_{ij}\left(\{\mathbf{X}_{ij}^{z}(t)\}_{z=1}^{m_{ij}}\right) = 1 - e^{-a_i \cdot \sum_{z=1}^{m_{ij}} \left(X_{ij}^{z}(t)\right)} \tag{3.3}$$

where a_i serves as an inter-class distinguishing parameter in order to prioritize different classes of traffic, and larger values of a_i result in higher class prioritization. This is because larger values of a_i make the utility function more sensitive to any increase or decrease in the QoS measures (i.e., larger values of a_i increase the slope of the utility

function). As explained later, users with steep utility function result in the highest rate of change in it, and hence they maximize the social welfare of the system.

It is imperative to point out that at every scheduling decision, the variations in the users' QoS measures can be computed whether the users are served or not. Therefore, a solution to Eq. (3.1) can be found by computing the aggregate utility of the system if user j is scheduled and all other users are not and then finding the set of users (i.e., $\mathbf{N}^*$) with the highest aggregate utility (in descending order[13]) provided that they satisfy the constraints of Eq. (3.1). In other words, a solution to Eq. (3.1) can be found by choosing the a set $\mathbf{N}^*$ of users for transmission such that

$$\text{Objective: } \underset{(i,j)\in\mathbf{N}^*,\mathbf{N}^*\subseteq\mathbf{N}}{\arg\max}\left(\sum_{i\in\mathbf{N}^*}^{K}\sum_{j\in\mathbf{N}^*}^{N_i}1-e^{-a_i\cdot\sum_{z=1}^{m_{ij}}\left(X_{ij}^{z}(t)\right)}+\sum_{i=1}^{K}\sum_{y=1,y\notin\mathbf{N}^*}^{N_i}1-e^{-a_i\cdot\sum_{z=1}^{m_{iy}}\left(X_{iy}^{z}(t)\right)}\right)$$

$$\text{Subject to: } \nu_{ij}^{z,\min}\le X_{ij}^{z}(t)\le\nu_{ij}^{z,\max},\forall j\in\mathbf{N},\forall z,1\le z\le m_{ij}$$

$$\left(\sum_{(i,j)\in\mathbf{N}^*}R_{ij}(t)\right)\le C$$

$$OC_{\mathbf{N}^*}(t)\le H \qquad (3.4)$$

where all users (j) in set $\mathbf{N}^*$ are selected to transmit and all other users ($y\notin\mathbf{N}^*$) are not. Since Eq. (3.4) requires computing the aggregate utility of the system if every user is served (and the other users are not), then the run-time complexity of our scheduling

[13] That is, the user with the highest aggregate utility is scheduled to transmit. If this user does not have enough data in his queue to fill the frame, then the user with the next highest aggregate utility is added to the set of selected users and so forth until the frame is filled. The base station in BWASs can send to multiple users simultaneously using code multiplexing as in HSDPA [2] or frequency multiplexing as in WiMAX [3] and [4].

scheme is $O(N^2)$, where N is the number of active users in each cell. This is reasonable in BWASs, since packet scheduling is implemented at the base stations of these systems, where there is enough processing capabilities. In addition, the size of the input (i.e., the number of active users N) is relatively small since the base station is in charge of only tens or hundreds of active users within its coverage area.

Since Eq. (3.4) involves the summation of user utilities given that each user is selected to transmit and all others are not, then clearly the users who result in the highest rate of change in the utility function, are actually the ones that are going to maximize the social welfare of the system. This implies that the steeper the slope of the user's utility, the greater his chance of getting scheduled to transmit. The slope of the utility function in Eq. (3.3) is steeper at low values for the QoS measures. This implies that the users with low QoS measures result in the highest rate of change in the utility function. Hence, these users are given more priority for transmission in order to improve their QoS. This property, which is known in economics as diminishing marginal utility [39], is very important because it can be used to ensure fairness of the scheduling scheme. More discussions about this property are in Section 3.3.3.

In the following lemmas, we show that our proposed packet scheduling scheme reduces to the Maximum Carrier to Interface Ratio (Max CIR) [9] and the Proportional Fairness (PF) [10] schemes as a special case regardless of the QoS measures.

Lemma 1: If p_{ij} is set to 1 for every user (i.e., the price is ignored), minimum and maximum bound constraints on the QoS constraints are ignored and H is set to 0 in Eq.

(3.4), then our packet scheduling scheme reduces to Max CIR. Proof is provided in Appendix A.

Lemma 2: Let, $\max_{ij} \overline{S_{ij}(t)}$ be the maximum throughput achieved among all users at time t. If a_i is set to $-\ln\left(1-\ln\left(\overline{S_{ij}(t)}/\max_{ij}\overline{S_{ij}(t)}\right)\right)/\sum_{z=1}^{m_{ij}}\left(X_{ij}^{z}(t)\right)$, minimum and maximum bound constraints on the QoS constraints are ignored and the opportunity cost constraint is ignored (by setting H to $\mathrm{Re}\, v_{Max}$), then our packet scheduling scheme reduces to PF as a special case. Proof is provided in Appendix A.

3.3.2 Dynamic Computation of Opportunity Cost

It is imperative to point out that in some cases, the optimization problem in Eq. (3.4) may not have a feasible solution. This is because the scheduling scheme may have to serve certain users to guarantee certain levels of QoS (e.g., minimum data rate or maximum delay) even though these users do not satisfy the opportunity cost constraint. Therefore, to satisfy both constraints, the bound on opportunity cost (i.e., H) has to be dynamically computed in order to ensure the existence of a feasible solution of Eq. (3.4) as follows. Let:

- $\mathrm{Re}\, v_{\mathbf{n}^*} = \sum_{(i,j)\in\mathbf{n}^*} p_{ij} \cdot R_{ij}(t)$, where $\mathbf{n}^* \in \mathbf{N}^*$ is the set of users that must be served at time t (i.e., current time frame) in order to guarantee their QoS requirements. That is, $\mathrm{Re}\, v_{\mathbf{n}^*}$ is the obtainable revenue from users that require QoS guarantees.

In this case, the opportunity cost of serving the users in $\mathbf{n}^*$ is given by $OC_{\mathbf{n}^*}(t) = \mathrm{Re}\, v_{Max} - \mathrm{Re}\, v_{\mathbf{n}^*}$, where $\mathrm{Re}\, v_{Max}$ is defined in Section 3.3. Therefore, to avoid infeasibility in Eq. (3.4), we must have $H \geq OC_{\mathbf{n}^*}(t)$. The network operator could, for example, set a predefined value for H, say ϑ, and use it only when $H \geq OC_{\mathbf{n}^*}(t)$ is satisfied as follows:

$$H = \begin{cases} OC_{\mathbf{n}^*}(t), \ if \quad \vartheta \leq OC_{\mathbf{n}^*}(t) \\ \vartheta, \quad otherwise \end{cases} \tag{3.5}$$

3.3.3 Scheduling Different Types of Traffic

In this section, we define the QoS measures that are used in the utility function to support best-effort traffic, where the user's average throughput is the main QoS, traffic with minimum data rate requirements, and traffic with maximum delay requirements. The QoS measures are chosen so that the scheduling scheme achieves the objectives outlined in Section 3.1. We make the following definitions:

- $\overline{S_{ij}(t)} \triangleq$ average throughput for user j of class i up to time t.

- $\max_{ij} \overline{S_{ij}(t)} \triangleq$ maximum average throughput achieved among all users up to time *t*.
- $S_{ij}^{\min} \triangleq$ minimum required average data rate of user *j* of class *i*.
- $S_{ij}^{\max} \triangleq$ maximum required average data rate of user *j* of class *i*.
- $D_{ij}^{\max} \triangleq$ maximum tolerable average packet delay of user *j* of class *i* at time *t*.
- $\overline{D_{ij}(t)} \triangleq$ actual average packet delay of user *j* of class *i* at time.

To achieve our design objectives, we let $m_{ij} = 2$ in Eq. (3.3) and let:

- $X_{ij}^1(t) = \mu_{ij}(t) = \left(\mathrm{P}_{ij}^1 - \frac{R_{ij}(t)}{C} \right)$, where $0 \le \mathrm{P}_{ij}^1 \le 1$. We define this measure in order to exploit the user channel quality conditions in the scheduling decision, and hence maximize the users' individual data rates and the system throughput. This is because the higher the instantaneous data rate of the user (normalized by the system capacity *C*), the lower $\mu_{ij}(t)$, which results in a higher rate of change in the utility function in Eq. (3.3) due to its diminishing marginal property as mentioned earlier. Therefore, users with good channel quality conditions will have higher priority to transmit. In addition, when $\frac{R_{ij}(t)}{C} > \mathrm{P}_{ij}^1$, $\mu_{ij}(t)$ becomes negative, and consequently the utility function in Eq. (3.3) sharply decreases (i.e., its slope becomes steeper). This is shown in Figure 3.2, which plots the utility as a function of $X_{ij}^1(t)$ for $a_i = 4$ and P_{ij}^1 = 0.3, 0.5 and 0.7 (the graphs with P_{ij}^1 = 0.7 and 0.3 are shifted on the X-axis by 0.2 and -0.2, respectively to better show the

differences between them). Therefore, P_{ij}^{l} can be interpreted as a "penalty" incurred from not serving users with good channel quality conditions, where smaller values of P_{ij}^{l} increase the penalty, and hence give more weight to the users' channel quality conditions in the scheduling decisions.

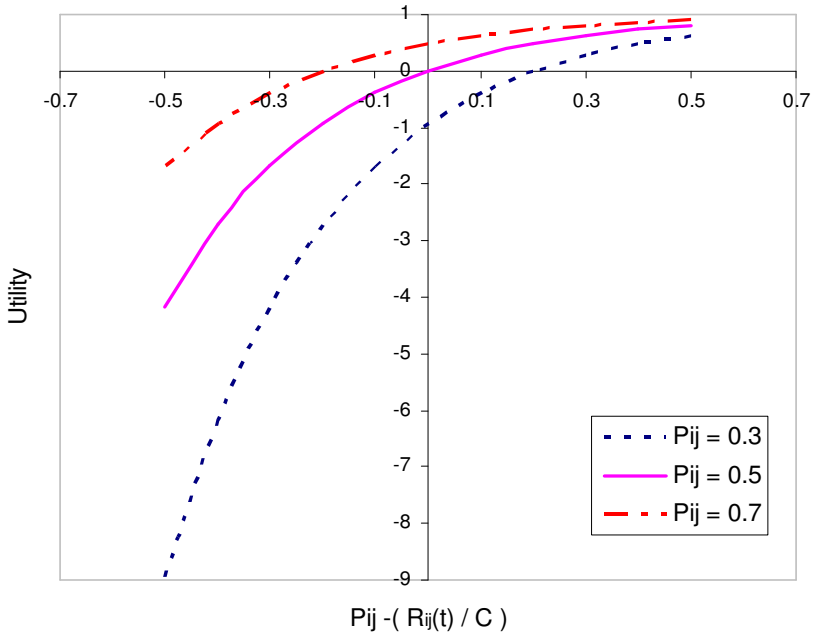


Figure 3.2: Effect of P_{ij}^{l} on the shape of the utility function

In our utility function, we use the same $X_{ij}^{1}(t)$ for all traffic types to increase the system throughput. However, we provide different definitions for $X_{ij}^{2}(t)$ for the different traffic types. For presentation purposes, let the class index i in $X_{ij}^{2}(t)$ be e, r, d for best-effort traffic, traffic with maximum data rate requirements and traffic with delay requirements, respectively.

For best-effort traffic, we define $X_{ej}^2(t)$ as follows

- $X_{ej}^2(t) = \alpha_{ej}(t) = \left(\dfrac{\overline{S_{ej}(t)}}{\max\limits_{ej} \overline{S_{ej}(t)}} - \mathrm{P}_{ej}^2 \right)$, $0 \le \mathrm{P}_{ej}^2 \le 1$. We define this measure to provide fairness for best effort traffic. Using this measure, if the user is receiving significantly lower average throughput compared to the one with the maximum average throughput, his fairness will be low indicating his dissatisfaction for the unfairness of the system. In this case, the scheduler will be forced to serve this user to increase his fairness measure. This is because, if a user with high average throughput is served, though his utility will increase, the social welfare of the system will not be maximized because of the rapid decrease of the utilities of those users with low average throughputs as a result of the diminishing marginal property of our proposed utility function. The role of P_{ej}^2 in determining the weight of this measure is similar to the role of P_{ij}^1 in $X_{ij}^1(t)$. However, in this case, larger values of P_{ej}^2 give more weight to $X_{ej}^2(t)$.

For traffic with minimum data rate requirements, we define $X_{rj}^2(t)$ as follows

- $X_{rj}^2(t) = \sigma_{rj}(t) = \left(\dfrac{\overline{S_{rj}(t)}}{S_{rj}^{\max}} - \mathrm{P}_{rj}^2 \right)$, $0 \le \mathrm{P}_{rj}^2 \le 1$. We define this measure in order to satisfy the users by granting them their required data rates. $\sigma_{rj}(t)$ also represents a fairness measure. This is because, if the user is receiving a low average throughput compared to other users who request the same data rate, the rate of

decrease in his utility function will be higher than the other users. The scheduler in this case will be forced to serve the user to increase his utility, and hence maximize the social welfare of the system. Larger values of P_{rj}^2 can be used to give more weight to $X_{rj}^2(t)$.

Finally, for traffic with maximum delay requirements, we define $X_{dj}^2(t)$ as follows

- $X_{dj}^2(t) = \varphi_{dj}(t) = \left(\mathrm{P}_{dj}^2 - \dfrac{\overline{D_{dj}(t)}}{D_{dj}^{\max}} \right)$, $0 \le \mathrm{P}_{dj}^2 \le 1$. We include this measure in order to satisfy the users' required average packet delays. $\varphi_{dj}(t)$ also represents a fairness measure similar to the case of traffic with data rate requirement. In this case, however, small values of P_{dj}^2 can be used to provide higher weight on $X_{dj}^2(t)$.

Using the above definitions, the scheduling problem becomes

$$\text{Objective: } \max_{(i,j)\in \mathbf{N}^*, \mathbf{N}^* \subseteq \mathbf{N}} \sum_{i=1}^{K} \sum_{j=1}^{N_i} 1 - e^{-a_i \cdot \left(\left(X_{ij}^1(t) \right) + \left(X_{ij}^2(t) \right) \right)}$$

$$\begin{aligned} \text{Subject to: } & S_{rj}^{\min} \le \overline{S_{rj}(t)} \le S_{rj}^{\max}, \quad \forall j \in \mathrm{N}, \forall z, 1 \le z \le m_{rj} \\ & \overline{D_{dj}(t)} \le D_{dj}^{\max}, \ \forall j \in \mathrm{N}, \forall z, 1 \le z \le m_{dj} \\ & \left(\sum_{(i,j)\in \mathbf{N}^*} R_{ij}(t) \right) \le C \\ & OC_{\mathbf{N}^*}(t) \le H \end{aligned} \tag{3.6}$$

The first constraint ensures that the users' average throughputs lie between their minimum and maximum requirements. The second constraint ensures that the users' average packet delays do not exceed their maximum delay.

While P^1_{ij}, P^2_{ej}, P^2_{rj} and P^2_{dj} are used to determine the weights of the QoS measures, a_i plays an important role in determining the shape of the utility function, and hence the level of inter-class prioritization. Larger values of a_i increase the slope of the utility function, and thus result in higher class prioritization. This is shown in Figure 3.3, which plots the utility function in Eq. (3.3) for different values of a_i (and penalty, i.e., P_{ij} of 0.5). a_i, along with other parameters ($P^1_{ij}, P^2_{ij} = P^2_{ej}$, P^2_{rj} and P^2_{dj}), therefore, should be set appropriately by the network operator as to achieve its desired level of inter- and intra-class prioritization, and hence its desired level of fairness. In the following section, we show the effect of some of these parameters on the system performance. Additional results are provided in Section B.4 in Appendix B.

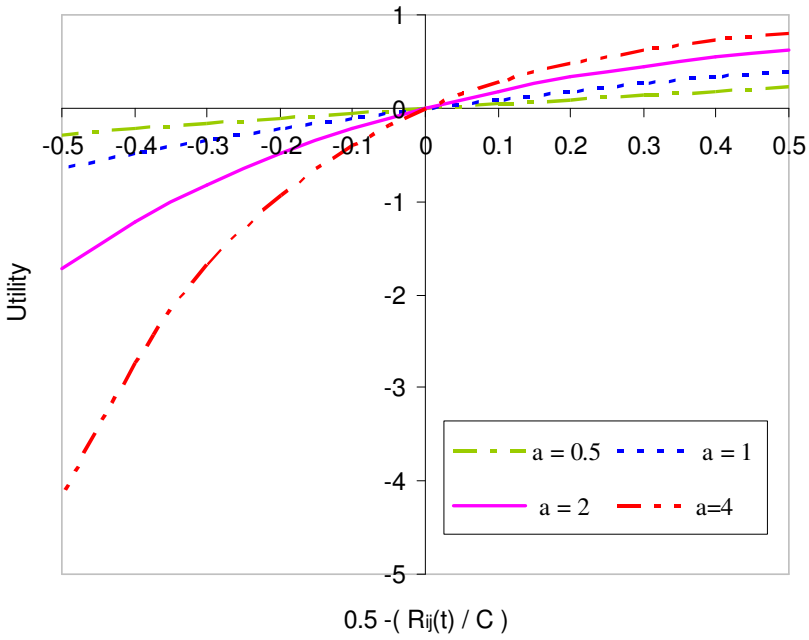


Figure 3.3: Effect of a_i on the shape of the utility function

3.4 Performance Evaluation

In this section, we evaluate the performance of our proposed packet scheduling scheme by means of dynamic discrete event simulation written in the Java programming language. We tested our scheme on HSDPA [2]. We first begin by describing the simulation model and the traffic model. We then discuss the channel model. Next we present the test cases and performance metrics, followed by detailed discussions and comparisons of the simulation results. All the relevant simulation parameters are included in Appendix B.

3.4.1 Simulation Model

We consider a single-cell scenario in our simulation (though we consider inter- and intra-cell interference as discussed in Section 3.4.3). The base station is located at the center of the cell. The cell radius is 1 Km and the base station's transmission power is 38 dBm.

Users are uniformly distributed in the cell. The Pedestrian A environment [56] is used in our experiments, where every mobile user moves inside the cell with a constant speed of 3 km/hr. This speed is the recommended value for the Pedestrian A environment by the 3GPP [56]. A total of 10 channel codes are used, which correspond to a total system capacity of 7.2 Mbps [7]. User connection arrivals are modeled as a Poisson process. The simulation time step is one time frame, which is 2 ms in HSDPA [2], and the simulation time is 400s.

3.4.2 Traffic Model

To demonstrate the ability of our scheme to support different classes with users having different QoS requirements, we consider three different classes with four different services, namely VoIP (class 1), audio streaming (class 2), video streaming (class 2) and FTP (class 3). In addition, to demonstrate the ability of our scheme to prioritize different classes (i.e., inter-class prioritization), we assume that class 1 has the highest priority and class 3 has the lowest priority. Moreover, we assume that audio streaming has a higher priority than video streaming in order to demonstrate the ability of our scheme to prioritize services with different QoS within the same class (i.e., intra-class prioritization).

To achieve such prioritizations, we choose appropriate values for a_i, P_{ij}^1, P_{ej}^2, P_{rj}^2 and P_{dj}^2 according to their role in the utility function as explained in Sections 3.3.1 and 3.3.3. These values can be found in Appendix B. Additional simulation results showing the effect of choosing different values for these parameters on the system performance are presented in Section B.4 in Appendix B. Furthermore, for demonstration purposes, we assume that p_{ij} = 6, 4, 2 and 1 units of money for VoIP, audio streaming, video streaming and FTP users, respectively.

For VoIP traffic, we adopt the model in [57], which assumes Adaptive Multi-Rate (AMR) codec. In this model, packets are generated using a negative exponentially distributed ON-OFF traffic source to simulate the talk and silence spurts, where the mean duration of both ON and OFF periods is 3s. During the ON periods, a voice packet of 244 bits is generated every 20 ms, corresponding to a source bit rate of 12.2 Kbps, which is comparable to one of the AMR bit rates [58]. The compressed IP/UDP/RTP header increases the bit rate to 13.6 kbps [59]. The ITU E-model [60] states that when the one-way mouth-to-ear delay exceeds 250 ms, the voice quality rating rapidly deteriorates. About 80 to 150 ms remain for the base station processing and connection reception when the delay induced by the voice encoder/decoder and other components in the system is subtracted [61]. Therefore, we set the maximum average packet delay threshold for VoIP traffic to a value between 80 and 150 ms, specifically 100 ms.

Audio streaming is modeled with a minimum rate of 12 Kbps, mean rate of 38 Kbps, maximum rate of 64 Kbps, maximum packet delay of 150 ms and a packet size uniformly distributed between 244 and 488 bits. These values are chosen from within the range of

specific QoS requirements defined by 3GPP in order to provide adequate service to mobile users [62], [63] and [64]. Video streaming is modeled with a minimum data rate of 64 Kbps, mean rate of 224 Kbps, maximum data rate of 384 Kbps and a packet size uniformly distributed between 1,200 and 2,400 bits [62], [63] and [64]. FTP traffic is simulated by a constant rate of 128 Kbps and a fixed packet size of 1,200 bits. Durations of VoIP and video streaming user connections are modeled by an exponential distribution with a mean value of 50s. Whereas, in case of FTP users it is assumed that each user requests one FTP file of size 50 MB and terminates his connection after the file download is complete.

3.4.3 Channel Model

The channel model describes the attenuation of the radio signal on its way from the base station to the user, and therefore, it describes how the channel condition of the user changes with time depending on the user's environment and speed. In our simulation, the channel model consists of five parts: distance loss, shadowing, multi-path fading, intra-cell interference and inter-cell interference. Details about each one of these parts are provided in Appendix B.

3.4.4 Test Cases and Performance Metrics

To provide QoS guarantees (e.g., minimum data rates or maximum packet delays), the scheduling scheme must be supported by a CAC scheme in order to block users when there is not enough capacity to provide such guarantees. In this chapter, we focus on

packet scheduling in order to show its performance independently from the CAC scheme. We, therefore, do not consider the case of guaranteed QoS in our experiments. Such a case is considered in Chapter 5 when we introduce our CAC scheme. In addition, since existing packet scheduling schemes cannot effectively support different types of traffic with different QoS requirements at the same time, we distinguish between two cases. In the first case, all users in the system belong to only one traffic type (i.e., VoIP, audio streaming, video streaming or FTP). For VoIP and audio streaming, we compare the performance of our proposed Fair Class-based Packet Scheduling scheme (denoted by FCBPS) to that of the Modified Largest Weighted Delay First (denoted by M-LWDF) [14] and [15], Fair Modified Largest Weighted Delay First (denoted by FM-LWDF) [16] and the Maximum CIR with Early Delay Notification (denoted by EDN) [17] schemes, since these schemes are designed for real-time traffic with delay requirements. For video streaming and FTP, we compare the performance of our scheme with that of the Maximum CIR (denoted by Max CIR) [9], Proportional Fairness (denoted by PF) [10] and the Fast Fair Throughput (denoted by FFT) [13] schemes, since these schemes are designed for non-real-time traffic with throughput requirements only.

In the second case, we evaluate the performance of our scheme under a multiplexed scenario in which users can request any of the four traffic types considered in our simulation. Such a case is designed to show the ability of our scheme to simultaneously serve different users with different QoS requirements in addition to show its ability to provide inter- and intra-class prioritization. In this case, the total arrival rate to the system is equally divided among the three classes of traffic.

The following performance metrics are used:

- Average packet delay: the average amount of time the packet spends in the queue at the base station in addition to the transmission time (delays of discarded packets and dropped connections are not counted).
- Average throughput: average number of successfully delivered bits over the lifetime of the user's connection (throughputs of dropped user connections are not counted).
- Channel utilization: percentage of the number of transmitted bits to the maximum number of bits that could be transmitted depending on the channel quality conditions of the users.
- Cell Throughput: average number of transmitted bits by the base station. It equals to the total number of transmitted bits over the number of servings (i.e., number of transmissions), measured over the simulation time.
- Service coverage: percentage of users who achieve their required QoS with a certain outage level. For audio streaming, a user's connection is dropped if his average packet loss (due to packet discarding, transmission errors and/or buffer overflow) exceeds 5% [65], [66] and [67]. For video streaming, a user's connection is dropped if his achieved average throughput is less than his minimum required rate. Finally, for FTP traffic, a user's connection is dropped if his achieved average throughput is less than 9.6 Kbps [13] and [16].

- Percentage of revenue loss: ratio of revenue loss to the maximum amount of revenue that could be earned, where the maximum revenue is equal to $\mathrm{Re}\, v_{Max}$ as defined in Section 3.3 and revenue loss is calculated from Eq. (3.2).

- Jain Fairness Index (JFI) [68]: a fairness index used to calculate fairness among users that belong to the same class (i.e., intra-class fairness). Let ψ_{ij} be the performance metric for user j of class i, where ψ_{ij} is set to the user's average packet delay for VoIP and audio streaming, and it is set to the user's average throughput for video streaming and FTP. Then the JFI is calculated as follows

$$JFI = \frac{\left(\sum_{z=1}^{N_{ij}} \psi_{ij} \right)^2}{N_{ij} \sum_{z=1}^{N_{ij}} (\psi_{ij})^2}, \ \psi_{ij} \geq 0 \ \ \forall j \tag{3.7}$$

where N_{ij} is the number of class i users who request the same QoS. Note that if all users who request the same QoS achieve the same ψ_{ij}, then JFI=1. Lower JFI values indicate that users have high variances in their achieved QoS, which reveals unfairness in distributing the wireless resources among them.

3.4.5 Simulation Results

In this section, we show and discuss the simulation results for the two cases considered in our experiments. The simulation results obtained in all experiments in this thesis have a 95% confidence level with 10% confidence intervals based on 10 independent runs.

Case 1: Single Traffic Class

In this section, we discuss the performance results of the evaluated schemes for VoIP and video streaming traffic only. The performance results of audio streaming and FTP are similar to those of VoIP and video streaming, and hence they are not shown here.

VoIP

Figure 3.4 depicts the average packet delay for VoIP traffic as a function of the arrival rate to the system. The figure shows that M-LWDF achieves the best packet delay under most network loads, whereas FM-LWDF has the worst packet delay. FM-LWDF performs poorly compared to the other schemes because of its fairness measure (i.e., the equalizer term), which is in terms of throughput and not in terms of delay. Hence, more resources are given to users with bad average throughput at the expense of those users with high packet delays. FCBPS (with maximum tolerable revenue loss of $\mathrm{Re}v_{Max}$; i.e., opportunity cost is ignored) achieves reasonably low packet delays at different network loads (within 5% of the performance of M-LWDF). This is due to the fact that as the user's average packet delay increases, the sharp decrease in his utility forces the

scheduler to serve him, and hence improve his packet delay. The average packet delay achieved by EDN is worse than our scheme and M-LWDF because as the network load increases (i.e., arrival rate ≥ 0.5), the packet delays of users exceed the threshold in EDN, and hence users are only served based on their packet delays without exploiting their channel quality conditions. Such users require more resources to transmit, causing more packet delays to users with good channel quality conditions. The average packet delays achieved by FCBPS with three different maximum tolerable revenue losses, namely $\mathrm{Re}v_{Max}$, $0.5 \cdot \mathrm{Re}v_{Max}$ and 0 are shown in Figure 3.5. As the maximum tolerable revenue loss decreases, the average packet delay increases. This is because when the maximum tolerable revenue loss is low, only high-revenue-generating users are served by FCBPS, and hence the packet delays of other users in the system increase causing an increase in the overall average packet delay.

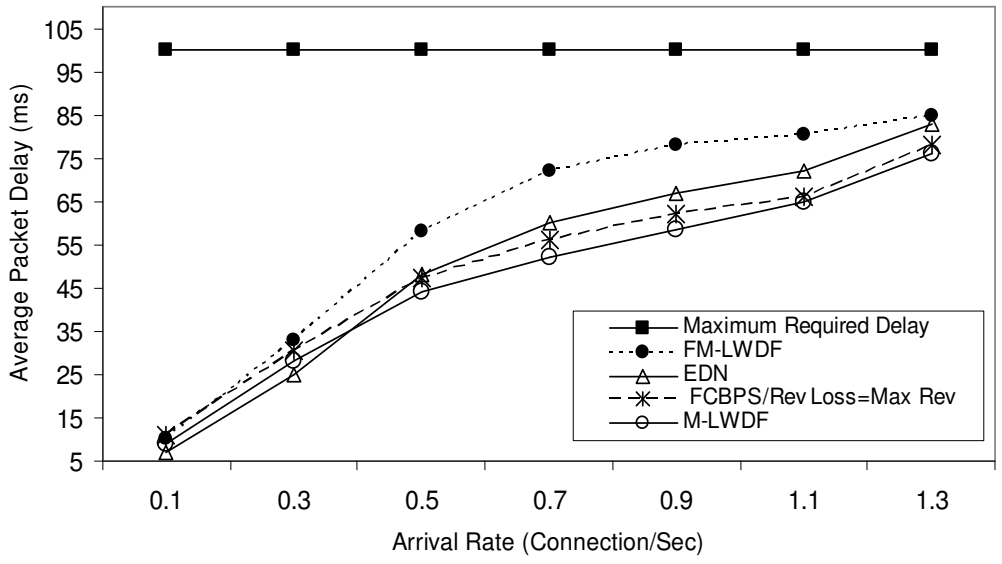


Figure 3.4: Average packet delay for VoIP traffic

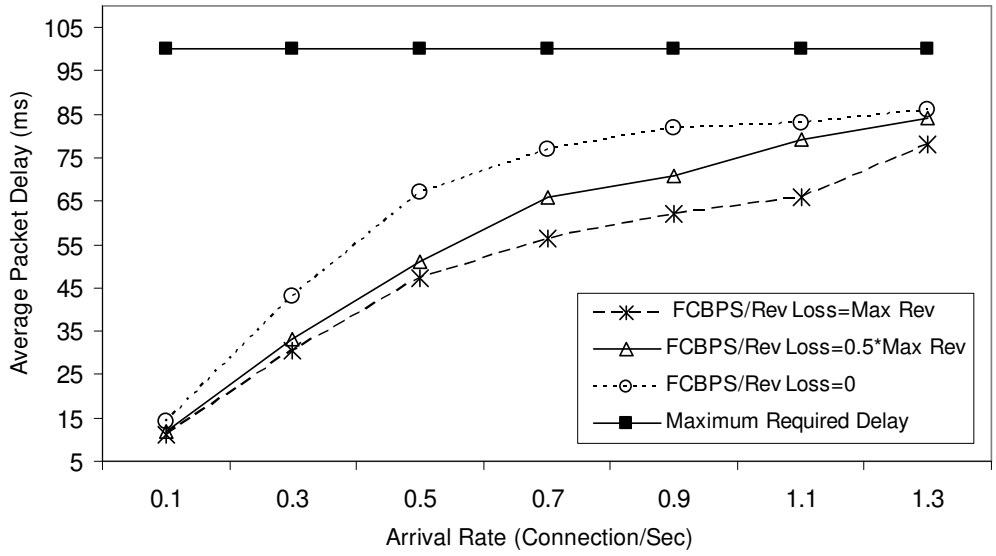


Figure 3.5: Average packet delay of FCBPS with different revenue losses for VoIP traffic

Figure 3.6 depicts the percentage of channel utilization for the evaluated schemes. Cleary, FCBPS achieves the best channel utilization even when the maximum tolerable revenue loss is set to $\mathrm{Re}v_{Max}$. This shows the ability of our scheme to exploit the variations of channel quality conditions of users to maximize the throughput of the network. An interesting result that is revealed from Figure 3.6 is that EDN achieves the lowest channel utilization despite the fact that it uses Max CIR in its scheduling decisions. The reason for this is that, at high arrival rates, EDN serves users only based on their packet delays as mentioned earlier, and therefore Max CIR is not really utilized. As Figure 3.7 shows, when the maximum tolerable revenue loss is decreased, the channel utilization of FCBPS increases because in this case users with good channel quality conditions are favored for transmission over those with less favorable channel quality conditions. This is due to the fact that good channel quality conditions allow for higher

bit rate transmissions, and consequently higher collected revenues.

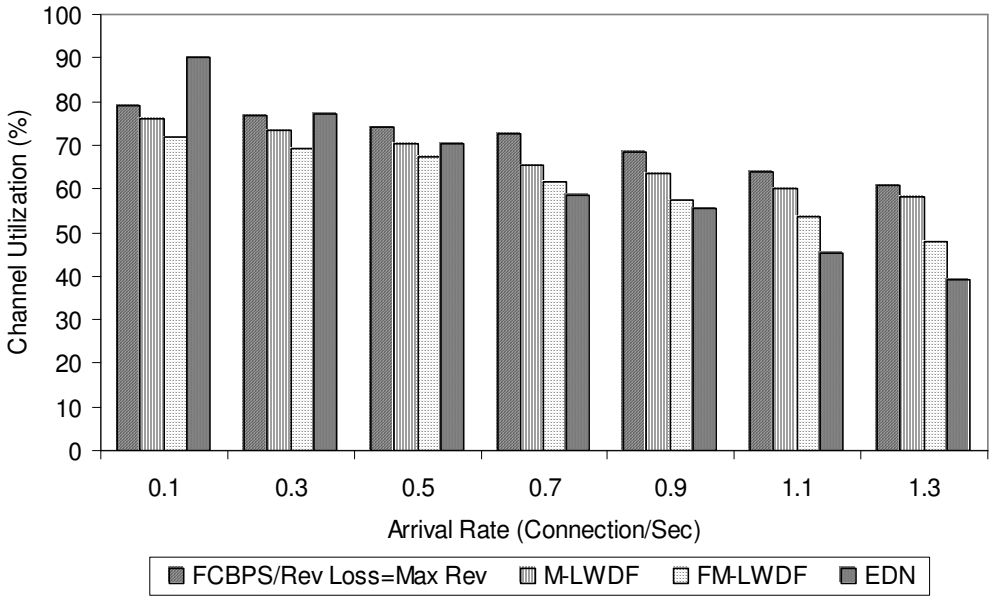


Figure 3.6: Percentage of channel utilization

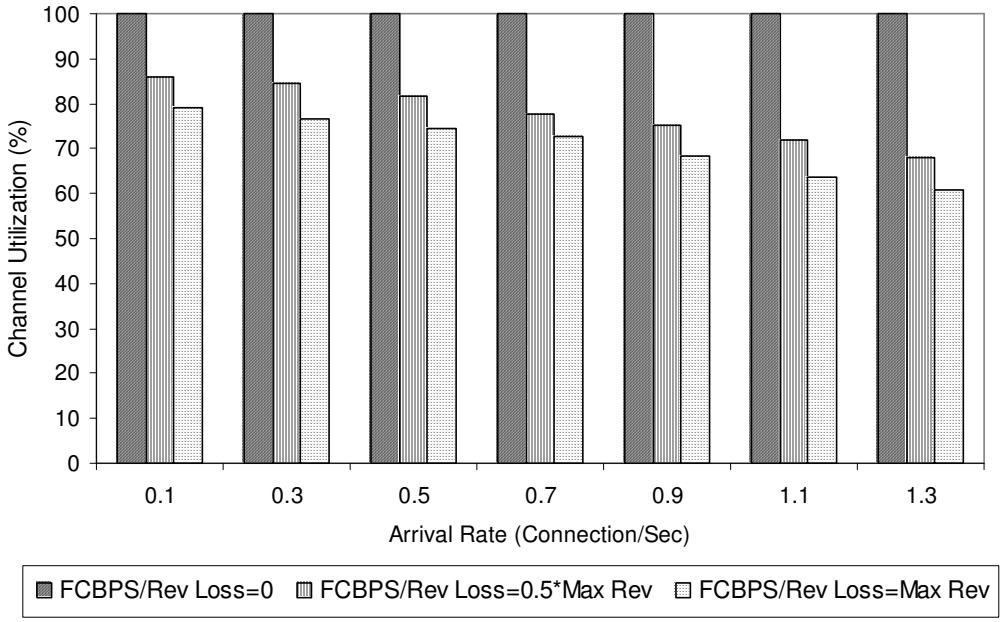


Figure 3.7: Percentage of channel utilization of FCBPS with different revenue losses

FCBPS also demonstrates superior performance in terms of service coverage and revenue loss compared to other schemes as shown in Figures 3.8, 3.9, 3.10 and 3.11. The exponential decrease in our proposed utility function when a user experiences high average packet delays forces the scheduler to serve him, and hence more users are covered by FCBPS. When the maximum tolerable revenue loss decreases, the revenue loss of FCBPS decreases, however, at the expense of service coverage. Therefore, using our scheme, the network operator can determine the level of revenue loss and the corresponding level of service coverage to maximize its revenues. We can also see that revenue loss is related to channel utilization, as expected. In fact, the higher the channel utilization, the lower the revenue loss.

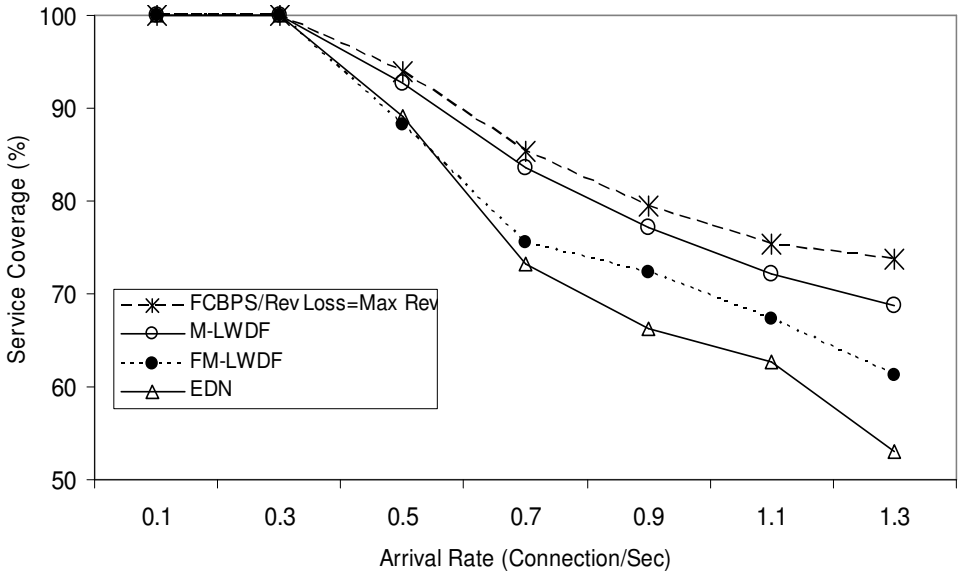


Figure 3.8: Percentage of service coverage for VoIP traffic

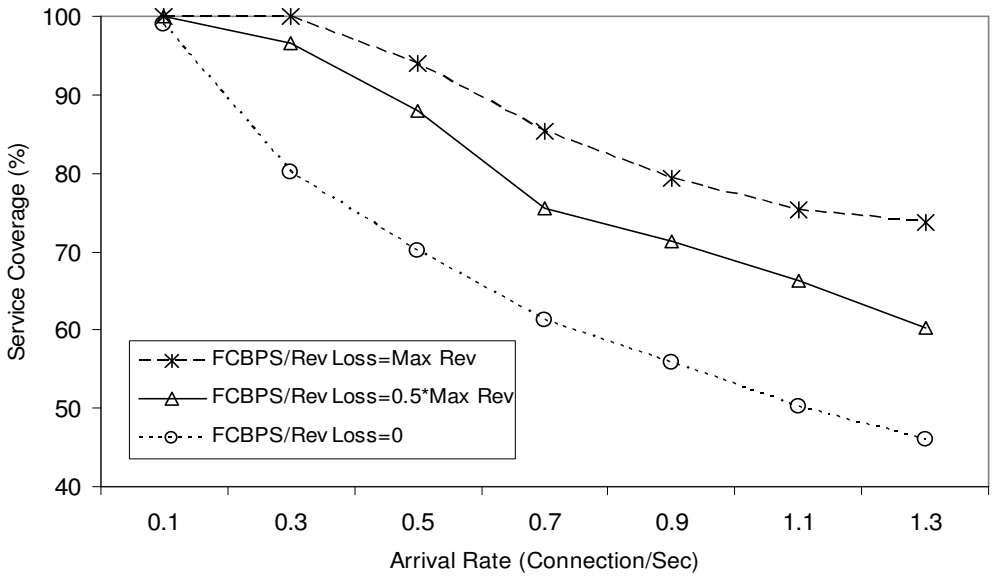


Figure 3.9: Percentage of service coverage of FCBPS with different revenue losses

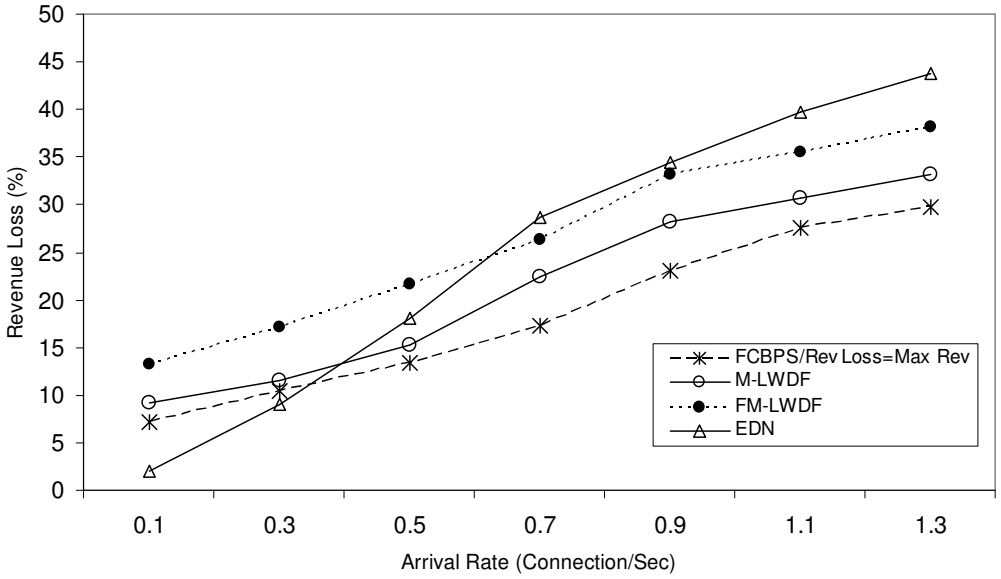


Figure 3.10: Percentage of revenue loss

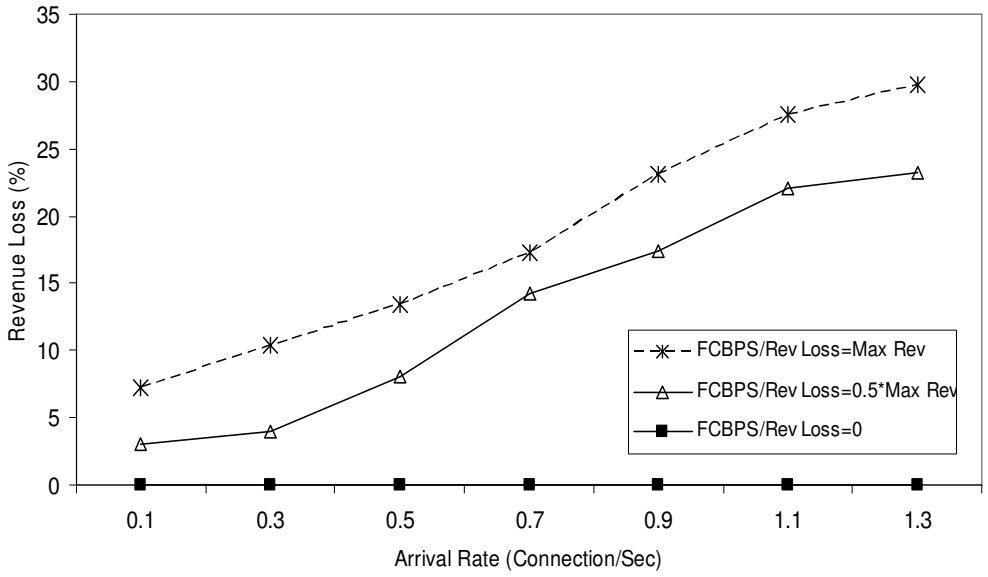


Figure 3.11: Percentage of revenue loss of FCBPS with different revenue losses

Figure 3.12 depicts the JFI of the evaluated schemes, which shows that FCBPS achieves the best fairness performance. This is due to the use of fairness measures in our proposed utility function, which allow the scheme to distribute the wireless resources fairly among users while exploiting the variations in their channel quality conditions. This results in increased fairness as well as increased user throughput. However, when the maximum tolerable revenue loss is decreased, the fairness of FCBPS deteriorates as shown in Figure 3.13. This behavior is expected as users are selectively scheduled to transmit based on the revenue they generate to the network operator.

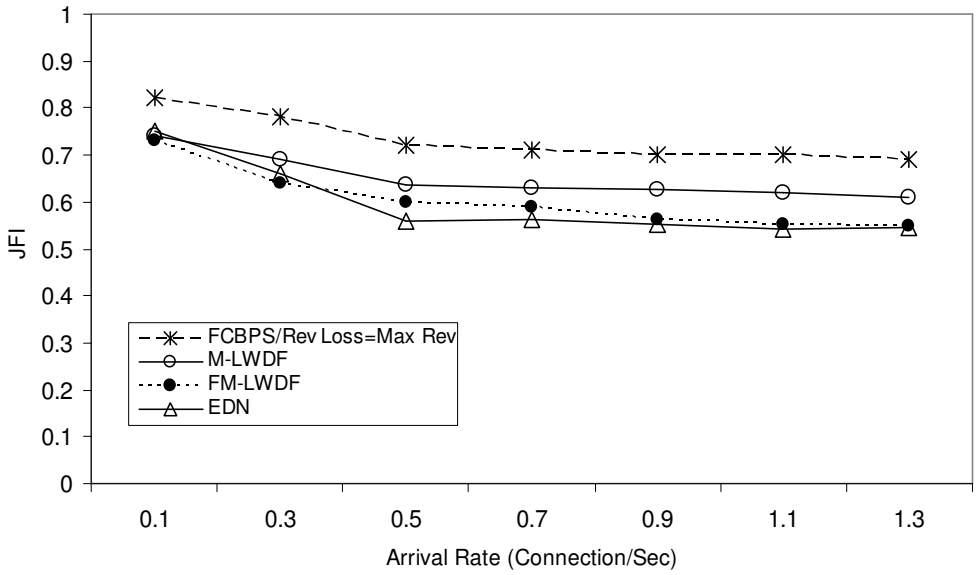


Figure 3.12: The Jain Fairness Index

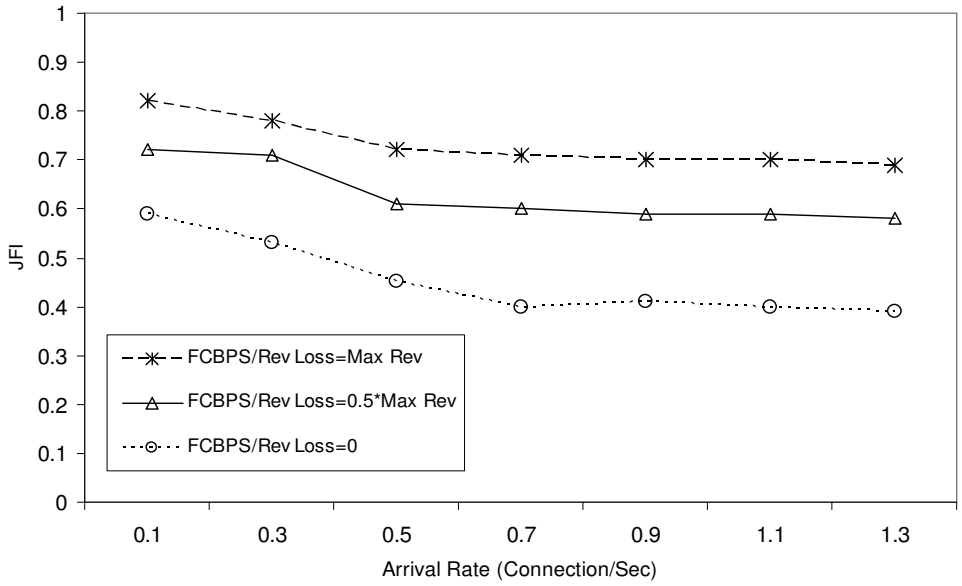


Figure 3.13: The Jain Fairness Index of FCBPS with different revenue losses

Video Streaming

The average throughput for video streaming users is shown in Figure 3.14. Max CIR achieves the best performance since it schedules the users based on their best channel quality conditions. FCBPS (with maximum tolerable revenue loss of $\mathrm{Re}v_{Max}$) outperforms PF. FFT on the other hand, has the lowest average throughput because of the equalizer term in FFT, which forces it to achieve long-term fairness at the expense of exploiting the channel quality conditions of different users. In addition, the average throughputs of users increase as the maximum tolerable revenue loss is decreased in FCBPS as shown in Figure 3.15. This is because high-revenue-generating users (from the network operator's perspective) are those with good channel quality conditions since more bits could be transmitted in this case. Therefore, as the maximum tolerable revenue loss is decreased, the performance of FCBPS approaches that of Max CIR. Figure 3.16 depicts the percentage of channel utilization. FCBPS achieves good utilization levels compared to PF and FFT. Max CIR, however, achieves the best channel utilization (100% under all arrival rates) because it only serves the users with the best channel quality conditions, and hence the channel is fully utilized. Moreover, the performance results of FCBPS in terms of the percentage of channel utilization, cell throughput, percentage of revenue loss and fairness for different levels of maximum revenue losses are similar to the case of VoIP traffic, and hence they are omitted.

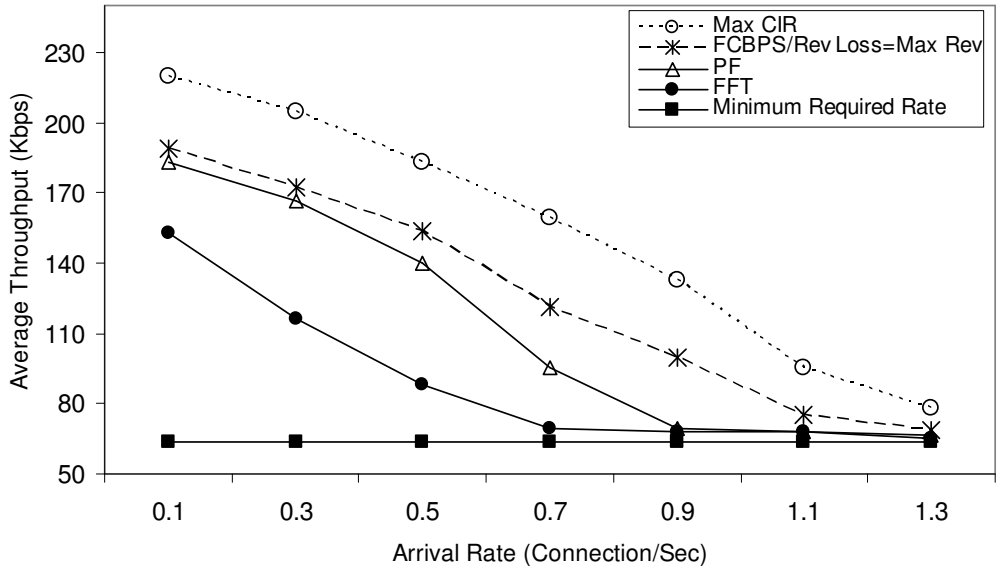


Figure 3.14: Average throughput for video streaming traffic

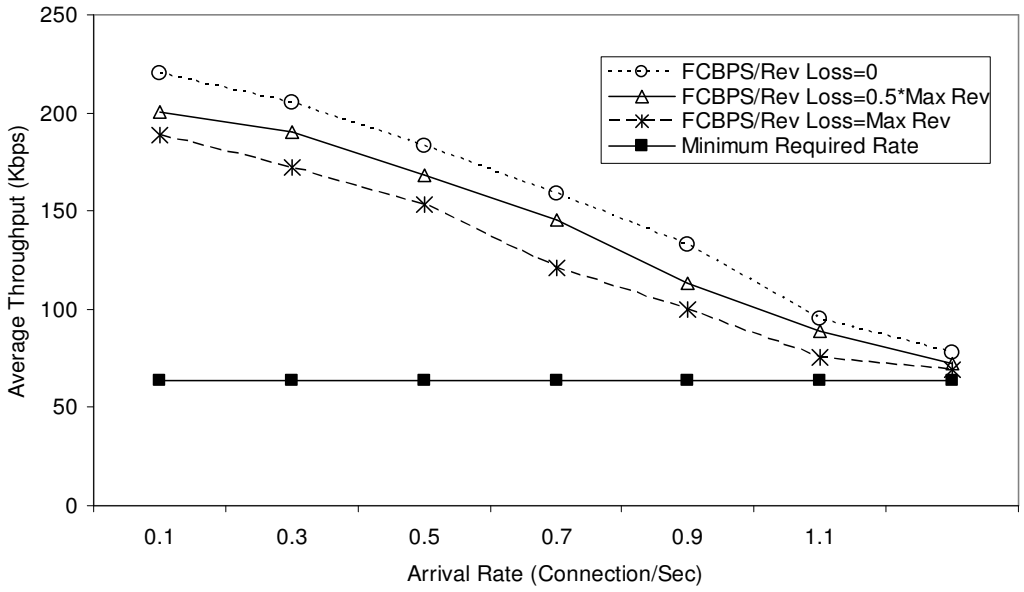


Figure 3.15: Average throughput of FCBPS with different revenue losses for video streaming traffic

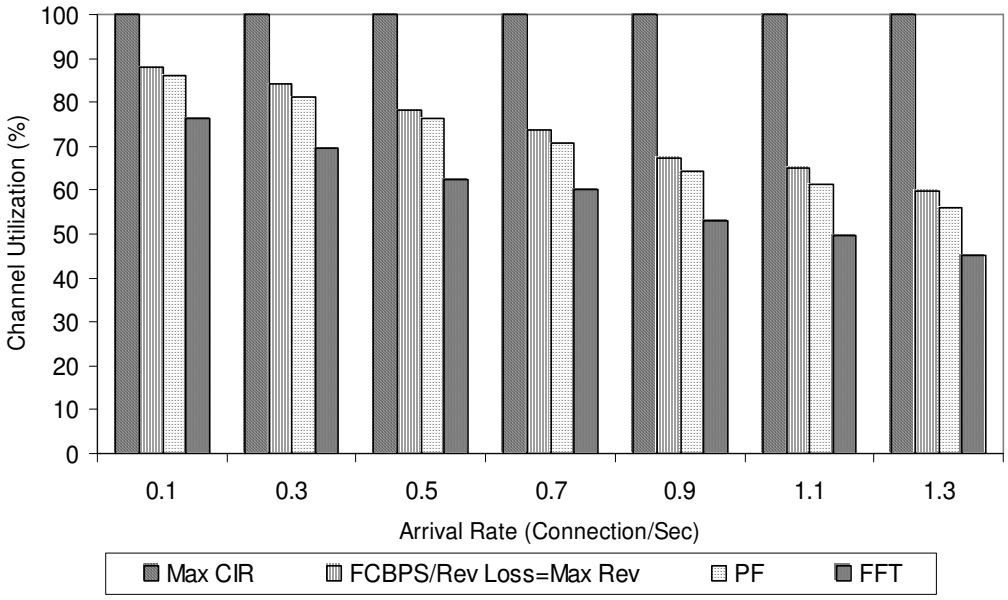


Figure 3.16: Percentage of channel utilization

The good channel utilization levels of FCBPS result in good cell throughputs compared to PF and FFT as shown in Figure 3.17. FCBPS also achieves good levels of service coverage compared to PF and Max CIR as shown in Figure 3.18. The best service coverage, nevertheless, is achieved by FFT, as expected due to the equalizer term. This happens, however, at the expense of low channel utilization and low cell throughput as mentioned earlier. Figure 3.19 shows the percentage of service coverage of FCBPS for different revenue losses. It can be seen that as the maximum tolerable revenue loss decreases, the service coverage decreases until it reaches that of Max CIR. This confirms our argument that with low maximum tolerable revenue loss, users with good channel quality conditions are favored for transmission over those with bad channel quality conditions since more bits can be transmitted, and hence greater revenues can be earned.

Therefore, packet scheduling schemes that better exploit the channel quality conditions of users result in lowest revenue losses as confirmed by Figure 3.20. This, however, comes at the expense of fairness as is clearly shown in Figure 3.21.

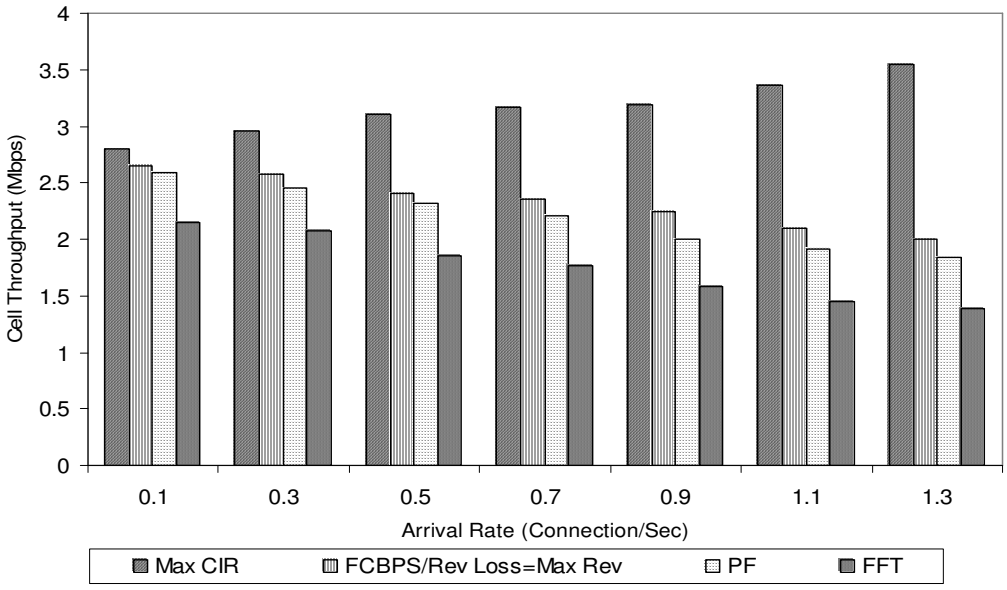


Figure 3.17: Cell throughput

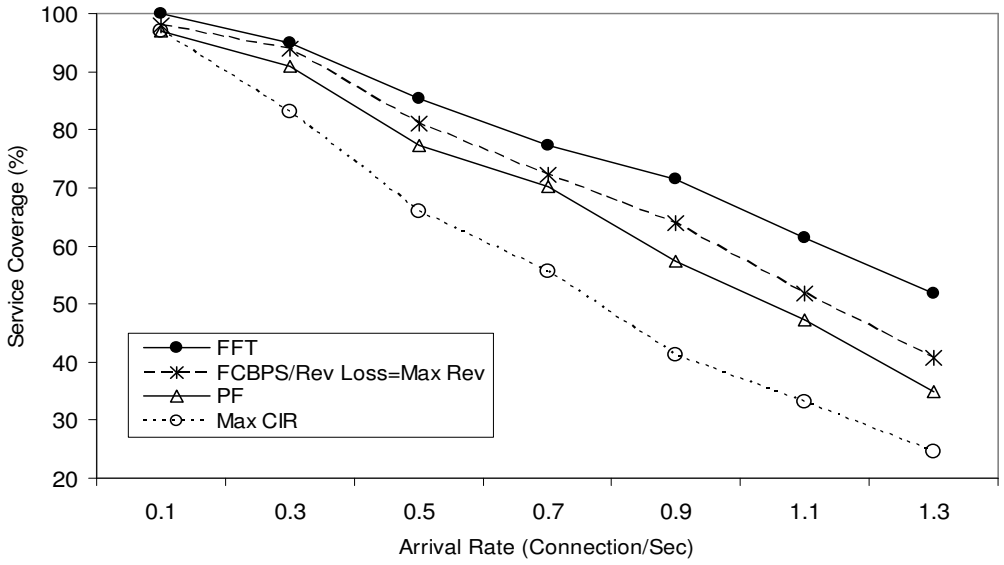


Figure 3.18: Percentage of service coverage for video streaming traffic

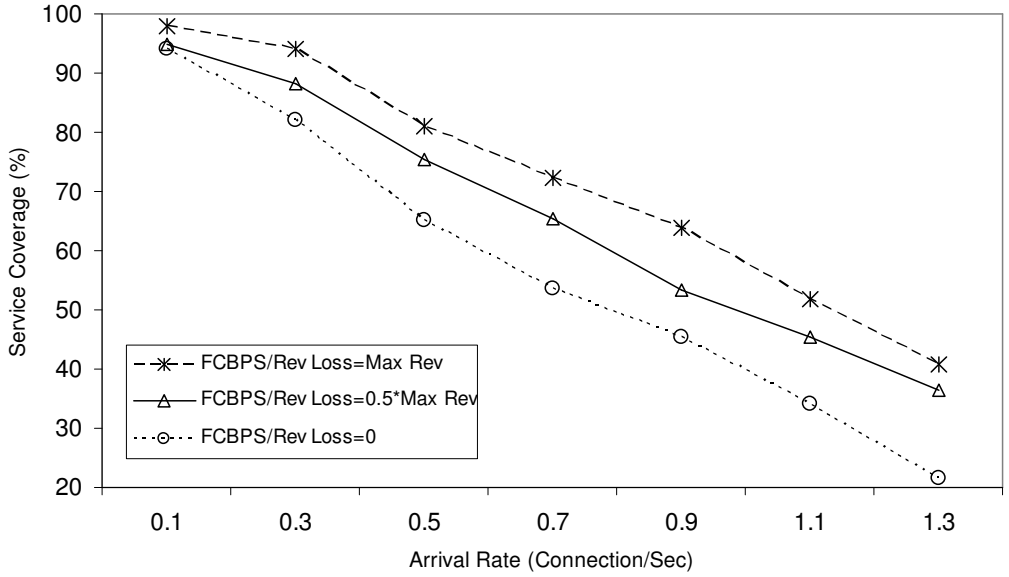


Figure 3.19: Percentage of service coverage of FCBPS with different revenue losses for video streaming traffic

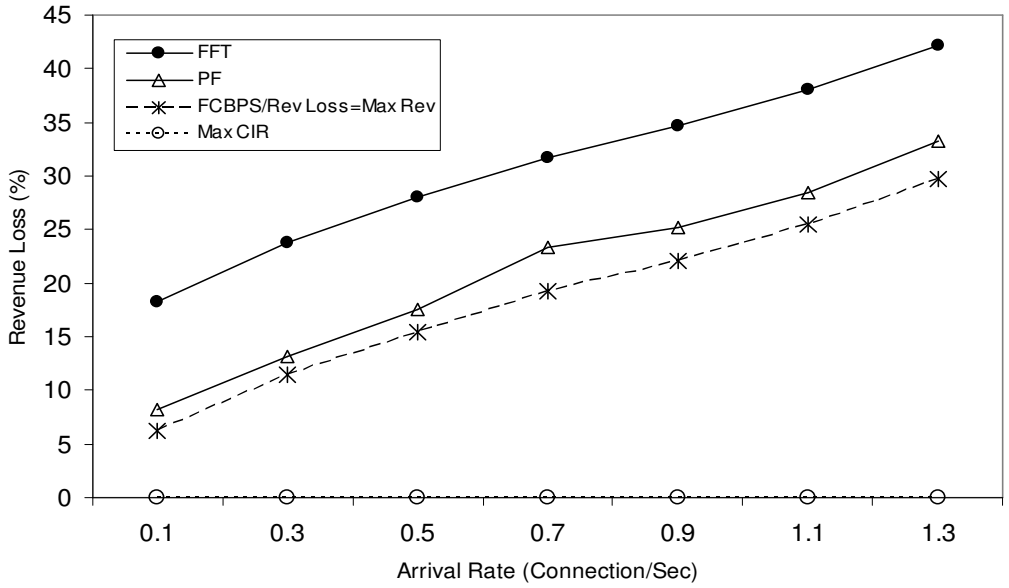


Figure 3.20: Percentage of revenue loss

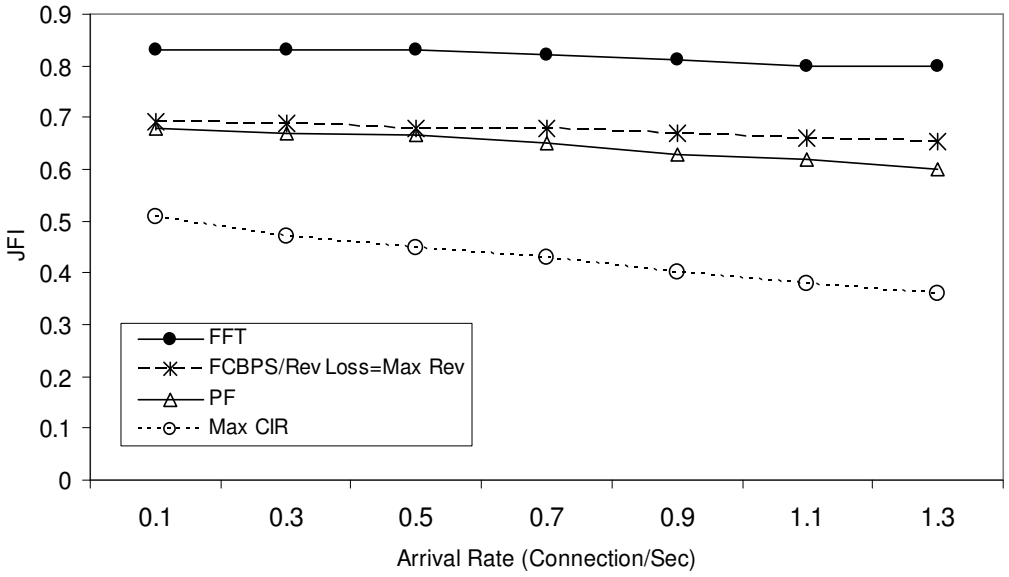


Figure 3.21: The Jain Fairness Index

Case 2: Multiplexed Traffic

In this section, we discuss the performance results of our scheme with maximum tolerable revenue loss of $\mathrm{Re}v_{Max}$ under a multiplexed traffic case in order to show its effectiveness in supporting multiple traffic types simultaneously. Figures 3.22 and 3.23 show the average packet delay for VoIP and audio streaming users, respectively. In general, both types of users achieve acceptable average packet delays under different network loads. It should be noted that the VoIP traffic outperforms audio streaming since the later has lower priority. Moreover, the performance results of VoIP traffic are better than the single traffic case, since the total arrival rate in multiplexed traffic is equally divided between the three classes of traffic, and hence there are fewer VoIP users in this case than the single traffic case (the arrival rate for class 2 is also equally divided

between audio and video traffic).

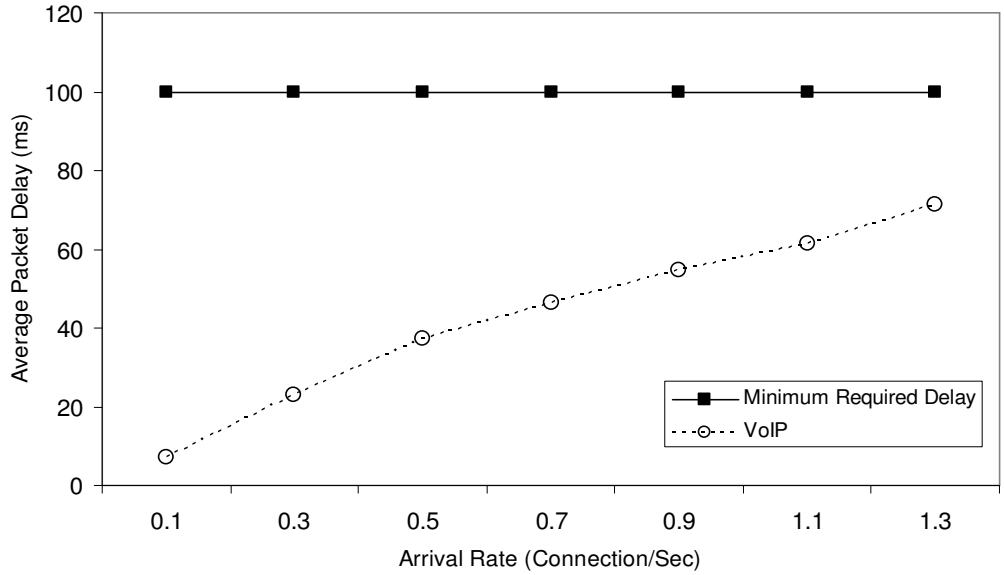


Figure 3.22: Average packet delay for VoIP

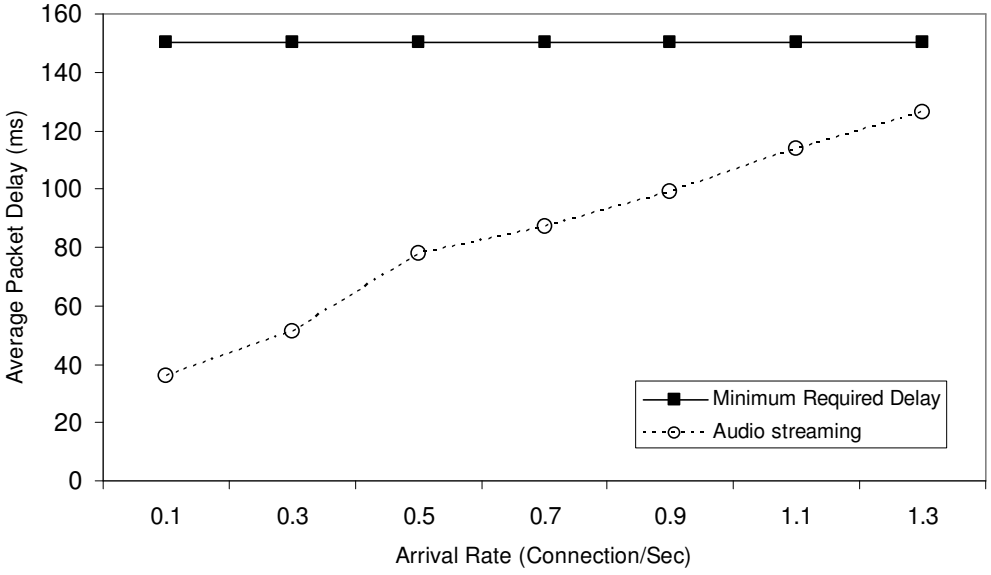


Figure 3.23: Average packet delay for audio streaming

Figures 3.24 and 3.25 depict the average user throughput for video streaming and FTP users, respectively. The figures show that video streaming users achieve higher average throughputs because they have higher priority than FTP users. The percentage of service coverage is shown in Figure 3.26. In general, our scheduler achieves acceptable service coverage for different types of traffic at different network loads, where traffic of higher priorities receives higher coverage. Figure 3.27 shows the JIF for each traffic type. The JIF of lower priority traffic (i.e., video streaming and FTP) is less than that of higher priority traffic (i.e., VoIP and video streaming). This is because lower priority traffic is assigned fewer time frames than higher priority traffic, hence, not allowing enough time for our defined fairness measures to make an impact.

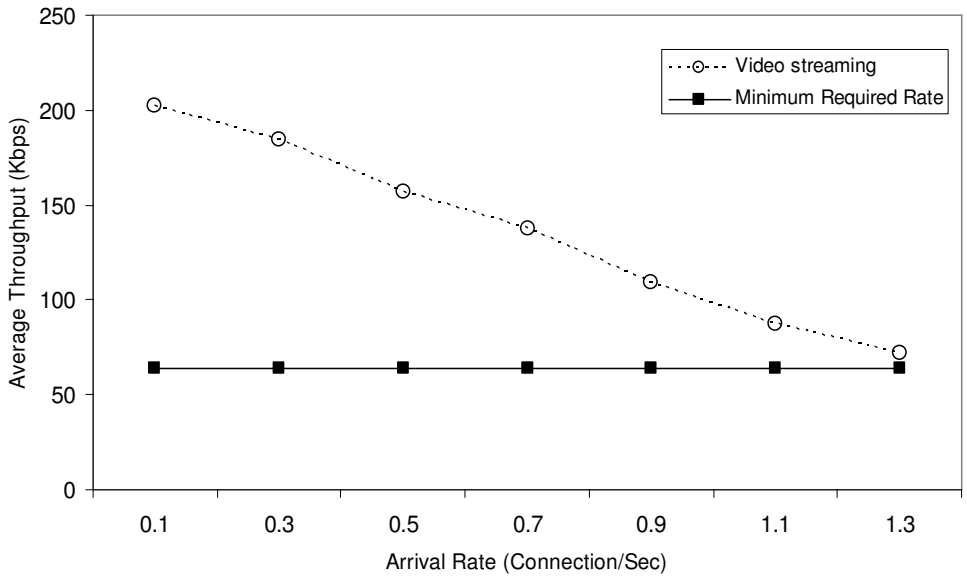


Figure 3.24: Average throughput for video streaming

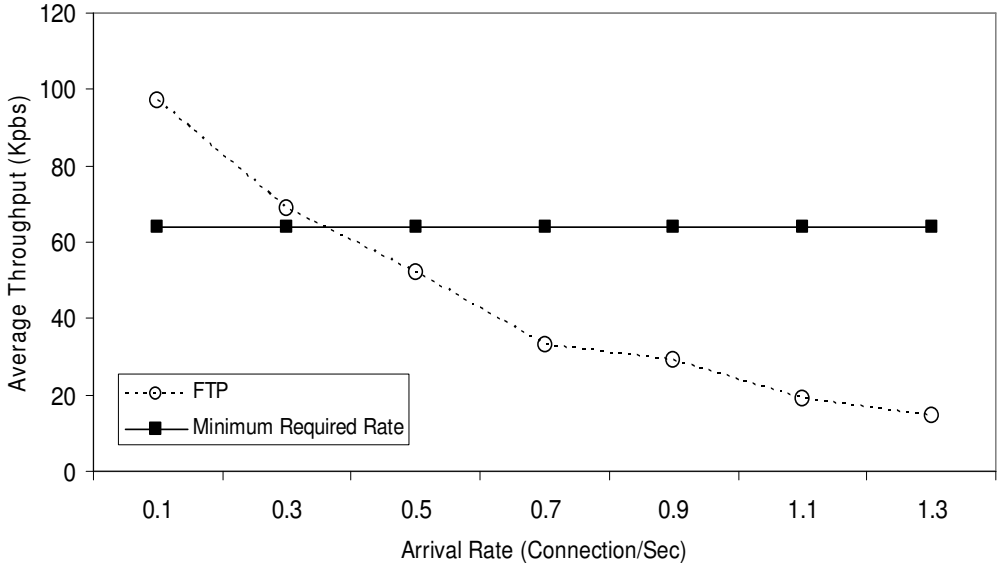


Figure 3.25: Average throughput for FTP

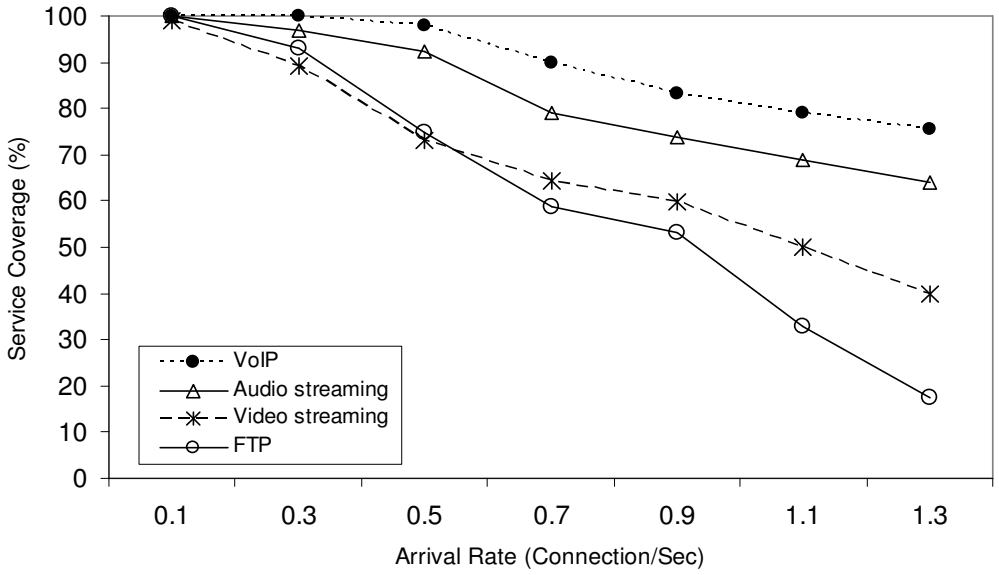


Figure 3.26: Percentage of service coverage for all traffic types

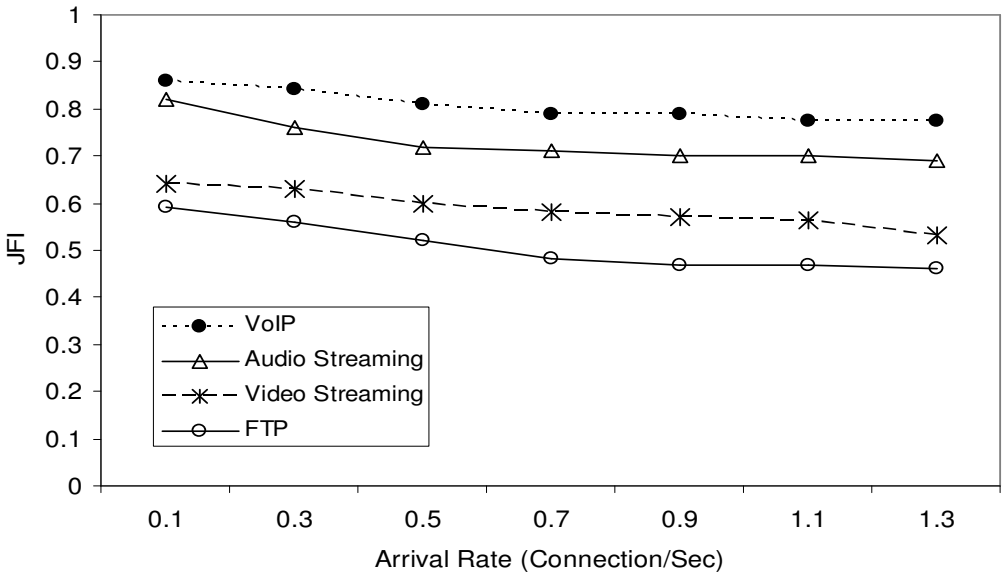


Figure 3.27: The Jain Fairness Index for all traffic types

The results shown in this section confirm that using our proposed scheme, the network operator can simultaneously support different types of services with different QoS requirements, prioritize different types of traffic within the same class (e.g, audio and video streaming), prioritize different classes and bound the revenue loss of serving users, hence, determining the appropriate level of fairness in the system.

3.5 Summary

In this chapter, a novel fair class-based downlink packet scheduling scheme for BWASs was proposed. The proposed scheme employs practical economic models through the use of novel utility and opportunity cost functions to simultaneously satisfy the diverse QoS

requirements of mobile users and maximize the revenues of network operators. Unlike most existing schemes, the proposed scheme is general and can support multiple classes of traffic with users having different QoS and traffic demands. To demonstrate its generality, we provide definitions for the proposed utility function in order to support three different types of traffic, namely best-effort traffic, traffic with minimum data rate requirements and traffic with maximum packet delay requirements. In addition, the proposed scheme uniquely incorporates fairness in its formulation in order to prevent users with good channel quality conditions from monopolizing the wireless shared channel, and hence leading to starvation to other users. We have shown mathematically that the two well-known scheduling schemes, Max CIR and PF are special cases of our scheme, which gives the network operators more flexibility in choosing between different scheduling schemes. Simulation results based on HSDPA show that the proposed scheme can enhance the performance of the wireless system by satisfying the QoS of users, bounding the revenue loss of serving them and ensuring fairness among them.

Chapter 4

Dynamic Bandwidth Provisioning Scheme

Packet scheduling, as mentioned earlier, will have a great impact on the performance of BWASs as, in essence, it is in charge of distributing their shared channel's resources among mobile users. Packet scheduling, however, is only a short-term bandwidth management scheme because it makes its decision on a frame by frame basis. In BWASs, the amount of resources (e.g., time slots) that each user requires may change from time to time due to his varying channel quality conditions. This problem is aggravated by the fact that, in some BWASs such as WiMAX networks, users are allowed to change their bandwidth requirements during the lifetime of their connections. Therefore, packet scheduling must be coupled with longer-term bandwidth management schemes. As explained in Chapter 2, there are two types of such management schemes, pre-admission,

which we refer to as admission-level bandwidth management and post-admission, which we refer to by class-level bandwidth management. The main focus of this chapter is on class-level bandwidth management. The main objective of class-level bandwidth management, which we refer to as "bandwidth provisioning", is to maintain acceptable levels of QoS throughout the lifetimes of user connections. This is achieved by spanning multiple time frames and deciding to optimally distribute them among the different classes of traffic, and hence their corresponding users. In doing so, some problems need to be carefully addressed. For example, bandwidth provisioning should be able to satisfy the bandwidth requirements of classes depending on the requirements of their admitted users, adapt to changes in bandwidth requirements of classes due to new admitted users or completed connections, support inter-class fairness, and consider the revenues of network operators in the bandwidth provisioning process.

In this chapter, we present a bandwidth provisioning scheme that is designed to provide efficient bandwidth management at the class level. We show how the concept of opportunity cost introduced in the previous chapter can be used at the class level to limit the revenue loss resulting from serving low-revenue-generating classes. We also present a dynamic weight update scheme, which aims at maximizing inter-class fairness while ensuring service differentiations between different classes.

The rest of this chapter is organized as follows. Section 4.1 outlines our proposed packet scheduling scheme and discusses its objectives. Section 4.2 describes the system model. Section 4.3 presents our proposed dynamic bandwidth provisioning and weight

update schemes. Section 4.4 presents the performance evaluation of our proposed schemes. Section 4.5 summarizes the chapter.

4.1 Scheme Outline and Objectives

The main contributions of this chapter are the dynamic bandwidth provisioning and the weigh update schemes. The proposed schemes are to be implemented at the base stations of BWASs. These schemes are designed to achieve the following objectives:

1) Supporting different types of traffic with users who have different bandwidth requirements;
2) Adapting to the varying bandwidth requirements of traffic classes;
3) Supporting inter-class fairness;
4) Supporting service differentiations between classes; and
5) Considering the revenues of network operators.

Basically, the main idea of the proposed dynamic bandwidth provisioning scheme is to allocate a number of given time frames to different classes of traffic to achieve the objectives above. In our proposed bandwidth provisioning scheme, each class of traffic is assigned a weight to represent its priority in the frame allocation process. The scheme works as follows. Let *NF* be the number of time frames to be allocated among the different classes. At time *t* (where *t* is the beginning of the next *NF* frames), the base station will evaluate the performance history of the different classes and will use this information to update their weights to maximize inter-class fairness as described in

Section 4.2. These weights, in turn, are used as input parameters to the proposed bandwidth provisioning scheme, which will allocate the next *NF* frames among the different classes of traffic based on their weights, the bandwidth requirements, channel quality conditions of their users, and the expected revenues.

Once each class is assigned a number of frames, these frames will be distributed to admitted users according to the packet scheduling scheme, which is executed every time frame. The network operator may utilize any existing packet scheduling scheme for distributing the partitioned frames among connections as our bandwidth provisioning scheme is independent from packet scheduling. Figure 4.1 shows an abstract timeline data flow chart of the proposed dynamic bandwidth provisioning and weight update schemes.

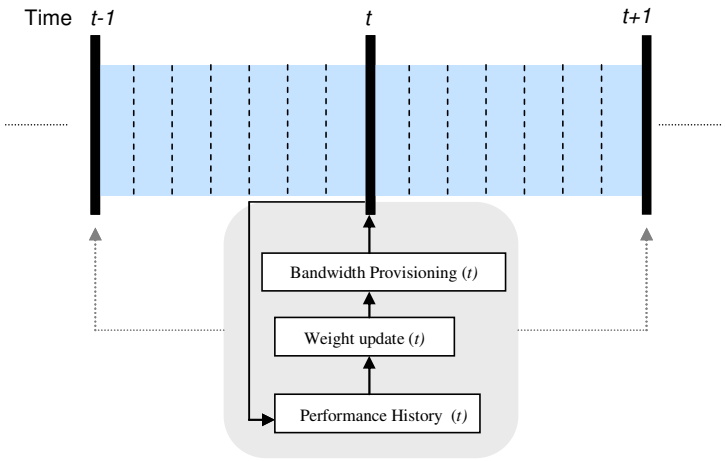


Figure 4.1: Dynamic bandwidth provisioning with the weight update scheme

4.2 System Model

We adapt a similar system model to that presented in Chapter 3, by considering a downlink time-slotted channel consisting of time frames. Data transmission is done at the base station on a frame by frame basis. We also consider K classes of traffic, where class i has higher priority than class i+1. Let N_i denote the number of class i users, and $N = \sum_{i=1}^{K} N_i$ is the total number of user connections in the system. We consider that users within each class can have different bandwidth requirements.

We consider that the network operator wants to allocate a total number of NF frames between the K classes of traffic. We assume that NF is given. In practice, simulation studies or real experiments can be used to determine empirically the appropriate value of NF that achieves the performance levels desired by the network operator.

4.3 Dynamic Bandwidth Provisioning Scheme

In this section, we present our proposed bandwidth provisioning scheme. We distinguish two cases of bandwidth provisioning. In the first no bandwidth guarantees are required for any class. We then extend our scheme to support minimum bandwidth guarantees. Next, we explain the weight update scheme to ensure fairness between classes.

4.3.1 Basic Bandwidth Provisioning

We make the following definitions. Let:

- $N_i \triangleq$ number of class i user connections.
- $N = \sum_{i=1}^{K} N_i \triangleq$ total number of user connections in the system.
- $S_{ij}^{\max} \triangleq$ maximum data rate required by user j of class i, $j = 1, ..., N_i$.
- $NF_i \triangleq$ number of frames allocated to class i.
- $\overline{R_i(t)} \triangleq$ effective average estimated data rate (per second) that the base station can transmit to class i users during the next NF frames. This data rate will depend on the estimated instantaneous channel quality conditions of the users as well as their bandwidth requirements. $\overline{R_i(t)}$ can be roughly estimated using a moving average (i.e., $\overline{R_i(t)} = \alpha \cdot \overline{R_i(t-2)} + (1-\alpha) \cdot \overline{R_i}(t-1)$, where $\overline{R_i}(t-1)$ is the average data rate that the base station was sending at to class i connections in the previous NF frames and $0 \le \alpha \le 1$) or using channel prediction schemes proposed in [69], [70] and [71].
- $p_{ij} \triangleq$ price per bit for user j of class i.
- $B_i^{\max} \triangleq$ total required maximum data rate per frame of all users in class i at the beginning of the NF frames. Let $\sum_{j=1}^{N_i} S_{ij}^{\max}$ be the total required maximum data rate per second of all users in class i, and let D_{frame} be the frame duration in seconds,

then $B_i^{\max} = \left(\left(\sum_{j=1}^{N_i} S_{ij}^{\max} \right) / 1 / D_{frame} \right)$. $B_i^{\max}$ determines the transmission rate the base station should be sending at per frame, in the next *NF* frames, to satisfy the maximum data rate requirements of class *i* users.

- $\overline{B_i} = \left((\overline{R_i(t)} / (1 / D_{frame})) \right) \triangleq$ actual (i.e., effective) total transmitted data rate per frame for class *i* users. That is, $\overline{B_i}$ determines the actual transmission rate per frame of the base station in the next *NF* frames for class *i* users.
- $\mathrm{Re}\, v_i^{\max} \triangleq$ total maximum revenue per frame of class *i* users at the beginning of the *NF* frames. Therefore, $\mathrm{Re}\, v_i^{\max} = \left(\left(\sum_{i=1}^{N_i} p_{ij} \cdot S_{ij}^{\max} \right) / 1 / D_{frame} \right)$. $\mathrm{Re}\, v_i^{\max}$ determines the revenue of the network operator per frame in the next *NF* frames if it grants all the users in class *i* their maximum required data rates. Therefore, $NF \cdot \mathrm{Re}\, v_i^{\max}$ is the upper bound of the total revenue of the network operator during the next *NF* frames.
- $\overline{\mathrm{Re}\, v_i} = \left(\left(\sum_{i=1}^{N_i} p_{ij} \cdot \overline{R_i(t)} \right) / 1 / D_{frame} \right) \triangleq$ effective total revenue per frame actually generated from serving all users in class *i*. Therefore, $NF_i \cdot \overline{\mathrm{Re}\, v_i}$ is the actual total revenue that the network operator earns from serving all users of class *i* provided that class *i* is allocated NF_i frames.
- $\{\overline{\mathbf{Rev}_i^z}\}_{z=1}^{K} \triangleq$ descending ordered set of the actual effective total revenue per frame resulting from serving the *K* classes.

To satisfy all users, the base station should allocate a data rate of $NF \cdot B_i^{\max}$ per *NF* frames to class *i,* $\forall i,\ 1 \le i \le K$. However, this may not be possible in practice due to the high demand of services that have high bandwidth requirements and also due to the limitations of the base station's capacity, which is determined by the channel quality conditions of the mobile users. Therefore, the main objective of our bandwidth provisioning scheme is to allocate the *NF* frames among the *K* classes of traffic such that $\sum_{i=1}^{K} NF_i = NF$ and the satisfaction of different users is maximized. To this end, our bandwidth provisioning scheme will distribute the *NF* frames amongst the *K* classes of traffic so that it maximizes the ratio between the data rate allocated to class *i* users, given that it is assigned NF_i frames, to the data rate that the base station should transmit at during *NF* frames to satisfy the maximum data rate requirement of class *i* (i.e., $\frac{NF_i \cdot \overline{B_i}}{NF \cdot B_i^{\max}}$). The frames allocated to class *i* (i.e., NF_i) should guarantee that no class is allocated more than its maximum required data rate (i.e., $NF_i \cdot \overline{B_i} \le NF \cdot B_i^{\max}$).

In addition, similar to the case of packet scheduling, it is important to realize that there is an opportunity cost of frame allocation at the class level. The opportunity cost (in terms of revenue) of frame allocation is the maximum revenue that the network operator will earn if it serves the highest revenue generating classes minus the revenue that it will earn by allocating the frames otherwise. To compute the maximum revenue that the network operator could earn in the next *NF* frames, we first need to know the number of frames needed by each class (NF_i^{req}) in order to achieve the maximum required data rate

by its users (i.e., $NF_i^{req} \cdot \overline{B_i} = NF \cdot B_i^{\max}$). Hence, $NF_i^{req} = \frac{NF \cdot B_i^{\max}}{\overline{B_i}}$. Therefore, the maximum revenue at the class level, $Max \operatorname{Re} v_c$, is equal to

$$Max \operatorname{Re} v_c = \sum_{z \in \{\operatorname{Re} v_i^z\}_{z=1}^K} NF_i^{req} . \operatorname{Re} v_i^z \text{ , given that } \sum_{z \in \{\operatorname{Re} v_i^z\}_{z=1}^K} NF_i^{req} \le NF \qquad (4.1)$$

The maximum revenue is obtained by allocating the frames to the class with the highest actual revenue. If this class can be served by fewer than *NF* frames, the remaining frames are allocated to the class with the second highest actual revenue, and so forth. Therefore, the opportunity cost ($OC(NF)$) of the frame allocation at the class level is equal to

$$OC(NF) = Max \operatorname{Re} v_c - \left(\sum_{i=1}^K NF_i \cdot \overline{\operatorname{Re} v_i} \right) \qquad (4.2)$$

This should be less than or equal to a predefined value H_c . For example, the network operator could restrict the revenue loss to be no more than 30% of the maximum obtainable revenue (i.e., $H_c = \zeta_c \cdot Max \operatorname{Re} v_c$, where $\zeta_c = 0.3$).

To summarize, in our multiple-frame bandwidth provisioning scheme, the following optimization problem will be solved

$$\text{Objective: } \max_{NF_i, 1 \le i \le K} \sum_{i=1}^{K} w_i \cdot \left(\frac{NF_i \cdot \overline{B_i}}{NF \cdot B_i^{\max}} \right)$$

$$\text{Subject to: } \sum_{i=1}^{K} NF_i = NF ,$$
$$NF_i \cdot \overline{B_i} \le NF \cdot B_i^{\max}, \quad \forall i, 1 \le i \le K \text{ , and}$$
$$OC(NF) \le H_c \tag{4.3}$$

where w_i is a weight assigned to class i to give it priority over class i+1 in the frame allocation process. Since the objective function and the constraints are linear, our bandwidth provisioning scheme can be solved using Linear Programming (LP) techniques.

The proposed bandwidth provisioning scheme is adaptive to the varying requirements of different classes of traffic, since the objective function is evaluated every *NF* frames. Therefore, if the required bandwidth (or frames) of class *i* changes during the current frames (due to new admitted connections and completed ones or bandwidth adaptive requests as it is the case in WiMAX), its new total required bandwidth will be reflected in the next *NF* frames.

4.3.2 Bandwidth Provisioning with Minimum Guaranteed Bandwidth

Even though the dynamic bandwidth provisioning scheme in Section 4.3.1 aims at maximizing the satisfaction of the different users, it does not provide bandwidth guarantees to traffic classes. The network operator may want to provide such guarantees.

Therefore, the bandwidth provisioning scheme should consider such a case[14]. Here, we extend our scheme to support minimum bandwidth guarantees. Let:

- $S_{ij}^{\min} \triangleq$ minimum required data rate of user j of class i, $j = 1, ..., N_i$.
- $B_i^{\min} \triangleq$ total required minimum data rate per frame of all users in class i at the beginning of the NF frames. That is, $B_i^{\min} = \left(\left(\sum_{j=1}^{N_i} S_{ij}^{\min} \right) / 1 / D_{frame} \right)$.

The dynamic bandwidth provisioning scheme should guarantee that no class of traffic is allocated less than its minimum required data rate (i.e., $NF \cdot B_i^{\min} \le NF_i \cdot \overline{B_i}$) or allocated more than its maximum required data rate (i.e., $NF_i \cdot \overline{B_i} \le NF \cdot B_i^{\max}$). Therefore, the same problem in Eq. (4.3) will be solved except that the bandwidth constraint changes to

$$NF \cdot B_i^{\min} \le NF_i \cdot \overline{B_i} \le NF \cdot B_i^{\max}, \forall i, 1 \le i \le K \tag{4.4}$$

As for packet scheduling, if the network operator wants to provide minimum bandwidth guarantees to some classes, the optimization problem in Eq. (4.3) may not have a feasible solution. This is because the bandwidth provisioning scheme may have to allocate a certain number of time frames to certain classes of traffic in order to satisfy their minimum bandwidth requirements even though they do not satisfy the opportunity

[14] To provide bandwidth guarantees, the bandwidth provisioning scheme must be supported by a CAC scheme in order to block users when there is not enough capacity to support their minimum bandwidth requirements. In this chapter, we focus only on bandwidth provisioning, and hence we only consider the basic case (Section 4.3.1) in our experiments. The case of bandwidth guarantees is considered in Chapter 5.

cost constraint. Therefore, to satisfy both constraints, the bound on opportunity cost (i.e., H_c) has to be dynamically computed in order to ensure the existence of a feasible solution of Eq. (4.3) as follows. Let:

- $\mathrm{Re}\, v_{\mathbf{K}^*} = \sum_{j \in \mathbf{K}^*} \overline{\mathrm{Re}\, v_i}$, where $\mathbf{K}^* \in \{1, 2, ..., K\}$ is the set of classes that require minimum bandwidth guarantees. That is, $\mathrm{Re}\, v_{\mathbf{K}^*}$ is the revenue the network operator will earn from serving these classes.

In this case, the opportunity cost of serving the classes in $\mathbf{K}^*$ is given by $OC_{\mathbf{K}^*}(NF) = Max\, \mathrm{Re}\, v_c - \mathrm{Re}\, v_{\mathbf{K}^*}$, where $Max\, \mathrm{Re}\, v_c$ is defined in Section 4.3.1. Therefore, to avoid infeasibility in Eq. (4. 3), we must have $H_c \geq OC_{\mathbf{K}^*}(NF)$. The network operator could, for example, set a predefined value for H_c, call it ϑ_c and use it only when $H_c \geq OC_{\mathbf{K}^*}(NF)$ is satisfied as follows

$$H_c = \begin{cases} OC_{\mathbf{K}^*}(NF), & if \quad \vartheta_c \leq OC_{\mathbf{K}^*}(NF) \\ \vartheta_c, & otherwise \end{cases} \tag{4.5}$$

4.3.3 Dynamic Weight Update Scheme

The weights in our bandwidth provisioning scheme determine the priority of each class, and hence they have a great impact on the frame allocation process and user satisfaction. In this section, we show how to update the weights to increase inter-class fairness of the

bandwidth provisioning scheme while maintaining a long-term service differentiation between them. Let:

- $U_i(t) = \dfrac{NF_i \cdot \overline{B_i^{efec}}}{NF \cdot B_i^{\max}} \triangleq$ utility of class *i* at the beginning of time *t*, where $\overline{B_i^{efec}}$ is the actual average data rate of class *i* (i.e., the effective rate class *i* was sending at during the previous frames) and *t* is the time at the end of the previous *NF* frames and the beginning of new ones (i.e., the beginning of a new bandwidth provisioning period). The higher the data rate assigned to class *i*, the higher its utility.
- $\overline{U_i(t)} = \alpha \cdot \overline{U_i(t-1)} + (1-\alpha) \cdot U_i(t) \triangleq$ average utility of class *i* at time *t*, computed as a moving average, where $0 \le \alpha \le 1$.
- $w_i(t) \triangleq$ weight of class *i* at time *t*.
- $\overline{w_i(t)} = \alpha \cdot \overline{w_i(t-1)} + (1-\alpha) \cdot w_i(t) \triangleq$ average weight of class *i*, where $0 \le \alpha \le 1$.
- $\{\mathbf{LA}^z(t)\}_{z=1}^{h_i} = \{LA^1(t), LA^2(t), ..., LA^{h_i}(t)\} \triangleq$ set of average utilities that are larger than the average utility of class *i* at time *t*, where h_i is the number of classes whose average utilities are larger than the average utility of class *i*.
- $\{\mathbf{LO}^q(t)\}_{q=1}^{l_i} = \{LO^1(t), LO^2(t), ..., LO^{l_i}(t)\} \triangleq$ set of average utilities that are lower than the average utility of class *i* at time *t*, where l_i is the number of classes whose average utilities are lower than the average utility of class *i*.

Three design features are taken into consideration in developing our weight update scheme. First, the weight of each class is gradually increased or decreased depending on its performance history and all other classes' performance histories. In particular, the weight of class *i* is increased or decreased depending on the difference between its average utility, the average utilities that are larger than it (i.e., set $\{\mathbf{LA}^z(t)\}_{z=1}^{h_i}$) and the average utilities that are smaller than it (i.e., set $\{\mathbf{LO}^q(t)\}_{q=1}^{l_i}$). To achieve this, the new weight of class *i* at time *t* is updated as follows

$$w_i(t-1)+\Delta w_i \tag{4.6}$$

where

$$\Delta w_i = \frac{\sum_{z\in\{\mathbf{LA}^z(t)\}_{z=1}^{h_i}} \left(LA^z - \overline{U_i(t)}\right) - \sum_{q\in\{\mathbf{LO}^q(t)\}_{q=1}^{l_i}} \left(\overline{U_i(t)} - LO^q\right)}{\sum_{i=1}^{K}\overline{U_i(t)}} \tag{4.7}$$

Note that Δw_i is the normalized difference between the average utilities that are larger than the average utility of class *i* and the average utilities that are less than it. Δw_i can be thought as a performance measure. It increases as the difference between the average utilities in $\{\mathbf{LA}^z(t)\}_{z=1}^{h_i}$ and average utility of class *i* increases, and it decreases as the difference between the average utility of class *i* and the average utilities in $\{\mathbf{Lo}^j(t)\}_{j=1}^{l_i}$

increases. Note that Δw_i is negative when $\sum_{z\in\{\mathbf{LA}^z(t)\}_{z=1}^{l_{\mathrm{lt}}}} \left(LA^z - \overline{U_i(t)}\right) < \sum_{q\in\{\mathbf{LO}^q(t)\}_{q=1}^{l_l}} \left(\overline{U_i(t)} - LO^q\right)$.

Δw_i is negative when the difference between class *i* and the classes of lower average utilities is higher than the difference between the classes of higher average utilities and class *i*. In this case, it is best to decrease the weight of class *i* to give a chance to classes of lower average utilities to be allocated more bandwidth, and hence increase inter-class fairness.

The second design feature is that the weights of lower priority classes are allowed to be temporarily higher than those of higher priority ones to further increase inter-class fairness. However, to ensure service differentiation between classes, we require that the ratio between the average weight of each class and the next class that has higher priority does not exceed a certain threshold $0 < \tau_i < 1$ (i.e., $\frac{\overline{w_i(t)}}{\overline{w_{i-1}(t)}} \le \tau_i$, where $\overline{w_1(t)} > \overline{w_2(t)} > ... > \overline{w_K(t)}$). This guarantees a long-term service differentiation between classes by ensuring that the long-term average weight of class *i* is less than or equal to $\tau_i \cdot \overline{w_{i-1}(t)}$.

An additional design feature is to restrict the weights to fall within a certain range as determined by the network operator (i.e., $W_{\min} \le w_i(t) \le W_{\max}$) in order to ensure that the weight update does not result in extremely high or low weight values.

Following our design features, the weight of each class is updated as follows

$$w_i(t) = \max\left(\min\left((w_i(t-1) + \Delta w_i), W_{\max}\right), W_{\min}\right) \tag{4.8}$$

That is, $w_i(t) = w_i(t-1) + \Delta w_i$ as long as $W_{\min} \le w_i(t-1) + \Delta w_i \le W_{\max}$. If $w_i(t-1) + \Delta w_i < W_{\min}$, then $w_i(t) = W_{\min}$. On the other hand, if $w_i(t-1) + \Delta w_i > W_{\max}$, then $w_i(t) = W_{\max}$. Note that Eq. (4.8) satisfies only the condition $W_{\min} \le w_i(t) \le W_{\max}$. Therefore, once $w_i(t)$ is computed, the condition $\frac{\overline{w_i(t)}}{\overline{w_{i-1}(t)}} \le \tau_i$ is checked. If it is not satisfied, then $w_i(t)$ is recomputed as follows

$$w_i(t) = \frac{\tau_i \cdot \left(\alpha \cdot \overline{w_{i-1}(t-1)} + (1-\alpha) \cdot w_{i-1}(t)\right) - \alpha \cdot \overline{w_i(t-1)}}{(1-\alpha)} \tag{4.9}$$

That is, $w_i(t)$ is computed such that $\frac{\overline{w_i(t)}}{\overline{w_{i-1}(t)}} = \tau_i$ as follows

$$\frac{\overline{w_i(t)}}{\overline{w_{i-1}(t)}} = \tau_i, \ \therefore \frac{\alpha \cdot \overline{w_i(t-1)} + (1-\alpha) \cdot w_i(t)}{\alpha \cdot \overline{w_{i-1}(t-1)} + (1-\alpha) \cdot w_{i-1}(t)} = \tau_i \tag{4.10}$$

Rearranging the terms

$$\alpha \cdot \overline{w_i(t-1)} + (1-\alpha) \cdot w_i(t) - \tau_i \cdot \left(\alpha \cdot \overline{w_{i-1}(t-1)} + (1-\alpha) \cdot w_{i-1}(t)\right) = 0 \tag{4.11}$$

Therefore,

$$w_i(t) = \frac{\tau_i \cdot \left(\alpha \cdot \overline{w_{i-1}(t-1)} + (1-\alpha) \cdot w_{i-1}(t)\right) - \alpha \cdot \overline{w_i(t-1)}}{(1-\alpha)} \tag{4.12}$$

4.3.4 Packet Scheduling

Once each class is allocated a number of frames, these frames will be distributed among users within each class according to the packet scheduling scheme, which is executed every time frame. These frames can be served in any order. For example, they could be served based on the delay or packet loss requirements of the different classes. In this thesis, the frames of the class with the highest priority are served first, then those of the class with the second highest priority and so forth. In the performance evaluation, which is the subject of next section, we adopt our packet scheduling scheme proposed in Chapter 3.

4.4 Performance Evaluation

In order to evaluate the performance of our proposed bandwidth provisioning and dynamic weight update schemes, we use the same simulation model, traffic model and channel model that we developed in Chapter 3. That is, we evaluate our schemes on HSDPA system, where we consider a single-cell scenario with Pedestrian A users [56]. In addition, we consider three different classes with four different types of services, namely VoIP (class 1), audio streaming (class 2), video streaming (class 2) and FTP (class 3).

Class 1 is assumed to have the highest priority and class 3 is assumed to have lowest priority. In addition, for demonstration purposes, we assume that p_{ij} = 6, 4, 2 and 1 units of money for VoIP, audio streaming, video streaming and FTP users, respectively. All the relevant simulation parameters are included in Appendix B.

The LP problem of Eq. (4.3) is solved using *lp-solve*, which is a free Linear/Integer Programming solver written in ANSI C programming language [72]. To access it in our Java-based simulator, we use the Java wrapper class provided by J. Ebert, which provides access to all the lp-solve API in Java format. Reference [72] provides detailed description on how to use this wrapper class in order to access the routines of lp-solve.

Unless otherwise specified, connection request arrivals are modeled as a Poisson process with a mean value of 0.5 connections per second. Users are uniformly distributed in the cell. Based on various simulation runs, we choose *NF*= 20 time frames (i.e. 20 x 2 ms) and we use a moving average to compute $\overline{R_i(t)}$ (i.e., $\overline{R_i(t)} = \alpha \cdot \overline{R_i(t-2)} + (1-\alpha) \cdot \overline{R_i(t-1)}$, with $\alpha = 0.99$ [12]). The simulation time step is one time frame, which is 2 ms in HSDPA [2], and the simulation time is 400s.

4.4.1 Test Cases and Performance Metrics

Three test cases are considered in our experiments. In the first case, we evaluate the performance of packet scheduling with and without dynamic bandwidth provisioning. This case is designed to show the advantage of using dynamic bandwidth provisioning along with packet scheduling. In the second case, we evaluate our dynamic bandwidth provisioning

scheme (with packet scheduling) under different fixed class weights and opportunity cost values for H_c in order to show their role in the bandwidth allocation process. In the third case, we evaluate our dynamic bandwidth provisioning scheme using our proposed weight update scheme. For this case, we set the minimum weight (i.e., $W_{\min}$) and maximum weight (i.e., $W_{\max}$) values to 1 and 10, respectively.

The following performance metrics are used to evaluate the performance of the proposed dynamic bandwidth provisioning and dynamic weight update schemes:

- Proportion of assigned frames ($\overline{P_i}$): the average ratio of assigned frames to class *i* to the total number of frames needed to satisfy its maximum data rate requirements.

- Service coverage: percentage of users who achieve their required QoS with a certain outage level. For audio streaming, a user's connection is dropped if his average packet loss (due to packet discarding, transmission errors and/or buffer overflow) exceeds 5% [65], [66] and [67]. For video streaming, a user's connection is dropped if his achieved average throughput is less than his minimum required rate. Finally, for FTP traffic, a user's connection is dropped if his achieved average throughput is less than 9.6 Kbps [13] and [16].

- Per-class weights: we report the temporal and average values of the dynamic weights per class, as well as the 10th and 90th percentile values, defined as the values, where 10% and 90% of measured weights are lower, respectively. These values are computed from the weights resulting from our weight update scheme.

4.4.2 Simulation Results

In this section, we show and discuss the simulations results for the three cases considered in our experiments.

Case 1: Scheduling with and without Dynamic Bandwidth Provisioning

Figures 4.2 (4.3) show percentage of service coverage for VoIP (audio streaming) before and after implementing our dynamic bandwidth provisioning scheme, with $w_1 = 6$, $w_2 = 4$ and $w_3 = 2$. As expected, the performance of VoIP and audio streaming is clearly improved when dynamic provisioning is implemented along with packet scheduling. This is because dynamic bandwidth provisioning aims at satisfying the users' bandwidth requirements over longer time intervals than packet scheduling, which only works in very small time intervals (i.e., single time frames). In other words, dynamic bandwidth provisioning improves the management of network resources, which results in more users meeting their minimum bandwidth requirements, hence, improving the overall service coverage. In addition, the figures show that, as the arrival rate to the system increases, the performance difference between the case of packet scheduling only and the case of packet scheduling with dynamic bandwidth provisioning increases. The reason for this is that at high arrival rates, more users compete for resources, and hence the performance difference between different bandwidth management techniques is clearly revealed.

The service coverage of video streaming is improved with dynamic bandwidth provisioning as shown in Figure 4.4. Such improvement is observed at arrival rates between

0.1 and 0.7 connections per second. At higher arrival rates, however, the service coverage of video streaming degrades to values below those of packet scheduling only. This is because, at high arrival rates, the demand for bandwidth is high and it is difficult to satisfy the bandwidth requests of all classes of traffic. In such a case, more time frames are allocated to VoIP and audio streaming than video streaming, since the former have higher priority. This leaves fewer frames for lower priority traffic, which negatively affects their performance in terms of service coverage. We remark, however, that the network operator would typically employ a CAC scheme in order to improve the satisfactions of users and prevent performance degradation at high arrival rates. In Chapter 5, we investigate the effect of utilizing CAC on the service coverage.

Similarly, the service coverage of FTP traffic is slightly lower with dynamic bandwidth provisioning as shown in Figure 4.5, since class 3 is assigned the lowest priority. Therefore, more time frames are allocated to classes 1 and 2 at the expense of class 3. The performance of class 3 can certainly be improved by increasing its priority as discussed in the next subsequent sections.

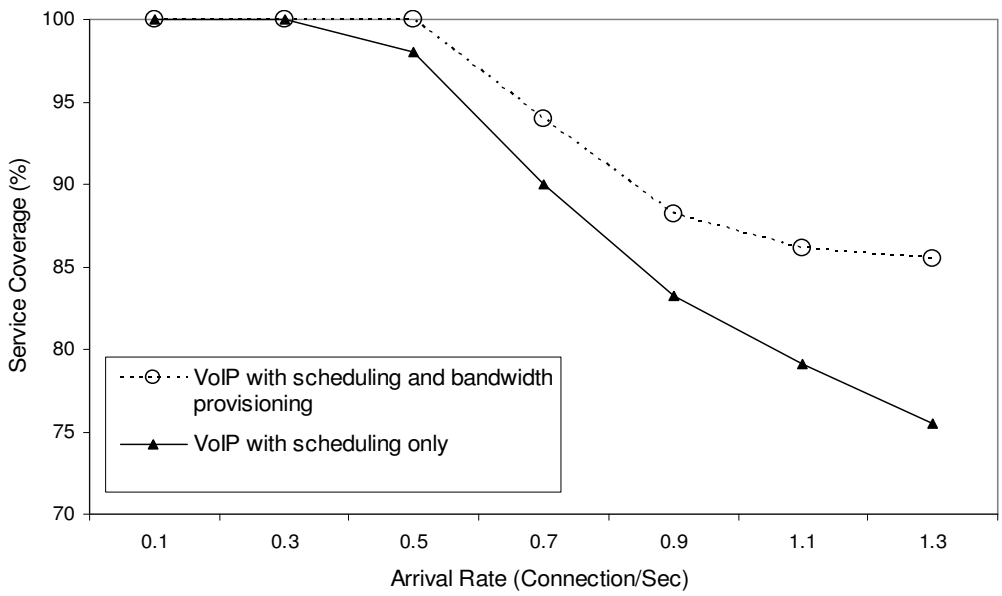


Figure 4.2: Service coverage for VoIP with/without bandwidth provisioning

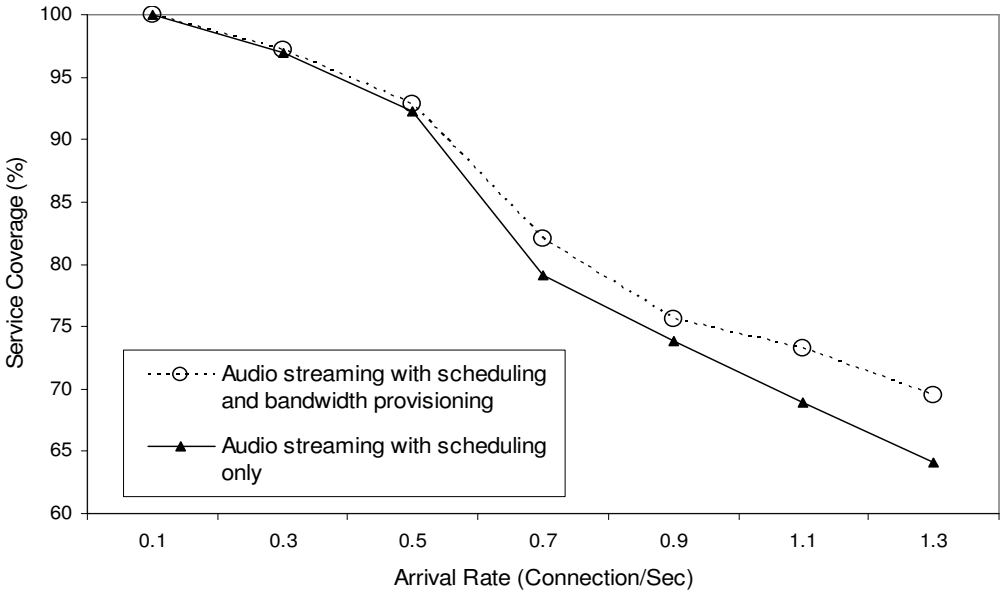


Figure 4.3: Service coverage for audio streaming with/without bandwidth provisioning

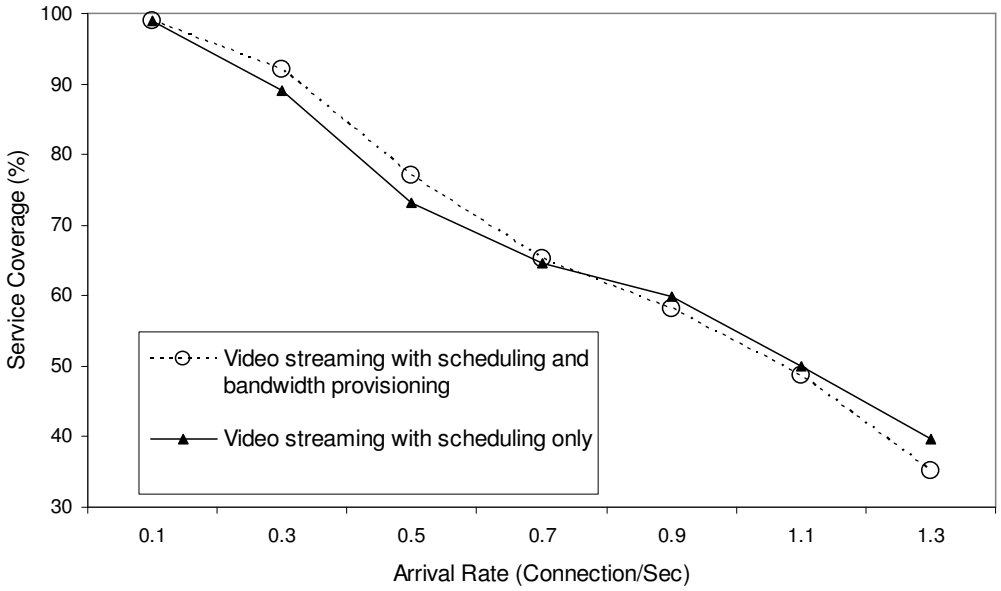


Figure 4.4: Service coverage for video streaming with/without bandwidth provisioning

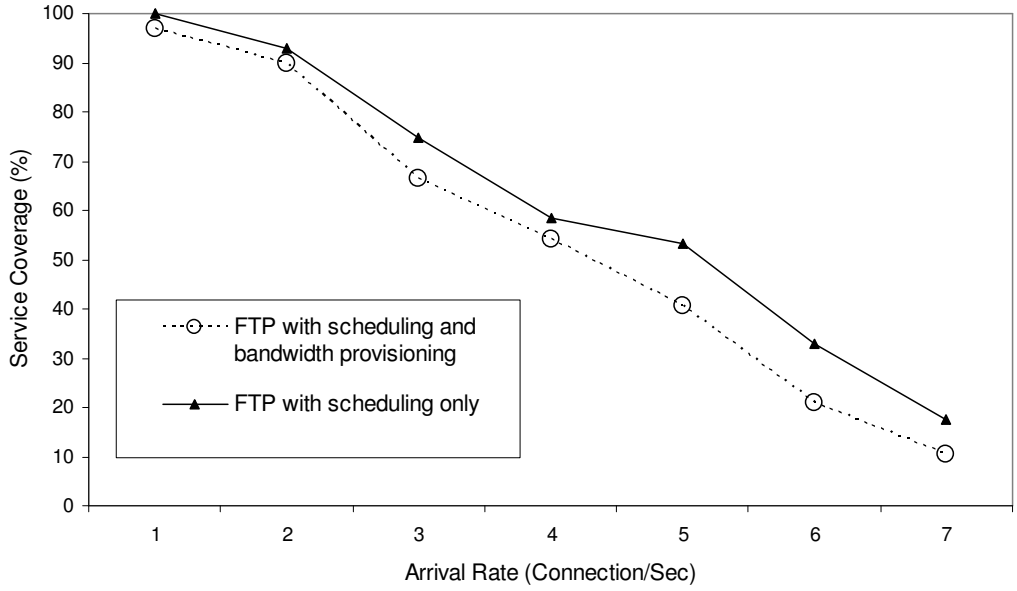


Figure 4.5: Service coverage for FTP with/without bandwidth provisioning

Case 2: Dynamic Bandwidth Provisioning with Fixed Weights

Table 4.1 shows the proportion of assigned frames for each traffic class with the corresponding fixed weights. Note that the proportion of assigned frames for class 3 can be increased by increasing its priority (through increasing its weight and decreasing the weights of higher priority classes). However, this occurs at the expense of decreasing the proportion of assigned frames for classes 1 and 2. In addition, the effect of opportunity cost of bandwidth provisioning can be controlled by controlling H_c, where we let $H_c = \zeta_c \cdot Max \operatorname{Re} v_c$. When $\zeta_c = 1$, this implies that the network operator can tolerate a revenue loss as high as the maximum revenue that could be obtained. That is, in this case, the opportunity cost of bandwidth provisioning is ignored. However, as ζ_c is decreased, then the network operator can tolerate less revenue loss, and thus more frames are given to the highest-revenue-generating classes (i.e., higher priority classes) as shown in Table 4.2. When $\zeta_c = 0$, then the network operator cannot tolerate any revenue loss, and hence only the classes that have the maximum revenue (i.e., $Max \operatorname{Re} v_c$ in Eq. (4.2)), which are classes 1 and 2 are assigned frames. Therefore, the network operator can choose the level at which it can tolerate revenue loss as a result of bandwidth provisioning by controlling ζ_c.

Table 4.1: Proportion of assigned frames with different fixed weights

w_1	w_2	w_3	$\overline{P_1}$	$\overline{P_2}$	$\overline{P_3}$	ζ_c
7	5	1	100%	62.1%	9.6%	1
6	4	2	100%	59.4%	16.5%	1
5	4	3	95.2%	51.7%	28.3%	1
1	1	1	91.7%	37.2%	45.8%	1

Table 4.2: Proportion of assigned frames with different opportunity cost values

w_1	w_2	w_3	$\overline{P_1}$	$\overline{P_2}$	$\overline{P_3}$	ζ_c
1	1	1	91.7%	37.2%	45.8%	1
1	1	1	97.1%	50.3%	31.2%	0.66
1	1	1	99.4%	59.1%	10.9%	0.33
1	1	1	100%	66.6%	0%	0

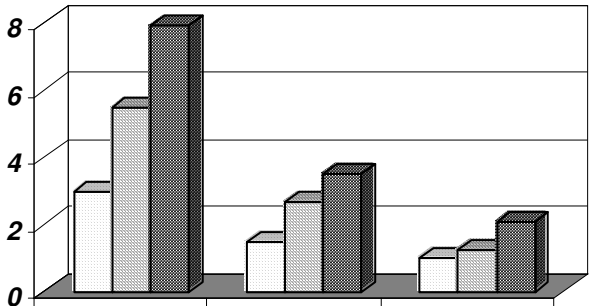
Case 3: Dynamic Bandwidth Provisioning with Dynamic Weights

The proportion of assigned frames for each class in case of dynamic weights is shown in Table 4.3. The weight ranges are chosen between 1 and 10 (i.e., $W_{\min} = 1$ and $W_{\max} = 10$). The importance of the weight update scheme is that it allows service differentiation between classes while at the same time it ensures inter-class fairness. The resulting fairness is more adaptive to the performance of classes since it is based on their performance history. Therefore, inter-class fairness can be better achieved using this scheme instead of setting fixed weights. The network operator can achieve different fairness levels by controlling τ_i, where small τ_i values result in less fairness. This is not possible with fixed weights since the performance of each class is not fixed due to the varying bandwidth requirements and channel quality conditions.

Table 4.3: Proportion of assigned frames with dynamic weights

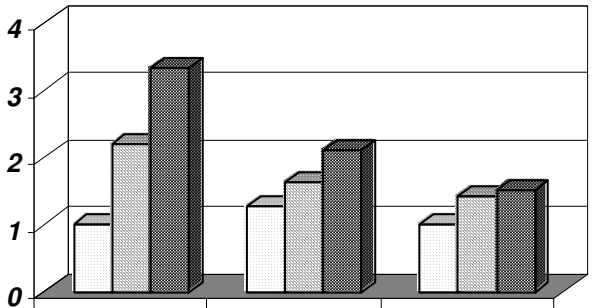
W_{min}	W_{max}	$\overline{P_1}$	$\overline{P_2}$	$\overline{P_3}$	ζ	τ_i
1	10	97.5%	63.4%	8.3%	1	0.5
1	10	86.8%	53.9%	28.7%	1	0.75
1	10	59.8%	42.3%	45.6%	1	1

The role of τ_i in controlling inter-class fairness is shown in Figures 4.6, 4.7 and 4.8, which depict the 10th, average and 90th percentile of the dynamic weights of each class for τ_i = 0.5, 0.75 and 1, respectively. The figures show that by increasing τ_i, the dynamic weight values for different classes are allowed to get closer to each other, hence, improving inter-class fairness. This behavior is also confirmed in Figures 4.9, 4.10 and 4.11, which show the instantaneous weights during the simulation time with τ_i = 0.5, 0.75 and 1 for all classes. The figures show that the instantaneous weights of low priority classes could be temporarily higher than those of higher priority classes. However, there is a clear separation, on average, between the weight of each class and that of the class of higher priority. This separation is due to the long-term service differentiation between classes that is achieved through the condition $\frac{\overline{w_i(t)}}{\overline{w_{i-1}(t)}} \leq \tau_i$.



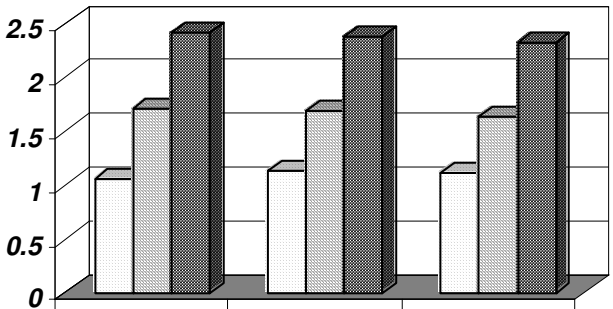
	class 1	class 2	class 3
10th percentile	2.97	1.49	1
avg	5.43	2.68	1.23
90th percentile	7.9	3.5	2.05

Figure 4.6: 10th, average and 90th percentile of dynamic weights with $\tau_i = 0.5$



	class 1	class 2	class 3
10th percentile	1	1.28	1
avg	2.19	1.62	1.42
90th percentile	3.32	2.1	1.5

Figure 4.7: 10th, average and 90th percentile of dynamic weights with $\tau_i = 0.75$



	class 1	class 2	class 3
10th percentile	1.05	1.13	1.11
avg	1.7	1.68	1.63
90th percentile	2.41	2.38	2.32

Figure 4.8: 10th, average and 90th percentile of dynamic weights with $\tau_i = 1$

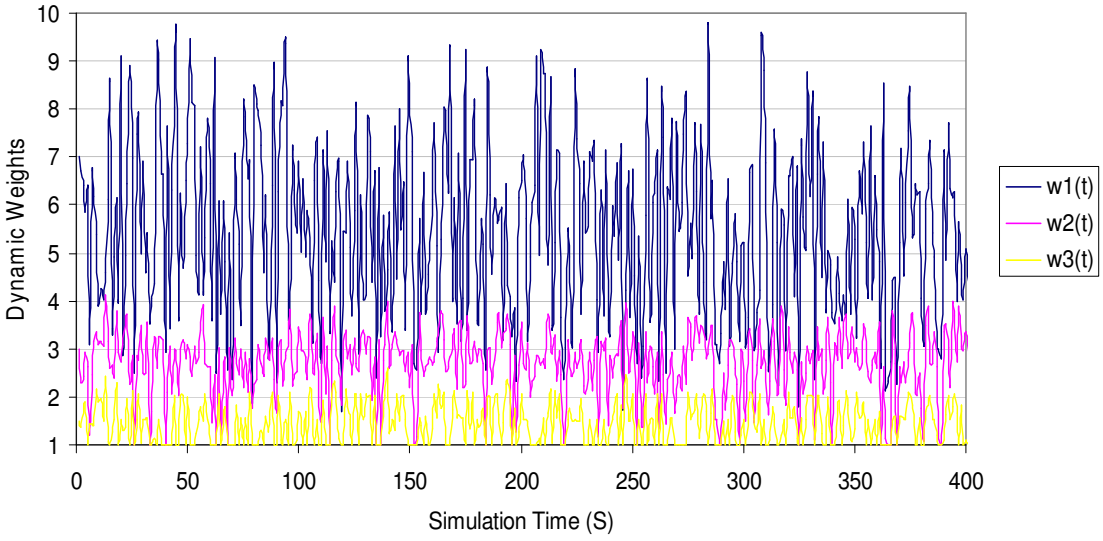


Figure 4.9: The dynamic weights with $\tau_i = 0.5$

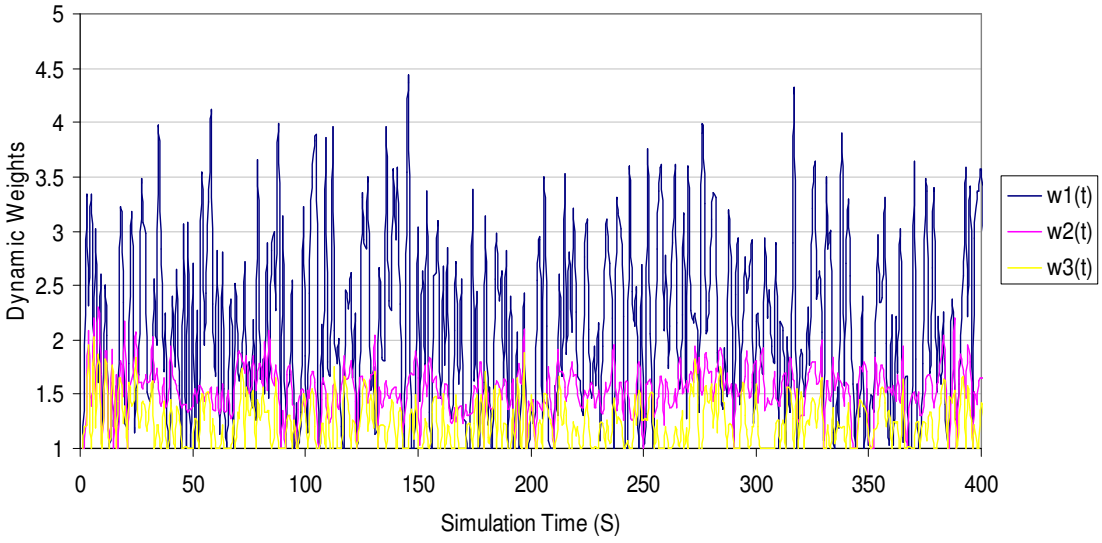


Figure 4.10: The dynamic weights with $\tau_i = 0.75$

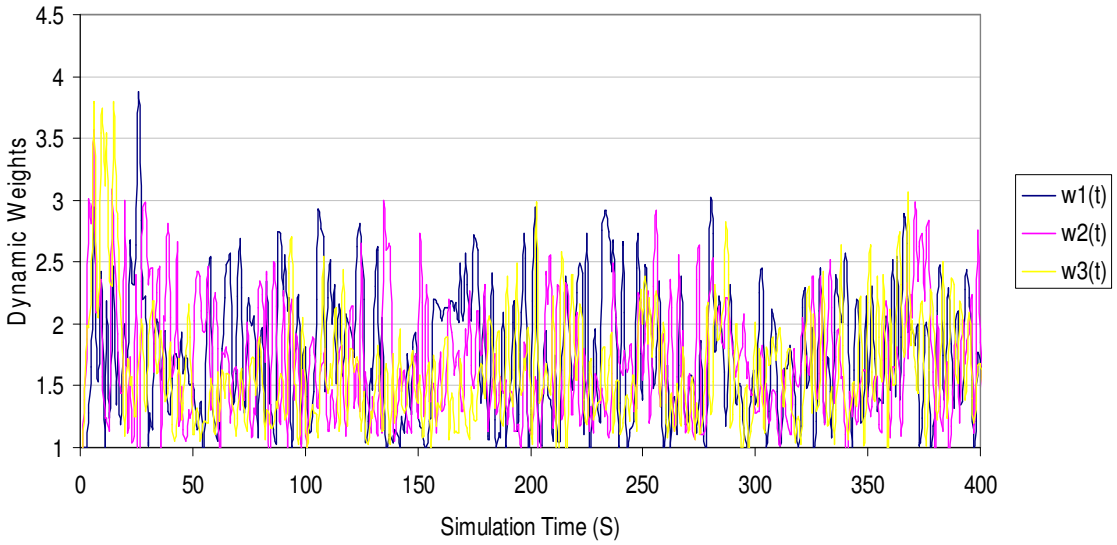


Figure 4.11: The dynamic weights with $\tau_i = 1$

4.5 Summary

This chapter has presented a novel bandwidth provisioning scheme for BWASs. The proposed scheme aims at providing efficient bandwidth management at the class level. Specifically, it spans multiple time frames and optimally decides how they are allocated among different classes of traffic based on their weights, the bandwidth requirements, the channel quality conditions of their users, and the expected revenue loss from each class. The scheme supports different classes of traffic supporting multiple users with different bandwidth requirements. It also incorporates a class-level opportunity cost function to bound the cost of allocating bandwidth to different classes to maintain a certain revenue level to the network operator. To maximize inter-class fairness, a weight update scheme is integrated with the bandwidth provisioning scheme to dynamically configure the weights of different classes based on their performance history to achieve a certain level of fairness as desired by the network operator. Simulation results show that the dynamic bandwidth provisioning scheme can improve the performance of packet scheduling in BWASs. This because the dynamic bandwidth provisioning scheme further improves the operation of bandwidth management as it considers the longer-term bandwidth requirements of users during their connections compared to packet scheduling, which only works in short time intervals. Results also confirm the effectiveness of the dynamic weight update scheme in improving inter-class fairness while ensuring service differentiations between different classes.

Chapter 5

Call Admission Control-based Dynamic Pricing Scheme

Chapters 3 and 4 introduced the packet scheduling and dynamic bandwidth provisioning schemes. These two schemes, however, are post-admission bandwidth management strategies, and therefore, cannot provide QoS guarantees to users, especially during congestion periods when the demand for bandwidth exceeds the supply. To provide such guarantees, admission-level bandwidth management is needed to limit the number of admitted user connections to maximize the utilization of the wireless network while maintaining the QoS of ongoing user connections at acceptable levels. This is typically achieved through employing a Call Admission Control (CAC) scheme. As explained in Chapter 2, although CAC is efficient in sustaining the packet-level QoS of ongoing user connections at acceptable levels amid congestion periods, it lacks the mechanism to

provide incentives to users to regulate their demand for wireless services. Hence, the connection blocking probabilities can reach high levels during congestion periods leading to user dissatisfaction and potential revenue loss. To overcome this problem, recent research proposals have been directed towards integrating CAC with dynamic pricing, so that the prices of the wireless network services are dynamically determined based on the load of the network. Dynamic pricing can certainly send the right signals to users to regulate their demand for wireless services. It encourages them to restrain their demand when the network is congested and increase their demand when it is underutilized. Dynamic pricing, therefore, is a promising solution to traffic control problems, which can help alleviate the problem of congestion and provide efficient bandwidth management.

In this chapter, we introduce our CAC-based dynamic pricing scheme that aims at providing efficient bandwidth management at the admission level. We show how it can be efficiently used to guarantee a congestion-free system given that user demand models are accurate in predicting their reaction to prices. We also extend our scheme to support different types of pricing mechanisms and examine the effect of inaccurate user demand models on the overall performance of the network.

Our scheme generalizes and enhances the work in [40], which was discussed in Chapter 2, in the following ways:

1) The scheme in [40] considers one type of connections with each user requesting one channel to use. Our scheme, on the other hand, considers multiple classes of traffic with users having multiple bandwidth requirements, which makes it more suitable for BWASs.

2) The scheme in [40] is designed to prevent congestion only, and hence dynamic pricing is only applied when the network is congested, where a flat rate pricing is assumed when the network is underutilized. Therefore, the scheme in [40] does not provide incentives for users to increase their demand for wireless services when the network is underutilized. Our proposed scheme, however, employs dynamic pricing during all network conditions (i.e., whether the network is underutilized or congested). This way, our scheme can maximize the utilization of BWASs when these systems are underutilized and prevent congestion when they are overloaded.
3) Our scheme is general and considers different variations of dynamic pricing (e.g., dynamic differentiated pricing, minimum price values, etc). In fact, the scheme in [40] can be considered as a special case of our scheme.
4) Dynamic prices in the scheme in [40] are computed based on assumptions about the users' utilities and not on the amount of available network bandwidth. The scheme, therefore, is incapable of capturing the dynamics of the network (i.e., changes in available bandwidth) and the varying bandwidth requirements of users. This explains the inability of the scheme to achieve zero connection blocking probabilities in spite of assuming an accurate user demand model. On the other hand, in our scheme prices are computed dynamically based on the amount of available bandwidth in the network. As a result, our scheme is shown to achieve zero connection blocking probabilities if an accurate user demand model is assumed.

The rest of this chapter is organized as follows. Section 5.1 provides an overview of our proposed CAC-based dynamic pricing scheme and discusses its objectives. Section 5.2 describes the system model. Section 5.3 presents our proposed CAC-based dynamic pricing scheme. Section 5.4 presents the performance evaluation of our proposed scheme. Section 5.5 summarizes the chapter.

5.1 Scheme Outline and Objectives

The main contribution of this chapter is the CAC-based dynamic pricing scheme. The proposed scheme is to be implemented at the base stations of BWASs or at any centralized component in these systems, where CAC is performed (e.g., the Radio Network Controller (RNC) in UMTS and HSDPA [2]). The main objectives of the proposed scheme are:

1) Supporting different classes of traffic, where each class can include different types of services each having its own bandwidth requirements;
2) Maximizing the utilization of the wireless network resources;
3) Preventing congestion; and
4) Supporting fairness between different types of services.

Our scheme consists of three components, namely a monitoring component, a CAC component and a dynamic pricing component as shown in Figure 5.1. Our scheme works as follows. At the end of the current time window and the beginning of new one, where the length of the time window is determined by the network operator, the monitoring component measures the amount of available bandwidth (i.e., unutilized bandwidth). If

the amount of available bandwidth is different from the one measured in the previous time window due to connection completion or new admitted connections, the monitoring component triggers the CAC component. The CAC component then computes the optimal number of new connection requests for each service within each class of traffic that would maximize the utilization of the new available bandwidth in the system and achieve certain fairness levels between different classes of traffic. The actual numbers of new connection requests for each service are, however, different from the optimal ones determined by the CAC component. In this case, the dynamic pricing component dynamically determines the prices of units of bandwidth for each service based on user demands to force the actual numbers of new connection requests during the new time window to be less than or equal to the optimal ones. The dynamic prices are computed independently from the optimal numbers of new connection requests. This simplifies the implementation of our scheme and provides network operators the flexibility to use different CAC and user demand functions without affecting the computation of prices.

As explained in Chapter 2, handoff connections are not affected by dynamic pricing when charging at admission level, since they were already charged at the cell where the connections were initiated. In this case, our scheme always admits handoff connections as long as there is enough bandwidth to support them. To prioritize such connections over new connections, the network operator can use a form of Guard Channel schemes in which a certain amount of bandwidth is exclusively reserved for handoff connections in order to maintain the handoff connection dropping blocking probability below a certain threshold [73], [74], [75], [76] and [77].

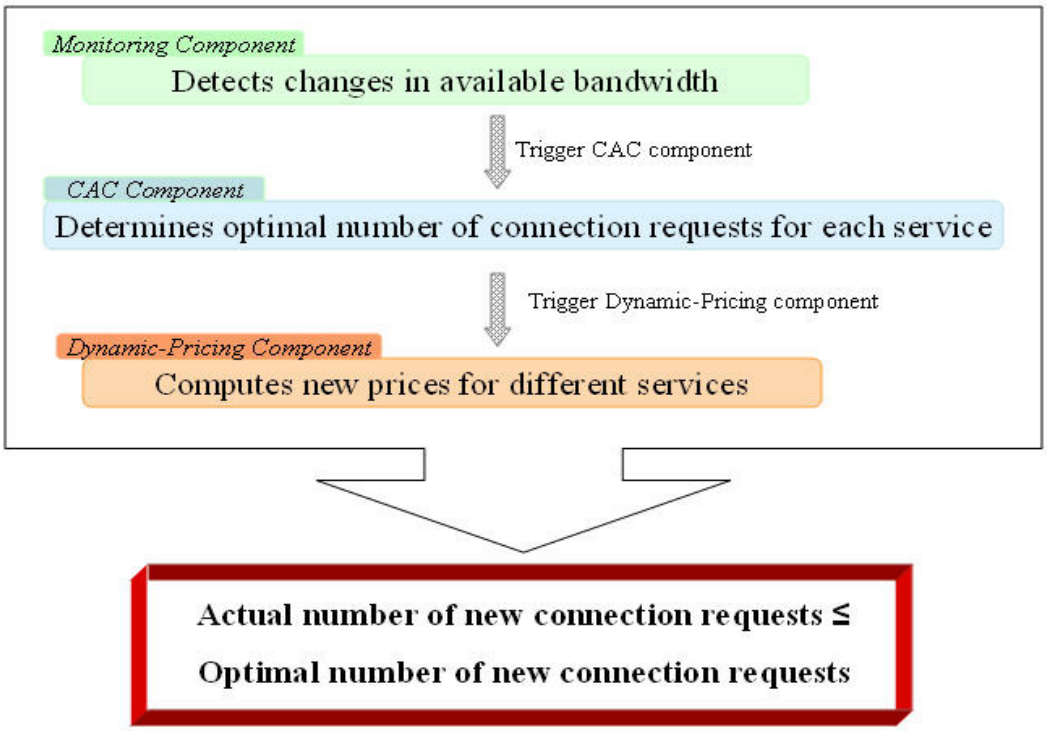


Figure 5.1: Components of CAC-based dynamic pricing scheme

5.2 System Model

Similar to the previous two chapters, we consider a BWAS consisting of a downlink time-slotted channel divided into time frames. Data transmission is done at the base station on a frame by frame basis. We also consider K classes of traffic, where class i has higher priority than class i+1. In addition we consider that each class includes a number of services, where service z of class i requires b_i^z units of bandwidth and service z has higher priority than service z+1. For example, the streaming class can include audio streaming and video streaming services each requesting different amounts of bandwidth.

Therefore, users within each class can have different bandwidth requirements depending on the services they are requesting.

5.3 CAC-based Dynamic Pricing Scheme

In this section, we describe each component of our proposed bandwidth management scheme. We first start with explaining each component in our scheme. Then we show how the dynamic pricing component is extended to support minimum price values and differentiated pricing. Next, we discuss the reduction in the optimal demand due to the use of differentiated pricing.

5.3.1 Components of CAC-based Dynamic Pricing Scheme

The monitoring component is simple. Its main function is triggering the CAC component if it detects a change in the available bandwidth at the beginning of the new time window. Most of the computation is done at the CAC and dynamic pricing components. Before proceeding with describing these two components, we make the following definitions. Let:

- $W_t \triangleq$ index of next time window.
- $WL \triangleq$ length in units of time of the next time window.
- $NS_i \triangleq$ total number of services in class i.
- $N_i^z \triangleq$ number of admitted users that request service z in class i.

- $N_i = \sum_{z=1}^{NS_i} N_i^z \triangleq$ number of admitted users in class *i*.
- $C \triangleq$ system capacity.
- $b_i^z \triangleq$ bandwidth request per unit of time of service *z* in class *i*.
- $B_{free}(W_t) \triangleq$ total available bandwidth in the next time window.
- $\eta_i^z(W_t) \triangleq$ number of connection requests for service *z* in class *i* in the next time window. Therefore, the maximum total demand of bandwidth by class *i* in the next time window is equal to $\left(\sum_{z=1}^{NS_i} b_i^z \cdot N_i^z\right) \cdot WL + \sum_{z=1}^{NS_i} \eta_i^z(W_t) \cdot \left(b_i^z \cdot WL\right)$, where $\left(\sum_{z=1}^{NS_i} b_i^z \cdot N_i^z\right) \cdot WL$ is the demand of class *i* already admitted user connections and $\sum_{z=1}^{NS_i} \eta_i^z(W_t) \cdot \left(b_i^z \cdot WL\right)$ is the maximum demand of new incoming users provided that they are admitted to the system.
- $\{\boldsymbol{\eta}_i\}_{i=1}^{K} = \left\{\{\boldsymbol{\eta}_1^z\}_{z=1}^{NS_1}, \{\boldsymbol{\eta}_2^z\}_{z=1}^{NS_2}, ..., \{\boldsymbol{\eta}_K^z\}_{z=1}^{NS_k}\right\} \triangleq$ vector of connection requests for each service in each class, where $\{\boldsymbol{\eta}_i^z\}_{z=1}^{NS_i} = \{\eta_i^1(W_t), \eta_i^2(W_t), ..., \eta_i^{NS_i}(W_t)\}$.
- $\eta_{total} \triangleq$ total number of users who could make connection requests in the next time window. η_{total} is equal to the total number of admitted users subtracted from the total number of users that could make connection requests[15] at the cell, where dynamic pricing is implemented.

[15] Total number of users that could make connection requests could be determined by calculating the number of network operator's subscribers residing in the cell. This can be easily determined given that

- $p_i^z(W_t) \triangleq$ price in terms of units of money per unit of bandwidth, which is charged to users requesting service *z* in class *i* during the next time window.
- $A_i^z \triangleq$ ratio of users (of the total number of users) who have sufficient *Willingness to Pay* (WTP) to make connection requests to service *z* of class *i*. Clearly, A_i^z is a function of the price (i.e., $A_i^z = f_i^z\left(p_i^z(W_t)\right) \rightarrow [0,1]$, where $p_i^z(W_t) = f_i^z\left(A_i^z\right)^{-1}$). It is reasonable to assume that A_i^z is a monotonically decreasing function of the price. A_i^z can be constructed from the system's history by observing users' responses to changes in the price. It should be noted that the computation of A_i^z is a pure economic topic that is outside the scope of this thesis. However, we utilize a well-known demand function in Section 5.4.2 to model A_i^z, although our scheme can work with any function for A_i^z as explained next.

The main objective of our CAC component is to find the optimal number of connection requests for each service in each class in the next time window so that the utilization of available bandwidth is maximized. To achieve this objective, the CAC component will solve the following optimization problem

---

mobile devices continuously communicate with the base stations that are covering the areas where the users are.

$$\text{Objective: } \max_{\{\boldsymbol{\eta}_i\}_{i=1}^{K}} \sum_{i=1}^{K}\sum_{z=1}^{NS_i} \eta_i^z(W_t)\cdot\left(b_i^z \cdot WL\right)$$

$$\text{Subject to: } \sum_{i=1}^{K}\sum_{z=1}^{NS_i} \eta_i^z(W_t)\cdot\left(b_i^z \cdot WL\right) \le B_{free}(W_t),$$

$$\sum_{i=1}^{K}\sum_{z=1}^{NS_i} \eta_i^z(W_t) \le \eta_{total},$$

$$\left(b_i^z \cdot N_i^z \cdot WL + \eta_i^z(W_t)\cdot\left(b_i^z \cdot WL\right)\right)/\left(C.WL\right) \le \upsilon_i^z, \forall z, 1 \le z \le NS_i, \text{ and}$$

$$\eta_i^z(W_t) \in \mathbf{Z}^+, \forall z, 1 \le z \le NS_i \tag{5.1}$$

where $0 \le \upsilon_i^z \le 1$, $\sum_{i=1}^{K}\sum_{z=1}^{NS_i} \upsilon_i^z = 1$ and $\mathbf{Z}^+$ is the set of positive integers. The first constraint ensures that the maximum demand of all classes in the next time window does not exceed the total available bandwidth (i.e., supply). The second constraint ensures that the resulting total number of connection requests to the system (i.e., $\sum_{i=1}^{K}\sum_{z=1}^{NS_i} \eta_i^z(W_t)$) is realistic and does not exceed the total number of subscribers. The last constraint is used to ensure fairness among different services, and hence ensure fairness among different classes of traffic by restricting each service's share of the total bandwidth (i.e., the average bandwidth of admitted users requesting that service plus the bandwidth of new user connections) not to exceed a predefined ratio (υ_i^z) determined by the network operator. For example, to achieve absolute fairness (i.e., an equal bandwidth share for each service) between different services, υ_i^z should be set to $1/\sum_{i=1}^{K} NS_i$. Besides ensuring fairness, the second constraint can be used to promote certain services or increase revenues by

assigning more bandwidth to services that are expected to yield higher revenues (e.g., services belonging to higher priority classes). It should be noted that the objective function and the constraints in Eq. (5.1) do not include the connection blocking probabilities of users. This is because our pricing component, as described below, can guarantee to force the actual number of connection requests in the next time window to be less than or equal to the optimal ones computed in Eq. (5.1). Hence, the system is guaranteed to be congestion-free. In addition, the objective function and the constraints in Eq. (5.1) are linear and the optimal number of users is integer. Hence, the optimal numbers of connection requests $\{\boldsymbol{\eta}_i^*\}_{i=1}^{K} = \left\{\{\boldsymbol{\eta}_1^{z^*}\}_{z=1}^{NS_1}, \{\boldsymbol{\eta}_2^{z^*}\}_{z=1}^{NS_2}, ..., \{\boldsymbol{\eta}_K^{z^*}\}_{z=1}^{NS_k}\right\}$ can be found using Integer Linear Programming (ILP) techniques.

The actual numbers of connection requests during the next *WL* window before dynamic pricing is implemented can, however, be different from the optimal ones computed by our CAC component (i.e., $\{\boldsymbol{\eta}_i\}_{i=1}^{K} = \left\{\{\boldsymbol{\eta}_1^{z}\}_{z=1}^{NS_1}, \{\boldsymbol{\eta}_2^{z}\}_{z=1}^{NS_2}, ..., \{\boldsymbol{\eta}_K^{z}\}_{z=1}^{NS_k}\right\} \neq \{\boldsymbol{\eta}_i^*\}_{i=1}^{K} = \left\{\{\boldsymbol{\eta}_1^{z^*}\}_{z=1}^{NS_1}, \{\boldsymbol{\eta}_2^{z^*}\}_{z=1}^{NS_2}, ..., \{\boldsymbol{\eta}_K^{z^*}\}_{z=1}^{NS_k}\right\}$). Therefore, the dynamic pricing component will adjust the prices of units of bandwidth for each service in each class so that the actual number of connection requests are less than or equal to the optimal ones computed in Eq. (5.1) (i.e., $\{\boldsymbol{\eta}_i\}_{i=1}^{K} = \left\{\{\boldsymbol{\eta}_1^{z}\}_{z=1}^{NS_1}, \{\boldsymbol{\eta}_2^{z}\}_{z=1}^{NS_2}, ..., \{\boldsymbol{\eta}_K^{z}\}_{z=1}^{NS_k}\right\} \leq \{\boldsymbol{\eta}_i^*\}_{i=1}^{K} = \left\{\{\boldsymbol{\eta}_1^{z^*}\}_{z=1}^{NS_1}, \{\boldsymbol{\eta}_2^{z^*}\}_{z=1}^{NS_2}, ..., \{\boldsymbol{\eta}_K^{z^*}\}_{z=1}^{NS_k}\right\}$) as follows. We know from the actual number of connection requests to service z in class i during the next time window (i.e., $\eta_i^z(W_t)$) that it constitutes the following ratio of the total users that could request the service

$$\frac{\eta_i^z(W_t)}{\eta_{total}} \tag{5.2}$$

From Eq. (5.2) we know that the ratio of users that have sufficient WTP to make connection requests for service z in class i is at least $\frac{\eta_i^z(W_t)}{\eta_{total}}$ (there could be other users who have sufficient WTP, but choose not to make connection requests in the next time window). However, the optimal ratio should equal to

$$\frac{\eta_i^{z^*}(W_t)}{\eta_{total}} \tag{5.3}$$

Therefore, to achieve this optimal ratio, the price of service z in class i should be computed such that

$$A_i^z = f_i^z\left(p_i^z(W_t)\right) = \frac{\eta_i^{z^*}(W_t)}{\eta_{total}}, \forall z, 1 \le z \le NS_i \tag{5.4}$$

There are two cases, and hence two implications for price setting. The first case occurs during congestion periods when the numbers of connection requests (before dynamic pricing is implemented) typically exceed the optimal ones. According to Eq. (5.4), the prices are increased so that $A_i^z = \frac{\eta_i^{z^*}(W_t)}{\eta_{total}} \Rightarrow \frac{\eta_i^z(W_t)}{\eta_{total}} = \frac{\eta_i^{z^*}(W_t)}{\eta_{total}}$. In this case, if A_i^z is accurate in modeling the users' WTP, then the ratio of incoming users who have

sufficient WTP to make connection requests is guaranteed to equal the optimal ratio. The second case occurs during network underutilization periods when the number of connection requests (before dynamic pricing is implemented) is typically lower than the optimal ones. According to Eq. (5.4), prices are lowered so that $A_i^z = \frac{\eta_i^{z^*}(W_t)}{\eta_{total}} \Rightarrow \frac{\eta_i^z(W_t)}{\eta_{total}} \le \frac{\eta_i^{z^*}(W_t)}{\eta_{total}}$. In this case, the prices are lowered so that enough users have sufficient WTP to make connection requests. It is imperative to point out that users with sufficient WTP may not actually make connection requests in the next time window depending on their preferences. Using our scheme they are, however, encouraged to make such requests due to low prices. In this case, the incoming number of connection requests is guaranteed to be less than or equal to the optimal ratio.

Based on the above discussion and from Eq. (5.4), the dynamic pricing component will set the new prices to class *i* services as follows

$$\{\mathbf{p}_i^z(W_t)\}_{z=1}^{NS_i} = \left\{ f_i'^1\left(\frac{\eta_i^{1^*}(W_t)}{\eta_{total}}\right), f_i'^2\left(\frac{\eta_i^{2^*}(W_t)}{\eta_{total}}\right), ..., f_i'^{NS_i}\left(\frac{\eta_i^{NS_i^*}(W_t)}{\eta_{total}}\right) \right\} \qquad (5.5)$$

where $f_i'^z$ is the inverse function of f_i^z. As explained in Chapter 2, the dynamic prices for different services can be announced to users after they make connection requests and before they are admitted into the system or they can be broadcasted to them periodically. For example, the prices can be broadcasted to users at the beginning of each time window.

Note that the price equation, Eq. (5.5), is computationally inexpensive and is independent of the objective function in Eq. (5.1). Such independence allows the network operator to use different objective functions, if desired, in the CAC component without affecting the computations of prices and vice versa. In addition, based on the aforementioned discussion, the actual numbers of connection requests are guaranteed to be less than or equal to the optimal values computed in Eq. (5.1). Thus, using our pricing scheme, the system is guaranteed to be congestion free. It is imperative to point out that, in our scheme, all connection requests are accepted, since as mentioned earlier, our scheme guarantees that the amount of bandwidth required by new incoming connections will not exceed the amount of available bandwidth.

It is imperative to point out that pricing functions in general are designed to cover the costs and make profits. Since dynamic pricing in theory may lower the price below the profit margins of the network operators, it is essential that dynamic pricing supports minimum price values. In other words, dynamic pricing should not lower the price below a certain minimum value determined by the network operator in order to ensure its viability. In addition, network operators typically maintain price differentiation between different services, assigning higher prices for higher priority services. For instance, VoIP is usually charged at a higher rate than FTP. Therefore, such a case should also be supported. In next subsections, we show how these two cases can be supported by our dynamic pricing component.

5.3.2 Dynamic Pricing with Minimum Price Values

Suppose that the network operator wants to maintain the price for each service above a certain predetermined value ($p_i^{z,\min}$) in order to cover its costs and make profits. That is,

$$p_i^z(W_t) \geq p_i^{z,\min}, \forall z, 1 \leq z \leq NS_i \tag{5.6}$$

From Eq. (5.4) and Eq. (5.6) we get

$$p_i^z(W_t) \geq \left(f_i'^z\left(\frac{\eta_i^{z^*}(W_t)}{\eta_{total}} \right) , p_i^{z,\min} \right), \forall z, 1 \leq z \leq NS_i \tag{5.7}$$

Therefore, the new prices to class *i* services should be set as follows

$$\{\mathbf{p}_i^z(W_t)\}_{z=1}^{NS_i} = \left\{ \max\left(f_i'^1\left(\frac{\eta_i^{1^*}(W_t)}{\eta_{total}} \right) , p_i^{1,\min} \right), \max\left(f_i'^2\left(\frac{\eta_i^{2^*}(W_t)}{\eta_{total}} \right) , p_i^{2,\min} \right), \right.$$
$$\left. ..., \max\left(f_i'^{NS_i}\left(\frac{\eta_i^{NS_i^*}(W_t)}{\eta_{total}} \right) , p_i^{NS_i,\min} \right) \right\} \tag{5.8}$$

5.3.3 Dynamic Differentiated Pricing

In this case, suppose that the network operator wants to differentiate between different services by requesting that the price of service z is higher than the price of a lower priority service (i.e., service z+1) by at least a certain value. That is,

$$p_i^z(W_t) \geq \chi_i^z \cdot p_i^{z+1}(W_t), \forall z, 1 \leq z,\ z+1 \leq NS_i \tag{5.9}$$

where χ_i^z is a predetermined value for service z of class i (e.g., 5%). From Eq. (5.4), Eq. (5.7) and Eq. (5.9), the prices should be set as

$$\{\mathbf{p}_i^z(W_t)\}_{z=1}^{NS_i} = \left\{ \max\left(f_i'^1\left(\frac{\eta_i^{1*}(W_t)}{\eta_{total}} \right), \chi_i^1 \cdot p_i^2(W_t), p_i^{1,\min} \right), \max\left(f_i'^2\left(\frac{\eta_i^{2*}(W_t)}{\eta_{total}} \right), \chi_i^2 \cdot p_i^3(W_t), p_i^{2,\min} \right), \right.$$
$$\left. ..., \max\left(f_i'^{NS_i}\left(\frac{\eta_i^{NS_i*}(W_t)}{\eta_{total}} \right), p_i^{NS_i,\min} \right) \right\} \tag{5.10}$$

5.3.4 Demand Reduction with Dynamic Differentiated Pricing

If the network operator implements dynamic differentiated pricing, then $p_i^z(W_t)$ may not yield the optimal number of connection requests. This is because, when the network is underutilized, the price should sometimes be lowered below the differentiated price in order to increase the number of users who have sufficient WTP, and hence encourage them to initiate more connections and increase the network utilization. In this case, an

upper bound on the reduction of the number of connection requests to service z of class i ($\eta_i^{z,red}(W_t)$) is equal to

$$\eta_i^{z,red}(W_t) \le \eta_i^{z^*}(W_t) - \eta_{total} \cdot A_i^z \tag{5.11}$$

where $\eta_i^{z^*}(W_t)$ is computed using the optimal price in Eq. (5.5) when there are no constraints on the price and $\eta_{total} \cdot A_i^z$ is the actual number of connection requests resulting from the setting the price according to Eq. (5.10). Note that even if the price is set at its optimal value in Eq. (5.5), the actual number of connection requests is not guaranteed to equal to $\eta_i^{z^*}(W_t)$ when the network is underutilized, hence, the inequality relationship in Eq. (5.11). This is because, as aforementioned, users may not initiate connection requests even if they have sufficient WTP to do so. Therefore, an upper bound on the ratio of reduction in demand during the next time window in case dynamic differentiated pricing is used, is given by

$$\frac{\sum_{i=1}^{K}\sum_{z=1}^{NS_i}\eta_i^{z^*}(W_t)\cdot\left(b_i^z \cdot WL\right) - \sum_{i=1}^{K}\sum_{z=1}^{NS_i}\eta_i^{z,red}(W_t)\cdot\left(b_i^z \cdot WL\right)}{\sum_{i=1}^{K}\sum_{z=1}^{NS_i}\eta_i^{z^*}(W_t)\cdot\left(b_i^z \cdot WL\right)} \tag{5.12}$$

Therefore, network operators should carefully choose χ_i^z in Eq. (5.9) to ensure appropriate service differentiation, while maintaining the reduction in demand within

acceptable limits. The effect of choosing different values for χ_i^z on demand reduction is investigated in the following section.

5.4 Performance Evaluation

In this section, we evaluate the performance of our CAC-based dynamic pricing scheme. Our scheme is evaluated on HSDPA system by means of discrete-event simulation written in Java programming language. As in Chapter 4, the ILP problem of Eq. (5.1) is solved using *lp-solve* [72]. We study a homogeneous system in statistical equilibrium, in which any cell is statistically the same as any other cell. In this system, the mean arrival and departure rates are the same in each cell. Therefore, we can decouple a cell from the rest of the system and evaluate the system performance based on the performance of the cell [31]. As explained in section 5.1, our scheme can utilize the technique of Guard Channel to prioritize handoff connections over new ones. Because such technique has been extensively studied in the literature [73], [74], [75], [76] and [77] and since handoff calls are not affected by dynamic pricing as mentioned earlier, we, therefore, focus our analysis on new calls only. The same simulation model and the channel model that we developed in Chapter 3 are adopted here in our simulation. The traffic model is, however, modified in this chapter in order to evaluate our scheme under varying arrival rates (see Section 5.4.1). All the relevant simulation parameters are included in Appendix B.

In this thesis and for demonstration purposes only, the total available bandwidth in the next time window ($B_{free}(W_t)$) is roughly estimated as follows

$$B_{free}(W_t) = \left(C - \sum_{i=1}^{K} \sum_{z=1}^{NS_i} b_i^z \cdot N_i^z \right) \cdot WL \tag{5.13}$$

where C is the effective system capacity and $b_i^z \cdot N_i^z$ is the demand of admitted users requesting service z in class i. Based on different simulation runs with Pedestrian A environment, we found that the maximum load that the system can support and achieve 100% service coverage is 2.4 Mbps. We, therefore, set C in Eq. (5.13) to 2.4 Mbps.

The duration of each user's connection is modeled by an exponential distribution with a mean value of 50 s. Users are uniformly distributed in the cell. Pedestrian A (Ped A) environment is used in our simulation, which is recommended by 3GPP [56]. Mobile users in Ped A environment move at a fixed speed of 3 km/hr.

5.4.1 Traffic Model

To demonstrate the ability of our scheme to support different classes with different types of services, we consider two different classes with three different services, namely audio streaming (class 1), video streaming (class 1) and FTP (class 2). As explained in Chapter 3, audio streaming is modeled by a minimum rate of 12 Kbps, mean rate of 38 Kbps and a maximum rate of 64 Kbps. Video streaming is modeled by a minimum rate of 64 Kbps,

mean rate of 224 Kbps and a maximum rate of 384 Kbps. FTP is modeled by a constant data rate of 128 Kbps. Since audio streaming and video streaming services have minimum and maximum data rates, the network operator has to choose the rate at which it will base its admission for users requesting any of these services. For example, the network operator could admit users based on their minimum data rates and assign them higher data rates only when the amount of available capacity permits so. Another option for the network operator is to admit users based on their maximum data rates. In this case, the users can receive the highest QoS guarantees. The network operator, however, may risk wasting its wireless resources if the users do not transmit at their maximum data rates. In our simulation and for demonstration purposes only, we admit audio and video streaming users based on their mean rates. That is, we consider that audio users request a data rate of 38 Kbps and video users request a data rate of 224 Kbps. In addition, we set υ_i^z in Eq. (5.1) to 1/4, 1/4 and 1/2, for audio streaming, video streaming and FTP, respectively, in order to achieve an equal share of bandwidth between all classes. We also set the length of the time window to 10 s.

Actual connection request arrival rates to the system normally vary over time, and therefore, we adopt a 24-hour model for the arrival rates. In this model, the day is divided into 24 hours starting at midnight, with different arrival rates are assigned to different hours of the day based on observation of the connection request arrivals in a typical business day [78], [79]. It is observed in [79] that the peak hours (maximum connection arrivals) occur around 11:00 AM and 16:00 PM. In our simulation, each hour of the day is simulated by 400 s and the performance results are collected at end of each simulated

hour. Connection arrivals are modeled by a Poisson process, where the mean total arrival rates to the system for each hour of the day are shown in Figure 5.2. The total arrival rate to the system is equally divided between the three services. The arrival rates in Figure 5.2 constitute the actual arrival rates before dynamic pricing is implemented. When dynamic pricing is implemented, the actual arrival rates will depend on the prices. In this case, during congestion periods, our pricing component guarantees that the actual numbers of user connection requests will equal the optimal ones as described in Section 5.3.1. On the other hand, when the network is underutilized, which occurs in early morning hours (00:00-05:00 AM) and at night (21:00-24:00 PM), our pricing component guarantees to provide incentives to users to use the network services. However, as discussed in Section 5.3.1, not every user who has a sufficient WTP to make a connection request at a certain time is willing to make such a request at that time. In this case, the arrival rate to the system may stay at its low level or it may increase up to the optimal one depending on the preferences of users. To evaluate such a case, we test our proposed scheme with no increase in the number of connection requests (i.e., the actual arrival rate stays at its low value and does not increase as a result of lower prices) and with a 10%, 30% and 50% increase in the number of connection requests, respectively (i.e., 10%, 30% and 50% of the users who have sufficient WTP to make connection requests as a result of lowering the prices will make such requests, respectively).

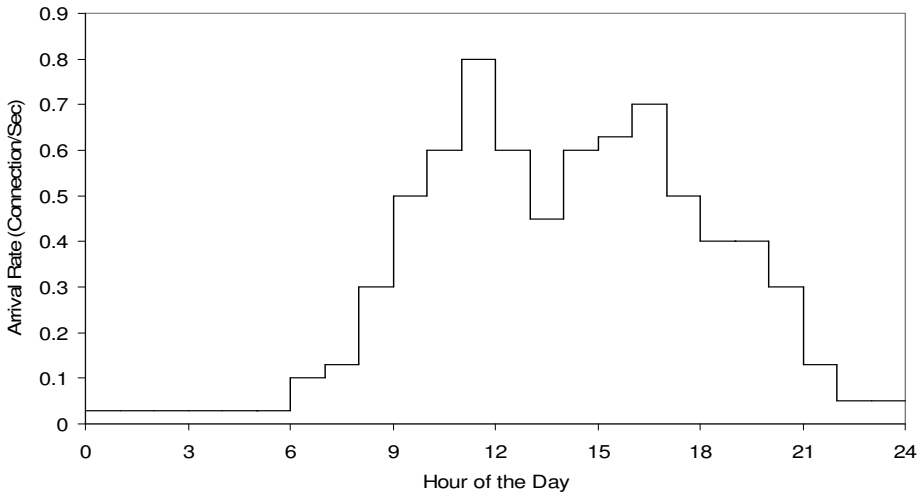


Figure 5.2: Arrival rates in a typical business day [79]

5.4.2 Demand Model

As aforementioned, our pricing scheme is general and can work with any demand model. To test our scheme, however, we utilize the following well-known demand model [43], [80],

$$A_i^z = f_i^z\left(p_i^z(W_t)\right) = a_i(W_t)\cdot e^{-c_i(W_t).p_i^z(W_t)} \tag{5.14}$$

where at any given time window W_t, $a_i(W_t)$ is the demand shift constant for class *i* users and $c_i(W_t)$ is the price elasticity of demand (i.e., the change in demand for a certain product or service due to a change in its price). The reason for using this particular demand model is that it can support different classes of traffic and different user demand

behaviors by considering their price elasticity of demand and their demand shift constants, which can assume different values for different times of the day. To ensure that $A_i^z = f_i^z\left(p_i^z(W_t)\right) \rightarrow [0,1]$, we set $a_i(W_t)$ to 1. In addition, for demonstration purposes, we set $c_i(W_t)$ to 1, 2, 3 for audio streaming, video streaming and FTP, respectively. These values are chosen so that users of higher priority service (e.g., audio streaming) are less responsive to price changes than those of lower priority services (i.e., FTP). This way, users requesting higher priority services are charged more than those requesting lower priority services. It should be noted that the actual values of $c_i(W_t)$ should be determined by market studies on real demand behaviors for the different users.

5.4.3 Test Cases and Performance Metrics

We evaluate our scheme under four cases. In the first case, we test the basic dynamic pricing when no constraints are imposed on the dynamic price values. In this case, we compare the performance of our Call Admission Control-based Dynamic Pricing scheme denoted by (CAC-bDP+x%, where x% = 0%, 10%, 30%, 50% denote the increase in user connection requests when network is underutilized as a result of lowering the prices as discussed in the previous section) with a Conventional CAC scheme denoted by (CCAC). In CCAC, no dynamic pricing is implemented. Instead, users are charged fixed prices and connection requests are always accepted as long as there is enough bandwidth to support them. In this case, we fix the prices to 0.35, 0.17 and 0.11 units of money per units of bandwidth for audio streaming, video streaming and FTP services, respectively. These

values are chosen so that at least 70% of users have sufficient WTP to make connection requests according to the demand model in Eq. (5.14). In practice, fixed prices are determined so that the majority of people have sufficient WTP to make connection requests, which is one of the main causes of congestion.

Not explicit comparison is made with the scheme in [40] as it reduces to a special case of our scheme, namely CAC-bDP+0%. In this case, users are not motivated to increase their demand in off-peak hours in spite of lowering the prices. Such a case is precisely the same as charging fixed prices during off-peak hours, which is the idea of the scheme in [40]. Thus, there is no need to duplicate the results.

In the second case, we test our dynamic differentiated pricing and measure its effect on demand reduction as explained in Section 5.3.3. Two values of χ_i^z are used in this case, 1.05 and 1.15 (i.e., 5% and 15% increase in prices, respectively). In both cases, we assume that the user demand function (i.e., A_i^z) is 100% accurate in modeling the users' responses towards price changes.

In the third case, we test the basic dynamic pricing but we consider that A_i^z has an error probability of 5%, 10% and 15%, respectively. That is, users will correctly react to price changes by lowering or increasing their demands with probabilities 95%, 90% and 85%, respectively. Evaluating such as case is of practical importance, since dynamic pricing may lead to undesirable results when the user demand models are inaccurate. To the best of our knowledge, such a case has never been considered before.

In the fourth case, we integrate all of our framework's components (i.e., packet scheduling, bandwidth provisioning and CAC-based dynamic pricing). In this case, our framework is denoted by (CAC-bDP+*x*%[16], where *x*% = 0%, 10%, 30%, 50%). We compare the performance of our framework with the case of Packet Scheduling and Bandwidth Provisioning without CAC (denoted by PSBPwoCAC). This is done to show the performance improvements of integrating our CAC-based dynamic pricing scheme with our packet scheduling and dynamic bandwidth provisioning scheme in providing efficient bandwidth management during the lifetime of the users' connections. Refer to Appendix C for flowcharts of the framework components.

We use the following performance metrics

- Percentage of bandwidth utilization: the percentage of the utilized bandwidth to the total bandwidth.
- Connection blocking probability: the probability that a users' connection is blocked due to insufficient bandwidth to meet his requirements.
- Percentage of bandwidth share: the percentage of used bandwidth for each class to the total utilized bandwidth. This metric is used to test our fairness measure in Eq. (5.1).
- Percentage of demand reduction: the percentage of utilized bandwidth when dynamic differentiated pricing is used to the utilized bandwidth when basic

[16] In this case, CAC-bDP+*x*% refers to all of our framework integrated components with a x% increase of the users' connections requests as a result of lower prices as aforementioned.

dynamic pricing is used instead. This metric is used to evaluate the extent of introducing dynamic differentiated pricing in reducing the demand.

- Revenue: the amount of money earned during the day. It is calculated by multiplying the total amount of data transmitted to the user with the price per bit, summed over all users in the system in a certain time interval, which is the simulation time in our experiments.

- Average packet delay: the average amount of time the packet spends in the queue at the base station in addition to the transmission time (delay of discarded packets and dropped users' connections is not counted).

- Average throughput: average number of successfully delivered bits over the lifetime of the user's connection (average throughputs of dropped user connections are not counted).

- Service coverage: percentage of users who achieve their required QoS with a certain outage level. For audio streaming, a user's connection is dropped if his average packet loss (due to packet discarding, transmission errors and/or buffer overflow) exceeds 5% [65], [66] and [67]. For video streaming, a user's connection is dropped if his achieved average throughput is less than his minimum required rate. Finally, for FTP traffic, a user's connection is dropped if his achieved average throughput is less than 9.6 Kbps [13] and [16].

5.4.4 Simulation Results

In what follows, we show the performance results of the aforementioned test cases.

Case 1: Basic Dynamic Pricing

Figure 5.3 shows the percentage of bandwidth utilization for our scheme and the CCAC scheme. The figure shows that our scheme can significantly increase the bandwidth utilization of the system as more users (i.e., 10%, 30% and 50%) decide to make connection requests as a result of lowering the prices during off-peak hours. For the case where users are not affected by the low prices (i.e., case with 0% increase), the bandwidth utilization of our scheme is similar to that of CCAC, which is expected since our scheme is distinguished by its ability to increase the utilization of the network when the demand is low. We remark, however, that since most users are price-sensitive, they will try to make their connection requests when prices are lower. Hence, the case of 0% is not common in practice. Therefore, using our scheme, the network operator can increase the usage of the network when it is underutilized, hence, increasing its revenues. In addition, our scheme can efficiently prevent network congestion, and therefore, achieving 0% blocking probabilities as shown in Figure 5.4. This is because our scheme optimally determines the prices of units of bandwidth as to encourage enough users to make connection requests, hence, ensuring that the system is never congested. The results confirm the superiority of our scheme compared to the CCAC scheme, where users are not provided any incentives to regulate their usage of the network. Such a scheme can result in very high blocking

probabilities during peak hours, and therefore, resulting in user dissatisfaction and potential revenue loss. For instance, Figure 5.4 shows that at peak hours (e.g., 11 AM), the blocking probability of the CCAC scheme can reach up to 18.6%.

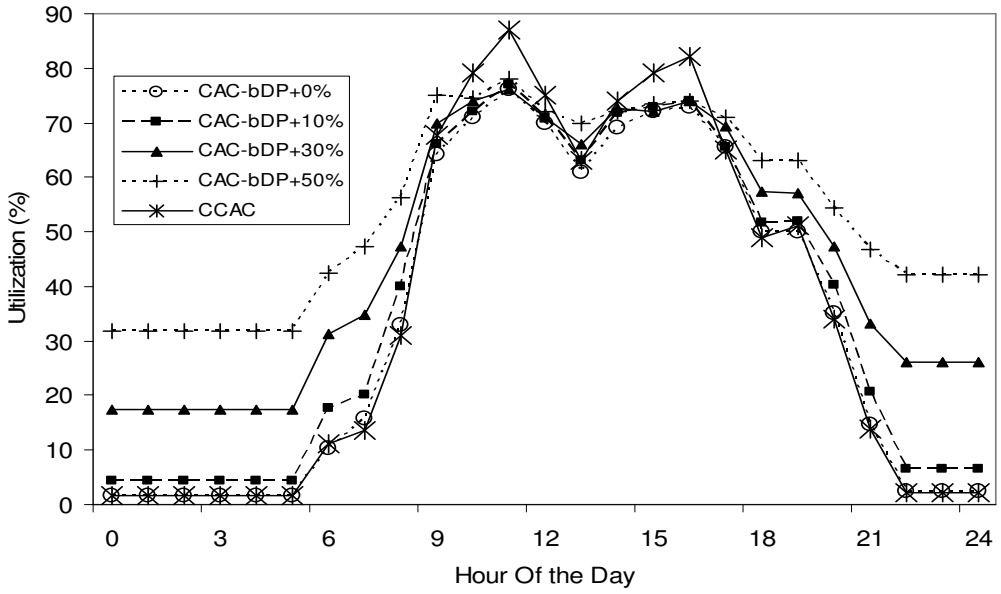


Figure 5.3: Percentage of bandwidth utilization at different hours of the day

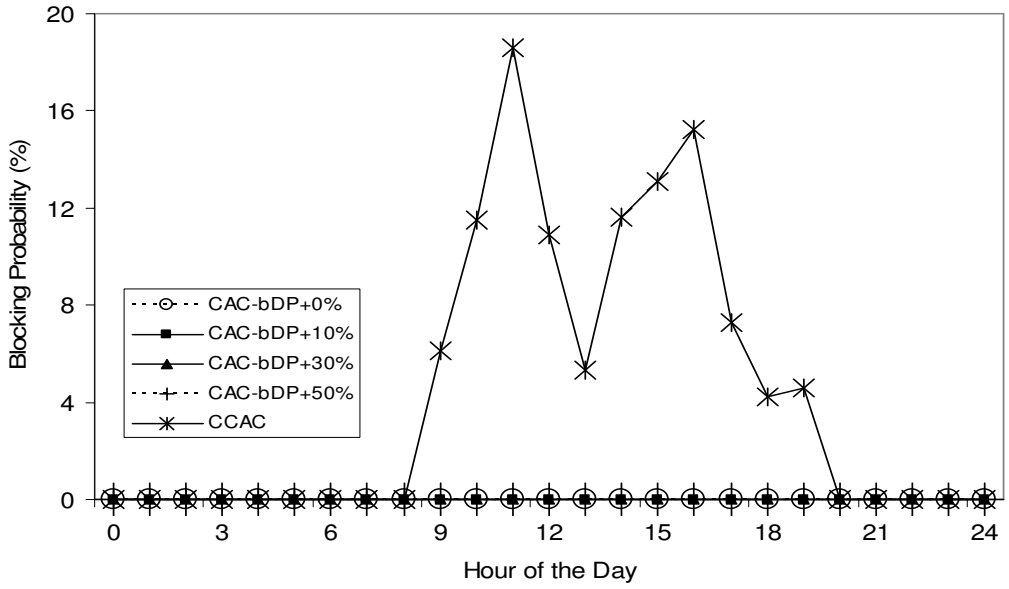


Figure 5.4: Blocking probability at different hours of the day

Table 5.1 shows the percentage of bandwidth share for each class. As mentioned earlier, we set υ_i^z in our objective function to respectively 1/4, 1/4 and 1/2, for audio streaming, video streaming and FTP so that each class gets the same share of bandwidth. The table shows that our scheme achieves better bandwidth share than CCAC. The reason for the unfair bandwidth share in CCAC is that, according to our traffic model, the actual arrival rate (before dynamic pricing is implemented) is equally divided between the three services. Class 1 users, on average, request more bandwidth compared to class 2 users, this results in a higher bandwidth share for class 1. Our scheme, on the other hand, determines the dynamic prices of units of bandwidth so that the incoming connection requests for the different services achieve the maximum possible bandwidth utilization while maintaining a certain fairness level (i.e., absolute fairness between the two classes

in this case). Hence, our scheme achieves better fairness as shown in Table 5.1. An interesting result revealed from Table 5.1 is that even though v_i^z is set to achieve absolute fairness in our scheme, the bandwidth share of class 1 is still higher than that of class 2. This is due mainly to the high bandwidth requirements of class 1 users in off-peak hours, where users are not affected by our dynamic prices. This is clearly shown in Figure 5.5, which depicts the bandwidth share of class 1 in each hour of the day. In peak hour periods, users are more affected by dynamic prices, and hence their connection requests tend to approach the optimal ones, therefore, achieving better fairness. In fact, as more users decide to make connection requests as a result of lower prices, the bandwidth share of class 1 tends to approach that of class 2 because, as mentioned earlier, prices are designed to achieve an equal share of bandwidth in our experiments. This explains the increased fairness of our scheme in Figure 5.5 as more users tend to make connection requests during off-peak hours.

Table 5.1: Percentage of bandwidth share

Scheme	Class 1	Class 2
CAC-bDP+0%	64.84%	35.16%
CAC-bDP +10%	61.048%	38.952%
CAC-bDP +30%	58.212%	41.788%
CAC-bDP +50%	57.16%	42.84%
CCAC	73.708%	26.292%

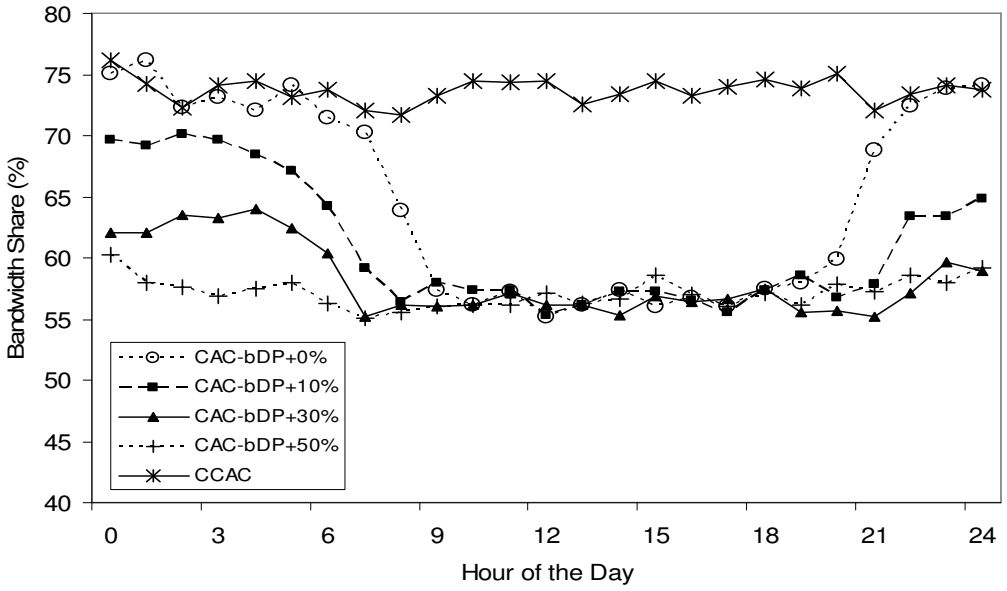


Figure 5.5: Percentage of bandwidth share for class 1 at different hours of the day

Table 5.2 shows the total revenue collected throughout the day for our scheme and CCAC. Our scheme clearly outperforms CCAC in terms of revenues. This is because our scheme charges users higher prices during peak hours. In addition, as more users decide to make connection requests, more revenues can be collected. The revenue collected from class 1 users is higher than that from class 2 users because the former pay higher prices for class 1 services in addition to requesting higher amount of bandwidth. It should be noted that more revenue can be earned if more bandwidth is assigned to class 1 (i.e., if class 1 is assigned more than 0.5 bandwidth share). Therefore, the fairness constraint in Eq. (5.1) can also be used to increase revenues by assigning more bandwidth to classes that are expected to yield higher revenues.

Table 5.2: Total revenue earned during the day (units of money)

Scheme	Class1	Class 2	Total
CAC-bDP+0%	297×10^4	54×10^4	351×10^4
CAC-bDP +10%	305×10^4	59×10^4	364×10^4
CAC-bDP +30%	312×10^4	68×10^4	380×10^4
CAC-bDP +50%	321×10^4	83×10^4	404×10^4
CCAC	282×10^4	27×10^4	309×10^4

Case 2: Differentiated Dynamic Pricing

Tables 5.3 and 5.4 show the percentage of reduction in demand with 5% and 15% price differentiation, respectively. The tables show that, as the percentage of price differentiation increases, the shift of the price from its optimal value increases, and hence the reduction in demand increases. In addition, the more users tend to make connection requests when the price is set at its optimal values, the more the demand decreases, since in this case more users are affected by the shift in price. Furthermore, all the reduction in demand is suffered by class 1 only, where class 2 is not affected by the price differentiation. This is expected because FTP service has the lowest priority. Therefore, its price is always set at its optimal value. Whereas, audio streaming and video streaming services have higher priorities. Hence, price differentiation is applied to them. The value of price differentiation (i.e., χ_i^z), therefore, should be carefully determined as to ensure appropriate service differentiation, while keeping the reduction in demand below acceptable levels.

Table 5.3: Percentage of reduction in demand when $\chi_i^z = 1.05$

Scheme	Class1	Class 2	Total
CAC-bDP+0%	2.30%	0%	2.30%
CAC-bDP +10%	2.74%	0%	2.74%
CAC-bDP +30%	3.32%	0%	3.32%
CAC-bDP +50%	4.01%	0%	4.01%

Table 5.4: Percentage of reduction in demand when $\chi_i^z = 1.15$

Scheme	Class1	Class 2	Total
CAC-bDP+0%	4.21%	0%	4.21%
CAC-bDP +10%	5.64%	0%	5.64%
CAC-bDP +30%	7.29%	0%	7.29%
CAC-bDP +50%	9.25%	0%	9.25%

Case 3: Inaccurate Demand Model

Figures 5.6, 5.7 and 5.8 depict the blocking probabilities with 5%, 10% and 15% error probability in user demand model, respectively. The figures show that, when the user's demand model is only partially accurate in modeling their behaviors towards price changes, the network operator can no longer guarantee a congestion-free system. This is expected as the network operator cannot ensure that, at peak hours, the right number of users will have sufficient WTP to make connection requests. Hence, the demand for bandwidth might exceed the system capacity especially during peak hours, which explains the blocking

probabilities shown in the figures. The figures also show that, as the error probabilities in the user demand model increase, the blocking probabilities increase as well. This is because of the increased number of users who decide to make connection requests when there is not enough bandwidth to support their connections. These users are erroneously classified by the demand model as being incapable of making connection requests (due to insufficient WTP), where in fact they actually can make such requests. In addition, the figures do not show significant difference in the blocking probabilities of our scheme when more users decide to make connections requests as a results of lowering the prices (i.e., cases of +10%, +30% and +50% increase in user connection requests). This is because, during off-peak hours, there is enough bandwidth to support many users. During these hours, the blocking probabilities are zero, since the increase of connections requests as a result of lower prices does not exceed the system capacity anyway.

Despite the partial accuracy of the user's demand model, the blocking probabilities are still much lower than the blocking probabilities of the CCAC scheme. For example, the blocking probability of CAC-bDP+50% at hour 11 PM with 15% probability of error in the user's demand model is 3.15% compared to 18.6% with CCAC. This is due to the fact that even in the presence of errors in the user's demand model, the majority of users still react correctly to the price incentives of our scheme. Thus, the total demand for bandwidth is still lower compared to the total demand when CCAC is implemented. Therefore, our CAC-based dynamic pricing scheme can still improve the system performance and achieve very low blocking probabilities even with inaccurate user's demand model.

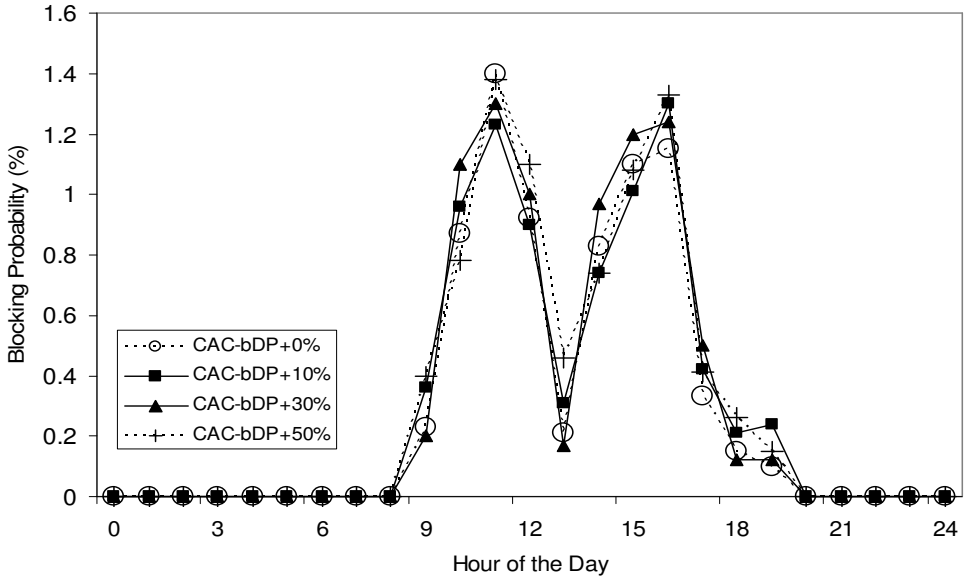


Figure 5.6: Blocking probability with 5% error probability at different hours of the day

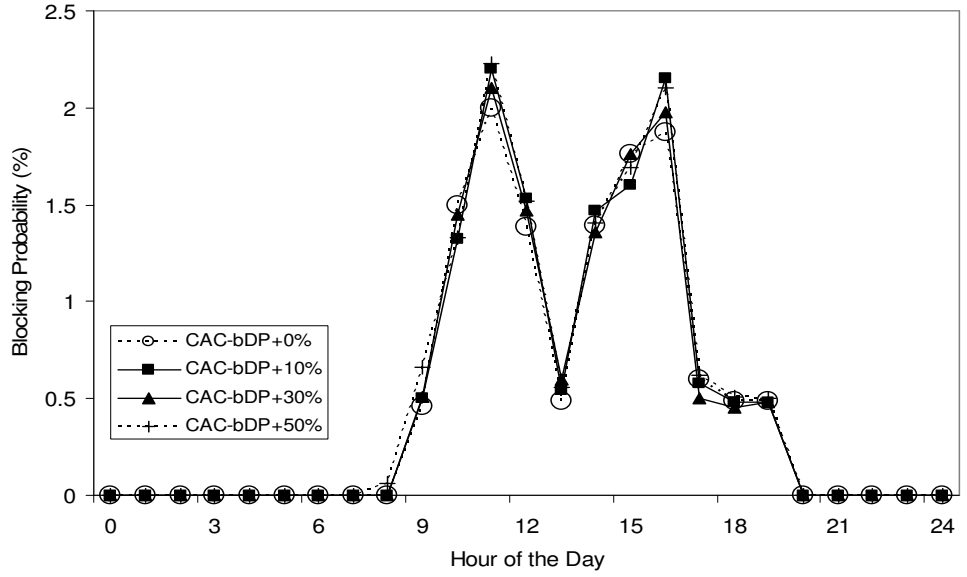


Figure 5.7: Blocking probability with 10% error probability at different hours of the day

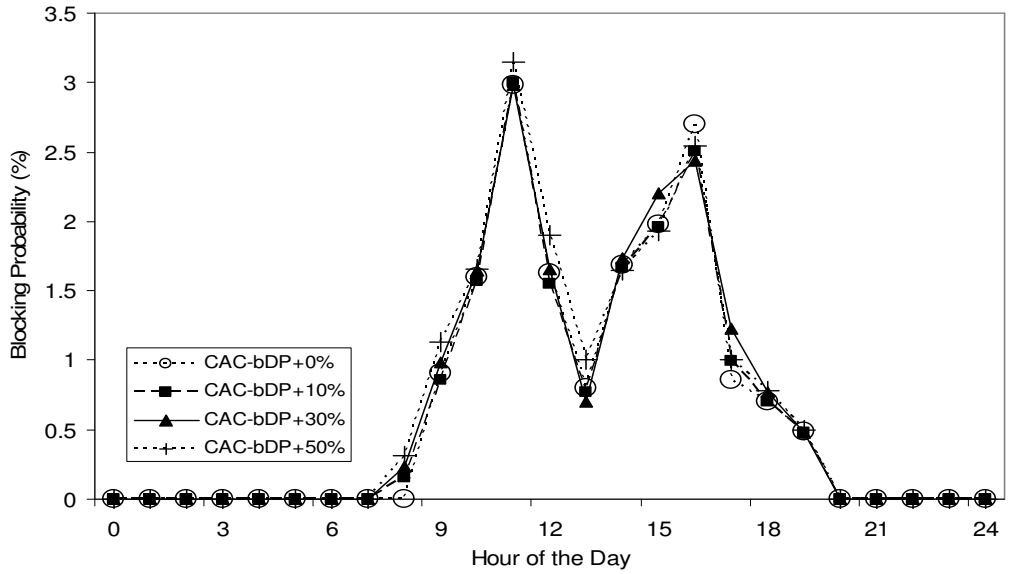


Figure 5.8: Blocking probability with 15% error probability at different hours of the day

Case 4: Integrated Framework Components

Figure 5.9 shows the average packet delay for audio users. The figure shows that our framework can maintain the average packet delay of users below their delay thresholds. This is because user connections are admitted only when there is enough bandwidth to support them. Hence, our framework achieves lower average packet delay than PSBPwoCAC during peak hours. An interesting observation from Figure 5.9 is that although the framework utilizes our CAC-based dynamic pricing scheme, the average packet delay of audio users increases as more users are admitted to the system (i.e., during peak hours and/or as a result of lower prices). This is due to the varying channel

quality conditions of users, which may allow the support of less bandwidth than that initially assigned by our CAC scheme. Another interesting observation is that the average packet delays of CAC-bDP+*x*%, *x*=10, 30 and 50 are higher than that of PSBPwoCAC during off-peak hours. This is expected as our CAC-based dynamic pricing scheme motivates more users to make connection requests, and hence results in increased system utilization and increased revenues. This is an enviable feature as long as the average packet delays of users are kept below their maximum delay thresholds, which is clearly achieved by our framework. A similar trend is also observed in Figures 5.10 and 5.11, which show the average throughput for video and FTP users, respectively.

Figures 5.12, 5.13 and 5.14 show the percentage of service coverage for audio, video and FTP users, respectively. The figures show that our framework can achieve full service coverage at all hours of the day due to using our CAC-based dynamic pricing scheme, which motivates users to make connection requests only when there is enough bandwidth to support them. Whereas, in PSBPwoCAC, users are admitted into the system regardless of the availability of bandwidth, and hence they are not provided any minimum bandwidth guarantees. This can lead to dropped user connections, and hence lower service coverage especially during peak hours.

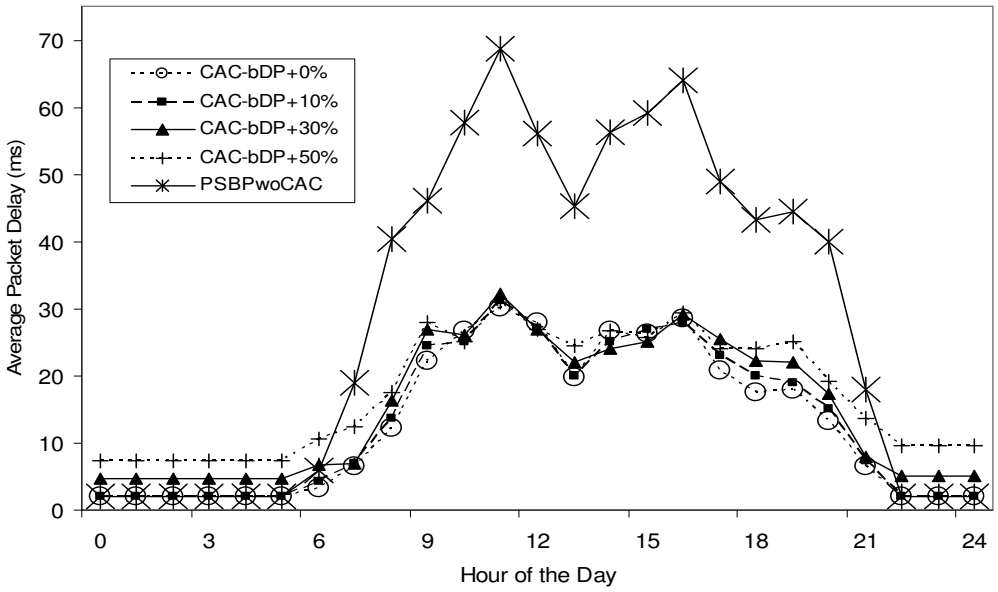


Figure 5.9: Average packet delay for audio users at different hours of the day

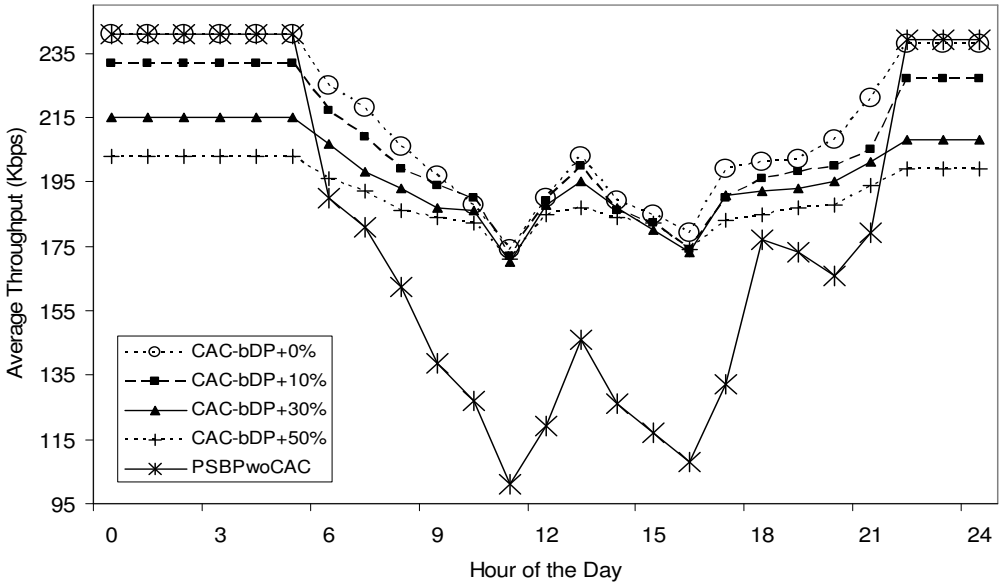


Figure 5.10: Average throughput for video users at different hours of the day

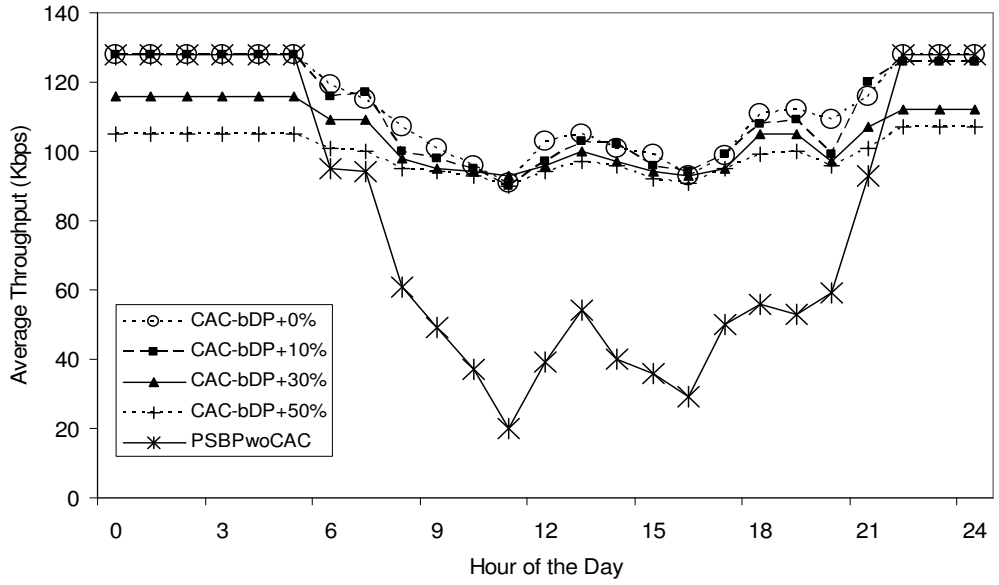


Figure 5.11: Average throughput for FTP users at different hours of the day

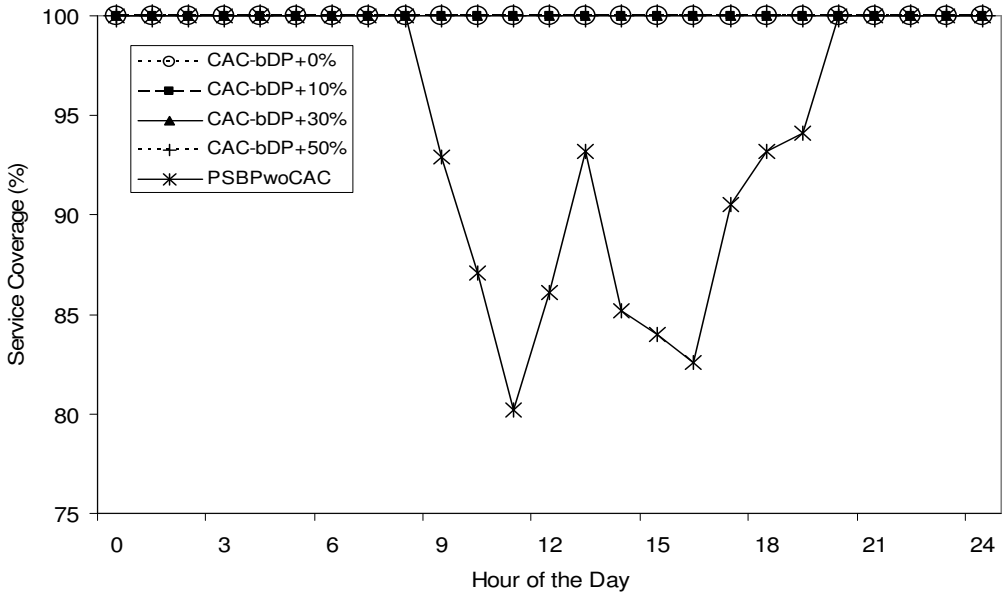


Figure 5.12: Percentage of service coverage for audio users at different hours of the day

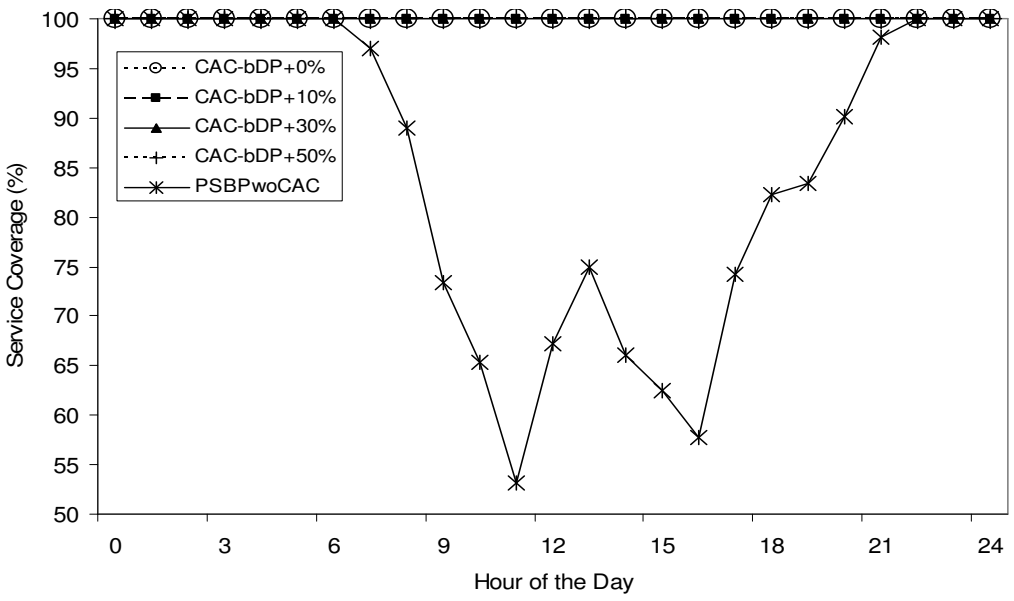


Figure 5.13: Percentage of service coverage for video users at different hours of the day

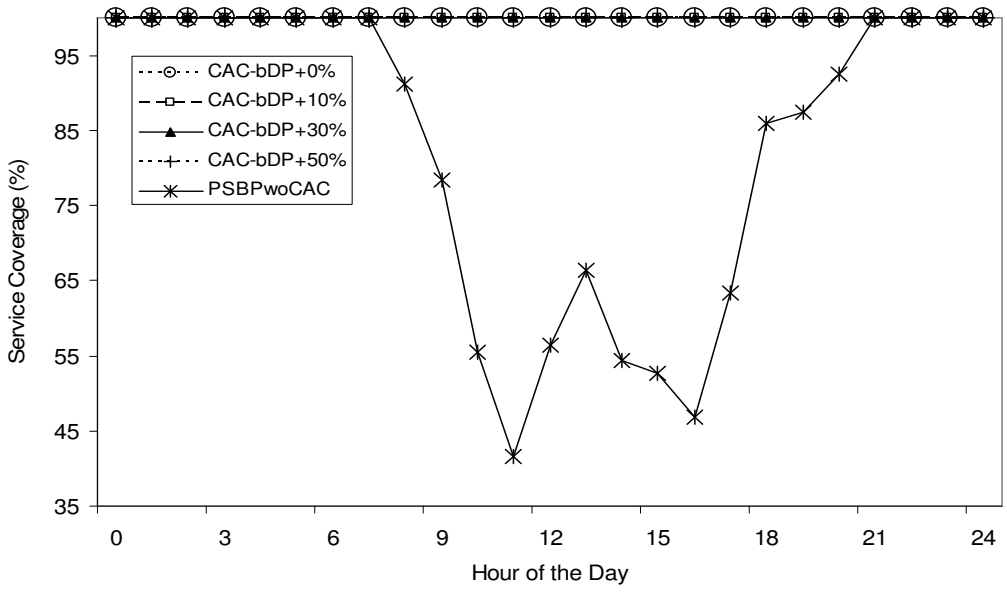


Figure 5.14: Percentage of service coverage for FTP users at different hours of the day

5.5 Summary

This chapter presented a CAC-based dynamic pricing scheme for BWASs. The scheme consists of three components, namely a monitoring component, a Call Admission Control component and a dynamic pricing component. The main objective of our proposed scheme is to provide monetary incentives to users to use the wireless resources efficiently and rationally, hence, allowing efficient bandwidth management at the admission level. By dynamically determining the prices of units of bandwidth according to the network load, the proposed scheme can guarantee that the number of connection request arrivals to the wireless system are less than or equal to the optimal ones computed dynamically, and hence guaranteeing a congestion-free system. Simulation results show that our scheme can significantly improve the utilization of the wireless system and increase the revenues of network operator. In addition, the scheme can guarantee zero blocking probabilities provided that the user's demand model is accurate in modeling their behaviors towards price changes. Simulation experiments for the case of inaccurate user's demand model show that our scheme can still achieve significant performance gains in terms of blocking probabilities compared to conventional schemes, where users are not provided any incentives to regulate their demand for wireless services.

.

Chapter 6

Conclusions and Future Work

Broadband Wireless Access Systems (BWASs) promise to revolutionize the mobile user's wireless experience by offering high data rates that are much beyond the capabilities of 2.5G and 3G wireless systems. In order to support as many users as possible, these systems exploit the bursty nature of the users' data traffic by utilizing shared channels for data delivery instead of dedicated ones. In addition to accommodating more users, the use of shared channels reduces the per-bit cost of transmission, hence, decreasing of cost of providing different wireless services. Despite their momentous advantages, the use of shared channels complicates the task of resource sharing and bandwidth management. This is because it requires more intelligent and sophisticated bandwidth management schemes to distribute the wireless resources among mobile users who have diverse and sometimes stringent QoS requirements. The problem of bandwidth management is even more aggravated given the fact that mobile users

constantly require varying amount of resources to satisfy their QoS requirements due to their varying channel quality conditions.

This thesis studied the problem of designing and developing efficient downlink bandwidth management schemes for BWASs and made solid contributions to the ongoing research in this area. In this chapter, we summarize and discuss the conclusions from this thesis and provide directions for future research work.

6.1 Summary of Contributions

We classified bandwidth management for BWASs into three related levels, namely packet-level, class-level and admission-level. We then proposed a bandwidth management framework for BWASs consisting of a number of novel economic-based approaches to provide efficient downlink bandwidth management at these three levels. The framework was designed to simultaneously achieve and balance between the following objectives:

1) Supporting different classes of traffic with users having different QoS requirements;
2) Maximizing the throughput of the wireless system;
3) Ensuring a fair distribution of wireless resources by supporting inter- and intra-class fairness;
4) Maximizing the network operator's revenues by limiting the revenue loss incurred from serving low-revenue generating users; and

5) Providing monetary incentives to the users to use the wireless resources efficiently and rationally in order to prevent network congestion.

In Chapter 3, we developed a packet scheduling scheme to achieve bandwidth management at the packet level. The scheduling scheme is responsible for scheduling the packets of different users for transmission at the base station, where scheduling occurs at every transmission time frame. The proposed scheme employs practical economic models through the use of novel utility and opportunity cost functions to simultaneously satisfy the diverse QoS requirements of mobile users and maximize the revenues of network operators. A general utility function was proposed to represent users with various QoS requirements. To demonstrate its generality, we showed how the utility function can be used to support three different types of traffic, best-effort traffic, traffic with minimum data rate requirements and traffic with maximum packet delay requirements. We then showed that the two well-known scheduling schemes, the Maximum Carrier to Interface Ratio (Max CIR) [9] and Proportional Fairness (PF) [10] are special cases of our proposed scheme. This gives the network operator more flexibility to choose between different scheduling disciplines. The main advantage of the opportunity cost function is that it limits the revenue loss resulting from serving low revenue generating users. It therefore allows the network operator to determine the level of revenue-fairness trade-off. To maximize the system throughput, the proposed packet scheduling scheme exploits multi-user diversity by utilizing the information of the channel quality conditions of the users in its scheduling decisions. This is done, nevertheless, in a way that ensures a fair

distribution of the wireless resources among users and prevents starvation of users with bad channel quality conditions.

In Chapter 4, we proposed a novel dynamic bandwidth provisioning scheme to provide efficient bandwidth management at the class level in BWASs. The proposed scheme was designed to improve the performance of packet scheduling and enhance the QoS of users throughout the lifetime of their connections. The proposed scheme spans multiple time frames and optimally allocates them to the different classes of traffic depending on their weights, the real-time bandwidth requirements, channel quality conditions of their users and the expected obtained revenues. Once each traffic class is allocated its optimal number of time frames, the packet scheduling scheme can then be used to distribute the frames among its admitted users. Therefore, the bandwidth provisioning scheme can be thought as a longer-term bandwidth management scheme compared to packet scheduling, which only checks one time frame at a time. Similar to the packet scheduling scheme proposed in Chapter 3, the bandwidth provisioning scheme utilizes an opportunity cost function to limit the revenue loss at the class level that results from serving low revenue generating classes (e.g., low priority classes).

To maximize inter-class fairness, we proposed a dynamic weight update scheme for the bandwidth provisioning scheme. The weight update scheme dynamically computes the weights of different classes of traffic based on their performance history to maximize inter-class fairness. A distinctive feature of this scheme is that it allows the weights of lower priority classes to be temporarily higher than those of higher priority classes while ensuring long-term service differentiation between them at the same time. The resulting

fairness is more adaptive to the performance of classes since it is based on their performance history. Therefore, inter-class fairness can be better achieved using this scheme compared to using fixed weights.

In Chapter 5, we proposed a Call Admission Control-based dynamic pricing scheme that aims at providing efficient bandwidth management at the admission level. The proposed scheme combines the benefits of both Call Admission Control (CAC) and dynamic pricing in order to achieve the best system performance. Specifically, the scheme aims at efficiently managing the bandwidth of BWASs in order to simultaneously satisfy the bandwidth requirements of users, maximize the utilization of these systems and prevent congestion. The scheme consists of three components, namely the monitoring component, the CAC component and the dynamic pricing component. The monitoring component continuously monitors the amount of available bandwidth in the system over a window of some time interval. If it detects any changes in the available bandwidth, it triggers the CAC component, which then determines the maximum number of connection requests for each wireless service that the system can support as to maximize its utilization and ensure fairness among different services, and hence fairness among different classes of traffic. The dynamic pricing component then determines the new prices of different wireless services so that exactly enough users have sufficient willingness to pay to make such connection requests. That is, dynamic prices are used to force the number of connection requests to different wireless services towards the optimal ones that are dynamically determined by the CAC component. This way, the proposed scheme is able to guarantee a congestion-free system. This guarantee, however, is only

valid if the user demand model is accurate in capturing their behaviors towards price changes. In case of inaccurate demand models, the scheme cannot provide such guarantees. Nevertheless, as shown in Chapter 5, it can still achieve substantial performance gain over conventional schemes that do not utilize dynamic pricing. A very distinctive feature of the proposed scheme is that the CAC and pricing functions are executed independently. This simplifies their implementation and provides network operators with the flexibility to use different CAC and user demands functions without affecting the computation of prices.

The proposed scheme was then extended to support dynamic pricing with minimum price values and differentiated pricing. We showed that if dynamic differentiated prices were implemented, the system could suffer some reduction in demand due to shifting the prices above their optimal values. Such reduction would depend on the level of differentiations between prices for different services. Refer to Appendix C for flowcharts of our framework components.

6.2 Future Research Directions

There are several directions by which the work in this thesis can be extended. In this section, we highlight some of these directions.

We have shown that the scheduling scheme in Chapter 3 can simultaneously serve multiple users in each time frame. We remark, though, that the scheduling decision is only optimized in the time domain. This is because multiple users are scheduled for simultaneous transmission each time only to fill the frame. Hence, if the user with the

highest aggregate utility has enough data to send, he will then utilize the whole frame. To further enhance the performance of the scheduling scheme, code multiplexing (in HSDPA) or frequency multiplexing (in WiMAX) should be considered as another domain besides the time domain in the scheduling decision. In this case, even if the user with the highest aggregate utility has enough data to fill the frame, the scheduling scheme may assign him only a fraction of the frame (by assigning him appropriate number of codes or frequencies), where the rest of the frame is assigned to other users to further enhance their QoS. Therefore, using the time and code/frequency domains in the scheduling decisions requires not only finding the optimal set of users for transmission but also the optimal number of codes/frequencies for each one of them. This, however, may complicate the scheduling problem as it will result in a combinatorial optimization problem.

The dynamic bandwidth provisioning scheme proposed in Chapter 4 assumes that the total number of frames to be allocated among different classes of traffic is given. This number can be empirically determined by simulations or real experiments as to achieve the desired system performance by the network operator. However, users continuously experience varying channel quality conditions throughout the lifetime of their connections. In addition, users have different, and sometimes changing, traffic demands. Therefore, a possible extension to the dynamic bandwidth provisioning scheme is to determine the number of frames based on the dynamics of the environment on which the dynamic bandwidth provisioning scheme is implemented in order to further optimize its performance.

In Chapter 5, we introduced a CAC-based dynamic pricing scheme that can achieve significant performance gains. There are still some issues that need to be addressed though. One issue is the consideration of social fairness[17]. The scheme may raise the prices of wireless resources (i.e., bandwidth) to very high levels especially during congestion periods. Such prices may not be affordable by many users. Dynamic pricing may, therefore, be viewed as promoting social unfairness as only rich people can afford to make connection requests. The issue of social fairness is closely related to the fact that the scheme guarantees that wireless resources are given to the users who value them the most but it does not provide any guarantees that the wireless resources are given to those who need them most. For example, a user may become unable to make an emergency call (for instance, to the police or the hospitable) due to high prices. The network operator can utilize a number of solutions to deal with these issues. For instance, the network operator can set a maximum threshold that dynamic prices are not allowed to exceed. This solution is similar to the case of dynamic pricing with minimum price values discussed in Section 5.3.2. The network operator, however, will have to choose the value of the maximum threshold appropriately to improve social fairness while at the same time keep the benefits of dynamic pricing. This is not a trivial task given that the optimal dynamic prices during congestion periods might be above the maximum threshold. Hence, users will not be discouraged to decrease their demand for wireless services. This implies that the network operator will not enjoy a congestion-free system.

[17] Social fairness refers to the state of economy where the majority of people are able to buy certain products regardless of their incomes. In the context of this thesis, it refers to the ability to use the wireless network services.

Another possible solution is to allow the users to make a certain number of calls at fixed low prices every given time period (e.g. 5 calls every month). The users in this case, can choose to activate their low-priced calls any time they desire. This way, the network operator can still use dynamic pricing and guarantee a certain number of calls to users, which they can make at affordable prices.

In addition, CAC-based dynamic pricing schemes usually ignore the effect of prices on user mobility. This may impact the performance and planning of the network because some cells might become congested in the long-run due to user mobility as a result of high dynamic prices in neighboring cells. This comes as a result of users avoiding making connection requests in some congested places such as downtown areas that are known to charge higher prices, and choosing instead to make such requests in other less congested nearby areas.

Bibliography

[1] H. Kaaranene, A. Ahtiainen, L. Laitinen and S. Naghian, "UMTS Networks, Architecture, Mobility, and Services", 2nd edition, John Wiley & Sons, 2005.

[2] 3GPP TS 25.308, "High Speed Downlink Packet Access (HSDPA); Overall Description", Release 5, March 2003.

[3] IEEE 802.16-2004, "Air Interface for Fixed Broadband Wireless Access Systems", October 2004.

[4] IEEE 802.16e, "Air Interface for Fixed and Mobile Broadband Wireless Access Systems", February 2005.

[5] B. Al-Manthari, N. Nasser and H. Hassanein, "Packet Scheduling in 3.5G High Speed Downlink Packet Access Networks: Breadth and Depth", *IEEE Networks Magazine*, vol. 21, no. 1, pp. 41-46, February 2007.

[6] C. Courcoubetis and R. Weber, "Pricing Communication Networks: Economic, Technology and Modeling", John Wiley & Sons, May 2003.

[7] 3GPP TS25.214 V5.5.0, "Physical Layer Procedures", Release 6.9.0, June 2006.

[8] H. Jiang, W. Zhuang, X. Shen and Q. Bi, "Quality-of-Service Provisioning and Efficient Resource Utilization in CDMA Cellular Communications", *IEEE Journal on Selected Areas in Communications*, vol. 24, no.1, pp. 4-15, January 2006.

[9] S. Borst, "Connection-level Performance of Channel-aware Scheduling Algorithms in Wireless Data Networks", *Proceedings of the Annual Joint Conference of the IEEE Computer Societies (INFOCOM)*, San Francisco, U.S.A., vol. 1, pp.321-331, March 2003.

[10] A. Jalali, R. Padovani and R. Pankaj, "Data Throughput of CDMA-HDR a High Efficiency-High Date Rate Personal Communication Wireless System", *Proceedings of the IEEE Vehicular Technology Conference (VTC)*, Boston, U.S.A., pp. 1854-1858, May 2000.

[11] A. Haider and R. Harris, "A novel Proportional Fair Scheduling Algorithm for HSDPA in UMTS Networks", *Proceedings of the Australian International Conference on Wireless Broadband and Ultra Wideband Communications (AusWireless)*, Sydney, Australia, pp. 43-49, August 2007.

[12] T. Bonald, "A Score-Based Opportunistic Scheduler for Fading Radio Channels", *Proceedings of the European Wireless Conference (EW)*, Florence, Italy, pp. 2244–2248, September 2004.

[13] G. Barriac and J. Holtzman, "Introducing Delay Sensitivity into the Proportional Fair Algorithm for CDMA Downlink Scheduling", *Proceedings of the IEEE International Symposium on Spread Spectrum Techniques and Applications(ISSSTA)*, Prague, Czech Republic, vol. 3, pp. 652-656, September 2002.

[14] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, R. Vijayakumar and P. Whiting, "Providing Quality of Service over a Shared Wireless Link", *IEEE Communications Magazine*, vol. 39, no. 2, pp. 150-154, February 2001.

[15] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, R. Vijayakumar and P. Whiting, "CDMA Data QoS Scheduling on the Forward Link with Variable Channel Conditions", Bell Labs Technical Report, April 2000.

[16] P. Jose "Packet Scheduling and Quality of Service in HSDPA", Ph.D. Dissertation, Aalborg University, Denmark, October 2003.

[17] A. Golaup, O. Holland and A. Aghvami, "A Packet Scheduling Algorithm Supporting Multimedia Traffic over the HSDPA Link based on Early Delay Notification", *Proceedings of the International Conference on Multimedia Services Access Networks (MSAN),* Orlando, U.S.A, pp. 78-82, June 2005.

[18] G. Song and Y. Li, "Utility-Based Resource Allocation and Scheduling in OFDM-Based Wireless Broadband Networks", *IEEE Communications Magazine*, vol. 43, no. 12, pp. 127-134, December 2005.

[19] K-H. Liu, L. Cai and X. Shen, "Multiclass Utility-Based Scheduling for UWB Networks", *IEEE Transactions on Vehicular Technology*, vol. 57, no. 2, pp. 1176-1187, March 2008.

[20] B. Rong, Y. Qian and H.H. Chen, "Adaptive Power Allocation and Call Admission Control in Multiservice WiMAX Access Networks", *IEEE Wireless Communications Magazine*, vol. 14, no.1, pp. 14-19, February 2007.

[21] N. Nasser and H. Hassanein, "Optimized Bandwidth Allocation with Fairness and Service Differentiation in Multimedia Wireless Networks", *Journal of Wireless Communications and Mobile Computing*, vol. 8, no. 4, pp. 501-511, May 2008.

[22] B. Rong, Y. Qian and K. Lu, "Integrated Downlink Resource Management for Multiservice WiMAX Networks", *IEEE Transactions on Mobile Computing*, vol. 6, no. 6, pp. 621-632, June 2007.

[23] N. Zorba and A. I. Perez-Neira, "CAC for Multibeam Opportunistic Schemes in Heterogeneous WiMax Systems under QoS Constraints", *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM)*, Washington, D.C., U.S.A, pp. 4296-4300, November 2007.

[24] E. B. Rodrigues and J. Olsson, "Admission Control for Streaming Services over HSDPA", *Proceedings of the Advanced Industrial Conference on Telecommunications/Service Assurance with Partial and Intermittent Resources Conference (AICT/SAPIR)*, Lisbon, Portugal, pp. 255-260, July 2005.

[25] J. Gadze, N. Pissinou, K. Makki and G. Crosby "On Optimal Slot Allocation for Reservation TDMA MAC Protocol in Shadow Fading Environment", *Proceedings of the EEE International Symposium on Wireless Communication Systems (ISWCS)*, Trondheim, Norway, pp. 809-813, October 2007.

[26] N. A. Ali, H. Hayajneh and H. Hassanein, "Cross Layer Scheduling Algorithm for IEEE 802.16 Broadband Wireless Networks", *Proceedings the IEEE International Conference on Communications (ICC)*, Beijing, China, pp. 3858-3862, May 2008.

[27] Y. Afek, M. Cohen, E. Hallman and Y. Mansour, "Dynamic Bandwidth Allocation Policies", *Proceedings of the Annual Joint Conference of the IEEE Computer Societies (INFOCOM)*, San Francisco, U.S.A, vol. 2, pp. 880-887, March 1996.

[28] X. Xu, D. Liang, H. Jiang and X. Lin,, "Dynamic Bandwidth Allocation in Fixed BWA Systems", *Proceedings of the International Conference on Communication Technology (ICCT)*, Beijing, China, vol.2, pp.1000-1003, April 2003.

[29] S. Choi and K.G. Shin, "A Comparative Study of Bandwidth Reservation and Admission Control Schemes in QoS-Sensitive Cellular Networks", *ACM Wireless Networks*, vol. 6, pp. 289–305. 2000.

[30] Y.Ma, J.J. Han and K.S. Trivedi, "Call Admission Control for Reducing Dropped Calls in Code Division Multiple Access (CDMA) Cellular Systems", *Proceedings of the Annual Joint Conference of the IEEE Computer Societies (INFOCOM)*, NewYork, U.S.A, pp. 1481–1490, March 2000.

[31] C.W. Leong, W. Zhuang, Y. Cheng and L. Wang, "Optimal Resource Allocation and Adaptive Call Admission Control for Voice/Data Integrated Cellular Networks", *IEEE Transactions on Vehicular Technology*, vol. 55, no. 2, pp. 654-669, March 2006.

[32] M. Kazmi, P. Godlewski and C. Cordier, "Admission Control Strategy and Scheduling Algorithms for Downlink Packet Transmission in WCDMA", *Proceedings of the IEEE Vehicular Technology Conference (VTC)*, Boston, U.S.A, vol. 2, pp. 674–680, September 2000.

[33] D. Kim, "Efficient Interactive Call Admission Control in Power-Controlled Mobile Systems", *IEEE Transactions on Vehicular Technology*, vol. 49, pp. 1017–1028, May 2000.

[34] B. Li, L. Li, B. Li, K.M. Sivalingam and X-R. Cao, "Call Admission Control for Voice/Data Integrated Cellular Networks: Performance Analysis and Comparative Study", *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 4, pp. 706-718, May 2004.

[35] S. Sun and W.A. Krzyman, "Call Admission Policies and Capacity Analysis of A Multi-Service CDMA Personal Communication System with Continuous and Discontinuous Transmission", *Proceedings of the IEEE Vehicular Technology Conference (VTC)*, Ottawa, Canada, vol. 1, pp. 218–223, May 1998.

[36] B. Hjelm, "Admission Control in Future Multi-Service Wideband Direct-Sequence CDMA (WCDMA) Systems", *Proceedings of the IEEE Vehicular Technology Conference (VTC)*, Boston, U.S.A, vol. 3, pp. 1086–1093, September 2000.

[37] B.M. Epstein and M. Schwartz, " Predictive QoS-based Admission Control for Multi Class Traffic in Cellular Wireless Networks", *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 3, March 2000.

[38] N. Nasser and H. Hassanein, "Radio Resource Management Algorithms in Wireless Cellular Networks", *Handbook of Algorithms for Wireless Networking and Mobile Computing*, A. Boukerche (ed.), pp. 415-437, Chapman Hall, CRC Press, 2006.

[39] H. Varian, "Intermediate Microeconomics: A Modern Approach", 7th edition, W.W. Norton & Company, 2005.

[40] J. Hou, J. Yang and S. Papavassilliou, "Integration of Pricing with Call Admission Control to Meet QoS Requirements in Cellular Networks", *IEEE Transactions on Parallel and Distributed Systems*, vol. 13, no. 9, pp. 898-910, September 2002.

[41] S. Yaipairoj and F.C. Harmantzis, "Congestion Pricing with Alternatives for Mobile Networks", *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC)*, Atlanta, U.S.A, vol. 4, pp. 671-676, March 2004.

[42] S.L. Hew and L. B. White, "Optimal Integrated Call Admission Control and Congestion Pricing with Handoffs and Price-Affected Arrivals", *Proceedings of the Asian-Pacific Conference on Communications (APCC)*, Perth, Australia, pp. 396-400. October 2005.

[43] S. Mandal, D. Saha and A. Mahanti, "A Technique to Support Congestion Pricing Strategy for Differentiated Cellular Mobile Services", *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM)*, St. Louis, U.S.A, vol. 6, pp. 3388-3392, December 2005.

[44] S. Yaipairoj and F.C. Harmantzis, "Auction-based Congestion Pricing for Wireless Data Services", *Proceedings of the IEEE International Conference on Communications (ICC)*, Istanbul, Turkey, pp. 1059-1064, June 2006.

[45] S. Mandal, D. Saha and M. Chatterjee, "Pricing Wireless Network Services Using Smart Market Models", *Proceedings of the IEEE Consumer Communications and Networking Conference (CCNC)*, Las Vegas, U.S.A., vol. 1, pp. 574-578, January 2006.

[46] S. Mandal, D. Saha and M. Chatterjee, "Dynamic Price Discovering Models for Differentiated Wireless Services", *Journal of Communications*, vol. 1, no. 5, pp. 50-56, August 2006.

[47] E. Viterbo and C.F. Chiasserini, "Dynamic Pricing for Connection-Oriented Services in Wireless Networks", *Proceedings of the IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, San Diego, U.S.A, vol. 1, pp. 68-72, September 2001.

[48] S.W. Han and Y. Han, "A Simple Congestion Pricing in Wireless Communication", *Proceedings of the IEEE Vehicular Technology Conference (VTC)*, Dallas, U.S.A, vol. 5, pp. 795-798, September 2005.

[49] J.W. Lee, R.R. Mazumdar and N.B. Shroff, "Downlink Power Allocation for Multi-class CDMA Wireless Networks", *Proceedings of the IEEE Joint Conference of Computer and Communications Societies (INFOCOM)*, New York, U.S.A, pp. 1480-1489, June 2002.

[50] J.W. Lee, R.R. Mazumdar and N.B. Shroff, "Downlink Power Allocation for Multi-class CDMA Wireless Systems", *IEEE/ACM Transactions on Networking*, vol. 13, no. 4, pp. 854-867, August 2005.

[51] P. Liu, M.L. Honig and S. Jordan, "Forward-Link CDMA Resource Allocation Based on Pricing", *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC)*, Chicago, U.S.A, vol. 3, pp. 1410-1414, September 2000.

[52] P. Liu, S. Jordan and M.L. Honig ,"Single-Cell Forward Link Power Allocation Using Pricing in Wireless Networks", *IEEE Transactions on Wireless Communications*, vol. 3 no. 2, pp. 533-543, March 2004.

[53] P. Zhang, S. Jordan, P. Liu and M.L. Honig, "Power Control of Voice Users Using Pricing in Wireless Networks" *Proceedings of SPIE ITCOM 2001 Conference on Modeling and Design of Wireless Networks*, Denver, U.S.A, vol. 4531, p. 155-165, August 2001.

[54] B. Al-Manthari, N. Nasser and H. Hassanein, "Dynamic Pricing in Wireless Cellular Networks", submitted to the *IEEE Communications Surveys and Tutorials*, June 2008.

[55] V. Pandey, D. Ghosal and B. Mukherjee, "Pricing-Based Approaches in the Design of Next-Generation Wireless Networks: A Review and A Unified Proposal", *IEEE Communications Surveys and Tutorials*, vol. 9, no. 2, pp. 88-101, May 2007.

[56] Deliverable D3. 2v2, "End-to-end Network Model for Enhanced UMTS", Available: http://www.ti-wmc.nl/eurane/

[57] Y.S. Kim "Capacity of VoIP over HSDPA with Frame Bundling" *IEICE Transaction of Communications*, vol. E89-B, no.12, pp.3450-3453, December 2006.

[58] 3GPP TS 26.071 V6, "AMR Speech Codec; General Description," December 2004.

[59] C. Bormann, C. Burmeister, M. Degermark, H. Fukushima, H. Hannu, L. Jonsson, R. Hakenberg, T. Koren, K. Le, Z. Liu, A. Martensson, A. Miyazaki, K. Svanbro, T. Wiebke, T. Yoshimura and H. Zheng, "RObust Header Compression (ROHC): Framework and four profiles: RTP, UDP, ESP, and uncompressed", IETF RFC Standards Track 3095, July 2001.

[60] ITU-T, "One-way transmission time" G.114, May 2003.

[61] B. Wang, K. I. Pedersen, T. E. Kolding and P. E. Mogensen, "Performance of VoIP over HSDPA", *Proceedings of the IEEE Vehicular Technology Conference* (*VTC*), vol. 4, pp. 2335–2339, May 2005.

[62] R. Lloyd-Evan, "QoS in Integrated 3G networks", 1st edition, Artech House, 2002.

[63] 3GPP TS 23.107 V5.12.0, "Quality of Service (QoS) Concept and Architectures", Release 5, March 2004.

[64] 3GPP TS 22.105 V 6.4.0, "Services and Service Capabilities", Release 6, September 2005.

[65] G. Rittenhouse and H. Zheng, "Providing VoIP Service in UMTS-HSDPA with Frame Aggregation", *Proceedings of the IEEE International Conference on* Acoustics, Speech, and Signal Processing (ICASSP), Philadelphia, PA, U.S.A, vol. 2, pp.1157-1160, March 2005.

[66] P. Lunden and M. Kuusela, "Enhancing Performance of VoIP over HSDPA", *Proceedings of the IEEE Vehicular Technology Conference* (*VTC*), Dublin, Ireland, pp. 825-829, April 2007.

[67] P. Lunden, J. Aijanen, K. Aho and T. Ristaniemi, "Performance of VoIP over HSDPA in Mobility Scenarios", *Proceedings of the IEEE Vehicular Technology Conference* (*VTC*), Marina Bay, Singapore, pp. 2046-2050, May 2008.

[68] R. Jain, D. Chiu and W. Hawe, "A Quantitative Measure of Fairness and Discrimination for Recourse Allocation in Shared Computer Systems", DEC Research Report, TR-301, September 1984.

[69] A. Duel-Hallen, S. Hu and H. Hallen, "Long-Range Prediction of Fading Signals", *IEEE Signal Processing Magazine*, vol. 17, pp. 62–75, May 2000.

[70] R. Vaughan, P. Teal, and R. Raich, "Short-Term Mobile Channel Prediction Using Discrete Scatterer Propagation Model and Subspace Signal Processing Algorithms", *Proceedings of the IEEE Vehicular Technology Conference (VTC)*, Boston, MA, USA, September 2000.

[71] T. Ekman, "Prediction of mobile radio channels, modeling and design", Ph.D. thesis, Uppsala University, Sweden, 2002.

[72] http://lpsolve.sourceforge.net/5.5/

[73] D. Hong and S.S. Rappaport, "Traffic Model and Performance Analysis for Cellular Mobile Radio Telephone Systems with Prioritized and None-prioritized Handoff Procedures," *IEEE Transactions of Vehicular Technology*, vol. 35, no. 3, pp. 77-92, August 1986.

[74] J. Hou and S. Papavassiliou, "Influence-Based Channel Reservation Scheme for Mobile Cellular Networks," *Proceedings of the IEEE Symposium on Computers and Communications (ISCC)*, Hammamet, Tunisia, pp. 218-223, July 2001.

[75] M. Salamah, "An Adaptive Multi-Guard Channel Scheme for Mutli-Class Traffic in Cellualr Networks", *Proceedings of the IEEE International Conference on Computer Systems and Applications (AICCSA)*, Dubai/Sharjah, U.A.E, pp. 716-723, March 2006.

[76] H. K. Pati, "A Distributed Adaptive Guard Channel Reservation Scheme for Cellular Networks", *International Journal of Communication Systems*, vol. 20, no. 9, pp. 1037-1058, September 2007.

[77] N.S. Khalil, H. Lababidih and M. Salamah, "Dynamic Guard Channel Allocation Scheme for Calls in WONs", *Electronic Letters*, vol. 43, no. 3, pp. 170-171, February 2007.

[78] D. Liu and Y. Zhang, "A Self-Learning Adaptive Critic Approach for Call Admission Control in Wireless Cellular Networks", *Proceedings of the IEEE International Conference on Communications (ICC)*, Anchorage, U.S.A., vol. 3, pp.1853-1857, May 2003.

[79] R. L. Freeman, "Telecommunication System Engineering", 3rd edition, Wiley, 1996.

[80] E. D. Fitkov-Norris, A Khanifar, "Congestion pricing in Cellular Networks, A Mobility Model with a Provider-Oriented Approach", *Proceedings of the IEEE International Conference on 3G Mobile Communication Technologies (3G)*, London, UK, pp. 63-67, March 2001.

[81] P. A. Hosein, "QoS Control for WCDMA High Speed Packet Data", *Proceedings Of the IEEE International Workshop of Mobile and Wireless Communication Networks (MWCN)*, Stockholm, Sweden, pp. 169-173, September 2002.

[82] F. Brouwer, I. Bruin, J. Silva, N. Souto, F. Ceras and A. Correia, "Usage of Link-Level Performance Indicators for HSDPA Network-Level Simulations in E-UMTS", *Proceeding of the IEEE International Symposium on Spread Spectrum Techniques and Applications (ISSSTA*), Sydney, Australia, pp. 844-848, August 2004.

Appendix A

Reduction Proofs of the Packet Scheduling Scheme

In this Appendix, we present the proofs for lemmas 1 and 2 in Chapter 3.

A.1 Proof of Lemma 1:

If p_{ij} is set to 1 for every user (i.e., the price is ignored), then the set $\{\mathbf{Rv}^g\}_{g=1}^{N}$ becomes $\{\mathbf{Rv}^g\}_{g=1}^{N} = \{Rv_{ij}^1,\ Rv_{ij}^2, ..., Rv_{ij}^N \mid Rv_{ij}^g \geq Rv_{ij}^{g+1}\}$, where $Rv_{ij}^g = R_{ij}(t)$. That is, the set $\{\mathbf{Rv}^g\}_{g=1}^{N}$ in this case will contain all users in the system in descending order by their instantaneous

supportable data rates. Therefore, $\mathrm{Re}\, v_{Max} = \sum_{g \in \{\mathbf{Rv}^g\}_{g=1}^{N}} Rv^g$, will be the sum of the maximum instantaneous supportable data rates of users that could send in the current time frame without exceeding the system capacity. Therefore, when H is set to 0, we get $OC_{\mathrm{N}^*}(t) = \mathrm{Re}\, v_{Max} - \sum_{j \in \mathrm{N}^*} p_{ij} \cdot R_{ij}(t) \le 0$. That is, the only set of users that satisfy $OC_{\mathrm{N}^*}(t) \le 0$ is the set of users that constitutes $\mathrm{Re}\, v_{Max}$. These are the users with the maximum instantaneous supportable data rates (in descending order) in the system. Therefore, at each scheduling decision, our scheduler will choose the set of users with the maximum instantaneous supportable data rates. This is equivalent to Max CIR.

A.2 Proof of Lemma 2:

As we will show later, the PF algorithm requires that $U_{ij}(t) = \ln\left(\overline{S_{ij}(t)} / \max_{ij} \overline{S_{ij}(t)}\right)$. Therefore, we first need to find the value of a_i such that $1 - e^{-a_i \cdot \sum_{z=1}^{m_{ij}} \left(X_{ij}^z(t)\right)} = \ln\left(\overline{S_{ij}(t)} / \max_{ij} \overline{S_{ij}(t)}\right)$. We get this by solving

$$e^{-a_i \cdot \sum_{z=1}^{m_{ij}} \left(X_{ij}^z(t)\right)} = 1 - \ln\left(\overline{S_{ij}(t)} / \max_{ij} \overline{S_{ij}(t)}\right) \qquad \text{(A.1)}$$

Which implies that,

$$-a_i \cdot \sum_{z=1}^{m_{ij}} \left(X_{ij}^{z}(t) \right) = \ln\left(1 - \ln\left(\overline{S_{ij}(t)} / \max_{ij} \ \overline{S_{ij}(t)} \right) \right) \tag{A.2}$$

Therefore, if we set $a_i = -\ln\left(1 - \ln\left(\overline{S_{ij}(t)} / \max_{ij} \ \overline{S_{ij}(t)} \right) \right) / \sum_{z=1}^{m_{ij}} \left(X_{ij}^{z}(t) \right)$, ignore the QoS constraints and ignore the opportunity cost constraint (by setting H is set to $\mathrm{Re}\, v_{Max}$), then the utility function in our scheduling scheme becomes $U_{ij}(t) = \sum_{i=1}^{K} \sum_{j=1}^{N_i} \ln\left(\overline{S_{ij}(t)} / \max_{ij} \ \overline{S_{ij}(t)} \right)$. Since the term $\max_{ij} \ \overline{S_{ij}(t)}$ is common to every user, we can then take it off, and hence the utility function becomes $\ln\left(\overline{S_{ij}(t)} \right)$. Therefore, our scheduling scheme will find the set of users so that:

$$\text{Objective: } \mathbf{N}^* = \text{Maximize} \sum_{i=1}^{K} \sum_{j=1}^{N_i} \ln\left(\overline{S_{ij}(t)} \right)$$

$$\text{Subject to: } \left(\sum_{(i,j) \in \mathbf{N}^*} R_{ij}(t) \right) \leq C \tag{A.3}$$

Maximizing the aggregate utility of the system is equivalent to maximizing the objective function F, where F is a function of $\vec{S}(t)$, and $\vec{S}(t)$ is a vector of the users' average throughputs at time t. That is, if we find a vector $\vec{S}(t)$ that maximizes F, then the aggregate utility function will also be maximized. Therefore the problem can be formulated as follows:

$$\text{Objective: Maximize} \quad F\left(\vec{S}(t)\right) \equiv \sum_{i=1}^{K}\sum_{j=1}^{N_i} \ln\left(\overline{S_{ij}(t)}\right)$$

$$\text{Subject to:} \quad \left(\sum_{(i,j)\in \mathbf{N}^*} R_{ij}(t)\right) \le C \qquad \text{(A.4)}$$

where $\overline{S_{ij}(t)}$ can be calculated by using an exponentially smoothed filter as follows [16]:

$$\overline{S_{ij}(t)} = \begin{cases} (1-1/t_c)\cdot\overline{S_{ij}(t-1)} + 1/\mathrm{t_c}\cdot R_{ij}(t) & \textit{if user } i \textit{ is served} \\ (1-1/t_c)\cdot\overline{S_{ij}(t-1)} & \textit{Otherwise} \end{cases} \qquad \text{(A.5)}$$

where t_c is the time constant of the filter and $R_{ij}(t)$ is the current supportable data rate of user *j*.

Since $\ln\left(\overline{S_{ij}(t)}\right)$ is strictly concave and is differentiable then so is the objective function *F*. Also since the feasible region is bounded, then an optimal solution exits. Furthermore, the solution is unique and we can use a gradient ascent method to find it as explained in [81]. However, a global optimal solution cannot be found, since the number of users and the channel capacity are varying with time. Nevertheless, we can look for a locally optimal solution. That is, at each time frame, schedule the set of users that would result in a movement towards the optimal solution. Let $F'_{ij}(\vec{\mathrm{S}}\,(\mathrm{t}))$ be the gradient of the objective function in the direction of serving user *j* (we focus on one user here, the rest of users in the set $\mathbf{N}^*$ can be found one by one using the same method). We would like to find the value of *j* with the largest gradient and moving to the maximal point along that

direction. Since we know what the user's average throughput would be if served or not, then the optimization problem can be reduced to finding the maximum gradient in the direction of serving user j (i.e., maximize $F'_{ij}(\vec{S}(t))$). We first find the gradient in the direction of serving user j. We can do this by parameterizing the movement along the ray in the direction of serving user j by μ, and then F_{ij} can be written as a function of μ as follows:

$$F_{ij}(\mu) = \sum_{i=1}^{K}\sum_{j=1}^{N_i} \ln\left(\overline{S_{ij}(t)} + \mu(\overline{S_{ij}(t+1)} - \overline{S_{ij}(t)})\right) \tag{A.6}$$

Taking the derivative with respect to μ and evaluating it at $\mu = 0$ (to find the critical point, in this case maxima), we get

$$\begin{aligned} F'_{ij}(\mu) &= \sum_{i=1}^{K}\sum_{j=1}^{N_i} \frac{1}{\left(\overline{S_{ij}(t)} + \mu(\overline{S_{ij}(t+1)} - \overline{S_{ij}(t)})\right)} \cdot \left(\overline{S_{ij}(t+1)} - \overline{S_{ij}(t)}\right) \\ &= \frac{1}{\left(\overline{S_{ij}(t)} + \mu(\overline{S_{ij}(t+1)} - \overline{S_{ij}(t)})\right)} \cdot \left(\overline{S_{ij}(t+1)} - \overline{S_{ij}(t)}\right) \\ &+ \sum_{i=1}^{K}\sum_{y=1, y\neq i}^{N_i} \frac{1}{\left(\overline{S_{iy}(t)} + \mu(\overline{S_{iy}(t+1)} - \overline{S_{iy}(t)})\right)} \cdot \left(\overline{S_{iy}(t+1)} - \overline{S_{iy}(t)}\right) \end{aligned}$$

$$\begin{aligned} \therefore \quad F'_{ij}(0) &= \frac{1}{\overline{S_{ij}(t)}} \cdot \left(\overline{S_{ij}(t+1)} - \overline{S_{ij}(t)}\right) + \sum_{i=1}^{K}\sum_{j=1, y\neq i}^{N_i} \frac{1}{\overline{S_{iy}(t)}} \cdot \left(\overline{S_{iy}(t+1)} - \overline{S_{iy}(t)}\right) \\ &= \frac{1}{\overline{S_{ij}(t)}} \cdot \left(\frac{R_{ij}(t)}{t_c} - \frac{\overline{S_{ij}(t)}}{t_c}\right), \qquad \text{(user } j \text{ is served (Eq. A.5))} \end{aligned}$$

$$+\sum_{i=1}^{K}\sum_{y=1,y\neq j}^{N_i}\frac{1}{\overline{\overline{S_{iy}(t)}}}\cdot\left(-\frac{\overline{S_{iy}(t)}}{t_c}\right), \qquad \text{(user } y \text{ is not served (Eq. A.5))}$$

$$=\frac{1}{\overline{\overline{S_{ij}(t)}}}\cdot\left(\frac{R_{ij}(t)}{t_c}\right)-\frac{1}{\overline{\overline{S_{ij}(t)}}}\cdot\left(\frac{\overline{S_{ij}(t)}}{t_c}\right)-\sum_{i=1}^{K}\sum_{y=1,y\neq j}^{N_i}\frac{1}{\overline{\overline{S_{iy}(t)}}}\cdot\left(-\frac{\overline{S_{iy}(t)}}{t_c}\right)$$

Therefore, the gradient in the direction of serving user *j* can be written as:

$$\frac{1}{\overline{\overline{S_{ij}(t)}}}\cdot\left(\frac{R_{ij}(t)}{t_c}\right)-\sum_{i=1}^{K}\sum_{y=1}^{N_i}\frac{1}{\overline{\overline{S_{iy}(t)}}}\cdot\left(-\frac{\overline{S_{iy}(t)}}{t_c}\right) \qquad \text{(A.7)}$$

The summation term and the constant scalar t_c are common terms for all users, and can be ignored. Consequently, the maximum gradient direction is reduced into $\arg\max_j F'_{ij}(\vec{S}(t))=\arg\max_j \frac{R_{ij}(t)}{\overline{\overline{S_{ij}(t)}}}$, which is the same as the PF scheme. The rest of users in set $\mathbf{N}^*$ are chosen based on their descending order of the aggregate system utility if they are served so that the time frame of the wireless system is filled. That is, they are served based $\arg\max_j \frac{R_{ij}(t)}{\overline{\overline{S_{ij}(t)}}}$. That is, all users in the set $\mathbf{N}^*$ are served based on the PF scheme. Therefore, if a_i is set to $a=-\ln\left(1-\ln\left(\overline{S_{ij}(t)}/\max_{ij}\ \overline{S_{ij}(t)}\right)\right)/\sum_{z=1}^{m_{ij}}\left(X_{ij}^{z}(t)\right)$, the QoS constraints are ignored and $OC_{\mathbf{N}^*}(t)$ is ignored (i.e., H is set to $\mathrm{Re}\,v_{Max}$), then our packet scheduling scheme reduces to the PF scheme.

Appendix B

Simulation Parameters

In this Appendix, we present the relevant simulation parameters used in the performance evaluation of our framework's components in Chapters 3, 4 and 5.

B.1 Simulation Parameters

Table B.1 shows the relevant simulation parameters that are utilized in all our experiments in this thesis.

Table B.1: Utility function Parameters

Parameter	Value
Tested Environment	Pedestrian A
Mobile speed for Pedestrian A	3 km/hr
Cell diameter	1000 m

Number of cells	1
Base station transmission power	38 dBm
Distance loss at 1 Km	137.4
Path loss exponent	3.52
Intra cell interference	30 dBm
Inter cell interference	-70 dBm
Base station height	30 m
Frame period	2 ms
Time slots per frame	3
Number of connection requests per user	1
Block error probability	10%
Channalization codes	10
Maximum system capacity	7.2 Mpbs (with 10 codes)
HSDPA coverage	Full
Node B buffer size	500 Mb
Connection request arrival rate	Poisson process
Spatial distribution of the users in the cell	Uniform

B.2 Utility Function Parameters

B.2.1 Multiplexed Traffic Case

Table B.2 shows the parameters used in the utility function of the proposed packet scheduling scheme in the multiplexed traffic case studied in Chapter 3. It should be noted that the values of these parameters are chosen to achieve certain levels of inter- and intra class prioritization, and hence achieve different fairness levels as shown in the results of Chapter 3. Therefore, the network operator may use different values, if desired, to achieve different levels of prioritizations and fairness, where the role of each parameter in the utility function is discussed in Sections 3.3.1 and 3.3.3. In section B.4, we investigate the effect of some of these parameters on the system performance.

Table B.2: Utility function parameters for multiplexed traffic case

Traffic Type	a_i [18]	P_{ij}^1	P_{ij}^2
VoIP	4	0.5	0.4
Audio Streaming	3.5	0.5	0.45
Video Streaming	3	0.5	0.7
FTP	2.5	0.5	0.7

[18] As explained in Chapter 3, larger values for the a_i parameter result in higher class prioritization.

B.2.2 All Other Cases

Table B.3 shows the parameters used in the utility function of the proposed packet scheduling scheme in all cases considered in this thesis except the multiplexed traffic case of Chapter 3.

Table B.3: Utility function parameters for all other cases

Traffic Type	a_i	P_{ij}^1	P_{ij}^2
VoIP	4	0.5	0.4
Audio Streaming	4	0.5	0.4
Video Streaming	3.5	0.5	0.5
FTP	3.5	0.5	0.5

B.3. Channel Model

The channel model describes the attenuation of the radio signal on its way from the base station to the user, and therefore, it describes how the channel condition of the user changes with time depending on the environment and the speed of the user. In our simulation, we utilize the code provided in [56] in order to simulate the user's varying channel quality conditions. In this code, the channel model consists of five parts: distance loss, shadowing, multi-path fading, intra-cell interference, and inter-cell interference. Each one of these parts is considered independently and is expressed in dB.

Path loss degrades the radio signal and is proportional to the distance between the base station and the user. It is described by the Okamura-Hata model for suburban areas as follows:

$$L(d) = L_{init} + 10.\beta \log_{10}(d) \tag{B.1}$$

where L_{init} is the distance loss at 1 km and is equal to 137.4, d is the distance from the mobile user to the base station in kilometers, β is the path loss exponent and is equal to 3.52. Shadowing or slow fading is caused by obstacles between the user and the base station. In our simulation, a correlated slow fading model is used and is constructed in the following manner:

$$D(d + \Delta d) = a \cdot D(d) + b \cdot \tilde{\sigma} \cdot N \tag{B.2}$$

where

$$a = \exp\left(-\Delta d / D_{corr}\right) \tag{B.3}$$

and

$$b = \sqrt{1 - a^2} \tag{B.4}$$

where Δd is the change in the distance between two subsequent time samples, N is the standard normal distribution and has a random value, $\tilde{\sigma}$ is the standard deviation and has a typical value of 8 dB in suburban areas and D_{corr} is the correlation distance, which

depends on the environment of the user. In pedestrian A environment, D_{corr} is typically set to 40 meters.

As the radio signal travels from the base station to the user, it is diffracted by obstacles, which result in several copies of the same signal. Thus, the received signal is a sum of those copies and this phenomenon is known as multi-path fast fading. In our simulation, multi-path corresponds to 3GPP channel models for Pedestrian A environment. Intra-cell and inter-cell interference are assumed to be constants as real life fluctuations in interference have little impact on the end result of the end-to-end simulator when compared to the variations introduced by path loss, shadowing and multi-path fading. Intra- and inter-cell interference are set to 30 and -70 dBm, respectively. Then at the user side, the Signal-to-Noise Ratio (SNR) is extracted from the received signal from the base station to determine how strong the signal is according to the following formula:

$$\begin{aligned} SNR &= P_{tx} - L_{Total} - 10 \cdot \log_{10}(10^{\frac{I_{intra} - L_{Total}}{10}} + 10^{\frac{I_{inter}}{10}}) \\ &= P_{tx} - 10 \cdot \log_{10}(10^{\frac{I_{intra}}{10}} + 10^{\frac{I_{inter} + L_{Total}}{10}}) \end{aligned} \quad \text{(B.5)}$$

where P_{tx} is the transmitted code power in dBm, L_{Total} is the sum of the path loss, shadowing and multi-path fading in dB, I_{intra} and I_{inter} are the intra and inter cell interference respectively in dBm.

The SNR is then mapped to Channel Quality Indicator (CQI) that is used to determine the rate at which the user can support from the base station. CQI is approximated through a linear function, based on 3GPP standard as follows [82]:

$$CQI = \begin{cases} 0 & SNR \le -16 \\ \left\lfloor \frac{SNR}{1.02} + 16.62 \right\rfloor & -16 < SNR < 14 \\ 30 & 14 \le SNR \end{cases} \quad \text{(B.3)}$$

The HSDPA specification comes with tables that determine the data rates for each combination of CQI and channel codes. These tables are used in our simulation and can be found in [7]. Therefore, the data rates that the users can receive from the base station vary in time depending on their locations, speeds, and the environments.

B.4 Additional Simulation Results

In this section, we present additional simulation results for different values of a_i, P_{ij}^1 and P_{ij}^2 in order to show their effects on inter- and intra-class prioritization. We test these parameters with our proposed packet scheduling scheme with maximum tolerable revenue loss of $\mathrm{Re}\, v_{Max}$. Unless otherwise specified, we use the same simulation models described in Chapter 3 as well as the same simulation parameters in Table B.2.

B.4.1 Effect of a_i on Inter-Class Prioritization

To demonstrate the effect of a_i on inter-class prioritization, we consider two types of traffic, VoIP and audio streaming. We fix a_i at 4 for VoIP and choose different a_i values for audio streaming. Figures B.1, B.2 and B.3 show the percentage of service coverage for both types of traffic when a_i for audio streaming is set to 2, 3 and 3.5, respectively. Clearly, as a_i is increased, the percentage of service coverage for audio streaming increases. This is because larger values of a_i make the utility function of audio streaming users steeper. Hence, users have more chance of being scheduled to transmit, which leads to improved service coverage.

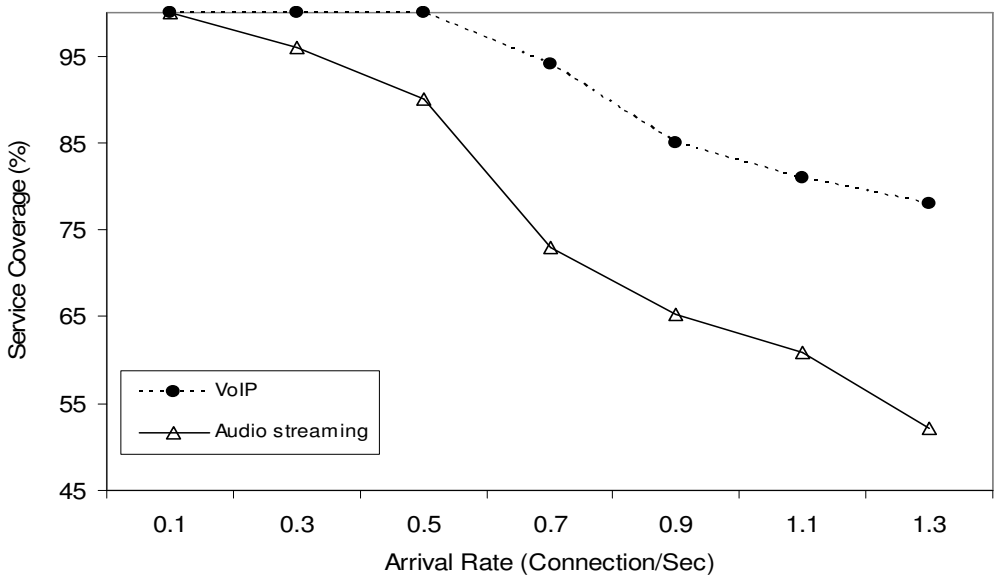


Figure B.1: Percentage of service coverage with $a_i = 2$ for audio streaming

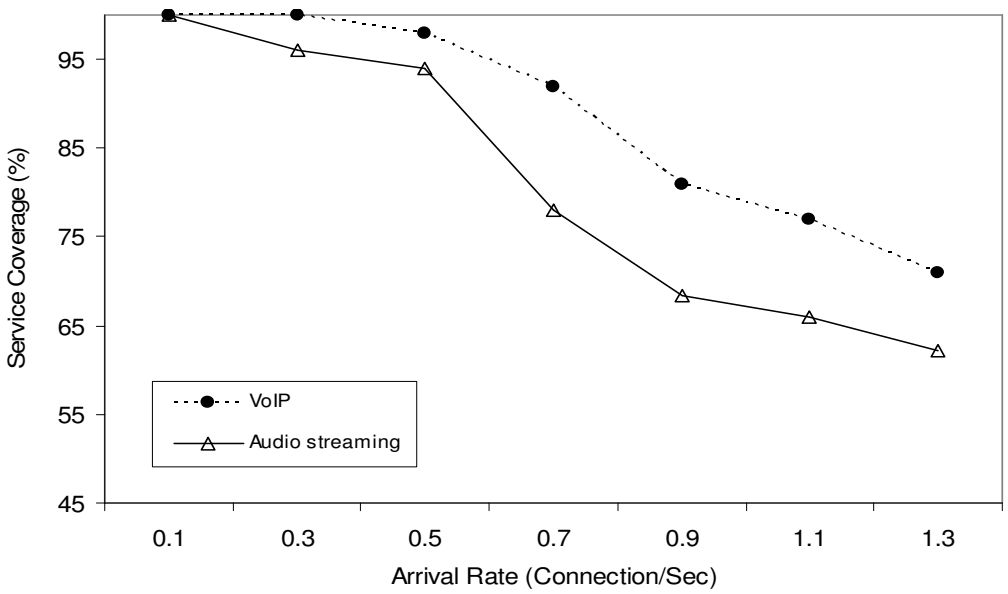


Figure B.2: Percentage of service coverage with $a_i = 3$ for audio streaming

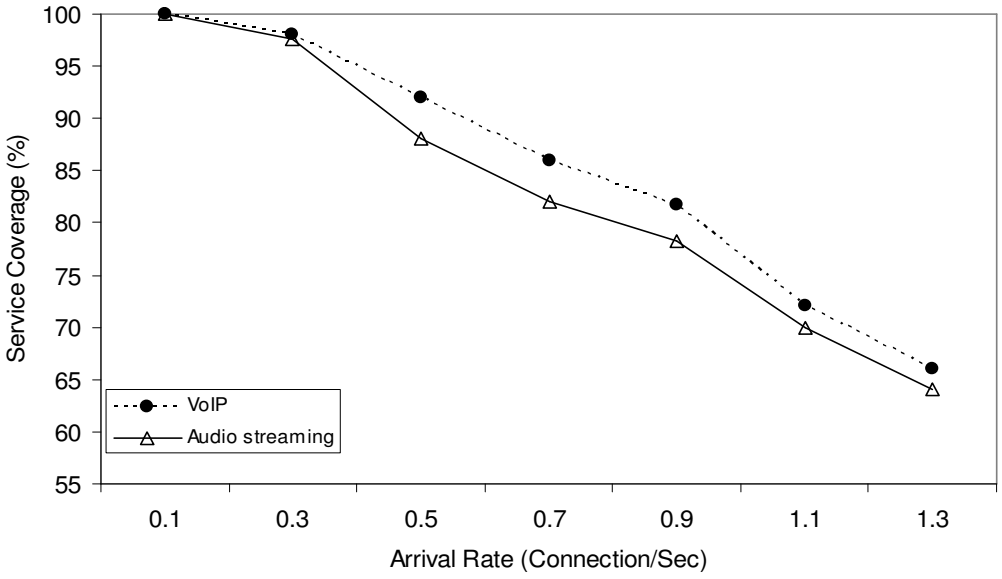


Figure B.3: Percentage of service coverage with $a_i = 3.5$ for audio streaming

B.4.2 Effect of P_{ij}^1 and P_{ij}^2 on Intra-Class Prioritization

To demonstrate the effect of P_{ij}^1 and P_{ij}^2 on prioritizing different QoS measures, we consider one type of traffic, VoIP, with a_i fixed at 4. In our first experiment, we fix P_{ij}^2 of VoIP at 0.4 and choose different values for P_{ij}^1 to show its effect on exploiting the users' channel quality conditions. Figure B.4 shows the percentage of channel utilization when P_{ij}^1 is set to 0.5, 0.4, 0.3 and 0.2, respectively. As P_{ij}^1 is decreased, the percentage of channel utilization increases. This is expected because as explained in Chapter 3, the smaller the values of P_{ij}^1, the higher penalty of not exploiting the users' channel quality conditions, and hence the higher their weights in the scheduling decision. However, as more weight is given to the users' channel quality conditions, the percentage of service coverage decreases. This is depicted in Figure B.5. The reason for this is that in this case only users with good channel quality conditions are scheduled to transmit. Hence, fewer users are covered. Therefore, the network operator should choose the appropriate value of P_{ij}^1 as to achieve its desired level of channel utilization and the corresponding service coverage.

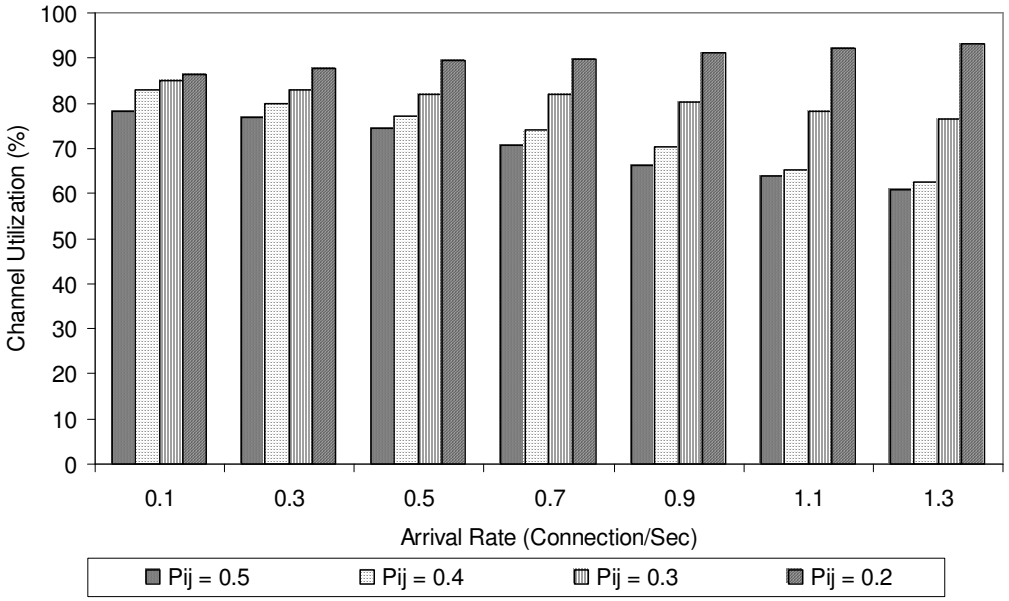


Figure B.4: Percentage of channel utilization for different values of P_{ij}^{1}

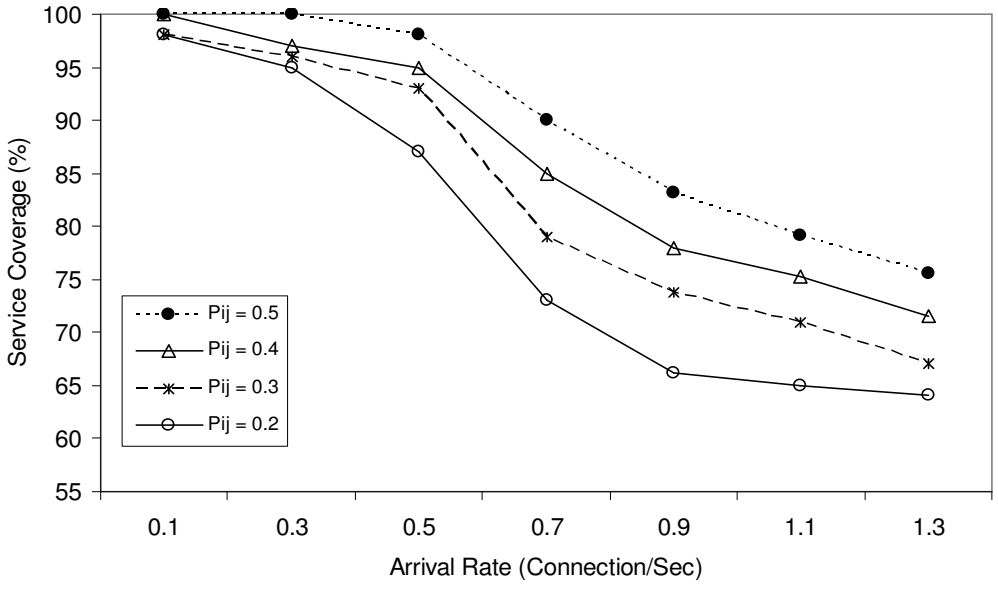


Figure B.5: Percentage of service coverage for different values of P_{ij}^{1}

In our second experiment, we fix P_{ij}^1 of VoIP at 0.5 and choose different values for P_{ij}^2 to show its effect on the delay QoS measure. As we decrease P_{ij}^2, the penalty of violating the delay measure increase, and hence more priority is given to users with larger average delays. This results in better service coverage as shown in Figure B.6. However, as more users are covered, the channel utilization decreases as already discussed in Chapter 3. Similar results can also be achieved when P_{ij}^2 is increased for best-effort traffic and traffic with minimum delay requirements, since in these two types of traffic, the larger the values of P_{ij}^2 are, the higher the penalty of violating the QoS measure. Therefore, the network operator should choose the appropriate values of P_{ij}^1 and P_{ij}^2 as to achieve its desired level of prioritization between the QoS measures.

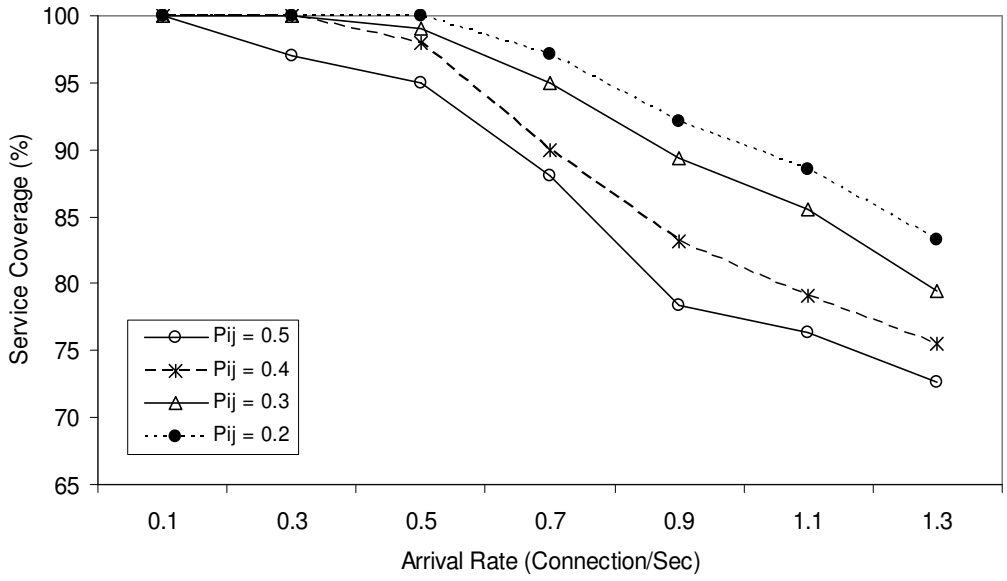


Figure B.6: Percentage of service coverage for different values of P_{ij}^2

Appendix C

Framework Flowcharts

This appendix presents the flowcharts of our bandwidth management framework. Figure C.1 shows the main flow of our framework passing through its major components. Figures C.2, C.3 and C.4 demonstrate the flowcharts of the CAC-based dynamic pricing scheme, dynamic bandwidth management scheme and the packet scheduling scheme, respectively.

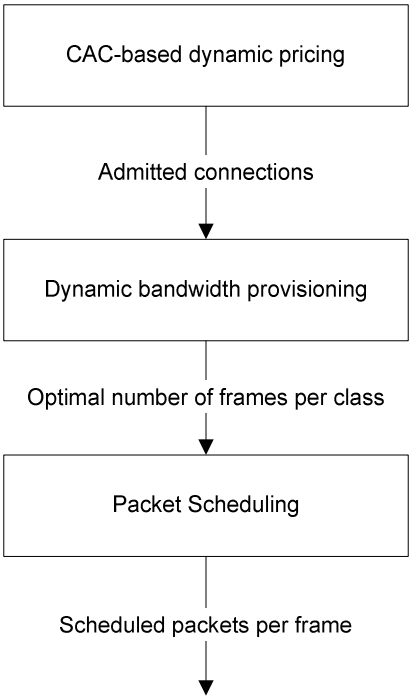


Figure C.1: Main flow of the framework

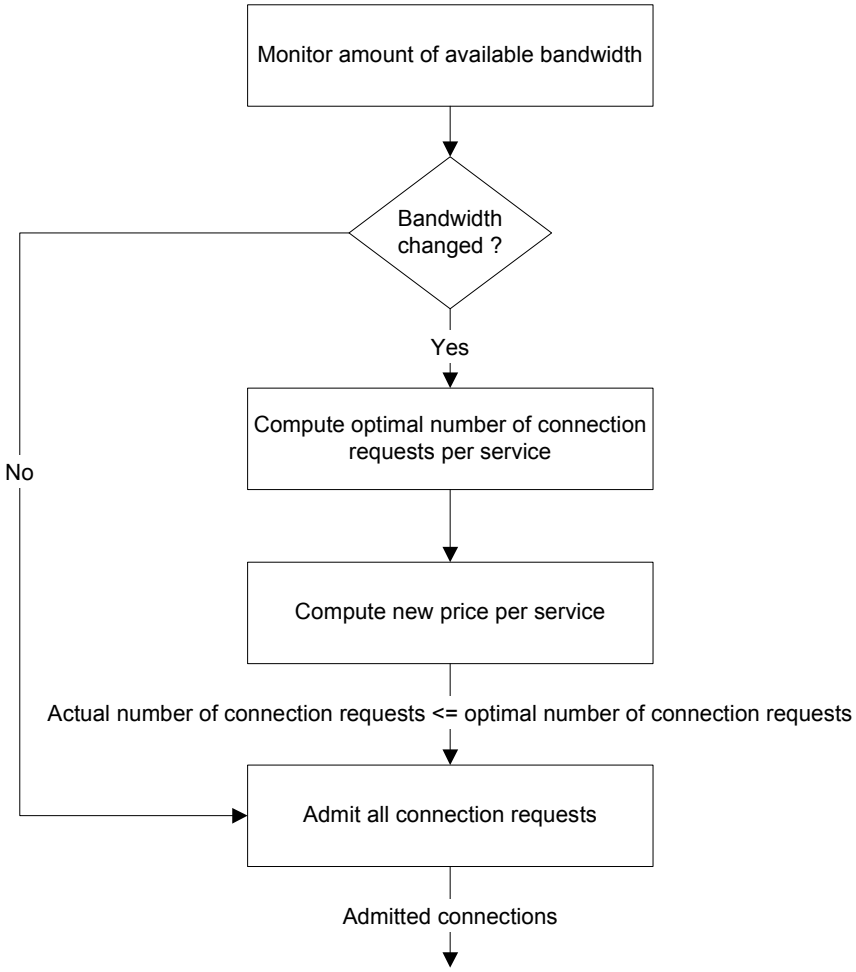


Figure C.2: CAC-based Dynamic Pricing Scheme flowchart

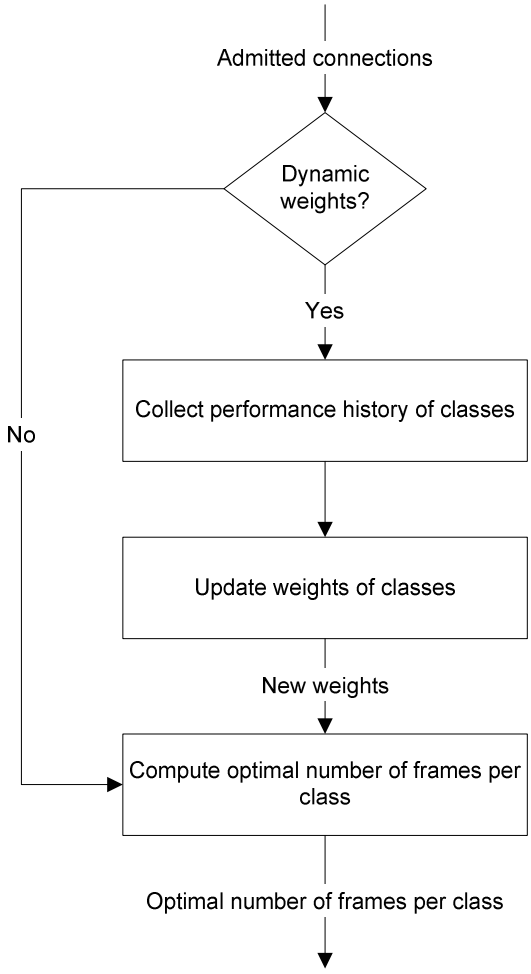


Figure C.3: Dynamic bandwidth provisioning scheme flowchart

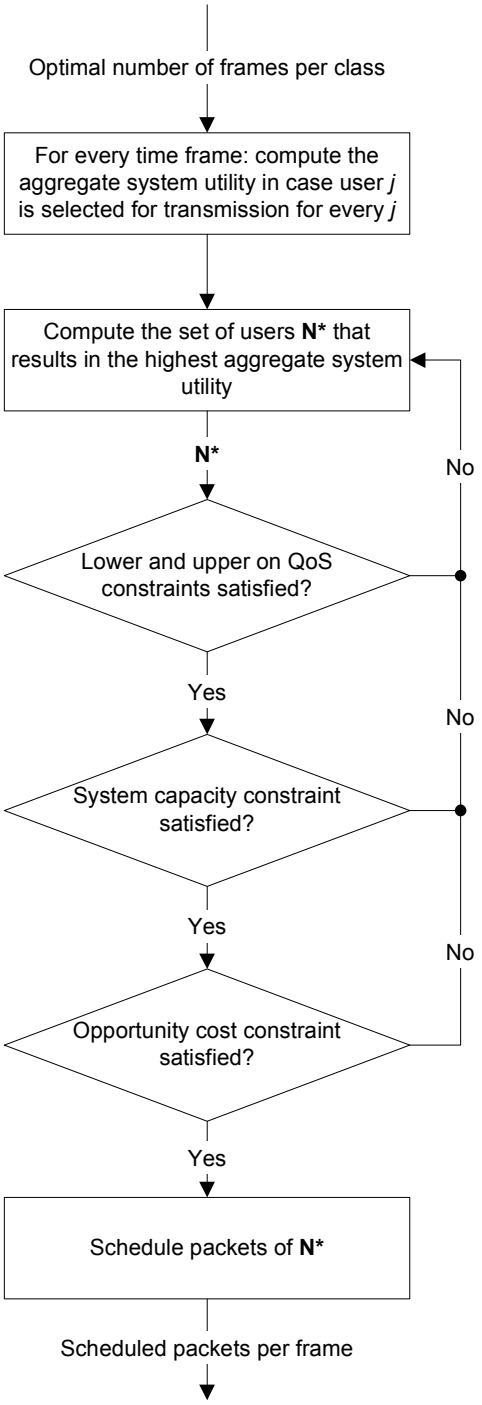


Figure C.4: Packet scheduling scheme flowchart