ABSTRACT

OPPORTUNISTIC TRANSMISSION SCHEDULING FOR NEXT GENERATION WIRELESS COMMUNICATION SYSTEMS WITH MULTIMEDIA SERVICES

by
Chengzhou Li

The explosive growth of the Internet and the continued dramatic increase for all wireless services are fueling the demand for increased capacity, data rates, and support of different quality of service (QoS) requirements for different classes of services. Since in the current and future wireless communication infrastructures, the performances of the various services are strongly correlated, as the resources are shared among them, dynamic resource allocation methods should be employed. With the demand for high data rate and support of multiple QoS, the transmission scheduling plays a key role in the efficient resource allocation process in wireless systems. The fundamental problem of scheduling the users' transmissions and allocating the available resources in a realistic CDMA wireless system that supports multi-rate multimedia services, with efficiency and fairness, is investigated and analyzed in this dissertation.

Our proposed approach adopts the use of dynamically assigned data rates that match the channel capacity in order to improve the system throughput and overcome the problems associated with the location-dependent and time-dependent errors and channel conditions, the variable system capacity and the transmission power limitation. We first introduce and describe two new scheduling algorithms, namely the Channel Adaptive Rate Scheduling (CARS) and Fair Channel Adaptive Rate Scheduling (FCARS). CARS exploits the channel variations to reach high throughput, by adjusting the transmission rates according to the varying channel conditions and by performing an iterative procedure to determine the power index that a user can accept by its current channel condition and transmission power. Based on the assignment of CARS and to overcome potential unfair service allocation, FCARS implements a compensation algorithm, in which the lagging users can receive

compensation service when the corresponding channel conditions improve, in order to achieve asymptotic throughput fairness, while still maintaining all the constraints imposed by the system.

Furthermore the problem of opportunistic fair scheduling in the uplink transmission of CDMA systems, with the objective of maximizing the uplink system throughput, while satisfying the users' QoS requirements and maintaining the long-term fairness among the various users despite their different varying channel conditions, is rigorously formulated, and a throughput optimal fair scheduling policy is obtained. The corresponding problem is expressed as a weighted throughput maximization problem, under certain power and QoS constraints, where the weights are the control parameters that reflect the fairness constraints. With the introduction of the power index capacity it is shown that this optimization problem can be converted into a binary knapsack problem, where all the corresponding constraints are replaced by the users' power index capacities at some certain system power index. It is then argued that the optimal solution can be obtained as a global search within a certain range, while a stochastic approximation method is presented in order to effectively identify the required control parameters. Finally, since some real-time services may demand certain amount of service within specific short span of time in order to avoid service delays, the problem of designing policies that can achieve high throughput while at the same time maintain short term fairness, is also considered and investigated. To this end a new Credit-based Short-term Fairness Scheduling (CSFS) algorithm, which achieves to provide short-term fairness to the delay-sensitive users while still schedules opportunistically the non-delay-sensitive users to obtain high system throughput, is proposed and evaluated.

OPPORTUNISTIC TRANSMISSION SCHEDULING FOR NEXT GENERATION WIRELESS COMMUNICATION SYSTEMS WITH MULTIMEDIA SERVICES

by
Chengzhou Li

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

Department of Electrical and Computer Engineering

August 2005

APPROVAL PAGE

OPPORTUNISTIC TRANSMISSION SCHEDULING FOR NEXT GENERATION WIRELESS COMMUNICATION SYSTEMS WITH MULTIMEDIA SERVICES

Chengzhou Li

Dr. Symeon Papavassiliou, Dissertation Advisor — Date
Associate Professor of Electrical and Computer Engineering, NJIT

Dr. Nirwan Ansari, Committee Member — Date
Professor of Electrical and Computer Engineering , NJIT

Dr. Sirin Tekinay, Committee Member — Date
Associate Professor of Electrical and Computer Engineering, NJIT

Dr. Roberto Rojas-Cessa, Committee Member — Date
Assistant Professor of Electrical and Computer Engineering, NJIT

Dr. Rajarathnam Chandramouli, Committee Member — Date
Assistant Professor of Electrical and Computer Engineering, Stevens Institute of Technology

BIOGRAPHICAL SKETCH

Author: Chengzhou Li

Degree: Doctor of Philosophy

Date: August 2005

Undergraduate and Graduate Education:

- Doctor of Philosophy in Electrical Engineering,
  New Jersey Institute of Technology, Newark, NJ, 2005
- Master of Science in Electrical Engineering,
  Beijing University of Posts and Telecommunications, Beijing, China, 1998
- Bachelor of Science in Electrical Engineering,
  Beijing University of Posts and Telecommunications, Beijing, China, 1995

Major: Electrical Engineering

Presentations and Publications:

Symeon Papavassiliou and Chengzhou Li,
"Efficient and fair bandwidth scheduling and allocation in next generation wireless networks with multimedia services,"
to appear in the upcoming book *Resource Allocation in Next Generation Wireless Networks*, Nova Science Publishers, 2005.

Chengzhou Li and Symeon Papavassiliou,
"Joint throughput maximization and fair uplink transmission scheduling in CDMA systems ,"
submitted to *IEEE Transactions on Vehicular Technology*, July 2005.

Chengzhou Li and Symeon Papavassiliou,
"Fair channel-adaptive rate scheduling in wireless networks with multirate multimedia services ,"
IEEE Journal on Selected Areas in Communications, Vol. 21, Issue 10, pp. 1604-1614, Dec. 2003.

Chengzhou Li and Symeon Papavassiliou,
"On the fairness and throughput tradeoff of multi-user uplink scheduling in WCDMA systems,"
to appear *in Proc. IEEE Vehicular Technology Conference (VTC), 2005.*

Chengzhou Li and Symeon Papavassiliou,
"Joint throughput maximization and fair scheduling in uplink DS-CDMA systems,"
in Proc. 2004 IEEE Sarnoff Symposium on Advances in Wired and Wireless Communication, pp. 193–196, April 2004.

Chengzhou Li and Symeon Papavassiliou,
"Opportunistic scheduling with short term fairness in wireless communication systems,"
in Proc. Conference on Information Sciences and Systems (CISS2004), pp. 167-172, March 2004.

Chengzhou Li and Symeon Papavassiliou,
"Dynamic fair bandwidth allocation in multiservice CDMA networks,"
in Proceedings of the 23rd IEEE International Conference on Distributed Computing Systems-Workshops pp. 850-855, May 2003.

Chengzhou Li and Symeon Papavassiliou,
"The link signal strength agent (LSSA) protocol for TCP implementation in wireless mobile ad hoc networks,"
in Proc. IEEE Vehicular Technology Conference (VTC), pp. 2528-2532, Oct., 2001.

To my parents, Guangfeng Li and Chunying Xu, without whom this would not have been possible

ACKNOWLEDGMENT

I would like to express my sincerest gratitude to my advisor, Dr. Symeon Papavassiliou, who is a great mentor and also a friend. His constant support, encouragement and insightful guidance throughout all these years has helped me in my research and on the road to the Ph.D. I would also like to thank Dr. Nirwan Ansari, Dr. Sirin Tekinay, Dr. Roberto Rojas-cessa and Dr. Rajarathnam Chandramouli for participating in my committee and giving me valuable comments on my dissertation and research work.

I am also very grateful to Ms. Brenda Walker for helping me in my work as a Teaching Assistant. I would like to thank my friends and colleagues in NJIT, Jin Zhu, Jie Yang, Jian Ye, Shen Xu, Jun Jiang for their help with my research work, and also Surong Zeng, Mizhou Tan, Jun Li, Wugang Xu, Haibo Zhang and his wife Po Hu for many years of warm friendship.

Finally, I want to thank my wife Yu Zhang for her love, support and encouragement. Without her, I could not have accomplished this. Very special thanks to my parents for their dedicated and selfless love.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES
(Continued)

Figure **Page**

CHAPTER 1

INTRODUCTION

1.1 Overview

The continuous growth in traffic volume and emergence of new services has begun to change the structure of wireless networks. Future mobile communication systems will be characterized by high throughput, integration of services and flexibility. Mobile users want to use wireless access for multimedia applications demanding much higher bandwidth than is available today. 3G systems, although not fully deployed, are supposed to offer at least 144-384 Kb/sec for high-mobility users with wide area coverage and 2Mb/sec for low mobility users with local coverage. In 4G network architectures much higher bandwidth (e.g 100 Mbps) could be provided. In addition to applications calling for higher bit rates, users will also want to use multiple services simultaneously [1] [2] [3] [4]. Recently, there has been a trend in wireless networking for increasing demand of network Quality of Service (QoS). As real time applications become more prevalent, guaranteeing these parameters is becoming increasingly important. Large-scale deployment of multimedia services over wireless networks depends heavily on the offered QoS, network reliability and cost effectiveness of the services.

Wideband CDMA (WCDMA) has been proposed as a key air interface technique for third generation (3G) wireless systems, and will continue to be adopted as a strong candidate for 4G systems that will provide differentiated services to multimedia traffic. With the capability of dynamically varying user channel rates, WCDMA systems can provide more flexibility in bandwidth allocation. Although the out-coming new radio technology will bring more bandwidth at air interface, if not managed properly, the bandwidth resources will not meet the requirements of future users. Since in the current and future communication infrastructures, both topologies and traffic evolve and fluctuate on widely different time scales and the performances of the various services are strongly correlated

as the resources are shared among them, dynamic resource allocation methods should be employed. With the demand for high data rate and support of multiple QoS, the transmission scheduling plays a key role in the efficient resource allocation process in wireless systems. The transmission scheduling determines the time instance a mobile user may receive service as well as how much resources should be allocated to support the requested service, to make the resource distribution efficient. The fundamental problem of scheduling the users' transmissions and allocating the available resources in a realistic CDMA wireless system that supports multi-rate multimedia services, with efficiency and fairness, is investigated and analyzed in this dissertation.

1.2 Transmission Scheduling in Wireless Networks

The resource allocation and transmission scheduling problems have been intensively studied in the wireline networks [5] [6] [7] [8] [9] [10] [11]. Those scheduling schemes are usually classified into two categories: work-conserving and non-work-conserving policies. Work-conserving policies attempt to transmit data as long as there are backlogged flows, but they may distort the traffic patterns. The non-work-conserving policies are devised to overcome these problems by assigning each packet an eligibility time. The most well known fair scheduling policy in the work-conserving category is the General Processor Sharing (GPS) [12]. This ideal fluid fair scheduling model has the property of firm bound on the service gap, which guarantees the required QoS profile. Based on this model, many approximate fair scheduling policies for packet switched wireline networks have been proposed. Examples include Packet General Processor Sharing (PGPS) [12], also known as Weighted Fair Queueing (WFQ), and Worst-case Weighted Fair Queueing (WF^2Q) [7] , etc.

Traditional scheduling schemes that have been applied in conventional packet switched networks cannot be applied directly in wireless networks, due to the different and challenging characteristics that these systems present. The factors that complicate the scheduling problem

in wireless networks refer to the time-varying channel conditions, the different physical layer technologies, as well as the multiple types of interrelated resources that need to be allocated and the corresponding quality constraints that need to be satisfied. Therefore the actual system capacity, as defined conventionally in wireline networks, is not fixed and known in advance, since it is a function of several parameters such as the number of users, the channel conditions, the transmission powers, etc. Furthermore, due to the varying channel conditions the utilized resources are not proportional to the achievable data rate and received throughput, and hence the concept of fairness presents a different meaning in wireless networks.

Time-varying channel conditions will cause the users to experience time-varying service quality. Suppose that the same resources, usually the service time in conventional packet switched networks, is assigned to two users. In wireline networks they are both expected to receive the same amount of service (transmitted or received data). However in wireless networks, their obtained service could be very different due to the various channel capacities under different channel conditions. Obviously, assigning more resources to compensate for the bad channel conditions will result in low resource utilization. Therefore, for fixed amount of resources, the different resource allocation schemes and transmission scheduling policies may result in different system throughput and QoS performance.

Besides the variable channel conditions, the physical layer techniques also impose their influence on the wireless transmission scheduling policies. For instance, the wireless networks with CDMA access technology have different available system resources than the wireline networks, and correspondingly increase the complexity of scheduling. The achievable throughput in a CDMA system depends not only on the service access time, but also on the transmit powers and the corresponding users' interference. Furthermore, multiple users can be scheduled in the same time slot, which is a major difference from the wireline and some TDMA-like scheduling schemes. Therefore, the simple service time fair scheduling fails to provide rational fairness in this case. The fair scheduling needs to

jointly consider multiple factors such as: access time, transmit power and the number of users to be scheduled at the same time. In addition, the conventional concept of capacity used in the wireline network, e.g. total bandwidth of the physical media, is not directly applicable in the CDMA systems. In CDMA systems the actual system capacity is not fixed and known in advance, since it is a function of several parameters such as the number of users, the channel conditions, the transmission powers, etc., which makes the general wireline scheduling policies inapplicable, and complicates the scheduling problem.

Due to the above, the conventional fair access time scheduling, or temporal fairness [13], is not sufficient to define the criteria of fair scheduling in the wireless networks. The temporal fairness attempts to guarantee the fair access time, which however cannot guarantee by itself the required QoS in wireless systems, due to the time-varying service quality. The throughput fairness on the other hand, aims to provide fair achieved throughput and solve the unfairness of the received service in the access time allocation which is introduced by the variable channel conditions. However, to reach this objective flows with weak channel conditions must obtain much more access time to compensate for the low throughput, which could result in low system throughput. Therefore, there exists a tradeoff between the fairness and achievable throughput, and to identify the optimal policy and point of operation that achieves maximum throughput, while maintaining fairness is of high practical and research importance.

Although, as mentioned above, the channel variation in the wireless environment makes the scheduling more complicated, on the other hand it allows the scheduler to exploit this characteristic and achieve possibly high system throughput. The improvement in throughput can be obtained by utilizing the multi-user diversity effect ([14], [15]) in wireless communications. Specifically, for a system with many users that have independent varying channels, with high probability there is a user with channel much stronger than its average SNIR requirement. Therefore the system throughput may be maximized by choosing the user with "relatively best" channel for transmission at a given time slot.

1.3 Dissertation Objective and Outline

With the capability of dynamically varying user channel rates, WCDMA systems can provide more flexibility in bandwidth allocation and transmission scheduling. Although the idea of dynamic allocation of bandwidth by varying the channel rate in WCDMA system has been recently studied for supporting multiple QoS [16] [17], the issue of fairness and adaptation to the changing conditions and vulnerabilities of the wireless channels has not been well addressed yet. The fundamental problem of allocating the available resources in the uplink of a wireless system that supports multi-rate multimedia services, with efficiency and fairness is investigated in this dissertation. Specifically, several wireless fair scheduling algorithms and policies that achieve high throughput and fairness are proposed and analyzed. The basic principle of these policies is the opportunistic transmission scheduling, taking into consideration various aspects of wireless networks in order to achieve high system throughput, while still maintaining fairness and satisfying the users' QoS requirements. The remaining of this dissertation is organized as follows.

In Chapter 2 two new scheduling algorithms are proposed, namely the Channel Adaptive Rate Scheduling (CARS) and the Fair Channel Adaptive Rate Scheduling (FCARS). First, through an iterative process CARS exploits the channel variations to reach high throughput, without however achieving fairness. To overcome the potential unfair service allocation, FCARS implements a compensation algorithm, in which the lagging users can receive compensation service when the corresponding channel conditions improve, in order to achieve asymptotic throughput fairness. In Chapter 3, the problem of opportunistic fair scheduling in the uplink transmission of CDMA systems, with the objective of obtaining the optimal throughput in the uplink scheduling under the long-term fairness constraint, is analyzed, and a throughput optimal fair scheduling policy is proposed. We first formulate the multiple constraint optimization problem and then we prove that it can be converted into a linear knapsack problem, where the final solution becomes a global search within a range. Furthermore, a stochastic approximation method is presented to estimate the

fairness control parameters. Since in wireless systems with different quality of service requirements some types of traffic may demand certain amount of service within specific short span of time in order to avoid service delays, in Chapter 4 a new Credit-based Short-term Fairness Scheduling (CSFS) algorithm, which achieves to provide short-term fairness to the delay-sensitive users while still schedules opportunistically the non-delay-sensitive users to obtain high system throughput, is proposed and evaluated. Finally, Chapter 5 concludes the dissertation by summarizing the main contributions and conclusions of this work, and discussing the directions for future work.

CHAPTER 2

SIMULTANEOUS OPPORTUNISTIC SCHEDULING IN CDMA SYSTEM

In this chapter, the problem of opportunistic scheduling with concurrent users in a CDMA system is studied and analyzed. We adopt the use of dynamically assigned data rate that matches the channel capacity in order to improve the system throughput. Users with better channel condition will obtain more bandwidth, while those with worse channel condition will get less bandwidth, which however may result to fairness violations. In order to properly compensate these users that lose part of their service we record the excess service each user receives, and at later time when the lagging user's channel condition improves it will receive more service to compensate for the lost part. The rest of this chapter is organized as follows. First we present some issues associated with the rate scheduling in CDMA networks. We provide some useful observations and definitions and prove some theorems that are used throughout this chapter. Then we describe the proposed channel-adaptive rate scheduling method and algorithm to achieve simultaneously the objectives of high overall throughput and fairness. Finally the performance evaluation of our proposed algorithm is presented along with some numerical results and discussions. The performance evaluation is achieved in terms of achievable fairness and throughput.

2.1 Related Work and Motivation

Due to the inert characteristic of wireless links, efficient and fair scheduling in the wireless system faces new challenges, and it has recently become an active research topic. For instance [18] [19] [20] studied the problems of wireless fair scheduling especially in TDMA systems. Most of the current work on scheduling in CDMA systems has mainly focused on how to transmit the packets effectively, while little work has been done on the aspects of fairness.

Various schemes have been proposed in the literature ([13] [21] [14] [15] [22] [23] [24]) to exploit the randomly time-varying channel states and obtain performance

improvements in wireless systems. In general in most of these schemes, the improvement is obtained by utilizing the multi-user diversity effect ([14], [15]) in wireless communications. This means that for a system with many users that have independent varying channels, with high probability there is a user with channel much stronger than its average SNIR requirement. Therefore the system throughput may be maximized by choosing the user with "relatively best" channel for transmission at a given slot. In [13] [21] a framework for opportunistic scheduling that maximizes the system performance by exploiting the time-varying channel conditions of wireless networks is presented. Three categories of scheduling problems - the temporal fairness, utilitarian fairness and minimum-performance guarantee scheduling - are studied and optimal solutions are given. Furthermore, for the EDGE/GPRS system the possibility of trading off efficiency for fairness when exploiting temporary fluctuations in channel conditions is studied in [24].

Despite the good performance demonstrated by various scheduling schemes that take into account the varying channel conditions, several problems still exist. In most of the schemes presented above only the downlink scheduling is considered. Although the downlink transmission rate assignment is important for several applications, the efficient uplink transmission scheduling plays an important role as well, especially with the prevailing of multimedia communications and applications. The uplink transmission scheduling problem is more complicated and requires further consideration of additional elements to make the corresponding scheduling policies feasible, due to the interference issues and power limitations. In addition, most of the scheduling schemes only studied the one server case, that is only a single user is served in a time-slot. Although it has been shown that scheduling users one-by-one can result in higher system throughput for high data rate traffic in the CDMA downlink [23], serving multiple users simultaneously can provide more flexible service [25] and are necessary to achieve high throughput in the CDMA uplink [26]. Furthermore, some real-time traffic may not tolerate the corresponding delay if they are

served one-by-one, especially when the scheduling cycle is relatively long compared with their delay requirements.

In this chapter we exploit and demonstrate the benefits that can be obtained in the realization of the fair scheduling in the uplink of CDMA systems, by utilizing their inherent advantage of serving several packets simultaneously, and thus making easier to follow the general principle of fluid Generalized Processor Sharing (GPS) Algorithm and the approximations of the Fluid Fair Queue (FFQ) model, which require that flows are served simultaneously each one with its deserved share of service [12] [7]. In [27] the authors devised a MAC protocol – WISPER (wireless multimedia access control protocol with BER scheduling) in order to accommodate more than one packet in a time slot. Furthermore, several attempts and methods on the assignment of system resources in order to achieve fair service allocation and prevent misbehaving users from seizing large portions of the system resources, have also been reported in the literature ([25] [28] [29] [30]). The basic idea of [25] [28] is to consider the system capacity as the only resource and schedule the users transmission data rate by their weight. However in practice the system capacity is not fixed and changes with the channel quality and number of backlogged users. To overcome some of these problems [30] [29] considers combinations of the power index and the chip rate as the system resources.

Although the scheduling methods of [29] [30] may provide a feasible data rate assignment and attempt to take into account the fairness when distributing the system resources, there are still two main problems: First the obtained resources may not be proportional to the data rate. When the allocated resources are converted into the transmission rate in [29] [30], the rate is not proportional to its weight any more. Moreover, power index assignment only shows the relationship of transmission power of all users. Second most of the proposed CDMA scheduling methods either assume perfect channel condition without considering the impact of the channel condition variations or simply model the channel by the two-state channel model [18] [30]. This model is not suitable for CDMA systems,

since the interference among users, the channel fading and the transmission rate can significantly affect the bit error probabilities. In a realistic CDMA system however we can take advantage of the variable rate scheduling capabilities by attempting to transmit at lower rate which could compensate for the variation of channel condition and meet the required user's requirements. In this paper we consider and address these two issues by presenting a technique that provides fair, channel-adaptive rate scheduling in CDMA wireless networks.

2.2 Some Useful Observations, Definitions and Theorems

Generally we use the signal to noise and interference ratio (SNIR) to measure the channel condition. Let W and r_i denote the bandwidth (also known as chip rate in CDMA system) and the transmission rate of user i respectively. Let also denote by h_i the corresponding channel gain, and by p_i user i's transmission power at a given slot. Then the received SNIR for user i is given by:

$$SNIR = \frac{h_i p_i}{\alpha \sum_{j=1, j\neq i}^{N} h_j p_j + W\eta_0} \quad i = 1, 2, \ldots, N \tag{2.1}$$

where η_0 is the one-sided power spectral density of Additive White Guassian Noise (AWGN) and α determines the proportion of the interference from other users. Without loss of generality, we assume $\alpha = 1$ for the CDMA system . Please note that the QoS requirement usually is given by the $\gamma = E_b/N_0$ (the ratio of bit energy to noise power spectrum density), which is related to the SNIR by

$$\frac{E_b}{N_0} = SNIR \cdot \frac{W}{r_i}$$

In the following analysis, for simplicity in the presentation, we use the requirement γ to present the corresponding SNIR requirement.

Let us consider the uplink of a CDMA system containing $B(t)$ backlogged users at time t. Each user is pre-assigned a weight ϕ_i according to its QoS requirements when admitted. When scheduling the transmissions of the various users during interval $[t, t +$

τ], where τ is the scheduling cycle, we consider only the backlogged users $B(t)$ at the beginning of the scheduling interval. Users that may become backlogged within the interval $[t, t+\tau]$ will be included in the next scheduling cycle. Let G_i be the spreading gain of mobile i and γ_i represent the minimal SNIR required to satisfy its QoS requirements. We assume that the chip rate W for all mobiles is fixed, and hence the spreading gain is determined by the bit rate r_i of mobile i, that is: $G_i = \frac{W}{r_i}$. The bit energy-to-equivalent noise spectral density ratio of general DS-CDMA system is expressed as

$$\gamma_i' = \frac{h_i p_i G_i}{\sum_{j=1, j\neq i}^{N} h_j p_j + W\eta_0} \quad i = 1, 2, \ldots, N$$

Given a transmission rate vector $\mathbf{r} = [r_1, r_2, \ldots, r_{B(t)}]$ there exists a power assignment $\mathbf{p} = [p_1, p_2, \ldots, p_{B(t)}]$ that meets the required QoS vector $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \ldots, \gamma_{B(t)}]$, i.e. $\gamma_i' \geq \gamma_i$ for all $i = 1, \ldots, B(t)$, if the following condition ([31]) is satisfied:

$$\sum_{i=1}^{B(t)} g_i \leq 1 \tag{2.2}$$

$$g_i = \frac{\gamma_i}{\gamma_i + G_i} \tag{2.3}$$

where g_i represents the power index of user i. In the rest of this paper we assume that the maximum transmission power is $p_{\max}$ and same for all users.

Therefore the previous condition (2.2), under which there exists a feasible power assignment $\mathbf{p} = [p_1, p_2, \ldots, p_{B(t)}]$ that meets the QoS requirements under the constraint that for each user i, $p_i < p_{\max}$, becomes([31]):

$$\sum_{i=1}^{B(t)} g_i \leq 1 - \frac{\eta_0 W}{\min\limits_{1\leq i\leq B(t)} \left\{ p_{\max} h_i \left(\frac{G_i}{\gamma_i} + 1 \right) \right\}} \tag{2.4}$$

For a system with $\sum_{i=1}^{B(t)} g_i = \Psi$, relation (2.4) requires that every user i in the system must have minimal level of channel gain: $h_i \geq \frac{\eta_0 W}{p_{\max}\left(\frac{G_i}{\gamma_i}+1\right)(1-\Psi)}$ to satisfy its required γ_i. However as we discuss later in subsection 2.2.2, in a realistic system this condition may not be always satisfied.

2.2.1 Fair Rate Scheduling in the Ideal System

In order to achieve fairness in a system with available uplink capacity R, the assigned data rate must satisfy the following relation for each mobile i, $i \in B(t)$: $r_i = R \cdot \phi_i / \sum_{j=1}^{B(t)} \phi_j$. The power index of mobile i is given by:

$$g_i = \frac{\gamma_i}{\gamma_i + G_i} = \frac{\gamma_i}{\gamma_i + \frac{W}{\phi_i' R}} \tag{2.5}$$

$$\text{where } \phi_i' = \frac{\phi_i}{\sum_{j=1}^{B(t)} \phi_j}$$

Let Ψ be the maximum available power index. If Ψ is shared by $B(t)$ backlogged users, the maximum power index that a single user can obtain in order to satisfy relation (2.4) is $g_{\max} = \frac{(1-\Psi)p_{\max}}{W\eta_0}$. Next we first give a definition of the reference system and based on it we prove the following proposition 1.

Definition 1 *The reference (ideal) system is defined as the one in which every user's channel gain is* $h_i = 1$.

Proposition 1 *For N backlogged users in the ideal system there exists a power vector $[p_1, p_2, \ldots, p_N]$, $p_i < p_{\max}$, and a rate assignment scheme that the corresponding rate vector $[\gamma_1, \gamma_2, \ldots, \gamma_N]$ satisfies the GPS service rule $\frac{r_k}{r_m} = \frac{\phi_k}{\phi_m}$, $1 \le k, m \le N$.*

Proof. Suppose the maximum available power index is Ψ, and let user k be the one with the maximum value of the product of weight and SNIR, i.e. $\phi_k \gamma_k = \max_{1 \le i \le N} \{\phi_i \gamma_i\}$, among the N users. Then user k will require the largest power index in the system to satisfy the GPS service rule. Let us assume that user k is assigned the maximum power index $g_{\max}$. Since under GPS rule the following should be satisfied: $\frac{r_k}{r_m} = \frac{\phi_k}{\phi_m}$, for every other user i we can find the corresponding power index $\widetilde{g}_i$ by the following relation $\frac{\widetilde{g}_k}{1-\widetilde{g}_k} \frac{1-\widetilde{g}_m}{\widetilde{g}_m} \frac{\gamma_m}{\gamma_k} = \frac{\phi_k}{\phi_m}$, and therefore we can obtain the corresponding power index vector: $[\widetilde{g}_1, \widetilde{g}_2, ..., \widetilde{g}_N]$. There are two possible cases for this power index vector: $\sum_{i=1}^{N} \widetilde{g}_i < \Psi$ or $\sum_{i=1}^{N} \widetilde{g}_i \ge \Psi$. In the first case since no user gets power index larger than $g_{\max}$ and $\sum_{i=1}^{N} \widetilde{g}_i < \Psi$,

condition (2.4) is clearly satisfied and thereby a feasible power vector $[p_1, p_2, ..., p_N]$ exists, and the corresponding calculated power index vector $[\widetilde{g}_1, \widetilde{g}_2, ..., \widetilde{g}_N]$ satisfies $\frac{r_k}{r_m} = \frac{\phi_k}{\phi_m}$. However in the second case since $\sum_{i=1}^{N} \widetilde{g}_i \geq \Psi$, the power index vector $[\widetilde{g}_1, \widetilde{g}_2, ..., \widetilde{g}_N]$ is not feasible. The feasible power index vector can be found by solving the system achievable rate R based on the following relations: $\sum_{i=1}^{N} g_i = \Psi$, where g_i is defined by (2.5). Let $f(R) = \sum_{i=1}^{N} g_i = \sum_{i=1}^{N} \frac{\gamma_i}{\gamma_i + (W/\phi_i R)}$ which is a continuous and increasing function with $\lim_{R \to 0} f(R) = 0$ and $\lim_{R \to \infty} f(R) = N$, there must exist a solution of R for function $f(R) = \Psi$, $0 < \Psi < 1$. Thus a feasible power vector exists for this case as well. ■

Since the chip rate is assumed to be fixed, we can use the spreading gain to represent the data rate throughout our analysis. In the following we denote the system throughput, the spreading gain vector and the corresponding power index vector of the ideal system by R^*, $[G_1^*, G_2^*, ..., G_N^*]$ and $[g_1^*, g_2^*, ..., g_N^*]$ respectively.

2.2.2 Non-Ideal System

In a realistic system due to path loss, channel fading, and shadowing, it is possible that some of the users' channel gains h_i may not be large enough. Under such environment, if the scheduler does not utilize any information about the channel quality, and the user rates as scheduled based on the ideal system are still applied, the throughput inevitably decreases due to the high error packet loss rate. For instance, let r_i^* and h_i^* be the rate and the channel gain of backlogged user i in the ideal system respectively. If we denote by p_i the corresponding transmission power, then the minimal received power necessary to meet the required SNIR γ_i for every user i is given by: $h_i^* p_i = \frac{\eta_0 W}{\left(\frac{W}{\gamma_i r_i^*} + 1\right)(1-\Psi)}$, in order to satisfy inequality (2.4). Once the channel gain of user i becomes very small, the received power may not reach the minimal required power even if the maximum transmission power is used, e.g. $h_i p_{\max} < h_i^* p_i$, and consequently the actual SNIR drops below γ_i. Thus the power index vector of the ideal system needs to be adjusted in order to obtain a new and feasible power index vector.

Suppose that in the non-ideal system the N backlogged users have the following channel gain vector $[h_1, h_2, ..., h_N]$. Without loss of generality, we assume that the last K users experience bad channel condition. This means that if the spreading gain vector (power index vector) of the ideal system is still applied, i.e. $\sum_{i=1}^{N} g_i^* = \Psi$, then for these K users relation (2.4) is violated. That is: $1 - \frac{\eta_0 W}{p_{\max} h_j (1+G_j^*/\gamma_j)} < \Psi$, $N - K + 1 \leq j \leq N$. We can increase the spreading gains of the K users till each one of them meets the condition: $1 - \frac{\eta_0 W}{p_{\max} h_j (1+G_j^*/\gamma_j)} = \Psi$. At that point we obtain a new spreading gain vector $[G_1^*, G_2^*, ..., G'_{N-K+1}, ..., G'_N]$. The loss of spreading gain translates to loss of total power index as well, which can be calculated as: $g^{lost} = \sum_{i=N-K+1}^{N} (\frac{\gamma_i}{\gamma_i + G_i^*} - \frac{\gamma_i}{\gamma_i + G'_i})$. However we may be able to increase the system throughput by redistributing the lost part to users with good channel condition. This technique is investigated and studied in detail in the next section. In the following we first introduce and define a new parameter, namely the Power Index Capacity (PIC).

Definition 2 *In a CDMA system with N backlogged users, given the total power index constraint $\sum_{i=1}^{N} g_i \leq \Psi$ and a feasible spreading gain vector $[G_1, G_2, ..., G_N]$, the maximum allowable power index of a user which does not violate the following condition: $\Psi \leq 1 - \frac{\eta_0 W}{p_{\max} h_i (1+G_i^*/\gamma_i)}$, is defined as the power index capacity (PIC) π_i of this user.*

Please note that the PIC of a user can be found by increasing its power index, while at the same time other users' power indices will be reduced or remain unchanged to keep the total power index $\sum_{i=1}^{N} g_i \leq \Psi$. Hence, when a user achieves its PIC, the newly adjusted power index assignment is feasible.

Theorem 1 *Given a feasible spreading gain vector $[G_1, G_2, ..., G_N]$ and the total power index constraint $\sum_{i=1}^{N} g_i \leq \Psi$, if we adjust the power index of users within their power index capacity $\boldsymbol{\pi} = [\pi_1, \pi_2, ..., \pi_N]$, the new assignment $[g'_1, g'_2, ..., g'_N]$ and corresponding spreading gain vector $[G'_1, G'_2, ..., G'_N]$ are still feasible.*

Proof. The sum of new power index assignment satisfies: $\sum_{i=1}^{N} g_i' \leq \Psi$, since they are adjusted within their power index capacity $\boldsymbol{\pi} = [\pi_1, \pi_2, ..., \pi_N]$. From the definition of power index, no user will violate relation (2.4) with the new power index vector $[g_1', g_2', ..., g_N']$ if $g_i' < \pi_i$. Therefore, there still exists a power vector $[p_1, p_2, \bullet\bullet\bullet, p_N]$, $p_i < p_{\max}$, that makes the assignment feasible. ■

Therefore based on the above arguments, the power indices can be adjusted within each user's PIC in order to improve the system throughput under the changing channel conditions, and as long as the constraint on the total power index is maintained, the corresponding power indices and power assignments are still feasible. Through the Channel Adaptive Rate Scheduling scheme that we introduce in the next section, the system throughput may increase if users with bad channel condition give up transmission opportunities while the corresponding lost power index g^{lost} can be absorbed by users with better channel conditions and high power index capacity.

2.3 Channel-Adaptive Rate Scheduling

In this section we first introduce and describe an algorithm – Channel Adaptive Rate Scheduling (CARS), which allocates the available system resources according to the corresponding channel conditions. In fact CARS tries to improve the system resource utilization under the constraint of each user's power index capacity π_i, without however considering the fairness issue. Then we provide a compensation algorithm (FCARS) and adjust the allocations achieved by CARS in order to achieve the objective of fairness. CARS exploits the time-varying channels to achieve high throughput, while guarantees that the power index capacity constraints are met and the scheduled transmission rates are feasible. In addition, the redistribution of part of the resources among some users that takes place in FCARS in order to achieve the fairness, still depends on the rules of CARS to keep the final assignment feasible, and therefore CARS is an integral component of FCARS.

2.3.1 Channel-Adaptive Rate Scheduling (CARS) Algorithm

Suppose that the total available power index at the beginning of each scheduling cycle is Ψ. Given the number of users N and their weights, we compute the reference power index by equation (2.5), which represents the ideal fair scheduling model. However, due to the variation of the channel conditions, this assignment scheme may result in high packet error rate for some users, and consequently waste the system resources. CARS uses (2.5) as a reference, and adjusts the final values by each user's power index capacity. In the following we describe the operation of the proposed CARS algorithm, while in Appendix A we provide the pseudo-code of the corresponding process.

Initially, no power index is distributed and therefore the remaining power index C_R is equal to Ψ. We use $[\Delta g_1, \Delta g_2, \ldots, \Delta g_N]^k$ and $[\Delta r_1, \Delta r_2, \ldots, \Delta r_N]^k$ to represent the power index increment vector and the corresponding rate increment vector in the k-th round of computation of the algorithm. Therefore, if the assigned power index after $(k-1)$-th round is $[g_1, g_2, \ldots, g_N]^{k-1}$, where $[g_i]^{k-1}$ denotes the power index assigned to user i after $(k-1)$-th round, the planned/estimated power index at k-th round is $[g'_1, g'_2, \ldots, g'_N]^k = [g_1, g_2, \ldots, g_N]^{k-1} + [\Delta g_1, \Delta g_2, \ldots, \Delta g_N]^k$. The actually accepted power index in the k-th round is $[g_i]^k = \min\left\{[g'_i]^k, \pi_i\right\}$, which depends on users' current power index capacities. The remaining power index after k-th round computation is $[C_R]^k = \Psi - \sum_{i=1}^{N}[g_i]^k$. Let us denote by $[\mathbf{H}]^k$ the set that contains the users that their assigned power index is $[g_i]^k = \pi_i$ after k-th round. The remaining power index is distributed only among users in set $\mathbf{B(t)} - [\mathbf{H}]^k$.

In the first round since the assigned power index is zero for all users, we compute the reference power index increment vector $[\Delta g_1, \Delta g_2, \ldots, \Delta g_N]^1$ according to the ideal system. Therefore the corresponding rate increment vector is $[\Delta r_1, \Delta r_2, \ldots, \Delta r_N]^1$, which satisfies the relation: $\frac{[\Delta r_i]^1}{[\Delta r_j]^1} = \frac{\phi_i}{\phi_j}$. Users whose power index capacities are not large enough, i.e. $\pi_i < [\Delta g_i]^1$ are also included in set $[\mathbf{H}]^1$. Then the unaccepted portion of power index in set $[\mathbf{H}]^1$ is collected in $[C_R]^1$ and re-distributed to users in set $\mathbf{B(t)} - [\mathbf{H}]^1$, during the

subsequent iterations. Our objective in the following rounds of assignment is to make the increment of users' rates proportional to their weights. Therefore we compute the reference power index vector and rate vector for users in set $\mathbf{B(t)} - [\mathbf{H}]^k$. However the increment of rate does not have linear relationship with the increment of power index. In order to provide a simple procedure to solve this problem without increasing the overall operation's complexity we adjust the weight of users such that the increment power index vector $[\Delta g_1, \Delta g_2, \ldots, \Delta g_N]^k, k \geq 2$, is obtained by:

$$[\Delta g_i]^k = \frac{[\phi_i']^k}{\sum_{j \in B(t)-H} [\phi_j']^k} [C_R]^{k-1} \quad \forall i \in \mathbf{B(t)} - [\mathbf{H}]^{k-1}$$

where $[\phi_i']^k = \phi_i \left(1 - [g_i]^{k-1}\right)^2$ and $[g_i]^{k-1}$ is the power index that has already been assigned to user i in the $(k-1)$-th round computation. The data rate increase that corresponds to the power index increase $[\Delta g_i]^k$, is

$$\begin{aligned} [\Delta r_i]^k &= \frac{[\Delta g_i]^k}{\left(1 - [g_i]^{k-1} - [\Delta g_i]^k\right)\left(1 - [g_i]^{k-1}\right)} \\ &= \frac{[\phi_i']^k \cdot M}{\left(1 - [g_i]^{k-1} - [\Delta g_i]^k\right)\left(1 - [g_i]^{k-1}\right)} \\ &= \frac{\phi_i \cdot M}{1 - \frac{[\Delta g_i]^k}{1-[g_i]^{k-1}}} = \frac{\phi_i \cdot M}{1 - \phi_i \left(1 - [g_i]^{k-1}\right) M} \end{aligned}$$

$$\text{where } M = \frac{[C_R]^k}{\sum_{j \in B(t)-H} [\phi_j']^k}$$

Thus the ratio of increased rates between users i and j is

$$\frac{[\Delta r_i]^k}{[\Delta r_j]^k} = \frac{\phi_i}{\phi_j} \cdot \frac{1 - \phi_j \left(1 - [g_j]^{k-1}\right) M}{1 - \phi_i \left(1 - [g_i]^{k-1}\right) M}$$

Therefore in most practical cases with large number of users the corresponding ratio of increased rates is close to the ratio of the original weights.

Please note that some small variations in the distribution of $[C_R]^k$ do not affect significantly the system performance. In fact the objective of the service distribution in

this subsection is that users with larger power index capacities get more service in order to improve throughput. The unfairness will be compensated in the following scheduling cycles as described in the next subsection.

2.3.2 Fair Channel-Adaptive Rate Scheduling (FCARS) Algorithm

CARS algorithm improves the throughput by giving better transmission chances to the users with better channel conditions. However, the fairness intended is not fulfilled because the allocation of resources depends on the randomly changing channel conditions. Here we propose a compensation algorithm—Fair Channel Adaptive Rate Scheduling (FCARS), which accomplishes the long-term fairness and at the same time maintains high throughput.

FCARS algorithm builds its operation on CARS scheme. The first step of FCARS is the rate allocation according to CARS (we refer to this as the initial scheduling). After the initial scheduling, the power index of users will be adjusted according to the amount of excess service they received. Let S_i^e denote the accumulated excess service that user i received with reference to the service that would be assigned by the ideal GPS algorithm in the reference system from the beginning of its session, and s_i denote the excess service that user i receives in the scheduling cycle under consideration. To distinguish the users that receive different amount of service with reference to the service provided by the ideal GPS algorithm, we classify the users into three types [18]: the lagging user, the leading user and the satisfied user. For a lagging user $S_i^e < 0$, since it receives less service than it deserves in an ideal system, while the leading user receives more service and the satisfied user receives the same amount of service.

Due to the limitation of the channel condition, not all lagging users are qualified to receive compensation in a scheduling cycle. Similarly in order to limit the impact on the leading users, not all leading users have to give up part of their service. We further divide the users into three groups: lagging group, leading group and normal group. Only users from the leading group may give up part of their service to be used for compensation to the

users of the lagging group. The classification of users into different groups depends on the results of the CARS algorithm and is described in the following definition. For a user-i let π_i be the power index capacity, g_i^* the power index of the reference (ideal) system, q_i and $q_i^{\max}$ the queue size and the maximum queue size respectively.

Definition 3 *For all backlogged users, considering the service scheduled by the CARS, the lagging group contains all users that received less service than the service that they would receive under the ideal GPS algorithm in the reference system, i.e.* $S_i^e < 0$, *and have capacity to accept more power index, i.e.* $\pi_i > g_i$. *The leading group contains those leading users whose excess service* $S_i^e > 0$, *and are assigned power index that is greater than* $\left[g_i^* \cdot f_q(\frac{q_i}{q_i^{\max}})\right]$ *in CARS in the current scheduling cycle. The rest of the users are included in the normal group. Function* $f_q(x)$ *is defined as* $f_q(1-x)\,\xi$, *where* ξ *is a constant, which can be adjusted to account for the different requirements of different classes of service (e.g. real-time vs. non-real-time services). In the following without loss of generality we assume* $\xi = 1$.

In other words users in the lagging group are those that receive less service while they have more power index capacity. Users in the leading group are those that can give up part of their service. Users that are neither in the leading group nor in the lagging group are included in the normal group. The leading group threshold $\left[g_i^* \cdot f_q(\frac{q_i}{q_i^{\max}})\right]$ ensures that the leading users only experience graceful performance degradation. Although [18] uses some measurement to provide limited degradation in service for leading users, it still introduces large delay for those users with big queue lengths. The FCARS algorithm uses a system parameter ξ and the leading user's queue length to control the amount of service that may be given up by a user. According to our proposed algorithm the leading user may give up no more than the service defined by the following relation in terms of power index:

$$g_i^g = \min\left\{f_q\left(\frac{q_i}{q_{\max}^i}\right) g_i,\; g_i - f_g^{-1}\left(f_g\left(g_i\right) - \frac{s_i^e}{\tau_m}\right)\right\} \tag{2.6}$$

where τ_m is the maximum scheduling cycle. Function $f_g(x)$ converts the power index to the corresponding data rate by the definition of power index. Since we have assumed that the chip rate is fixed for all users, the data rate can be determined by the assigned power index. Relation (2.6) limits the maximum power index the leading users will give up, which ensures that the leading users will not release more service than their excess service. $f_g(g_i)$ converts the reference power index into the reference rate, which is the rate that user i deserves in the ideal system.

Therefore all services in terms of power index will be collected and distributed among users in the lagging group according to their power index capacities and queue lengths. Once the channel condition of a user improves the user may catch up in the next one or multiple scheduling cycles. This depends on the service it lost before, and on its current queue length. The lagging users with larger queue length obtain more compensation services. To realize this process a compensation weight ϕ_i^c is defined.

Definition 4 *The compensation weight ϕ_i^c of lagging user i in current scheduling cycle τ_k is defined as: $\phi_i^c(\tau_k) = \frac{q_i(\tau_k)}{q_{\max}^i}\phi_i$, where $q_i(\tau_k)$ is the queue length of user-i at the beginning of cycle τ_k.*

The service that is given up by the leading users is distributed among the lagging users in proportion to their compensation weights following a similar procedure as the one used in CARS algorithm. In order to realize the compensation procedure the excess service of every user after each scheduling cycle is collected. The general idea is to compare the actual assigned service to that of the ideal system. Assuming that the achievable throughput is $R = \sum_{i \in B(t)} r_i$, the reference rate of the ideal system is: $r_i^* = \phi_i' \cdot R$. Let g^g denote the service given up by the leading group in terms of power index. It is possible that the lagging group may not absorb g^g completely. In this case the remaining power index is redistributed among the users of leading group and normal group in proportion to their weights in order to improve the system utilization. Appendix A gives the pseudo-code of the complete operation of the FCARS algorithm.

2.3.3 Algorithm's Complexity

In this section we discuss the complexity of our proposed algorithm. Since FCARS is built on CARS we first give the complexity of CARS. It can be easily seen that CARS algorithm is of iterative nature. At each iteration we perform the computation of the power index of the reference system, the comparison of users' estimated power index to their power index capacities, and the assignment of proper power index, which can be implemented in $O(n)$, where n is the number of unscheduled users in this round of operation. Considering a system with N backlogged users, we may need up to N rounds of the repeated operation to distribute the system resources, which makes the worst case algorithm complexity $O(N^2)$. In addition to CARS algorithm, FCARS adds the operations of excess service computation, classification of the three groups, and the computation of power indices that the leading users have to give up. Each of these operations can be implemented in $O(N)$, and therefore the overall worst case complexity of FCARS is still $O(N^2)$.

In practice in the general case if the number of backlogged users is large enough, there will be a considerable number of users that are in good channel state at a given time, and thus the iteration assignment of CARS will be completed within the first few rounds with high probability. In this case the complexity of CARS as well as of FCARS is reduced to $O(N)$.

It should be noted here that the proposed transmission scheduling algorithm is performed at the beginning of every scheduling cycle. Therefore the length of scheduling cycle τ is a parameter that may affect the system performance. The smaller the scheduling cycle, the more effective the scheduling algorithm in terms of adjusting the rate according to the changes of the channel conditions and keeping track of the sudden changes in the conditions of the users' channels. However, a smaller τ demands more computational power since it requires more frequent scheduling, and therefore if the changes are smooth a larger value for τ will be more efficient from operational point of view. As a result a dynamic scheme that adjusts the value of τ with the changes of the environment and the number of active

users is desired. In the next section we obtain an indication of the impact of the scheduling cycle τ on the achieved performance through a numerical study. However the optimal choice of the scheduling interval τ is part of our current research.

2.3.4 Design Issues and Modifications of FCARS

In this chapter the emphasis is mainly placed on the development of bandwidth/rate scheduling strategies and algorithms, in order to improve the system throughput while provide fairness among the various users of the network, given their corresponding weights. Each user is pre-assigned with a weight according to its QoS requirements. However the various flows in the network may present different QoS requirements, including differences in delay sensitivity. Although this paper does not aim to explicitly address the problem of user weight and/or class assignment, in this section we discuss how specific parameters and requirements, can be taken into account within the framework presented here, either explicitly or implicitly.

One implicit way to consider the delay sensitivity of the different flows within the formulation described in previous sections, is to utilize the concept of effective bandwidth [32] [33] which represents the amount of bandwidth required by each flow to statistically achieve the corresponding QoS. Effective bandwidth is a scalar that summarizes resource usage and that depends on the statistical properties and QoS requirements of a source. Based on this concept the required rate and thus the corresponding weight of a flow can be determined to achieve the desired QoS. Considering this and taking into account that our algorithm provides higher compensation to users with higher weights, implicitly we can consider the delay sensitivity and differences among different flows.

In addition in order to explicitly distinguish between delay-sensitive and non-delay-sensitive traffic, the basic FCARS algorithm can be modified to provide the compensation based on the user priorities, which can be defined according to the delay sensitivity of the corresponding flows. The Modified FCARS (M-FCARS) algorithm is similar to FCARS

except that the lagging users (as defined in FCARS) are divided into classes according to their delay-sensitivity, and the resources, i.e. power indices, given up by the leading users, are distributed among the lagging users based on these classes. It should be noted that here we do not focus on improving the user delay performance when they are experiencing bad channel conditions since this would decrease the efficient resource utilization and achievable throughput. Instead we perform a priority-based compensation procedure, which shortens the time required for the high priority user to obtain the lost service.

2.4 Performance Evaluation

In this section we evaluate the performance of the proposed algorithms in terms of achievable fairness and throughput. In order to obtain a better understanding of the improvements achieved by CARS and FCARS algorithms we compare their performance with the corresponding performances of three other rate scheduling algorithms, namely the WFQ with Ideal Channel (WFQ-IC), WFQ with Error Channel (WFQ-EC), Channel Adaptive WFQ (CA-WFQ). The WFQ-IC refers to the case where all channels are in perfect condition and each user is assigned the data rate according to its weight and the system achievable rate with the given power index. Obviously this is a non-realistic case, however we consider it here for comparisons purposes only. In the WFQ-EC although the actual state of the links may not be good and may vary with the time, the base station is not aware of that, and assumes them to be in good condition all the time, thus assigning to each user data rate according to its weight following the WFQ principle as in the ideal system. Therefore the high packet error rate becomes inevitable and the system throughput decreases. In the CA-WFQ scheme the scheduler is assumed to be aware of the channel condition variation of the mobile users and at the same time the strict fairness property of service allocation is enforced at all time instances. This scheme attempts to enforce the strict fairness rule and achieve the same fairness as the WFQ and as a result its throughput performance in some cases may be limited by the worst user's behavior [31]. Furthermore we present some

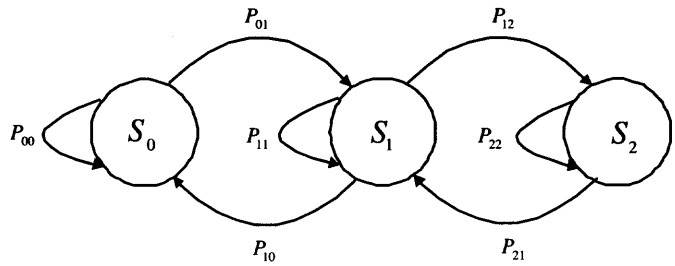

Figure 2.1 Channel state transition diagram

numerical comparative results of FCARS and a pure throughput maximization scheme [26], that achieves the maximum total uplink throughput by allowing only the best k users in terms of their received power to transmit.

In the following we first describe the model and assumptions used throughout our performance study, and then we present the results of the comparative study.

2.4.1 Model and Assumptions

In general, throughout our numerical study, unless explicitly indicated, we consider a single cell multi-rate DS-CDMA with seven users. All seven users are continuously backlogged during the simulation and generate packets with average size of 320 bytes. The available system power index is assumed to be 0.9, the system chip rate is 10^7chip/second and the required SNIR is 10dB. Perfect power control is used and hence a mobile user may increase its transmission power to overcome the channel fading until it reaches its pre-defined maximum power. With the limitation of each individual user's transmission power, the maximum power index a user can accept is 0.7, which corresponds to the maximum transmission rate of 2.33Mbit/s.

To compare the performances achieved by the various algorithms two different scenarios are considered. In scenario-1 (Table 2.1) all users have the same weight, which allows us to better understand and compare the achievable performances when users have different channel conditions. In scenario-2 the operation and effectiveness of FCARS algorithm is demonstrated in an environment where users present different weights. For demonstration

Table 2.1 The Weight Plans During Simulation

User ID	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Weight (Scenario 1)	3	3	3	3	3	3	3
Weight (Scenario 2)	3	6	6	1	6	2	10

purposes in most of the results presented in this section a scheduling cycle of 0.5 seconds is assumed. However numerical results that demonstrate the influence of different scheduling cycles on the achievable performance of FCARS algorithm are also presented.

To better study the effect of our algorithm on the service allocation throughout our simulation study five of the users are assigned good channels and are able to achieve low BER of 10^{-6}, while the rest two users have very unstable channels which makes the corresponding received power vary according to a three-state Markov channel model [34] shown in Figure 2.1. Table 2.2 lists each state's corresponding parameters. The state occupation time follows exponential distribution with average times for each state as listed in Table 2.2. When the channel is in one state, it means that the received power resides within a certain range if the maximum transmission power is used. This received power can be converted to this user's power index through Definition 2. Table 2.2 provides the range of the achievable power index capacity for each state. In order to simulate the abrupt channel condition change that enables us to observe the performance of compensation algorithm more clearly, the occupation time of state 1 is set to a small value, which means that the channel hardly stays at state 1 that has the middle range of the received power. Besides the channel state transition the channel also varies within the received power range of its current state.

Table 2.2 Channel State Parameters

	State Transition Probability	**Average State Occupation Time (sec)**	**Power Index Range**
State 0	$p_{00} = 0.997, p_{01} = 0.003$	1	[0.01, 0.1)
State 1	$p_{11} = 0.02, p_{12} = 0.49,$ $p_{10} = 0.49$	0.0006	[0.1, 0.5)
State 2	$p_{22} = 0.999, p_{21} = 0.001$	3	(0.5, 0.7]

2.4.2 Numerical Results and Discussion

Throughput Performance Results

We first present the corresponding results under scenario 1 where all the users have the same weight. Specifically Figure 2.2 demonstrates the achieved total service for all the five schemes under consideration as a function of time. We can easily see from this figure that the performance of CARS and FCARS approach that of the ideal WFQ-IC, as they are able to make full use of the available power index by adjusting the assigned transmission rates to the corresponding channel conditions. Please note that sometimes their values may even exceed the ideal case because of the uneven allocation of power index. From this figure we also observe that the performance achieved by both CA-WFQ and WFQ-EC is considerably degraded due to the reduced throughput to achieve fairness for the CA-WFQ, and the increased packet loss rate in WFQ-EC.

In Figure 2.3 we present the number of lost packets for five schemes as a function of time where we can see that the number of lost packets under WFQ-EC is very high. This happens because due to the ignorance of the channel condition variation in WFQ-EC the scheduler assigns transmission rates assuming good channel conditions, and thus the two bad users experience very high packet loss rate which reduces significantly the actual service received by them. The CA-WFQ taking into account the channel conditions and

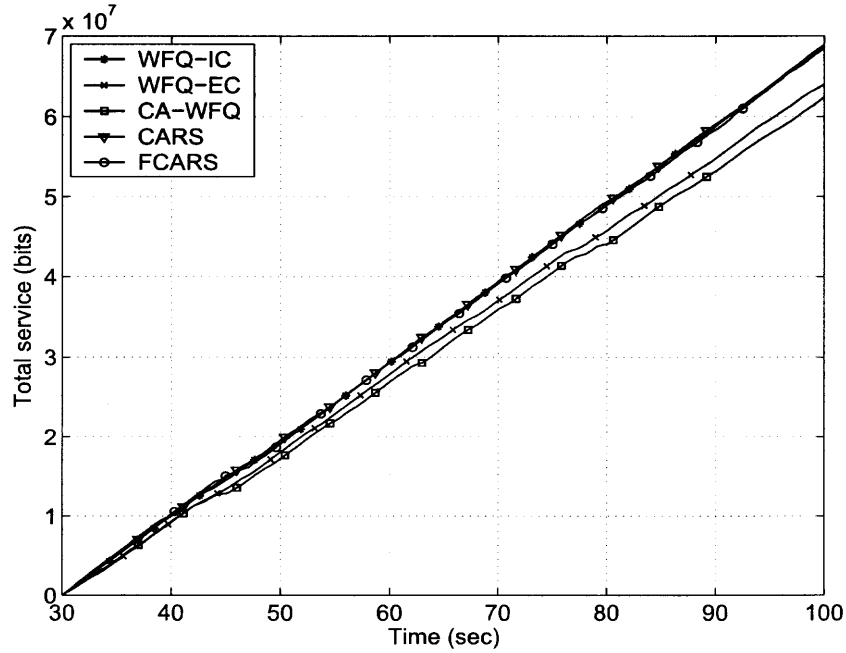

Figure 2.2 Total service (in bits) all users received under the different scheduling schemes

attempting to maintain the fairness in the system, schedules low rates not only for the bad users but also limits the throughput of the five good users. Due to its low throughput the number of lost packets in CA-WFQ is low, even less than the corresponding results in FCARS and CARS.

In Figure 2.4 we present the total achievable service (in bits) provided under FCARS algorithm for five different scheduling cycles. Since we have assumed that the channel changes states every 0.3 seconds, the performance for the cases with scheduling cycle of 0.1 and 0.3 are very close. Furthermore we notice that, as we expected, the achievable performance deteriorates as the scheduling cycle increases. This happens because as the scheduling length increases, given the fast and abrupt changes of the channel conditions, the scheduling policy does not react fast and does not adjust the scheduled rates in order to catch up the channel change, and this causes higher packet losses and inefficient resource allocation.

In Figure 2.5 and Figure 2.6 we present the overall achievable throughput as well as the corresponding individual users' throughput, under scenario 1, for FCARS and Maximum

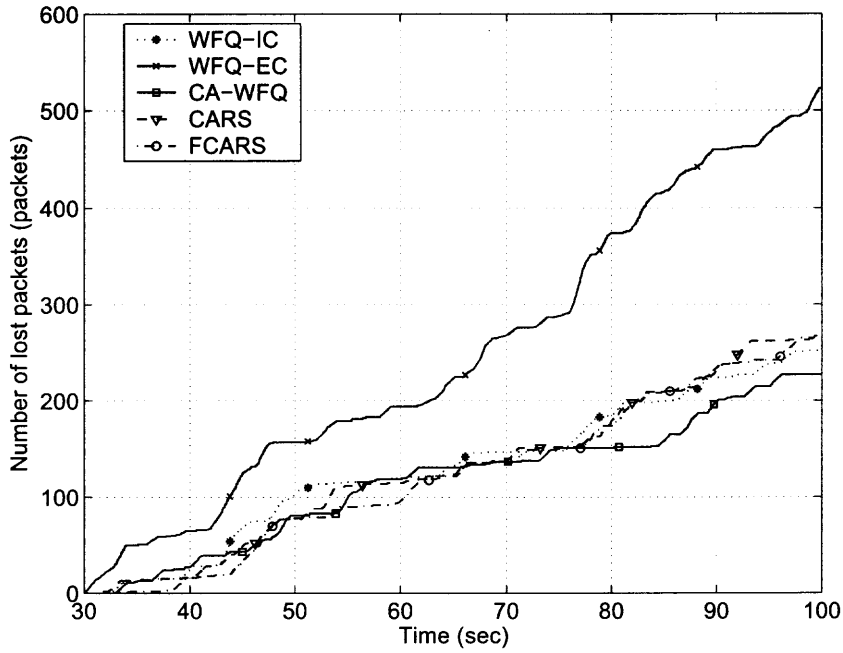

Figure 2.3 Number of lost packets under the different schemes

uplink throughput (MAX) scheme. The MAX scheme [26] achieves the maximum total uplink throughput by allowing only the best k users in terms of their received power to transmit. Parameter k is determined by iteratively comparing the throughput of best i users, $1 \leq i \leq N$. In this experiment, the channels of all users were modeled by the multi-state Markov fading channel models with different average received power (i.e. users 1 to 3 have higher average received power), which enables us to better interpret the influence of the different schemes on the achievable throughputs. From these figures we can see that MAX scheme achieves high throughput but biases against the group of users with low received average power. As we expected, the users with higher average power get more chances to transmit and they achieve higher throughput. On the other hand although the total throughput achieved by FCARS is slightly lower, due to the effectiveness of its compensation mechanism, FCARS achieves fairness and provides similar throughputs for all the users.

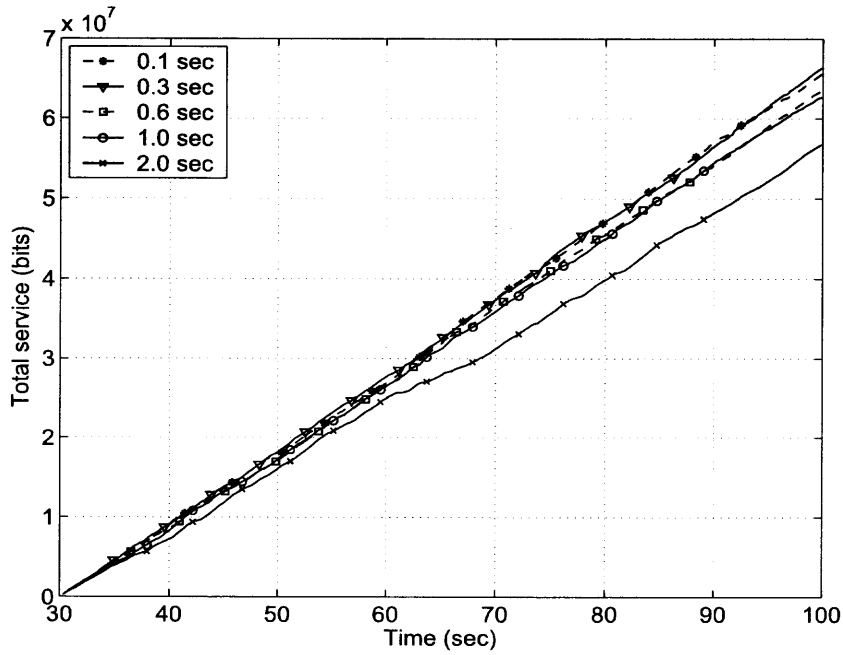

Figure 2.4 Total service received (in bits) by all users under FCARS for the different scheduling cycles

Fairness results

Despite the relatively large volume of traffic served by CARS, Figure 2.7 shows that this gain is obtained by taking advantage of the users with bad channels. In this figure we present the excess service received under CARS by the two bad users and a good user. The service lost by the bad users is not regained even if the corresponding channels improve and therefore the fairness property is not satisfied by CARS. FCARS on the other hand, as demonstrated in Figure 2.8 solves this problem successfully through its service compensation process. As we can see from this figure FCARS algorithm allows the bad user to gain its lost service at a later time when its channel condition improves and therefore eventually both the good and bad user receive comparable (according to their weights) throughput. Figure 2.9 presents the excess service received by the two bad users and a typical good user as time evolves under FCARS algorithm. Although the channel condition of bad users vary the excess service converges to zero. This happens because as explained before the bad user receives compensation from the good users for its service that was lost at previous times

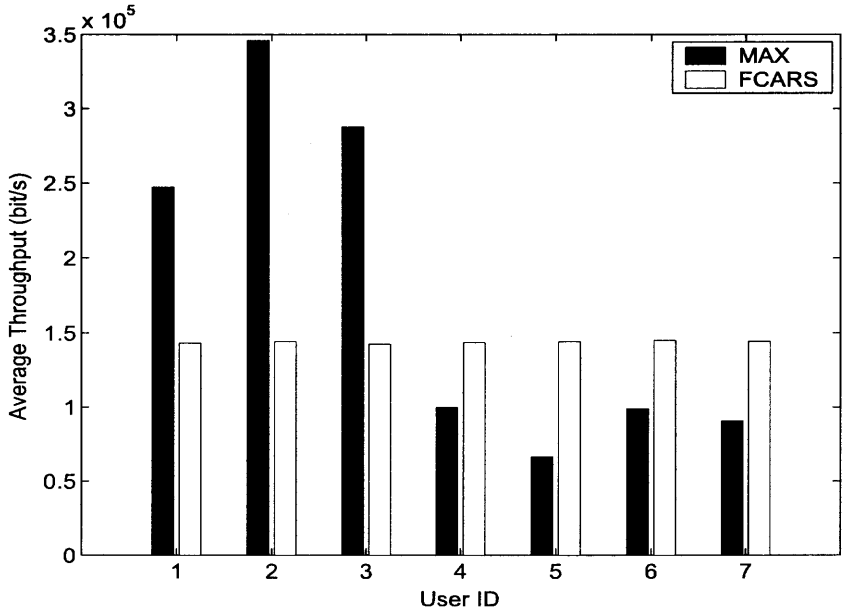

Figure 2.5 Throughput comparison of different users under FCARS and MAX algorithms

due to the bad channel conditions. Figure 2.10 clearly shows how the bad user acquires its lost service by receiving very high transmission rate after its channel recovers. Another advantage of FCARS, as can be confirmed by Figure 2.9 and Figure 2.10, is that the excess service of good users varies only slightly compared to the more abrupt variation of excess service of bad users. This is because the lost service by the bad users is fairly shared among all qualified good users, and similarly every good user gives up only small portion of its service at a later time. As a result the FCARS algorithm manages to provide smooth service to the good users, while at the same time achieves fairness and high throughput.

In order to obtain a more in-depth understanding of FCARS fairness operation and study the impact of the various users' different weights on FCARS scheme, in Figure 2.11 we present the excess service received by each user as the system evolves for scenario 2. We can see from this figure that bad user 2, who has larger weight than bad user 1, observes higher service loss than bad user 1. At the same time the good users that own larger weight also acquire more excess service accordingly (e.g. good user 5).

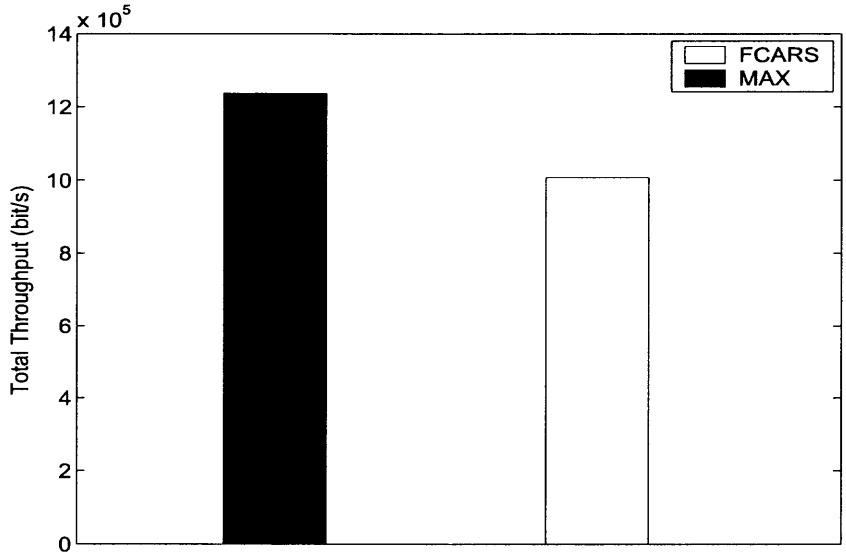

Figure 2.6 Total Throughput comparison of under FCARS and MAX algorithms

Table 2.3 Achieved Fairness Under FCARS For Different Number of Users and weights

	3 Users	**7 Users**	**13 Users**	**23 Users**
Weight:1	1	1.012	1.022	1
Weight:4	1.759	4.052	4	4.004
Weight:8	2.601	8.003	7.997	8.005

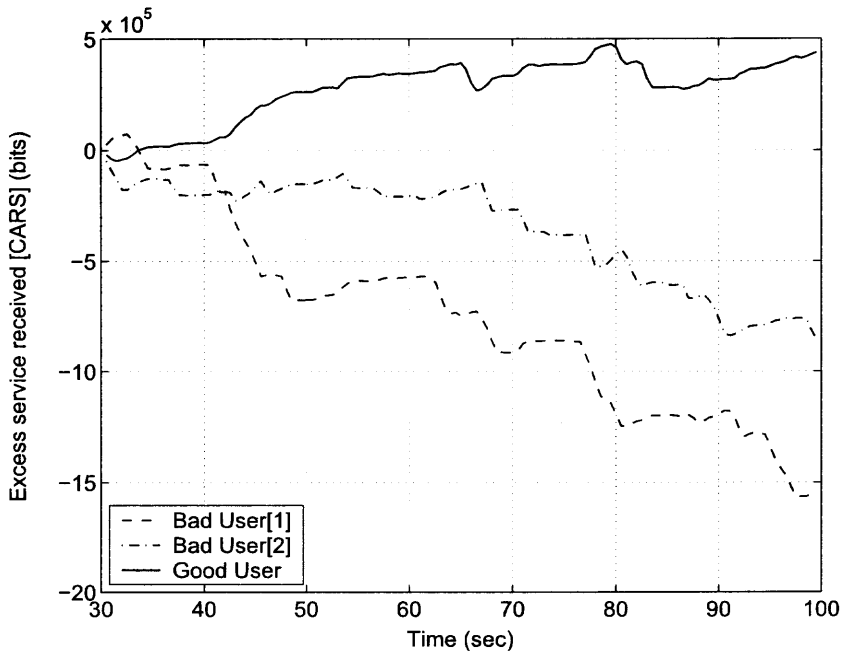

Figure 2.7 Excess service received (in bits) by good users and bad users under CARS

Furthermore Table 2.3 presents the achieved fairness under FCARS algorithm for different number of users in the system. The users were divided into three groups and different original weights were assigned in each group (the corresponding values are shown in the first column of the table). From this table we observe that the more the users in the system the better the fairness achieved by FCARS, since as the number of users increases the system has more flexibility on providing service adjustments. It should be noted, that in the case of 3 users the available system power index (resources) is much larger than the capacities of all users, and since the algorithm first attempts to achieve high throughput by exploiting the varying channel conditions, the obtained throughput mainly depends on the respective channel conditions, and thus in this case the scheduler has limited effect on the adjustment of service allocation. However as the number of users in the system increase the proposed algorithm achieves fairness and high throughput simultaneously, since it has more flexibility on the corresponding power index adjustments.

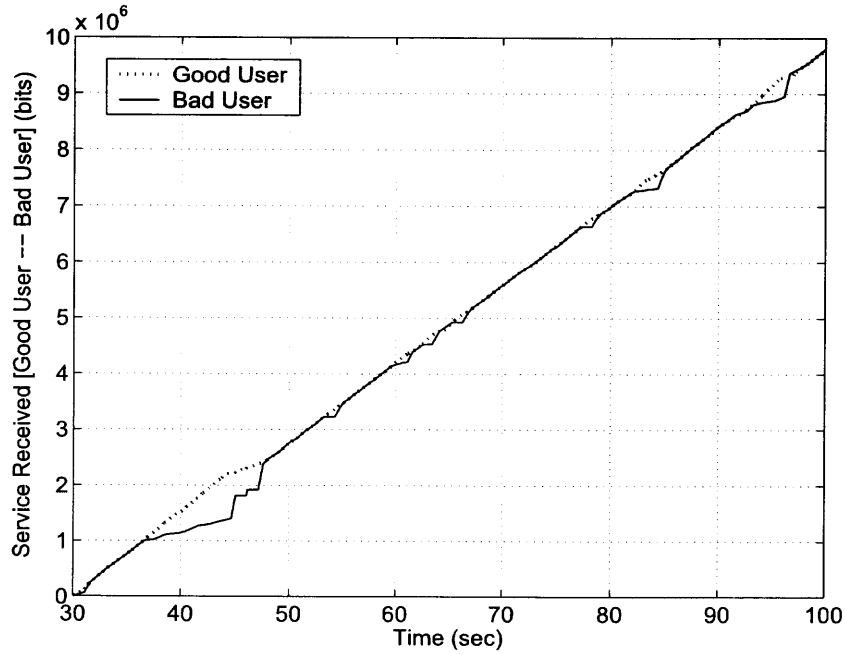

Figure 2.8 Service received (in bits) by good user and bad user under FCARS

Delay Results

Figure 2.12 presents the sample distribution of the delay experienced by the various packets under the five different schemes for scenario 1. As expected the ideal system (WFQ-IC) presents the best delay performance since it provides steady fixed rate service. We also observe that a large number of packets have larger delay in the WFQ-EC due to its high packet loss rate and retransmission. The CA-WFQ scheme achieves to implement the fairness at the cost of reduced system throughput and increased delay for most of the packets. In CARS we notice that more packets have smaller delay when compared to FCARS, however at the same time there is a larger number of packets that suffer from larger delay, and thus CARS provides an unbalanced service. This happens due to the unfairness in the service allocation and distribution introduced by CARS in an attempt to improve throughput. As a result some users receive more service and experience smaller delay, while others receive reduced service and have longer delays. FCARS by compensating the lagging users, reduces the delay caused by the bad channels and achieves to offer a more fair service.

As mentioned before in subsection 2.3.4 the M-FCARS, in order to explicitly distinguish between delay-sensitive and non-delay-sensitive traffic, performs a priority-based compensation procedure, which shortens the time required for the high priority user to obtain back its lost service due to the bad channel conditions. Figure 2.13 present the corresponding results of the case of two traffic classes – Class 0 represents the delay-sensitive traffic and gets higher priority than Class 1 in the M-FCARS service compensation procedure. For this experiment in order to observe more clearly the results, instead of using the three-state Markov channel model, we intentionally set the error channel model to be periodically good and bad with the respective power index capacities of 0.7 and 0.01. The two users with bad channel conditions are assigned to Class 0 and Class 1. They experience the exactly same channel conditions, and therefore compete for the service compensation once their channels become good. As we see in Figure 2.13 the Class 0 and the Class 1 users have the same delays when their channels are in bad state. However once their channels recover, the delay of Class 0 user drops quickly due to its higher priority in receiving service compensation. On the other hand the Class 1 user presents relatively slower delay improvement until the Class 0 user completely recovers.

2.5 Conclusion

Fair scheduling in CDMA system still faces a lot of challenges due to the unique characteristics of CDMA systems, such as: there exist variations in the available system capacity (rate); the limited transmission power constrains the link capacity; the channel experiences busty errors and location-dependent errors; multiple users are served simultaneously.

In this chapter we studied the fundamental problem of efficient and fair dynamic rate scheduling. We first defined an important parameter—Power Index Capacity (PIC), which indicates the possible power index a user can accept. The power index adjustment among users within the constraint of PIC was analyzed, which creates the basis for a feasible rate scheduling and service compensation process. We then proposed two new algorithms,

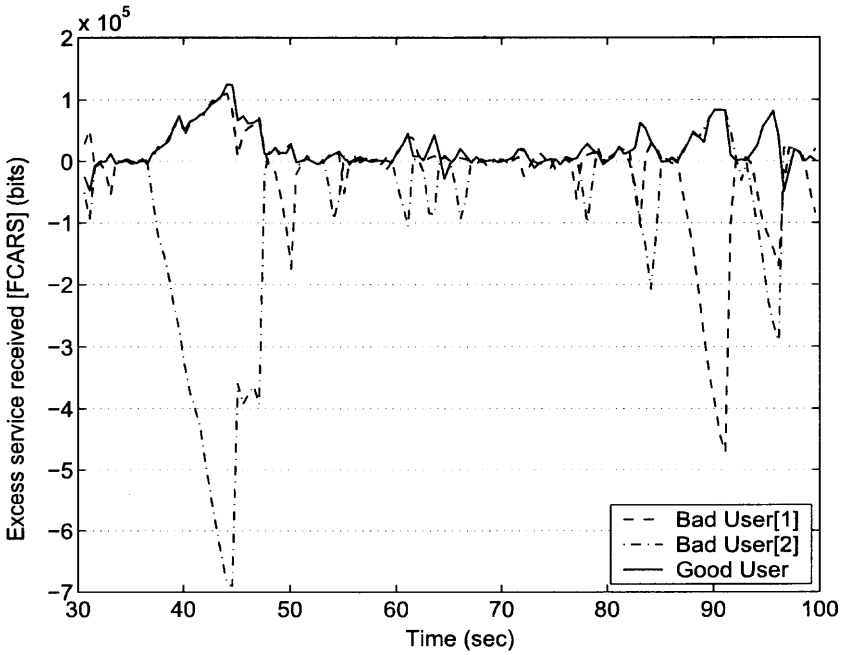

Figure 2.9 Excess service received (in bits) by the good user and bad user under FCARS

CARS and FCARS. Through an iterative process CARS estimates the power index of a single user starting with reference the ideal system, and then in the following iterations redistributes the unused power index to the users with better channel condition, in order to fully utilize the available system resources. Users with better channel condition obtain more bandwidth, while those with worse channel condition get less bandwidth, which however may result to fairness violations. To overcome the unfair service allocation FCARS implemented a compensation algorithm, in which the lagging users can receive compensation service when the corresponding channel conditions improve.

The performance of the proposed algorithms in terms of achievable fairness and throughput were obtained via modeling and simulation and were compared with the performances of other rate scheduling algorithms. The corresponding results demonstrated the significant improvements that can be achieved by FCARS that manages not only to offer performance that approaches the reference ideal system in terms of throughput, but also to achieve the long-term fairness by keeping the service received by each user proportional to its weight despite the users' channel variations.

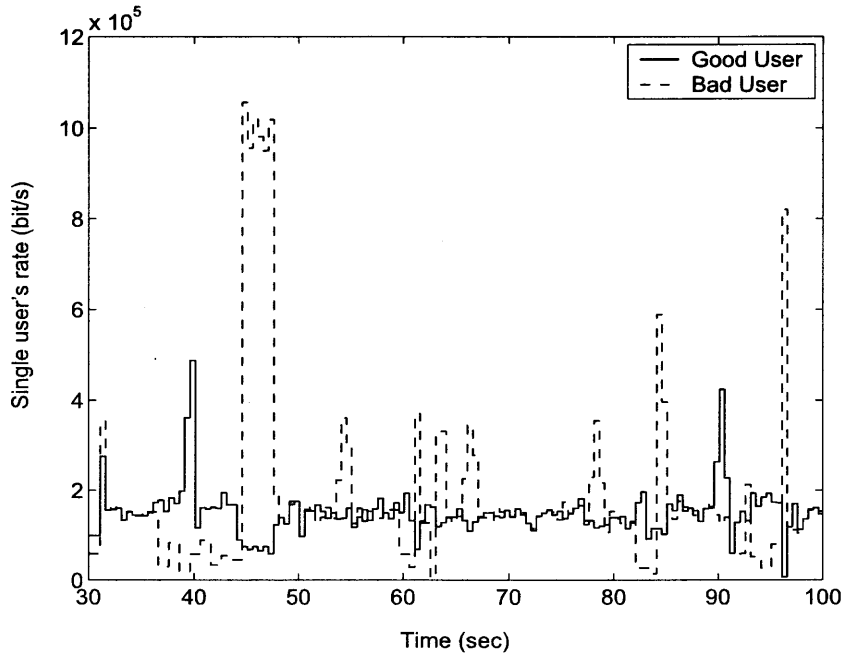

Figure 2.10 Scheduled transmission rates of a good user and a bad user under FCARS

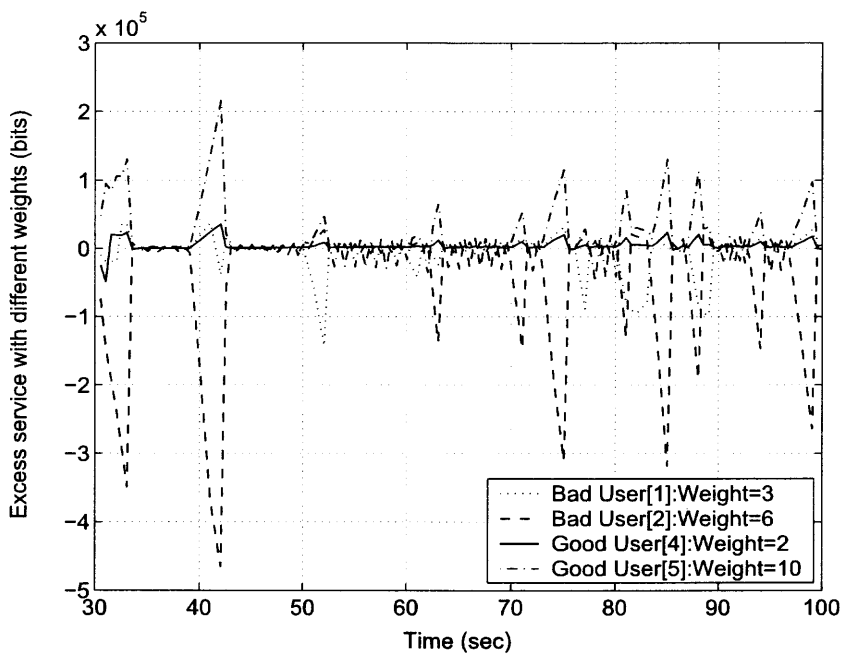

Figure 2.11 Excess service received (in bits) under FCARS for scenario-2 (users with different weights)

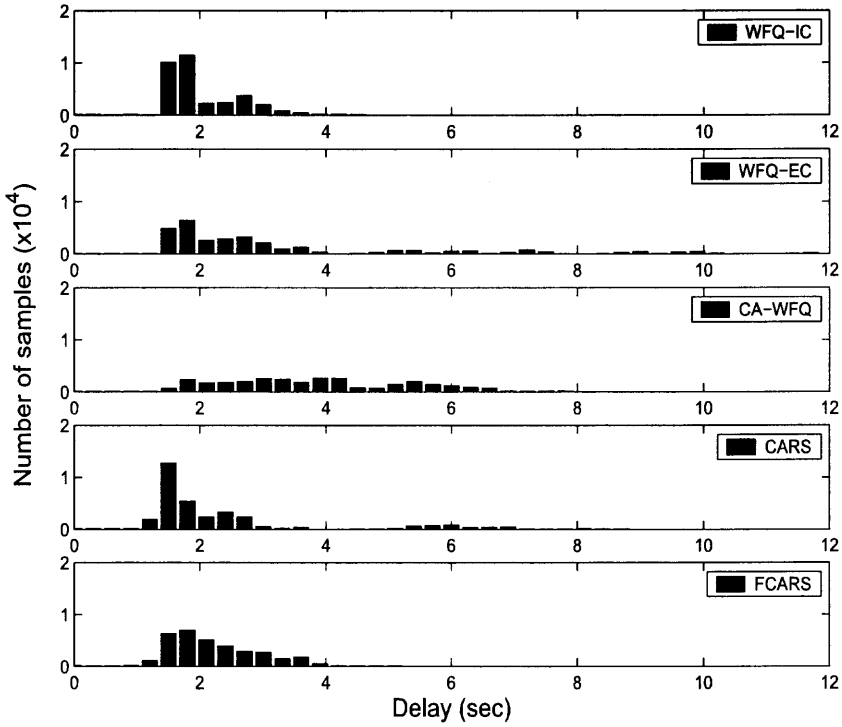

Figure 2.12 Sample distribution of delay in different cases (sample interval = 0.3 second)

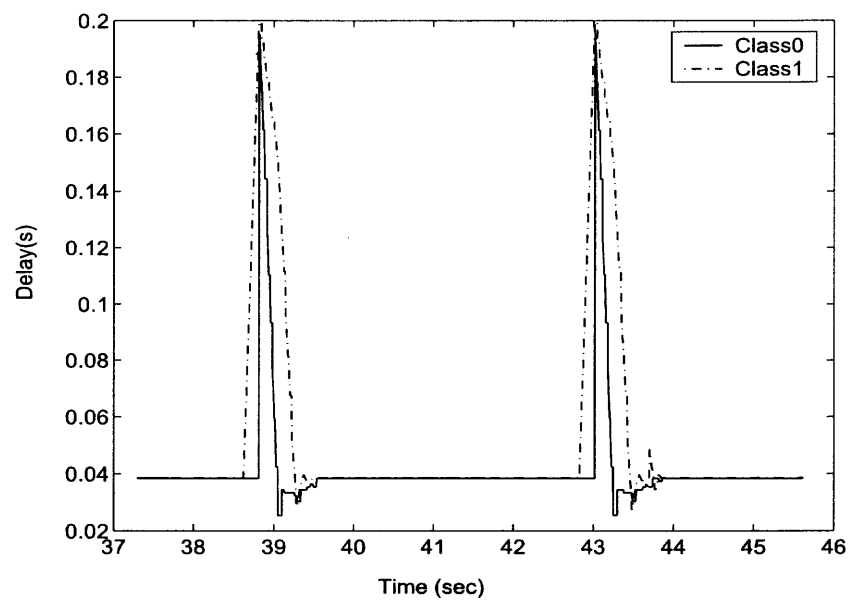

Figure 2.13 Delay differences of Class 0 and Class 1 users with bad channels under M-FCARS algorithm

CHAPTER 3

THROUGHPUT MAXIMIZATION FAIR SCHEDULING

In this chapter we study the problem of opportunistic fair scheduling in the uplink transmission of CDMA systems, with the objective of obtaining the optimal throughput in the uplink scheduling while still maintaining the long-term fairness. The FCARS algorithm described in the last chapter provides opportunities for transmission for all admitted users in each slot, despite their possible undesired channel conditions. The throughput improvement is achieved by limiting the transmit rate (or transmit power) of weak users. These weak users however are compensated at a later time when their channel conditions improve, in order to maintain the pre-specified desired fairness. However, the distribution of power index among all admitted users at the same time may lower the achievable system throughput due to the convexity of the relationship between the power index and the transmit rate.

Therefore, in this chapter the emphasis is placed on devising an optimal scheduling policy that reaches the maximum system throughput in the uplink CDMA system, while satisfying the long-term fairness property. The rest of the chapter is organized as follows. In section 3.1 the system model that is used throughout our analysis is described, and the problem of the uplink scheduling in CDMA systems is rigorously formulated as a multi-constraint optimization problem. It is demonstrated that this problem can be expressed as a weighted throughput maximization problem, under certain power and QoS constraints, where the weights are the control parameters that reflect the fairness constraints. Based on the concept of power index capacity, this optimization problem is converted into a simpler linear knapsack problem in section 3.2.1, where all the corresponding constraints are replaced by the users' power index capacities at some certain system power index. The optimal solution of the latter problem is identified in sections 3.2.2 and 3.2.3, while in section 3.2.4 a stochastic approximation method is presented in order to effectively identify

the required control parameters. Section 3.3 contains the performance evaluation of the proposed method, along with some numerical results and discussion.

3.1 System Model and Problem Formulation

In this study we consider a single cell DS-CDMA system with $B(k)$ backlogged users at time slot k. The users' channel conditions are assumed to change according to some stationary stochastic process, while the uplink transmission rate is assumed to be adjustable with the variable spreading gain technique [35]. Each user i is associated with some pre-assigned weight ϕ_i according to its QoS requirement. In the following for simplicity in the presentation we omit the notation of the specific slot k from the notations and definitions we introduce. Let us denote by r_i the transmission rate of user i in the slot under consideration. We assume that the chip rate W for all mobiles is fixed, and hence the spreading gain G_i of user i is defined as: $G_i = \frac{W}{r_i}$. Let us also denote by γ_i the required SINR (Signal to Interference and Noise Ratio) level of user i, by h_i the corresponding channel gain, and by p_i user's i transmission power at a given slot, which however is limited by the maximum power value $p_i^{\max}$. Therefore the received SINR γ_i' for a user i is given by:

$$\frac{h_i p_i G_i}{\alpha \sum_{j=1, j\neq i}^{B(k)} h_j p_j + W\eta_0} = \gamma_i' \quad i = 1, 2, \ldots, B(k) \tag{3.1}$$

where η_0 is the one-sided power spectral density of additive white Gaussian noise (AWGN) and α determines the proportion of the interference from other users' received power. Without loss of generality in the following we assume $\alpha = 1$. Obviously to meet the SINR requirement, the received SINR γ_i' has to be larger than the corresponding threshold γ_i, i.e. $\gamma_i' \geq \gamma_i$. In the following we assume perfect power control in the system under consideration, while users are scheduled to transmit at the beginning of every fixed-length slot. The objective of the optimal scheduling policy Q^* is to find the optimal number of allowable users and their transmission rates, which achieves the maximum system throughput while maintaining the fairness property.

3.1.1 Problem Formulation

Let $R(k) = \sum_{i=1}^{B(k)} r_i(k)$ denote the total throughput in slot k. Our objective function is to maximize the expectation of $R(k)$ by selecting the optimal transmit power vector $(p_1, p_2, ..., p_{B(k)})$ and transmit rate vector $(r_1, r_2, ..., r_{B(k)})$ That is:

$$\max \; E\left(\sum_{i=1}^{B(k)} r_i\right) \tag{3.2}$$

subject to specific SINR, maximum transmit power, and fairness constraints, as follows:

$$\begin{aligned} \frac{h_i p_i G_i}{\sum_{j=1, j\neq i}^{B(k)} h_j p_j + W\eta_0} &\geq \gamma_i, \quad for \;\; i = 1, 2, \ldots, B(k) \\ p_i &\leq p_i^{\max}, \; for \;\; i = 1, 2, \ldots, B(k) \\ \frac{\overline{r}_i}{\phi_i} &= \frac{\overline{r}_j}{\phi_j} \;\; \text{for } 1 \leq i, j \leq B(k) \end{aligned} \tag{3.3}$$

where $\overline{r}_i = E(r_i)$ denotes the mean throughput of user i in the corresponding backlogged period. It has been shown in [36] [37] that the above constrained optimization problem can be considered as equivalent to the following problem (3.4), where Z is the minimal value among all $\frac{\overline{r}_i}{\phi_i}$, i.e. $Z = \min_i\{\frac{\overline{r}_i}{\phi_i}\}$. In (3.4) we transform the objective function (3.2) into finding the optimal transmit powers and rates that maximize the minimal normalized average rate Z.Therefore:

$$\begin{aligned} & & \max \;\; Z & & \\ s.t. & \quad Z & \leq & \frac{\overline{r}_i}{\phi_i}, & 1 \leq i \leq B(k) \\ \frac{h_i p_i W/r_i}{\sum_{j=1, j\neq i}^{B(k)} h_j p_j + W\eta_0} & & \geq & \gamma_i & i = 1, 2, \ldots, B(k) \\ & p_i & \leq & p_i^{\max} & 1 \leq i \leq B(k) \end{aligned} \tag{3.4}$$

Apparently the solution of the above problem will finally make $Z = \frac{\overline{r}_i}{\phi_i}$ for $1 \leq i \leq B(k)$ since one can always reduce its throughput for the benefit of other users in order to

maximize Z. With the constraint $Z = \frac{\bar{r}_i}{\phi_i}$ the objective function then is generalized to

$$\max \sum_{i=1}^{B(k)} w_i \bar{r}_i \tag{3.5}$$

where w_i is an arbitrary positive number. Here the crucial observation [36] is that the optimal scheduling policy will be the one that maximizes the sum of weighted throughputs and equalizes the normalized throughput. The maximization of mean weighted rate in (3.5) is obtained by the maximization of the weighted rate in every slot, i.e. max $\sum_{i=1}^{B(k)} w_i r_i$ for every slot k. In conclusion, to obtain the optimal uplink throughput while keeping the fairness, we must solve the following problem,

$$\max \sum_{i=1}^{B(k)} w_i r_i \tag{3.6}$$

$$s.t. \quad \frac{h_i p_i W / r_i}{\sum_{j=1, j\neq i}^{B(k)} h_j p_j + W\eta_0} \geq \gamma_i \quad i = 1, 2, \ldots, B(k) \tag{3.7}$$

$$p_i \leq p_i^{\max} \quad 1 \leq i \leq B(k) \tag{3.8}$$

The fairness constraint, that is $\frac{\bar{r}_i}{\phi_i} = \frac{\bar{r}_j}{\phi_j}$, is represented by the choice of w_i. By adjusting the value of w_i, the user will get more or less opportunities to transmit data, and hence the corresponding normalized throughput is balanced. As we discuss later in this paper, the value of w_i can be approximated by a stochastic approximation algorithm, which has already found its application in [21] [37] under similar situations. Please note that since we assume perfect power control in the CDMA system under consideration, only the equality case of (3.7) is considered here.

The following proposition 2 states that the optimal solution is achieved when a user either transmits at full power or does not transmit at all.

Proposition 2 *The optimal solution that maximizes the weighted throughput of problem (3.6) is such that*

$$p_i(k) \in \{0,\ p_i^{\max}\} \quad \textit{for } i = 1, 2, ..., B(k)$$

Proof. : In order to minimize the multiple access interference, users transmit with the minimum required power to meet the required threshold γ_i. Therefore we consider the equality case of constraint (3.7). To maintain exactly the threshold γ_i for user i, the achievable transmit rate is represented as

$$r_i(k) = \frac{h_i p_i W}{\gamma_i \left(\sum_{j=1, j\neq i}^{B(k)} h_j p_j + W\eta_0 \right)} \tag{3.9}$$

The objective function then becomes

$$Z = \sum_{i=1}^{B(k)} w_i r_i = \sum_{i=1}^{B(k)} \frac{w_i h_i W}{\gamma_i} \frac{p_i}{\sum_{j=1, j\neq i}^{B(k)} h_j p_j + W\eta_0} \tag{3.10}$$

Differentiating twice with respect to the transmit power of a user m, we obtain:

$$\frac{\partial^2 Z}{\partial p_m^2} = 2 \sum_{i=1, i\neq m}^{B(k)} \frac{w_i h_i W}{\gamma_i} \frac{p_i h_m^2}{\left(\sum_{j=1, j\neq i}^{B(k)} h_j p_j + W\eta_0 \right)^3} \tag{3.11}$$

Since w_i is positive number, obviously (3.11) is nonnegative, while the objective function is a convex function of p_m. Hence the optimal solution of this problem is that the transmit power obtains the value of its boundary, i.e. either 0 or $p_i^{\max}$. ■

In section 3.2 the corresponding optimization problem is transformed to an equivalent problem of a simpler form, which facilitates the identification of the optimal solution. However, in the following we first introduce the concept of power index capacity which is used to represent the corresponding constraints, under the problem transformation.

3.1.2 Power Index Capacity

It has been shown in [31] that by solving the constraints (3.7)(3.8) the following inequality must be satisfied if there exists a feasible power assignment $\mathbf{p} = [p_1, p_2, \ldots, p_{B(k)}]$ that meets the QoS requirements:

$$\sum_{i=1}^{B(k)} g_i \leq 1 - \frac{\eta_0 W}{\min\limits_{1\leq i\leq B(k)} \left\{ p_i^{\max} h_i \left(\frac{G_i}{\gamma_i} + 1 \right) \right\}} = 1 - \frac{\eta_0 W}{\min\limits_{1\leq i\leq B(k)} \left\{ \frac{p_i^{\max} h_i}{g_i} \right\}} \tag{3.12}$$

where

$$g_i = \frac{\gamma_i}{\gamma_i + G_i} \tag{3.13}$$

is defined as the power index of user i. Relation (3.12) is the necessary and sufficient condition such that a power and rate solution is feasible under constraints (3.7)(3.8) [31].

Let us regard $\sum_i g_i$ as the actual system load, which is the sum of power indices assigned to all backlogged users, while we assume that there is a target system load ψ. It should be noted that ψ here is not fixed but has value $0 \leq \psi < 1$. If we consider $\sum_i g_i = \psi$, then (3.12) can be rewritten as follows:

$$\begin{aligned} \min_{1 \leq i \leq B(k)} \left\{ \frac{p_i^{\max} h_i}{g_i} \right\} &\geq \frac{\eta_0 W}{1-\psi}, \text{ and } g_i \leq \psi \\ \text{therefore } \frac{p_i^{\max} h_i}{g_i} &\geq \frac{\eta_0 W}{1-\psi} \text{ for all } i, 1 \leq i \leq B(k) \end{aligned} \tag{3.14}$$

Hence given the system load ψ the maximum possible power index g_i a user can accept in (3.14) is determined by the maximum transmit power $p_i^{\max}$ and the channel gain h_i.

Definition 5 *In a CDMA system with $B(k)$ backlogged users at time slot k, given the target system power index ψ, the maximum power index that does not violate (3.12) for a single user whose channel gain is h_i is defined as the power index capacity (PIC) $\pi_i(h_i, \psi)$ of this user.*

From (3.14) it can be easily found that the PIC of user i is:

$$\pi_i(h_i, \psi) = \min \left\{ (1-\psi) \frac{p_i^{\max} h_i}{\eta_0 W}, \ \psi \right\} \tag{3.15}$$

Note that in (3.15) the power index capacity is limited by the target system power index. This is reasonable since a power index capacity that is greater than ψ will have no practical meaning and application. Furthermore since our focus in this paper is to find an optimal

scheduling policy as well as the optimal system load ψ, the value of ψ in (3.15) is not determined in advance. Finally it should be noted that in some cases the chosen target system load ψ, e.g. when ψ is very close to 1, may not be achievable due to the maximum transmit power limitation. In this case, we still apply the power index capacity definition of (3.15) with the ψ set to the chosen target system load.

Intuitively, the power index represents the relationship between the transmission power and the corresponding interference that is caused to other users. If we considered that the total system power index is fixed to ψ, larger power index g_i for user i indicates that it has relatively higher signal to interference ratio compared to the other users with smaller power index, while at the same time it causes more interference to them. Accordingly, users with high power indices may lower their transmission power to reduce the interference they may cause, which in turn means that they will have smaller power index to limit the intracell interference of the system, and therefore satisfy relation (3.12) that guarantees the existence of a feasible transmission power solution.

3.2 Problem Transformation and Optimal Solution

3.2.1 Problem Transformation

The corresponding constraints in terms of the power index can be represented as follows:

$$\max\ Z = \sum_{i=1}^{B(k)} w_i f_r\left(g_i, \gamma_i\right) \tag{3.16}$$

$$\sum_{i=1}^{B(k)} g_i \leq \psi \tag{3.17}$$

$$g_i \leq \pi_i(h_i, \psi), \quad 1 \leq i \leq B(k) \tag{3.18}$$

$$0 \leq \psi < 1 \tag{3.19}$$

Please note that in the objective function we represent the rate $r_i = f_r(g_i, \gamma_i)$ as a function of power index g_i, where

$$f_r(g_i, \gamma_i) = \frac{g_i}{1 - g_i} \frac{W}{\gamma_i} \tag{3.20}$$

which converts the power index into transmission rate, and can be easily derived from (3.13) by replacing G_i with $\frac{W}{r_i}$.

In the following let $\mathbf{V} = \{v_1, v_2, ..., v_i, ...\}$ denote the set that contains all the power and rate vectors that satisfy constraints (3.7) and (3.8), and $v_i = \{p_{1,i}, p_{2,i}, ..., p_{B(k),i}, r_{1,i}, r_{2,i}, ..., r_{B(k),i}\}$. The elements $p_{j,i}$ and $r_{j,i}$ represent the transmit power and rate of the jth user in the ith vector. Similarly we define another set $\mathbf{V}'$ containing the power and rate vectors v_i' that satisfy constraints (3.17), (3.18) and (3.19). By definition it is obvious that any power and rate vector $v_i \in \mathbf{V}$ is feasible. However, since in constraint (3.19), ψ can be infinitely close to 1, the required transmit power could also approach infinite. The following proposition states that, if perfect power control is assumed, for any rate (or power index) vector that satisfies constraints (3.17), (3.18), and (3.19), there always exists a feasible transmit power vector.

Proposition 3 *If the power index assignment for all $B(k)$ backlogged users satisfies constraints (3.17), (3.18) and (3.19), there always exists a feasible transmit power assignment, i.e. $p_i < p_i^{\max}$ for $1 \leq i \leq B(k)$.*

Proof. Let vector $\mathbf{g} = \{g_1, g_2, ..., g_{B(k)}\}$ be the power index vector that satisfies constraints (3.17),(3.18),and (3.19). Denote $\psi = \sum_{i=1}^{B(k)} g_i$ the sum of all power indices in vector $\mathbf{g}$. From the definition of power index capacity, the power index capacity of each user is $\pi_i(h_i, \psi)$ and $g_i \leq \pi_i(h_i, \psi)$. Based on Definition 5 and equation (3.15) we have the following relation,

$$\psi \leq 1 - \frac{\eta_0 W \cdot \pi_i(h_i, \psi)}{p_i^{\max} h_i} \leq 1 - \frac{\eta_0 W \cdot g_i}{p_i^{\max} h_i}$$

Hence for any user i, the transmit rate may be chosen within range

$$p_i^{\max}\frac{g_i}{\pi_i(h_i,\psi)} \leq p_i \leq p_i^{\max}$$

which still satisfies the above inequality and proves this proposition. The power control of the CDMA system will choose the minimal transmit power, that meets the required SINR. ■

The following proposition proves that the two sets $\mathbf{V}$and $\mathbf{V}'$ contain the same elements, which means that (3.17)(3.18)(3.19) and (3.7)(3.8) impose the same constraints over our problem.

Proposition 4 *Any vector $v_i \in \mathbf{V}$ is also included in set $\mathbf{V}'$, while any vector $v_i' \in \mathbf{V}'$ is also included in set $\mathbf{V}$.*

Proof. Suppose that $v_i \in \mathbf{V}$, and therefore it satisfies constraints (3.7)(3.8). It is apparent that: $p_{j,i} \leq p_j^{\max}$. Since, as shown earlier, constraints (3.7) and (3.8) can also be represented by (3.12) [31], v_i also satisfies (3.12). Using function (3.20), we can convert the rate vector $\{r_{1,i},r_{2,i},...,r_{B(k),i}\}$ into the corresponding power index vector $\{g_{1,i},g_{2,i},...,g_{B(k),i}\}$. Let $\psi = \sum_{j=1}^{B(k)} g_{j,i}$. For a feasible power and rate vector, with known ψ ($0 \leq \psi < 1$ [31]), we can find each user's power index capacity $\pi_j(h_j,\psi)$. Since v_i satisfies (3.12), based on proposition 3 and the definition of power index capacity, we conclude that $g_{j,i} \leq \pi_j(h_j,\psi)$. That means that the assigned powers and rates in v_i also satisfy the constraints (3.17),(3.18) and (3.19). Therefore, $v_i \in \mathbf{V}'$.

Let us consider vector $v_i' = \{p_{1,i}',p_{2,i}',...,p_{B(k),i}',r_{1,i}',r_{2,i}',...,r_{B(k),i}'\} \in \mathbf{V}'$. As before, the rate vector part can be converted to corresponding power index vector $\{g_{1,i}',g_{2,i}',...,g_{B(k),i}'\}$. Let $\psi = \sum_{j=1}^{B(k)} g_{j,i}'$ and hence $g_{j,i}' \leq \pi_j'(h_j,\psi)$ due to constraints (3.17,3.18,3.19). Note that for the case where $\psi' > \sum_{j=1}^{B(k)} g_{j,i}'$, $\pi_j'(h_j,\psi) \geq \pi_j'(h_j,\psi')$. Based on the previous discussion we can easily conclude that the power vector is feasible. Therefore,

$$\psi \leq 1 - \frac{\eta_0 W \cdot g_{j,i}'}{p_{j,i}' h_j}$$

which satisfies (3.12), for user j, $1 \leq j \leq B(k)$. Therefore, $v_i' \in \mathbf{V}$. ■

The above proposition shows that the optimal solution can also be obtained with the new constraints since they define the same solution set. Please note that, as mentioned before, the fairness constraints in the original problem are replaced by parameters $w_i's$. The choice of the proper values of $w_i's$ that maintain the fairness is discussed in detail later in this paper.

Among the new constraints, the right hand side of inequalities (3.17)(3.18) are not fixed values, but are functions of the selected target system load ψ. Hence whether or not the final solution is feasible also depends on the choice of ψ. For any value of $\psi \in [0, 1)$, there could be many feasible solutions among which one will be the optimal. Moreover, there must exist an optimal system load ψ^* that can achieve the overall best solution. It is natural to regard the objective Z as the function of system load ψ, $Z = F(\psi)$, and thus Z is the local optimal result at some specific ψ. The maximum Z is achieved when $\psi = \psi^*$. The ultimate objective of the proposed method is to find this optimal ψ^*, and the optimal power index assignment vector under it.

3.2.2 Optimal Solution for a Given System Load

Before obtaining the best system load, we first discuss how to find the local best solution. Assuming that the value of $\psi \in [0, 1)$ is known, the right hand side of (3.17)(3.18) can be determined. Combining the two constraints together, we can express the optimization problem (3.16) by replacing g_i with $\pi_i(h_i, \psi)x_i$, $0 \leq x_i \leq 1$, as follows:

$$\max \ Z = \sum_{i=1}^{B(k)} w_i f_r\left(\pi_i(h_i, \psi)x_i, \gamma_i\right) \tag{3.21}$$

$$s.t. \quad \sum_{i=1}^{B(k)} \pi_i(h_i, \psi)x_i \leq \psi, \ 0 \leq x_i \leq 1$$

Note that (3.21) is a non-linear continuous knapsack problem with the x_i taking continuous values between 0 and 1. In general solving this type of problem is proven

to be difficult or even impossible in some cases [38]. However proposition 2 limits the transmit power of a user i, to either $p_i^{\max}$ or 0 for the optimal solution. This condition provides a possible method to solve the above non-linear knapsack problem. Without loss of generality we suppose that the optimal solution is when the first K users transmit at their maximum power, $p_i = p_i^{\max}$, $1 \leq i \leq K$. The optimal system load is $\psi^* = \sum_{i=1}^{K} g_i$. The following theorem states that the power index of an individual user is equal to its power index capacity under ψ^*, that is: $g_i = \pi(h_i, \psi^*)$.

Theorem 2 *Let the optimal solution allow K users to transmit at their maximum power and the system achieves the system load ψ^*. The power index that an individual user received in this case is equal to its power index capacity, that is: $g_i = \pi(h_i, \psi^*)$.*

Proof. For those users whose transmit powers are zero, the corresponding power index capacities are also zero. Therefore, their power indices are zero as well. Without loss of generality, we assume that the K users under consideration are identified as follows: $1 \leq i \leq K$. Based on proposition 2, we have

$$\frac{h_i p_i^{\max} G_i}{\sum_{j=1, j\neq i}^{B(k)} h_j p_j + W\eta_0} = \gamma_i \quad for\ 1 \leq i \leq K \tag{3.22}$$

Performing some manipulations in these K equations, we have

$$\frac{h_i p_i^{\max}}{g_i}(1 - \sum_{i=1}^{K} g_i) = W\eta_0 \quad for\ 1 \leq i \leq K \tag{3.23}$$

Letting $\psi^* = \sum_{i=1}^{K} g_i$, we obtain g_i as

$$g_i = \frac{h_i p_i^{\max}(1 - \psi^*)}{W\eta_0} \tag{3.24}$$

From the definition of power index capacity, we find that $g_i = \pi(h_i, \psi^*)$. ■

With reference to the optimal solution of problem (3.21) we can prove the following theorem.

Theorem 3 *The optimal solution of the constrained optimization problem (3.21) can be obtained by solving the following linear 0-1 knapsack problem*

$$\max\ Z = \sum_{i=1}^{B(k)} w_i \frac{W}{\gamma_i} \frac{\pi_i(h_i,\psi)}{1-\pi_i(h_i,\psi)} x_i \tag{3.25}$$
$$s.t. \quad \sum_{i=1}^{B(k)} \pi_i(h_i,\psi) x_i \leq \psi,\ x_i = \{0,1\}$$

Proof. Since $f_r(x,\gamma_i) = \frac{W}{\gamma_i}\frac{x}{1-x}$ for user i, we present the objective function of (3.21) as follows:

$$\max\ Z' = \sum_{i=1}^{B(k)} w_i \frac{W}{\gamma_i} \frac{\pi_i(h_i,\psi)}{1-\pi_i(h_i,\psi)x_i} x_i \tag{3.26}$$

Based on proposition 2 we know that the optimal solution is achieved when the transmit power of a user i is either $p_i^{\max}$ or 0. According to Theorem 2, in terms of power index that means that users are assigned either their power index capacity or 0 for the chosen system load ψ. In the above relation (3.26), the solution for x_i is either 1 or 0. Therefore, we can modify (3.26) as follows without changing the final optimal solution.

$$\max\ Z = \sum_{i=1}^{B(k)} w_i \frac{W}{\gamma_i} \frac{\pi_i(h_i,\psi)}{1-\pi_i(h_i,\psi)} x_i \tag{3.27}$$

where $x_i = \{0,1\}$. ■

Instead of solving for the optimal solution of the above integer knapsack problem (3.25), which is in principle NP-hard, we utilize a greedy algorithm (GA) in order to obtain an approximate solution. Let Z_a denote the result achieved by the approximate solution, while Z and Z_c denote the corresponding results of the optimal solutions for the integer and continuous knapsack problems, respectively. It has been proven that $Z_a \leq Z \leq Z_c$ [39]. Furthermore let:

$$\alpha_i \triangleq \frac{W}{\gamma_i\left(1-\pi_i(h_i,\psi)\right)}$$

which is a constant value for an individual user. Lets further suppose that all backlogged users are sorted in descending order according to $w_i(k)\alpha_i$, i.e. $w_i(k)\alpha_i \geq w_j(k)\alpha_j$, for $i < j$. If it is not the case, these values can be sorted in $O(n \log n)$ time through an efficient procedure. Thus the optimal continuous solution of problem (3.25) is given by

$$\begin{aligned} x_i &= 1, \text{ for } i < s \\ x_j &= 0, \text{ for } j > s \\ x_s &= \frac{\psi - \sum_{i<s} \pi_i(h_i, \psi)}{\pi_s(h_s, \psi)} \end{aligned} \tag{3.28}$$

An algorithm that finds the critical point s within $O(n)$ time in a system with n users, is provided in [39]. Based on solution (3.28), the greedy algorithm (GA) obtains the approximate solution U as follows:

$$U = \max\{U_1, U_2\} \tag{3.29}$$

where

$$U_1 = \begin{cases} x_i = 1, \text{ for } i < s \\ x_j = 0, \text{ for } j \geq s \end{cases}, \; U_2 = \begin{cases} x_i = 1, \text{ for } i = s \\ x_j = 0, \text{ for } i \neq s \end{cases}$$

It has been shown in [39] that in worst case $\frac{Z_a}{Z} = \frac{1}{2}$. Let Z represent the result that corresponds to the integer solution of (3.27) when ψ is assigned a value from $[0, 1)$, and Z^* be the result when $\psi = \psi^*$. From the definition of ψ^* we know that Z^* is the maximum value among all Z, i.e. $Z^* = \max_{\psi}\{Z\}$. Based on proposition 2 and the analysis in the previous subsection, it is easy to find that $\psi^* = \sum_i \pi_i(h_i, \psi^*)x_i, x_i = \{0, 1\}$. Therefore, when the optimal system power index ψ^* is chosen, $Z_a = Z = Z_c = Z^*$. Since $Z_a \leq Z \leq Z^*$ and the equality $Z_a = Z^*$ holds only when $\psi = \psi^*$, the simple GA can be used to obtain the optimal solution.

3.2.3 Optimal System Load

As we discussed in the last subsection the optimal solution of problem (3.21) depends on the selected system load ψ. Relation (3.15) shows that the power index capacity increases as ψ decreases. At the first point when $\pi_i = \psi$, the power index capacity reaches its largest value and then it decreases linearly following the value of ψ. Although a smaller value of ψ may increase the single user's power index capacity at some range, the finally achieved objective function could be low due to the small system load ψ. On the other hand, setting large ψ reduces the individual user's power index capacity as (3.15) indicates. The consequence of smaller power index capacity is that more users are required to share ψ, and probably a small objective function should be used due to the convexity of function $f_r(x, \gamma_i)$ that converts the power index to throughput. Therefore whether or not the objective function reaches its maximum value depends not only on the value of the system load ψ, but also on how it is shared among the candidate users. There must exist an optimal value of system load ψ^* that can achieve the maximum weighted rate.

Let the power index vector $\mathbf{g}$ denote the optimal solution, which can be found through the method described in the previous section for a given specific value of ψ. Apparently $\mathbf{g}$ is a function of ψ. The objective function (3.16) is the sum of individual weighted rates that are obtained from $\mathbf{g}$ using function $f_r(x, \gamma_i)$. Therefore Z can also be regarded as a function of ψ. Let $FZ(\psi)$ be the function that gives the maximum value of the sum of weighted rates at ψ. Then the original optimization problem can be rewritten as follows:

$$\begin{aligned} \max \ Z &= FZ(\psi) \\ s.t. \ 0 &\leq \psi < 1 \end{aligned} \tag{3.30}$$

The optimal solution ψ^* of the above problem and its corresponding power index assignment by (3.28) with $\psi = \psi^*$, provides the final optimal solution of (3.16).

Problem (3.30) is a simple unconstrained maximization problem that searches for the maximum Z within the interval $[0, 1)$. The disadvantage of (3.30) is that it does not have

an explicit expression. Hence algorithms that rely on the first or second order derivatives will not be applicable in this case. Therefore the searching process depends on the result of (3.28). Note that every time when a new value of ψ is chosen, the order of $w_i(k)\alpha_i$ will be different from that of previous ψ. The time of calculating the best result for a newly chosen ψ, including the time of re-ordering the users (if needed), is easily obtained as $O(n \log n) + O(n) = O(n \log n)$ if n is assumed to be large enough. Moreover, there are many possible local maximum points within the range $0 \leq \psi < 1$. The final optimal ψ must be a global best value. Although in [40] many searching algorithms on how to locate the minimum/maximum solution within a range are described, to make these algorithms effective there must be only one extreme point in the specified range. However in general it is not possible to know the range which contains only the global optimal value. Thus an exhaustive search within $[0, 1)$ would be needed. However the following proposition provides a lower bound ψ^0 with respect to the searching range instead of 0, in order to restrict the corresponding feasible searching range.

Proposition 5 *The lower bound of the feasible searching range is given by*

$$\psi^0 = \min_{1 \leq i \leq B(k)} \left(\frac{\zeta_i}{1 + \zeta_i} \right), \text{ where } \zeta_i \triangleq p_i^{\max} h_i / \eta_0 W$$

Proof. : With the decrease of the target system load ψ, the individual's power index, provided by (3.13), will keep increasing till ψ reaches the point ψ_i for user i, that is $(1 - \psi_i)\varsigma_i = \psi_i$. With respect to user i, if $\psi \leq \psi_i$ its power index $\pi_i(h_i, \psi) = \psi$. ψ_i is given by $\psi_i = \varsigma_i/(1 + \varsigma_i)$, which varies with different users since their ς_i are not likely the same. Let ψ^0 be the minimum among all ψ_i's. Once $\psi < \psi^0$ all backlogged users will have the same power index capacities $\pi_i(h_i, \psi) = \psi$. Define a small increment $\Delta\psi$ and let $\psi' = \psi + \Delta\psi < \psi^0$. Apparently for all users their power indices will all have small increment $\Delta\psi$, such that $\pi_i(h_i, \psi') = \psi + \Delta\psi$. Maintaining the previous power index assignment and giving $\Delta\psi$ to any backlogged user will help increase the objective function

(3.16). We hence can keep adding $\Delta\psi$ to ψ till it reaches $\psi^0 = \Delta\psi + \psi$, which proves this proposition. ■

Since the optimal ψ can reside between ψ^0 and 1, we need to calculate a series of sample values after every interval $\Delta\psi$. Apparently the smaller the $\Delta\psi$ the more samples we get and thereby the more accurate is the obtained result. On the other hand, it also increases the required computational time and power.

3.2.4 Fairness Conditions

As mentioned before, the fairness is controlled by the vector $\mathbf{w} = \{w_1, w_2, ..., w_{B(k)}\}$. When changing the values of w_i, we are actually pursuing a set of optimal fixed values $\mathbf{w}^*= \{w_1^*, w_2^*, ..., w_{B(k)}^*\}$ that balance the rate of users with varying channel conditions and hence keep the fairness. Since we do not know in advance the exact distribution of the channel conditions, and the number of users may also change, it is difficult to obtain vector $\mathbf{w}^*$ in advance. Therefore, a real time algorithm is required that is capable of converging w_i towards w_i^*, while maintaining the asymptotic fairness. Stochastic approximation algorithm has been proven to be effective in estimating such parameters. Note that this algorithm has been implemented in [21] [37] in order to solve similar problems. Generally the stochastic approximation algorithm is a recursive procedure for finding the root of a real-value function $f(x)$. In many practical cases the form of function $f(x)$ is unknown. Therefore the result with the input variable x cannot be obtained directly. Instead the observations of the results, sometimes with noise, will be taken. It has been proven that the root of $f(x)$ can be estimated with the observation $Y_n = f(x_n)$ by the following procedure,

$$x_{n+1} = x_n - \epsilon_n Y_n$$

where $\epsilon_n > 0$, $\epsilon_n \to 0$. We can simply let $\epsilon_n = 1/n$. In most situations the value of $f(x_n)$ may not be directly available, but instead the $f(x_n) + e_n$, where e_n is the observation

noise. In this case, the above approximation approach still applies, with the observed value replaced by $Y_n = f(x_n) + e_n$. The convergence of x_n to the root requires $E(e_n) = 0$.

Here we define our function $f(\mathbf{w}) = \{f(w_1), f(w_2), ..., f(w_{B(k)})\}$ as follows:

$$f(w_i) = \frac{E[r_i(n)]}{E\left[\sum_j r_j(n)\right]} - \frac{\phi_i}{\sum_j \phi_j} \tag{3.31}$$

whose root w_i^* will make $f(w_i) = 0$ which satisfies the fairness condition (3.3). The noise observation Y_n in our case is:

$$Y_n = \frac{r_i(n)}{E\left[\sum_j r_j(n)\right]} - \frac{\phi_i}{\sum_j \phi_j} \tag{3.32}$$

It is easy to prove that the mean of noise $E[e_n] = E[f(w_i) - Y_n] = 0$. Therefore, the value of w_i^* is then recursively obtained by

$$w_i(n+1) = w_i(n) - Y_n/n$$

However, Y_n need to know the mean of total system throughput $E\left[\sum_j r_j(n)\right]$. We use a smoothed value $\overline{R}(n)$ to approximate $E\left[\sum_j r_j(n)\right]$ and update $\overline{R}(n)$ as follows:

$$\overline{R}(n) = \overline{R}(n-1)\beta + (1-\beta)\sum_j r_j(n-1) \tag{3.33}$$

where β is the smooth factor which determines how the estimated $\overline{R}(n)$ follows the change of actual achieved system throughput. In the remaining of the paper, throughout the performance evaluation of our approach, the value $\beta = 0.999$ is chosen. The numerical results presented in subsections 3.3.2 and 3.3.2, with respect to the convergence of w_i's and the achievable fairness, demonstrate that such a method is very effective in approximating the optimal values of w_i^* and therefore controlling and maintaining the fairness.

3.3 Performance Evaluation

In this section we evaluate the performance of the proposed method in terms of the achievable fairness and throughput, via modeling and simulation. Furthermore to better understand the performance of the proposed scheduling algorithm - in the following we refer to as MAX-FAIR (Throughput Maximization and Fair Scheduling) - we compare it to the Maximum Throughput (MAX) scheme [26], which achieves the maximum total uplink throughput by allowing only the best k users in terms of their received power to transmit, and the HDR algorithm [15] [41], which is a downlink single user scheduling algorithm. In the MAX scheme parameter k is determined by iteratively comparing the throughput of best i users, $1 \leq i \leq N$, where N is the total number of users. The throughput achieved by MAX scheme is regarded as the upper bound throughput in the uplink CDMA scheduling. On the other hand, since HDR achieves temporal fairness, we consider it here to mainly observe the difference between temporal fairness and throughput fairness and their corresponding advantages in specific cases.

3.3.1 Model and Assumptions

Throughout our numerical study we consider a single cell DS-CDMA multi-rate system with multiple active users. All active users are continuously backlogged during the simulation and generate packets with average size of 320 bytes. The maximum transmission power is the same for all users, i.e. $p_i^{\max} = 2W$, while the system chip rate is $W = 1.2288 \times 10^6 chip/\text{sec}$ and the required SINR is $\gamma_i = 8\text{dB}$, same for all users. The transmission time is divided into $1ms$ equal length slots, while the simulation lasts for 1.7×10^5 slots.

To study the impact of the channel condition variations on the system throughput and fairness performance, we model the channels through an 8-state Markov Rayleigh fading channel model [34]. According to this model the channel has equal steady-state probabilities of being in any of the eight states. We also assume that the coherent time is much larger than the length of a time-slot, hence the channel state is assumed to be constant

	s=1	s=2	s=3	s=4
$p_{s,s}$	0.9304	0.8419	0.8170	0.8216
$p_{s,s-1}$	0	0.069	0.0879	0.0894
$p_{s,s+1}$	0.0696	0.0891	0.0951	0.089
	s=5	**s=6**	**s=7**	**s=8**
$p_{s,s}$	0.8349	0.8590	0.8945	0.9616
$p_{s,s-1}$	0.0876	0.0777	0.0637	0.0384
$p_{s,s+1}$	0.0775	0.0633	0.0418	0

Table 3.1 Channel State Transition Probability

within a time-slot. At the beginning of each time-slot, the channel model decides to transit to a new state, which can only be itself or one of its neighbor states, i.e. from state s to s, $s+1$ or $s-1$. Table 3.1 summarizes the state transition probabilities for all the eight states.

Furthermore four different cases with respect to the ranges of the average SNRs that are assigned to the various users are considered. Specifically, Table 3.2 presents the corresponding ranges, and lists the assignment of the average SNRs for each user for a seven user scenario, under all these cases. The four different cases represent four different scenarios with respect to the SNR as follows (from top to bottom): large SNR range with low SNR users, low SNR, middle SNR and high SNR. In the next subsection we evaluate the performance of MAX-FAIR, MAX and HDR methods under all four cases and compare their corresponding achieved throughput and fairness.

In most of the numerical results presented in the next subsection, unless otherwise is explicitly indicated, all users are assumed to have the same weight. Such a scenario allows us to better understand and compare the achievable performances of the various scheduling schemes, when users have different channel conditions. However the operation and effectiveness of the proposed MAX-FAIR policy is also demonstrated in an environment where users present different weights.

	1	2	3	4	5	6	7
case: $[-3,3]$	-3	-3	-3	0	0	0	3
case: $[-4,-2]$	-4	-4	-4	-3	-3	-3	-2
case: $[0,1]$	0	0	0	1	1	1	1
case: $[2,4]$	2	2	2	3	3	3	4

Table 3.2 Simulation Cases With Different SNR(dB) Distribution

3.3.2 Numerical Results and Discussion

The numerical results presented in subsections 3.3.2 and 3.3.2 refer mainly to the impact of some of the parameters associated with the proposed MAX-FAIR algorithm on its operation and achieveable performance, and allow us to obtain a better understanding of its operational characteristics and properties. Then in subsections 3.3.2 and 3.3.2 comparative results about the achievable throughput and fairness of the MAX-FAIR, MAX and HDR algorithms, are presented.

Finite System Power Index Samples

Fig. 3.1 shows the sensitivity of the weighted throughput achieved by the MAX-FAIR algorithm as a function of the number of samples used to obtain these values. The last point in the horizontal axis correspond to the optimal value. Moreover the different curves provided in this figure correspond to different combinations of the SNR ranges and the number of active users. As can be seen, the more samples we choose, the closer is the obtained maximum value to the optimal value, which clearly presents the tradeoff between the accuracy and the required computational power, as discussed before in subsection 3.2.3. For instance we observe that in the cases with small SNR range (e.g. [0,1]dB), even 20 samples are sufficient to get satisfactory results, while for the cases with larger SNR range, (e.g. [-3,3]dB), more samples may be required.

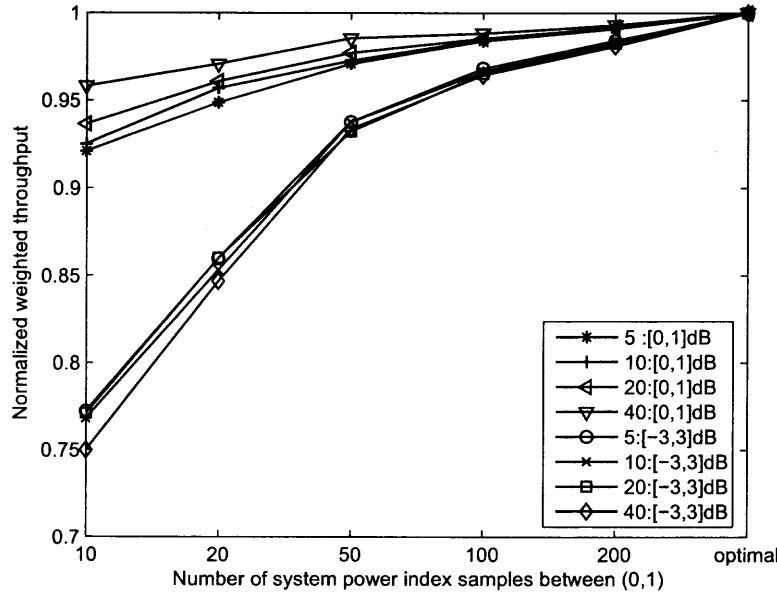

Figure 3.1 The impact of number of samples on the weighted throughput (MAX-FAIR)

Furthermore, as it can be observed from this figure, for the case of [0,1]dB, the larger the number of active users in the system the less sensitive is the achievable maximum result to the number of samples (i.e. the slope of the corresponding curve becomes smoother as the number of active users increases). On the other hand, when there are users with high SNR values (e.g. [-3,3]dB), the increasing number of active users makes the achieved throughput drop slightly for small number of samples. This difference in the system behavior is closely related to different number of simultaneously served users, under different SNR ranges and channel conditions, as depicted by the different observed service patterns in Fig. 3.2.

Specifically, in Fig.3.2, we present the probabilities of the number of simultaneously served users in each scheduling cycle. For this experiment we consider 40 backlogged users in the system and perform 200 trials. In each trial, users are randomly assigned the SNRs in the designated SNR range. We observe that when there are users having high SNR values, e.g. in the cases of [-3, 3]dB and [2,4]dB, only a small number of users (at

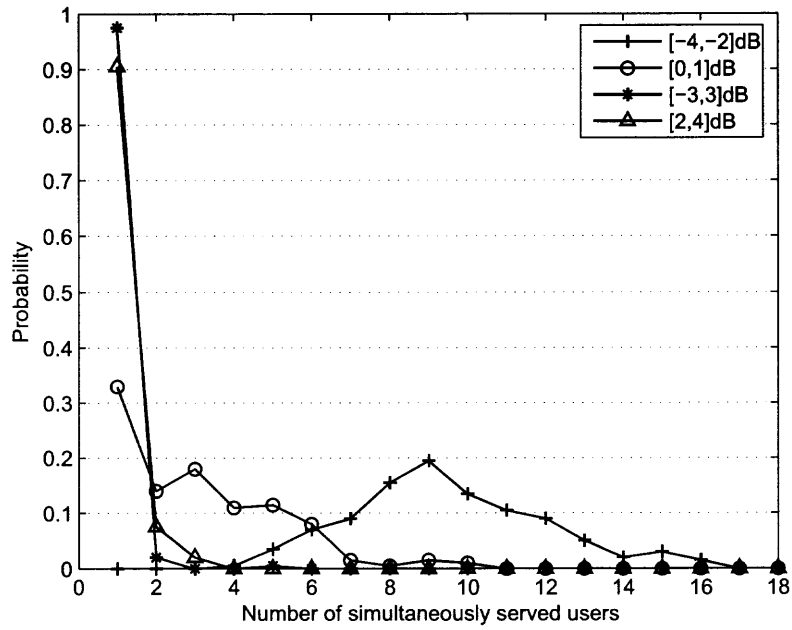

Figure 3.2 The service pattern under different channel conditions (i.e. SNRs) (MAX-FAIR)

most 2 in this experiment), are served concurrently. However in the case that all users have small SNR values, e.g. in the case of [-4,-2]dB, the number of simultaneously served users increases significantly (it is distributed between 4 and 17 in our case as can be seen by Fig.3.2). Such user distribution indicates that in the case that a single user can not consume all the system resources (e.g. the case where users have low SNR values), more users will be scheduled simultaneously in order to achieve a more efficient resource utilization and as a result increase the total system throughput. This also demonstrates the advantage of our proposed scheduling algorithm over the one-by-one scheduling algorithms that have been proposed in the literature. As a result, with respect to Fig. 3.1, for the case of [0,1]dB, multiple users are scheduled to reach the maximal throughput. Increasing the number of active users enables the system to schedule more available candidates to achieve higher throughput, and therefore the achievable result is less sensitive to the number of samples. However, for the case [-3, 3]dB at most only 1 or 2 users are scheduled for simultaneous

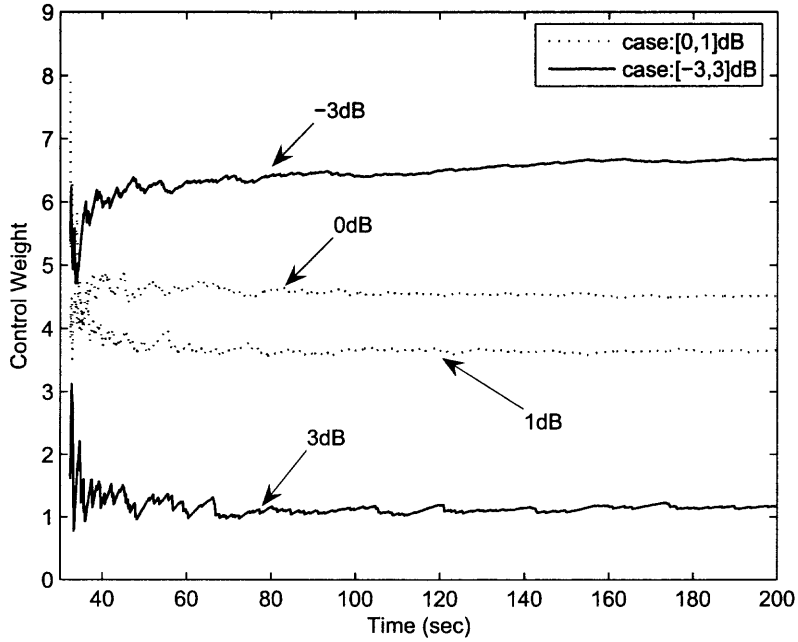

Figure 3.3 The convergence of w_i's for different users and different SNR ranges (MAX-FAIR)

transmission. In the following experiments and numerical results we adopt the accuracy of 100 samples, which is sufficient to reach 95% of the optimal weighted throughput.

Parameter Convergence by Stochastic Approximation

As described in sections 3.1.1 and 3.2.4 parameters $w_i's$ are used to represent the fairness constraints in our optimization problem formulation. Fig. 3.3 shows the dynamic change of parameters w_i's as the system and time evolves, for two different cases that correspond to two different SNR ranges. For each such case the corresponding values of two users - one user with strong channel and one user with weak channel - are presented. As mentioned before, all the users are assigned the same weight in order to more clearly demonstrate the influence of the channel conditions on w_i's. It can be seen by this figure that the converged values of w_i's has the effect of compensating users with the weak channels and reducing the priority of users with strong channels in the scheduling policy.

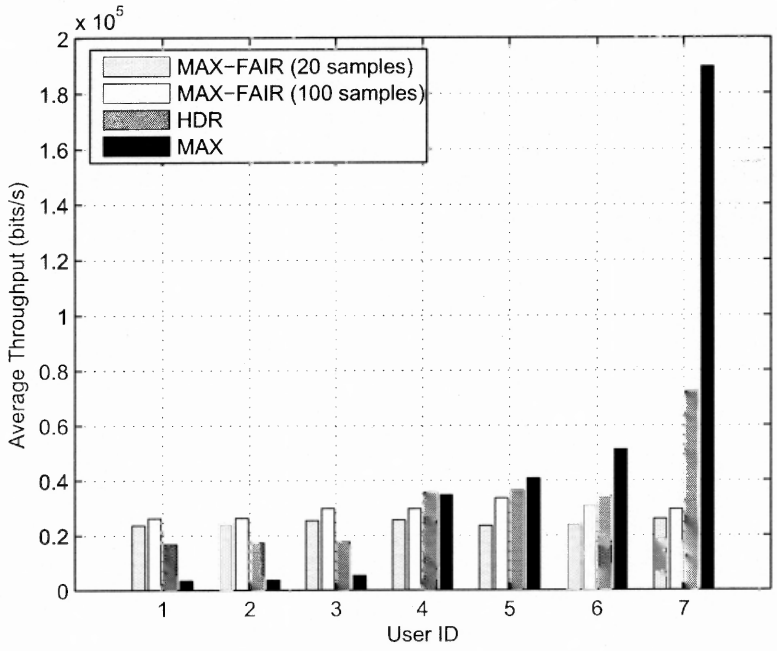

Figure 3.4 Average throughput for the [-3,3]dB case

In fact the converged values of w_i's will make both users (weak and strong) to gain proper system resources and therefore achieve fair throughput. Please note that it is the relative values of w_i's that control the priority of accessing the system resources, and not their absolute values. Furthermore it should be noted that the lower the average SNR of a weak user, the larger the gap between the weak user and a strong user, which has negative impact on the achievable system throughput, as we will see in the following subsection.

Throughput and Fairness Performance

Fig. 3.4 shows the average throughputs of all the users under the MAX-FAIR, MAX and HDR methods, for a seven user scenario where the average SNR range is [-3,3]dB and the corresponding average SNR assignments to the seven users are as shown in Table 3.2. In order to better demonstrate the tradeoff between the computational complexity and the achievable throughput of MAX-FAIR approach, we obtained the corresponding results under two different cases with respect to the number of power index samples (i.e. 20 and

100 samples). As observed in this figure the MAX-FAIR with 100 power index samples achieves slightly higher throughput, however it requires five times the computational power of the MAX-FAIR with 20 power index samples.

When compared to other two scheduling schemes, MAX-FAIR presents the best throughput-fairness performance (balances the achievable throughput of all users) despite the variable channel conditions of the different users, which indicates that the fairness is well maintained under the proposed scheduling algorithm. As mentioned before in the paper, the main objective of HDR is to achieve temporal fairness. Therefore, under HDR scheduling each user's throughput is closely related to its channel conditions. That is why in figure 3.4 we observe that users 1, 2 and 3 have smaller throughput than users 4, 5 and 6, while user 7 has the largest throughput under the HDR scheme. Under the MAX algorithm, user 7 consumes most of the system resources and achieves much higher throughput than the rest of the users due to the fact that the objective of MAX algorithm is to achieve the highest possible total system throughput, without however considering the fairness issue. In fig. 3.5 we further measure and evaluate the fairness performance by the standard deviation of the average throughput under all the four different SNR cases. Among the three algorithms, MAX-FAIR algorithm has the smallest deviation for all the different cases under consideration, while the corresonding values change only slightly from case to case. We also find that in general the standard deviation increases as the SNRs become higher. This happens because small fluctuation of w_i results in larger throughput change, if all the users have higher SNR levels.

Fig. 3.6 compares the corresponding average system throughputs of the three algorithms under evaluation, for the different SNR ranges (cases). As we expected, MAX-FAIR outperforms HDR in most cases due to the simultaneous scheduling of multiple users, as has been demonstrated in Fig. 3.2, and consequently results in higher resource utilization. However in the case of SNR range of [-3,3]dB, MAX-FAIR achieves slightly lower throughput than the HDR. The reason of that resides in the different fairness criterions

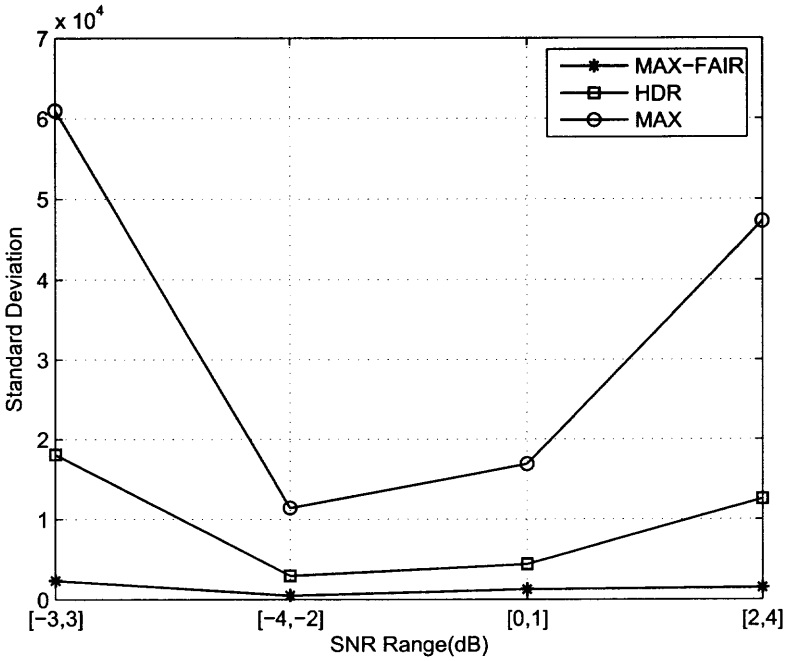

Figure 3.5 Standard deviation of achievable average throughputs

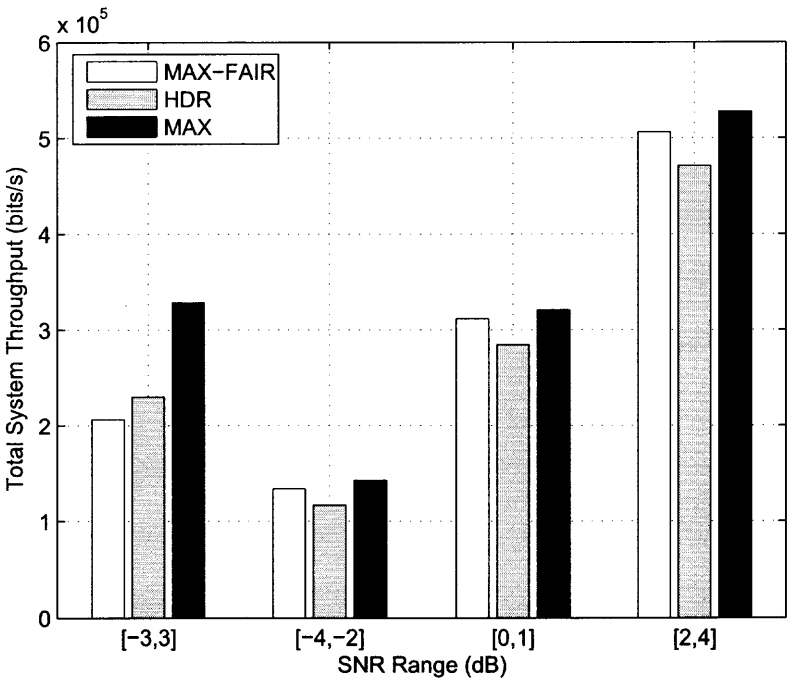

Figure 3.6 Achieved system throughput under different SNR ranges

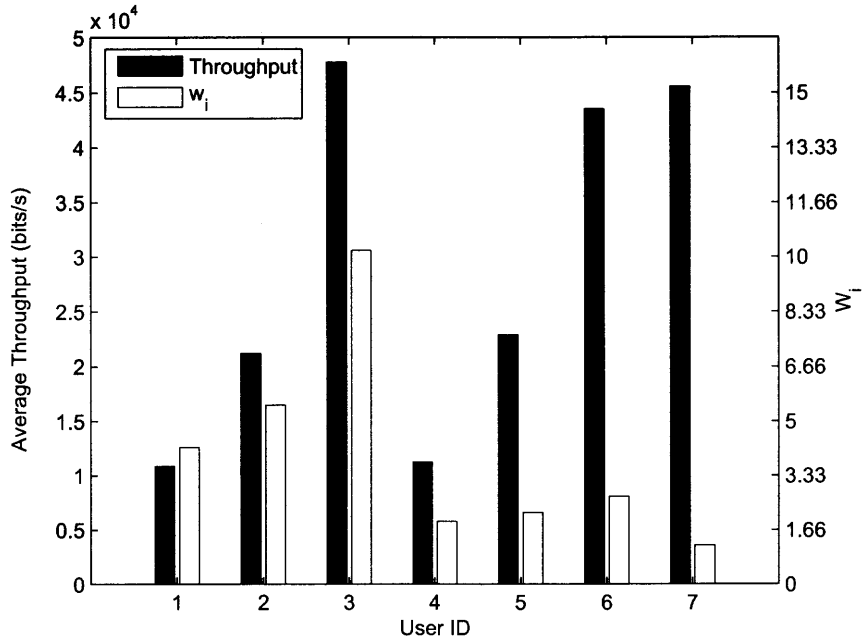

Figure 3.7 Average throughput under different QoS requirements (weights) by MAX-FAIR

considered and satisfied in these two algorithms, namely the throughput fairness and temporal fairness. If we examine again Fig. 3.3, we notice that users that have low average SNR (-3dB) (e.g. users 1,2 and 3) finally converge to a high w_i, which enables them to have equal opportunity to transmit under the MAX-FAIR scheduling policy. Due to their weak channel conditions, their average throughputs will be low and hence the total system throughput will become lower because of the satisfaction of the throughput fairness constraint. However, as explained before since access time is not the only resource to be shared among the users in these systems, considering throughput fairness instead of temporal fairness is more meaningful in these systems and environments, despite the slightly lower total throughput that can be achieved in some cases under this consideration. One possible alternative solution is to relax the fairness constraint if the QoS permits it. Our experiments have demonstrated that after relaxing the fairness to 85% of its original requirement, the MAX-FAIR catches up and outperforms the HDR.

In order to obtain a more in-depth understanding of the MAX-FAIR fairness operation, in the following fig. 3.7 we present the achieved average throughputs for all the seven users under MAX-FAIR scheme, for a scenario where the SNR range is assumed to be [-3, 3]dB and the users are assigned different weights. The different weights can be considered as the mapping of different QoS requirements. In this scenario, users 1 and 4 have weight 1, users 2 and 5 have weight 2, while users 3, 6 and 7 have weight 4. Fig. 3.7 demonstrates that the MAX-FAIR successfully schedules the transmissions and distributes the resources so that the various users achieve throughput according to their corresponding assigned weights. Specifically users with weight 2 and 4 obtain respectively two times and four times the throughput achieved by users with weight 1. In this figure we also present (on the right hand side vertical axis) the converged values of parameters w_i's. Here the different values of w_i's reflect both the channel condition variations and the weight differences. Please note that the relationship between w_i and weight is not linear due to the nonlinearity between the allocated resources and throughput.

Number of Users

Fig. 3.8 shows the achieved total system throughput under MAX and MAX-FAIR algorithms as a function of the number of backlogged users, for the case where the users' SNRs are located within [0,1]dB range. Please note that as mentioned before MAX algorithm provides the maximum uplink transmission throughput without considering the fairness property, and therefore is assumed to provide the upper bound throughput in uplink scheduling. From this figure we can clearly observe the great advantage of the proposed MAX-FAIR approach and its ability to achieve very high throughput, while still maintaining the fairness. When the number of backlogged users reaches a certain level, e.g. 35 in this experiment, the throughput becomes flat for both MAX-FAIR and MAX which means that the chances of improving the throughput by opportunistic scheduling with multiple users have been fully utilized.

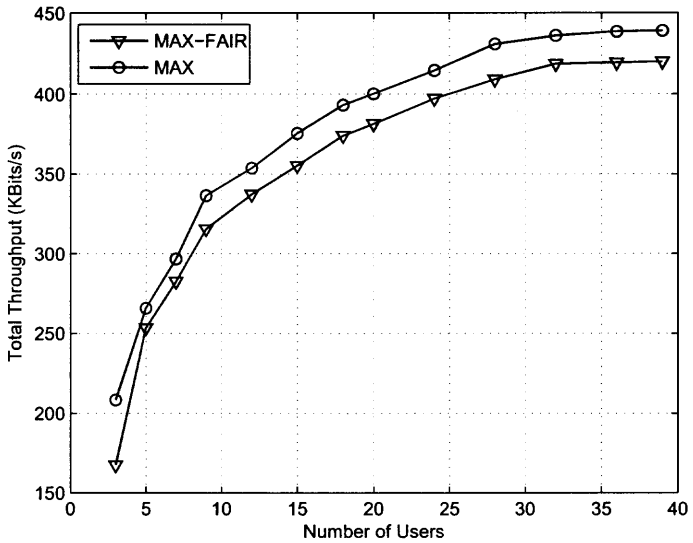

Figure 3.8 System throughput as a function of the number of backlogged users

3.4 Conclusion

In this chapter the CDMA uplink throughput maximization problem, while maintaining the throughput fairness among the various users was considered. It was shown that such a problem can be expressed as a weighted throughput maximization problem, under certain power and QoS requirements, where the weights are the control parameters that reflect the fairness constraints. A stochastic approximation method was presented in order to effectively identify the required control parameters. The numerical results presented in the paper with respect to the convergence of the the control parameters and the achievable fairness, demonstrated that this method is very effective in approximating the optimal values and therefore controlling and maintaining the fairness. Furthermore the concept of power index capacity was used to represent all the corresponding constraints by the users' power index capacities at some certain system power index. Based on this, the optimization problem

under consideration was converted into a binary knapsack problem, where the optimal solution can be obtained through a global search within a specific range.

The performance of the proposed policy in terms of the achievable fairness and throughput were obtained via modeling and simulation and were compared with the performances of other scheduling algorithms. The corresponding results revealed the advantages of the proposed policy over other existing scheduling schemes, and demonstrated that it achieves very high throughput, while satisfies the QoS requirements and maintains the fairness among the users, under different channel conditions and requirements.

CHAPTER 4

OPPORTUNISTIC SCHEDULING WITH SHORT-TERM FAIRNESS

In this chapter we propose a new Credit-based Short-term Fair Scheduling (CSFS) algorithm, which exploits the wireless channel variations to obtain high throughput via opportunistic scheduling, while at the same time introduces into the service scheduling policy the information of service interval, in order to provide more flexibility on the control of fairness in systems that support multiple classes of users with different requirements.

The combined problem of maintaining long-term fairness and still reaching optimal throughput has been addressed in the previous chapter using opportunistic scheduling policies that exploit the wireless channel variations. Specifically, the throughput performance improvement is obtained by utilizing the multi-user diversity effect in wireless communications. Hence there is probability that some users that keep having relatively bad channel conditions, may be prevented from being selected to receive proper service. However, some types of traffic demand certain amount of service within specific short span of time in order to avoid service delays. This objective is referred as short-term fairness. The optimal throughput by opportunistic scheduling can only provide the long-term fairness property. The high throughput of opportunistic scheduling is achieved by sacrificing the short-term fairness, i.e. it delays the transmission of users with bad channel condition temporarily. Although those users who lose service will be eventually compensated when their channels improve, their services within a short time interval are not guaranteed due to the randomly time-varying wireless channels. Such short-term unfairness in low level may result in the timeout of higher layer protocols and may cause the consequent system and service performance degradation.

Since the optimal solution to the transmission scheduling problem satisfying the short-term fairness constraints, is in principle difficult to be addressed analytically, in [42] the authors analyzed only some special cases and proposed a heuristic scheduling policy

that tries to satisfy the strict short-term fairness. However this approach is not suitable for traffic with multiple and different short-term fairness requirements. WCFQ (Wireless Credit-based Fair Queueing) opportunistic policy which is based on the CBFQ (Credit-based fair queueing) [43] approach of wireline fair scheduling, has been proposed in [37]. By mapping the channel conditions into a cost function, WCFQ trade-offs the fairness and throughput. It has been shown in [37] that this scheduler can provide temporal fairness with statistical fairness bound.

In this chapter we propose and evaluate a new algorithm - in the following we refer to as CSFS (Credit-based Short-term Fairness Scheduling) algorithm - which achieves to provide short-term fairness to the delay-sensitive users, while still schedules opportunistically the non-delay-sensitive users to obtain high system throughput. The remaining of this chapter is organized as follows. In section 4.1 we first present the system model and assumptions used throughout this chapter, and then we provide an overview of the WCFQ algorithm which is the basis for our proposed CSFS algorithm. In section 4.2 a detailed description of the CSFS algorithm is provided and its fairness property is discussed. Section 4.3 contains the performance evaluation of the proposed CSFS scheduling scheme and demonstrates the advantages in the scheduling flexibility that can be achieved by the proposed approach, especially in wireless systems that support users with different quality of service requirements. Finally section 4.4 concludes the chapter.

4.1 System Model

4.1.1 System Model

In the following we assume that the wireless channel is divided into equal length slots and shared by all active users. All users may access the channel in a time-division multiple access manner. In each time slot only one user is served. We also assume that the base station (BS) has perfect information about each user's channel condition, i.e. the feasible transmit rate to this user. The scheduler is assumed to reside in the BS and hence the BS can

Term	Definition
f_t	flow that receives service in slot t
$\alpha_i(t_1, t_2)$	service received by user i from slot t_1 to slot t_2
$K_i(t)$	credit for user i at the end of slot t
$U_i(t)$	cost for user i to transmit at slot t

Table 4.1 Summary of Notation

make decisions on which user should receive service at the current time slot. The wireless channel condition, which is affected by fast fading, shadow fading, and long-time-scale variation, has multiple states, and at a given time slot the channel can be in any of the states with the respective feasible transmit rates. The channel condition is assumed not to change within the duration of a time slot. Here we only consider the case of fixed packet length L and therefore the packet length has the same meaning as the slot length, which represents the service (access time) to users. Each user i is associated with some pre-assigned weight ϕ_i according to its Quality of Service (QoS) requirements, and $\sum_i \phi_i = 1$. The notations used throughout the remaining of this chapter are summarized in Table 4.1.

4.1.2 The Wireless Credit-based Fair Queueing Scheduling Algorithm

As explained earlier Users contending for the wireless medium will have different costs of transmission depending on their current channel condition. WCFQ provides a mechanism to exploit inherent variations in channel conditions and select low-cost users in order to increase the system's overall performance.

The following Table 4.2 presents the credit update rule according to the WCFQ scheduling algorithm.

Considering this credit update rule and the following scheduling policy $Q(t)$ [37]:

$$Q(t) = \arg\min_i \frac{L - K_i(t) + U_i(t)}{\phi_i} \tag{4.1}$$

for $(i = 1; i \leq N; i++)$

if $(i \neq f_{t+1})$

$$K_i(t+1) = K_i(t) + \max\left(\frac{L - K_{f_{t+1}}(p)}{\phi_{f_{t+1}}}, 0\right)\phi_i$$

else

$$K_i(t+1) = \max(0, K_i(t) - L)$$

end if

end for

Table 4.2 Credit Update Rule

where $Q(t)$ denotes the user selected for receiving service at slot t as shown in (4.1), WCFQ guarantees the statistical fairness bound as [37]:

$$\begin{aligned} & \Pr\left(\left|\frac{\alpha_i(t_1,t_2)}{\phi_i} - \frac{\alpha_j(t_1,t_2)}{\phi_j}\right| \geq \frac{L+x}{\phi_i} + \frac{L+x}{\phi_j}\right) \\ \leq\ & \Pr\left(\frac{U_i}{\phi_i} + \frac{U_j}{\phi_j} \geq \frac{x}{\phi_i} + \frac{x}{\phi_j}\right) \end{aligned} \tag{4.2}$$

where the cost function $U_i(t)$ is assumed to be independent identically distributed in each time slot t. Therefore, given the required statistical fairness bound $g(x)$ as

$$\Pr\left(\left|\frac{\alpha_i(t_1,t_2)}{\phi_i} - \frac{\alpha_j(t_1,t_2)}{\phi_j}\right| \geq \frac{L+x}{\phi_i} + \frac{L+x}{\phi_j}\right) \leq g(x)$$

it can be mapped to $U_i(t)$ as follows:

$$\begin{aligned} \Pr\left(\frac{U_i}{\phi_i} + \frac{U_j}{\phi_j} \geq \frac{x}{\phi_i} + \frac{x}{\phi_j}\right) &\leq g(x) \\ \Pr(U_i \leq x) &\geq \sqrt{1 - g(x)} \end{aligned}$$

The larger $U_i(t)$ that is chosen, the more opportunistic this scheduling policy would be, and as a result higher throughput could be achieved. However on the other hand smaller $U_i(t)$

makes the scheduler more fair. An extreme case is when $U_i(t) = 0$, which corresponds to the CBFQ fair scheduling policy [43].

It should be noted here that in most of the cases, the statistical fairness bound (4.2) may not be proper and/or accurate. To demonstrate this let us consider the case where only statistic fairness with intervals $(t_1, t_2) \geq M$ slots is assumed, that is:

$$\Pr\left(\left|\frac{\alpha_i(t_1,t_2)}{\phi_i} - \frac{\alpha_j(t_1,t_2)}{\phi_j}\right| \geq \frac{L+x}{\phi_i} + \frac{L+x}{\phi_j}\right) \tag{4.3}$$

$$where\ t_2 - t_1 + 1 \geq M$$

Such situations may occur in many delay non-sensitive applications that present only loose delay requirements, e.g. the timer of TCP connections. In such applications the unfairness in time range less than M slots is not important and hence the opportunistic scheduling policy can take advantage of it to improve the system throughput. Apparently, in opportunistic scheduling the larger the value of M, the smaller the probability of (4.3) will be. However, in WCFQ the cost function lacks the ability of reflecting various intervals M. Once the statistic fairness bound is determined, all the above cases will be treated equally. In fact, all fairness bounds presented in the literature have no restriction on the interval time (t_1, t_2). A possible solution for WCFQ is to make the right side of (4.2), i.e. $\frac{L+x}{\phi_i} + \frac{L+x}{\phi_j}$ to fit the M-slot short-term fairness requirement. However, such method is still unable to handle the situation where the users have different short-term fairness requirements.

However, with larger M we can utilize larger cost function in (4.2), which favors the channel condition in opportunistic scheduling and achieves higher throughput. At the same time we can still satisfy the statistic fairness bound if the new cost function is properly chosen. It should be noted here that although the weight ϕ_i determines the share of service a user may receive, and has some influence on the fairness bound within certain time interval, its ability of controlling fairness is limited under opportunistic scheduling policy.

In this chapter, we do not intend to give an explicit short-term fairness bound as in [42]. Our main objective is to bring the information of service interval (t_1, t_2) into the scheduling policy which provides more flexibility on the control of fairness and achieve higher system throughput than the WCFQ method. The new proposed approach is based on the WCFQ which enables us to inherit its flexibility between the tradeoffs of fairness and throughput.

4.2 Credit-based Short-term Fairness Scheduling

In this section, using the WCFQ and CBFQ algorithms as the basis, we propose the Credit-based Short-term Fairness Scheduling algorithm (CSFS). One of the key features of CSFS is that it achieves to provide short-term fairness to the delay-sensitive users, while still schedules opportunistically the non-delay-sensitive users to obtain high system throughput. CBFQ [43] has been introduced in the wireline scheduling to provide proportional fairness as WFQ [8] [12], but with lower computational complexity. The fairness is maintained by updating a credit counter for each active flow. It has also been shown in [43] that CBFQ has the same fairness and delay bounds as SCFQ [6] and PGPS [12].

4.2.1 Scheduling Algorithm Description

The credit update rule used in our scheme is shown in Table 4.2 and is the same one with that of [43] and [37], which ensures that all backlogged users have fair credit updates proportional to their weights.

In the following let us denote by W_i (measured in time slots) the observation window for user i. It should be noted here that different classes of users may have different values of W_i e.g. real-time traffic will have smaller observation window than the non-delay sensitive data traffic. Let $S_i(t, W_i)$ be the service (the number of slots accessed) received by user i in the observation window W_i slots from slot $(t - W_i + 1)$ to slot t, i.e. $S_i(t, W_i) = \alpha_i(t - W_i + 1, t)$. Thus the service deviation in the observation window ending at slot t

is given by $W_i\phi_i - S_i(t, W_i)$ since a user must receive $W_i\phi_i$ service to maintain absolute fairness. By normalizing the service, we can define $\Delta S_i(t, W_i) = 1 - \frac{S_i(t,W_i)}{W_i\phi_i}$. Therefore we define the new scheduling policy as:

$$\widehat{Q}(t) = \arg\min_i \frac{L - K_i(t) + \widehat{U}_i(t)}{\phi_i} \tag{4.4}$$

where the function $\widehat{U}_i(t)$ is non-negative and defined as

$$\widehat{U}_i(t) = \max(U_i(t) - F_i(\Delta S_i(t, W_i)), 0) \tag{4.5}$$

where $F_i(x)$ is a pre-defined function according to the requirements of the network. The exact role of this function is explained in more detail later in this chapter. The value of $\Delta S_i(t, W_i)$ depends on the length of interval W_i. For a larger W_i we expect smaller values of $\Delta S_i(t, W_i)$ on average. Hence by choosing different values of W_i we may change the role of $\Delta S_i(t, W_i)$ in the scheduling policy. The main principle behind the scheduling policy (4.4) is to introduce the fairness interval information W_i in the new cost function $\widehat{U}_i(t)$. $U_i(t)$ still plays its role in balancing the fairness and throughput as in WCFQ, however its control on the selection of the next user to receive service will change with the influence of $\Delta S_i(t, W_i)$ and function $F_i(x)$. Larger W_i will make the scheduling policy turn more towards the channel factor (more opportunistic factor), while smaller W_i will make the scheduling care more about fairness. $F_i(\Delta S_i(t, W_i))$ may grow till it cancels $U_i(t)$. At the point where $\widehat{U}_i(t) = 0$ the user i will be scheduled following the CBFQ scheme trying to maintain the fairness, independent of the channel conditions.

An example of the form of function $F_i(x)$ is shown in Figure 4.1, and is described as follows:

$$F_i(x) = \begin{cases} y(x) & x > A \\ 0 & x \leq A \end{cases}$$

where $y(x)$ consists of serval linear lines. By choosing the points A, B, C, D, etc. and the gradients in the respective range, $F_i(x)$ will control in what speeds and ranges $\Delta S_i(t, W_i)$

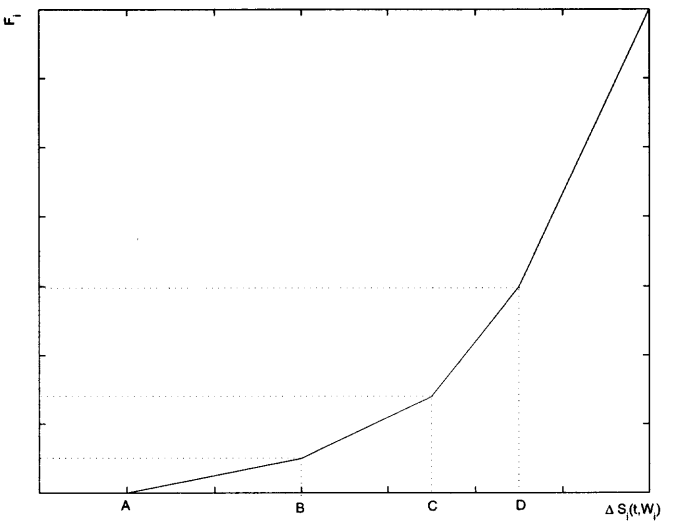

Figure 4.1 Example of function $F_i(x)$

will impose its effect on the scheduling policy. With reference to Figure 4.1, if the scheduler is to be insensitive to small unfairness, we can give large A and make the gradient of line AB small. Similarly, a small A and concave $F_i(x)$ with large gradient around A is suitable for applications with more strict short-term requirements.

4.2.2 Fairness Discussion

In this subsection we discuss the fairness property of the proposed CSFS algorithm. From the definition of the scheduling policy (4.4) we observe that function $\widehat{U}_i(t)$ in CSFS contains two elements. One of them, i.e. function $U_i(t)$, represents the transmission cost due to the channel conditions as in the WCFQ, while the second one reflects the short-term fairness cost considerations. Following similar steps with the proof in [37] we can obtain the following result regarding the accumulated credit count of each user.

Proposition 6 *For any backlogged user i, there exists a time slot $t' \leq t$ which bounds the accumulated credit count $K_i(t+1)$ as*

$$0 \leq K_i(t+1) \leq L + \widehat{U}_i(t')$$

Clearly, since the value of $\widehat{U}_i(t)$ is between 0 and $U_i(t)$, the accumulated credit count in CSFS has the same bound as WCFQ. We can also extend this result to the fairness bound, and therefore we can easily conclude that there exist slots p' and q' satisfying the following expression:

$$\left|\frac{\alpha_i(t_1,t_2)}{\phi_i}-\frac{\alpha_j(t_1,t_2)}{\phi_j}\right| \leq \frac{L+\widehat{U}_i(p')}{\phi_i}+\frac{L+\widehat{U}_j(q')}{\phi_j} \tag{4.6}$$

In fact we have not been able to derive the short-term fairness performance directly from the credit and fairness bounds. However the users' credit increases under the proposed CSFS and the WCFQ schemes can be used to reflect their corresponding performances in achieving the short-term fairness. Therefore the credit accumulation is treated as an indication of service deficit. Zero credit indicates that the user receives equal or more than its share of service. However those with large value of credit accordingly receive less service than their fair share. In WCFQ, such situation can only be mitigated by the changing of channel conditions, once the cost function $U_i(t)$ is determined. But in CSFS the large credit also means the large service deviation function $F_i(x)$ according to the respective function definition for the specific user. The trend of credit increment will cease or become slow in such situations. Therefore, in CSFS the credit has less probability of reaching the bound, which also indicates better short-term performance.

Based on inequality (4.6), the statistic fairness bound can be obtained as follows:

$$\begin{aligned}
&\Pr\left(\left|\frac{\alpha_i(t_1,t_2)}{\phi_i}-\frac{\alpha_j(t_1,t_2)}{\phi_j}\right| \geq \frac{L+x}{\phi_i}+\frac{L+x}{\phi_j}\right) \\
\leq\ &\Pr\left(\frac{\widehat{U}_i(p')}{\phi_i}+\frac{\widehat{U}_j(q')}{\phi_j} \geq \frac{x}{\phi_i}+\frac{x}{\phi_j}\right) \\
=\ &\Pr\left(\frac{U_i(p')-F_i(\Delta S_i(p'))}{\phi_i}+\frac{U_j(q')-F_j(\Delta S_j(q'))}{\phi_j}\right. \\
&\qquad\qquad \left.\geq \frac{x}{\phi_i}+\frac{x}{\phi_j}\right)
\end{aligned} \tag{4.7}$$

Therefore, since $F_i(x)$ is non-negative, for the same statistic fairness constraint, larger cost function $U_i(t)$ can be applied compared to (4.2) to improve the throughput.

4.3 Performance Evaluation

In this section we evaluate the performance of our algorithm through modeling and simulation. In order to better understand and evaluate the underlying principles, of the proposed CSFS algorithm we compare its performance results with the corresponding results obtained by the WCFQ algorithm.

4.3.1 Simulation Model and Assumptions

In the following we consider a single cell Time Division Multiple Access (TDMA) system with seven users that are assumed to be continuously backlogged. The service received by all users is measured in number of slots while the duration of a slot is $10ms$. Two classes of services are considered in this study: the delay-sensitive (real-time) service (in the following we refer to as class 1) and the delay-non-sensitive data service (in the following we refer to as class 2). Unless otherwise is explicitly indicated, a user i that belongs to class 1 (class 2) is assumed to have observation window size $W_i^1 = 20$ ($W_i^2 = 200$) and regulated by the service deviation function $F_{d1}(x)$ ($F_{d2}(x)$). The functions $F_{d1}(x)$ and $F_{d2}(x)$ are defined as follows:

$$F_{d1}(x) = \begin{cases} 0 & x < 0 \\ 400x & 0 \leq x \leq 0.05 \\ 20 & 0.05 < x \end{cases}$$

$$F_{d2}(x) = \begin{cases} 0 & x < 0.1 \\ 20(x - 0.1) & 0.1 \leq x \leq 1 \\ 20 & 1 < x \end{cases}$$

Obviously, function $F_{d1}(x)$ has large gradient and stringent short-term fairness property which is suitable for delay-sensitive traffic, while $F_{d2}(x)$ only keeps relatively loose short-term fairness which can be applied to non-delay-sensitive traffic.

The cost function $U_i(t)$ used by CSFS and WCFQ algorithms in this study is defined as:

$$U_i(t) = 20 - 20 \times \left(\frac{E_i(t)}{20}\right)^{1/\beta} \tag{4.8}$$

where $E_i(t)$ is the SNR of the pilot signal. Assuming that the pilot signal is transmitted at fixed power the value of $E_i(t)$ represents the instant channel quality. The maximal value of $E_i(t)$ is 20. Obviously a bad channel will result in higher cost, i.e. larger value of $U_i(t)$. For representation simplicity we use parameter β to control the value of the cost function. In this study we only consider cases where $\beta \geq 1$ and therefore with larger β, $U_i(t)$ takes smaller values for the same $E_i(t)$, and thus users are less affected by the variation of the channel conditions. The instant value of $E_i(t)$ depends on the channel condition of user i at that time. An eight-state Markov Rayleigh fading channel model [34] is used throughout our study. The received signal to noise ratio, from zero to 20, is divided into eight states in a way that a channel has equal steady state probabilities of staying at any state. To study the performance of various algorithms to the changing channel conditions users are assigned different average SNRs.

4.3.2 Observation Window Size

In this subsection and corresponding experiment our goal is to study the impact of the various observation window sizes on the achievable throughput. Therefore, for simplicity in the presentation of the results, in this experiment all seven users are assumed to belong to class 2 and are assigned accordingly function $F_{d2}(x)$ in the CSFS algorithm. Figure 4.2 presents the average throughput of CSFS with $\beta = 5$, WCFQ with $\beta = 5$ and WCFQ with $\beta = 40$ for different observation window sizes. As can be seen by this figure the effect of changing the window size on the achievable throughput is significant for the CSFS algorithm. When the window size is small, e.g. 10, in an attempt to satisfy the short-term fairness the freedom of opportunistic scheduling is limited under CSFS, and therefore the

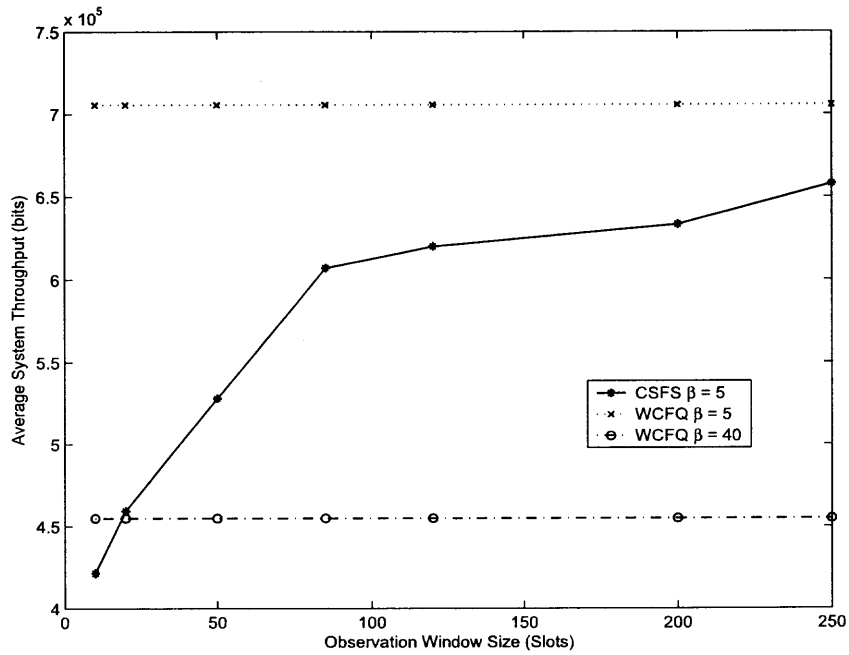

Figure 4.2 Average system throughput for different observation window sizes

throughput is low. This is actually reflected by the use of the service deviation function $F_i(x)$ in relation (4.5). When the window size becomes larger, the scheduler has more flexibility in providing the fairness which allows it to further exploit the channel variation and improve the system throughput.

However the achieved throughput under WCFQ is not sensitive to the changes of W_i since WCFQ contains no information of the observation window size. Please note that in the WCFQ case, the larger W_i will also improve the short-term fairness performance, which makes it possible to use smaller β in (4.8) and obtain higher throughput (e.g. WCFQ with $\beta = 5$). In Figure 4.2 we notice that the throughput performance of WCFQ is better than that of CSFS for the same $\beta = 5$. However this throughput improvement may be achieved only when all users have the same short-term fairness requirement, which is not suitable for multimedia traffic. In the situation that strict short-term fairness is required for some of the users only, WCFQ has to use large β, which will lower the achievable throughput as can be seen by Figure 4.2 (e.g. WCFQ with $\beta = 40$). On the other hand CSFS can still

operate with small β and still provide the required short-term fairness as it is demonstrated in the following experiment.

4.3.3 Fairness and Throughput

In this subsection we study the tradeoffs in the performance of throughput and fairness under CBFQ, CSFS ($\beta = 5$) and WCFQ scheduling algorithms. In CSFS algorithm, users 1 and 2 are assumed to belong to class 1, while the rest of the users are assumed to belong to class 2. In the WCFQ algorithm, five scenarios with $\beta = 5, 20, 40, 60, 80$ are simulated. Although WCFQ is not affected by the observation window size, for comparison purposes we still set a window size $W_i = 20$ for user 2, which is the same as that in CSFS, in order to measure the short-term fairness performance.

Figure 4.3 presents for user 2 (that belongs to class 1) the corresponding probabilities of unfairness in the observation window $W_i = 20$ under the different algorithms. Apparently CBFQ has the best short-term fairness performance, and is used here as benchmark. However as can be seen later in this subsection this happens at the cost of low system throughput. With increasing β, the short-term fairness of WCFQ is improving as well. In order to reach similar (un)fairness performance with the CSFS ($\beta = 5$) algorithm, the WCFQ has to raise its β to 80. As explained before the consequence of high β for WCFQ is the low system throughput as observed in Figure 4.4, where the total system throughput as well as the throughput of a class-2 user (e.g. user 5) are presented. This happens because under WCFQ users 3 to 7 (i.e all class-2 users) are also forced to have unnecessary (not required) short-term fairness similar to the class-1 users (e.g user 2). This point is also confirmed by the comparison of the throughput of user 5 (that belongs to class 2) presented in Figure 4.4. Specifically despite the small β in CSFS, the function $F_{d1}(x)$ guarantees the quality of service of user 2 while at the same time $F_{d2}(x)$ maintains the high throughput of user 5. This clearly demonstrates the advantages that can be achieved in the scheduling flexibility by CSFS mechanism compared to WCFQ scheduling algorithm.

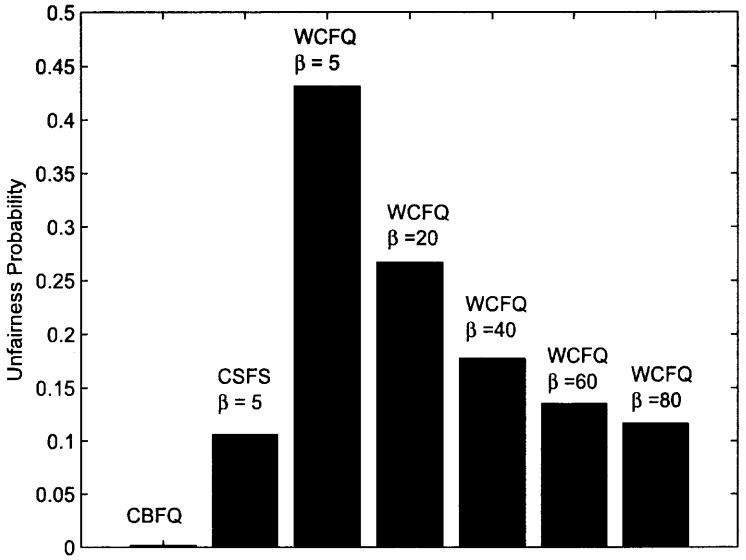

Figure 4.3 Probabilities of unfairness in observation window size of 20 slots for User 2 under different algorithms and choices of β

Finally from this figure we also observe that as mentioned before the achievable throughput under CBFQ is lower than the corresponding ones achieved under CSFS and WCFQ.

4.4 Conclusion

Due to the inner characteristics of wireless communication, the users may experience location-dependent and time-dependent errors, that will prevent them from receiving service and consequently break the fairness. Achieving short-term fairness while still maintaining maximum throughput could be extremely difficult due to the changing wireless channel conditions. In this chapter we studied the problem of providing short-term access time fairness with opportunistic scheduling while still maintaining high system throughput. We proposed a new Credit-based Short-term Fair Scheduling algorithm which exploits the wireless channel variations to obtain high throughput via opportunistic scheduling, while at the same time introduces the information of service interval into the service scheduling policy. With properly defined service deviation functions for different classes of users,

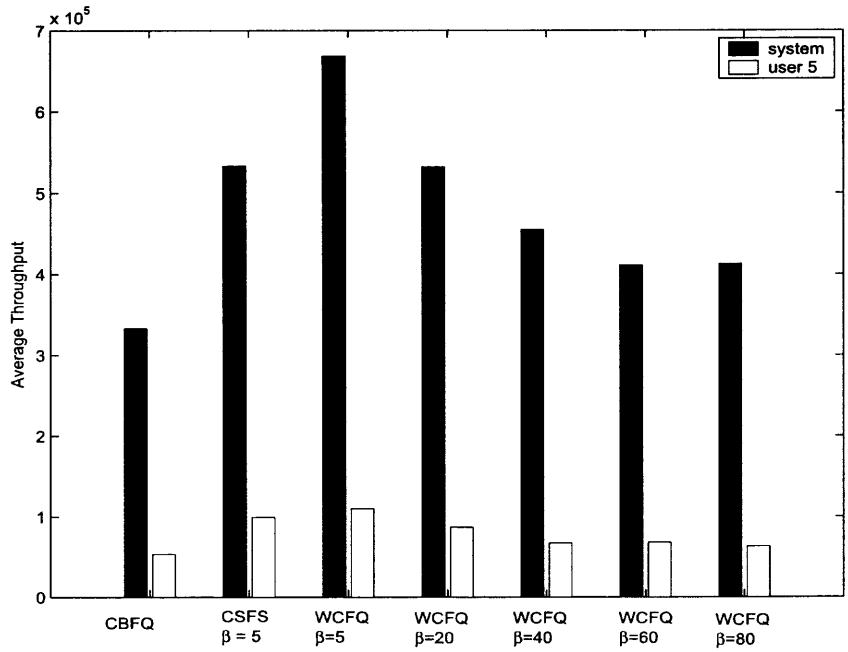

Figure 4.4 Average throughput of the whole system and user 5 (class 2)

we demonstrated that the proposed algorithm provides more flexibility in the control of fairness in systems with multimedia traffic, while still achieves high throughput.

CHAPTER 5

SUMMARY AND FUTURE WORK

5.1 Summary and Contributions

Wideband CDMA (WCDMA) has been proposed as a key air interface technique for third generation (3G) wireless systems, and will continue to be adopted as a strong candidate for 4G systems that will provide differentiated services to multimedia traffic. With the capability of dynamically varying user channel rates, WCDMA systems can provide more flexibility in bandwidth allocation. The integration of multimedia capabilities to wireless networks, requires the systems to support different QoS requirements and traffic characteristics. Since in the current and future mobile wireless communication infrastructures, both topologies and traffic evolve and fluctuate on widely different time scales, and the performances of the various services are strongly correlated as the resources are shared among them, dynamic resource allocation and scheduling methods should be employed.

Mobile users always experience time-varying channel and location-dependent errors, which could cause lower transmission speed and unexpected longer delay. In many cases the efforts (or system resources) taken to transmit data may not be proportional to the actual throughput, if the transmission is not properly scheduled, which may also result in waste of resources. Therefore, devising new scheduling algorithms play a key role for the operational effectiveness and efficiency of the next generation wireless communication systems.

In this dissertation, we first studied the fair scheduling problem in the slotted-CDMA systems. The difficulties of transmission scheduling in CDMA systems can be summarized into the following aspects: a) the actual system capacity, as defined conventionally, is not fixed and known in advance, since it is a function of several parameters such as the number of users, the channel conditions, the transmission powers etc.; b) multiple users must be simultaneously scheduled into a slot in order to efficiently and fairly utilize the system

resources; c) due to the varying channel conditions the utilized resources are not proportional to the achievable data rate and received throughput.

To overcome the above difficulties, the concept of power index capacity which indicates the possible power index a user can accept was initially proposed and studied. In our approach the system power index is regarded as a fixed resource, instead of the conventional system capacity, e.g. the total available transmit rate. The actual service to a single user is the combined effect of the assigned power index and the access time. The power index adjustment among users within the constraint of PIC was analyzed, which creates the basis for a feasible rate scheduling and service compensation process.

To improve the system performance in terms of throughput, in chapter 2 we adopted the opportunistic scheduling policy that further exploits the variation of channel conditions. Therefore we designed and analyzed two new scheduling algorithms, CARS and FCARS. Through an iterative process CARS estimates the power index of a single user starting with reference the ideal system, and then in the following iterations redistributes the unused power index to the users with better channel condition, in order to fully utilize the available system resources. Users with better channel condition obtain more bandwidth, while those with worse channel condition get less bandwidth, which however may result to fairness violations. To overcome the unfair service allocation FCARS implemented a compensation algorithm, in which the lagging users can receive compensation service when the corresponding channel conditions improve. The corresponding results demonstrated that our proposed approach improves the system throughput and achieves the longterm fairness by keeping the service received by each user proportional to its weight despite the users channel variations.

Throughout this part of the dissertation we assumed that all backlogged users are allowed to transmit simultaneously but with various transmit rates, despite their possible undesired channel conditions. The throughput improvement is achieved by limiting the transmit rate of weak users. Although such an approach maintains the fairness it is not

optimal with respect to the system throughput, due to the convexity of the relationship between the power index and the transmit rate. Therefore in chapter 3, we considered, studied and analyzed the optimal throughput opportunistic fair scheduling problem. It was shown that such a problem can be expressed as a weighted throughput maximization problem, under certain power and QoS requirements, where the weights are the control parameters that reflect the fairness constraints. A stochastic approximation method was presented in order to effectively identify the required control parameters. Furthermore the concept of power index capacity was used to represent all the corresponding constraints by the users' power index capacities at some certain system power index. Based on this, the optimization problem under consideration was converted into a binary knapsack problem, where the optimal solution can be obtained through a global search within a specific range. The corresponding results revealed the advantages of the proposed policy over other existing scheduling schemes, and demonstrated that it achieves very high throughput, while satisfies the QoS requirements and maintains the fairness among the users, under different channel conditions and requirements

The optimal throughput opportunistic scheduling policy described in 3 can only provide the long-term fairness property. The high throughput of opportunistic scheduling is achieved by sacrificing in many cases the short-term fairness. Although those users who lose service will be eventually compensated when their channels improve, their services within a short time interval are not guaranteed due to the randomly time-varying wireless channels. Such short-term unfairness in low level may result in the timeout of higher layer protocols and may cause the consequent system and service performance degradation. Therefore in chapter 4 we studied the problem of providing short-term access time fairness with opportunistic scheduling while still maintaining high system throughput. We proposed a new Credit-based Short-term Fair Scheduling algorithm which exploits the wireless channel variations to obtain high throughput via opportunistic scheduling, while at the same time introduces the information of service interval into the service scheduling

policy. With properly defined service deviation functions for different classes of users, we demonstrated that the proposed algorithm provides more flexibility in the control of fairness in systems with multimedia traffic, while still achieves high throughput.

5.2 Future Work

It should be noted that the proposed opportunistic scheduling approach presented in this dissertation, considered only the single cell scenario and did not account for the inter-cell interference. As a result, an optimal scheduling policy in terms of maximizing the system throughput while maintaining the fairness among the users has been devised, which aims to provide a in-depth analysis of the achievable performance, when considering the throughput fairness constraints in the uplink CDMA scheduling.

However in a realistic system with multiple cells, that may include a large number of voice and data users, the inter-cell interference would affect the scheduling policy. In the literature, studies about the scheduling problem in a multi-cell environment have been mainly conducted based on a large number of voice user scenario, which has uniformly loaded the cells. For traditional voice service, if the number of active users is large, the inter-cell interference is usually modeled as a Gaussian process [44], and the fast fluctuation of the individual inter-cell interference can be averaged out. However this may not be the case for the high-speed uplink data service where only a relatively small number of users may be allowed to transmit simultaneously in order to achieve high throughput while maintaining fairness. In [45] and [46] the authors studied some specific only scenarios where data users are present along with the voice users. Specifically, in [45] the admission control of a single data user when many voice users are present is studied, while the inter-cell interference caused by this data user is combined into that of voice users. The objective of [46] was to maximize the data throughput with fixed number of voice users. In that system model the inter-cell interference is treated as constant, in a similar way with the noise.

In principle, as argued in [46] and was demonstrated in this dissertation as well, it is not possible to achieve high-speed uplink transmission if a large number of data users transmit at the same time due to the high intracell interference. In case where a multi-cell scenario is considered the SINR of each active user in (3.1) is then represented as follows:

$$\frac{h_i p_i G_i}{\sum_{j=1, j\neq i}^{B(k)} h_j p_j + I_{inter} + W\eta_0} = \gamma_i \quad i = 1, 2, \ldots, B(k) \tag{5.1}$$

where I_{inter} denotes the inter-cell interference. In this case additional considerations are needed in order to accurately account for the inter-cell interference and address the scheduling problem, including the cooperation among cells, interference cancelation techniques, maximum allowable transmission power adjustment methods, etc. This problem is interesting and challenging, and is part of our current and future research.

Furthermore, as demonstrated in this dissertation there is a tradeoff between the short-term performance and the achievable total throughput performance. We expect that the fast channel variation will have better short-term performance, however, in the case of slow channel condition fluctuation, short-term fairness requires the scheduler to serve not the throughput optimal candidates but some urgent users with possibly slightly worse channel conditions. The algorithms that measure the packet deadlines own such properties and functions. However it is a challenging problem, and of high practical and research importance, to study to what extent the throughput performance is compromised by satisfying the short-term performance. For instance by simply allocating resources to the user with the earliest deadline may not always give satisfactory results. Therefore, as part of our future research, we will analyze this problem, and we will investigate strategies that can control this tradeoff for different applications with different QoS requirements.

APPENDIX A

Pseudo Code of CARS and FCARS Algorithms

This appendix gives the pseudo codes of CARS and FCARS algorithms, which are described in Chapter 2.

```
/* compute power index increment by GPS rule*/
power_index_increment(i)
  if ( A = ∅ )
   /*the power index at ideal system  */
     Δg_i = ideal_power_index_assignment (i);
 /*distribute the remained power index among the
   un-scheduled users of set U */
   else if (i ∉ A )

        Δg_i = φ_i(1 − g_i)^2 / Σ φ_j(1 − g_j)^2 · C_R ;

   end if

CARS_scheduler()
   if ( C_R ≠ 0 and A ≠ B(t)) repeat
      for ( all i, i ∉ A )
         Δg_i = power_index_increment (i);
 /* if the totally assigned power index exceeds its  PIC */
         if ((g_i + Δg_i) > π_i)
            g_i = π_i ;
            A = A ⋃ {i};
          else
 /* accumulate the power index assigned in this round */
            g_i = g_i + Δg_i ;
          end if
       end for
   end repeat
 C_R = Ψ − Σ_{i∈B(t)} m_i ; /* the remained power index */
```

Figure A.1 Pseudo-code of CARS algorithm

power_index_increment (i,U)

if ($\mathbf{A} = \varnothing$ **and** $\mathbf{U} = \mathbf{B}(t)$)

Δg_i = **ideal_power_index_assignment** (i);

*/*distribute the remained power index among the un-scheduled users in set U */*

else if ($i \notin \mathbf{A}$ **and** $i \in \mathbf{U}$)

$$\Delta g_i = \frac{\phi_i (1-g_i)^2}{\sum_{j \notin A, j \in U} \phi_j (1-g_j)^2} \cdot C_R \ ;$$

end if

/ allocate the power index ψ to users in set U */*

cars_scheduler(ψ, U)

$C_R = \psi$;

if ($C_R \neq 0$ **and** $\mathbf{A} \neq \mathbf{U}$) **repeat**

for (*all* i, $i \in \mathbf{U}$ **and** $i \notin \mathbf{A}$)

/ the reference power index for user i */*

Δg_i = **power_index_increment** (i,U);

if $((g_i + \Delta g_i) > \pi_i)$

$g_i = \pi_i$;

$\mathbf{A} = \mathbf{A} \bigcup \{i\}$;

else

$g_i = g_i + \Delta g_i$;

end if

end for

$$C_R = \psi - \sum_{i \in B(t)} m_i \ ;$$

end repeat

compute_excess_service ( );

/ the excess service each user receives by the assigned power index vector* **g** **/*

compute_excess_service(g)

R = **total_rate (g)**; */* find the total achievable rate */*

for (*all* i, $i \in \mathbf{B(t)}$)

$$r_i^* = \frac{\phi_i}{\sum_{j \in B(t)} \phi_j} \cdot R \ ; \ /^* \text{ calculate the reference rate } ^*/$$

$r_i = f_g(g_i)$; */* convert the power index to rate */*

$S_i^e = S_i^e + (r_i - r_i^*) \cdot \tau_k$; */* the received excess services */*

end for

Figure A.2 Pseudo-code of FCARS algorithm: Part One - Enhanced CARS

```
define_groups( )
  for ( all i, i ∈ B(t) )
  /* the reference power index assignment in the ideal system */
      g_i^* = ideal_power_index_assignment (i);  ;
      if ( s_i^e > 0 and g_i > g_i^* f_q(q_i / q_i^max) )
          G_lead = G_lead ∪ {i} ; /* add into leading group */
      else if ( s_i^e < 0 and π_i > g_i )
          G_lag = G_lag ∪ {i} ; /* add into lagging group */
      else
          G_normal = G_normal ∪ {i} ; /* add into normal group */
      end if
    end for

given_up_service( )
  for ( all i, i ∈ G_lead )
      /* user i gives up service in term of power index */
      g_i^g = min{ f_q(q_i / q_max^i) · g_i , g_i − f_g^{-1}( f_g(g_i) − s_i^e / τ ) } ;
      g^g = g^g + g_i^g ;  /* add to the total given up service */
      g_i = g_i − g_i^g ;  /* adjust user i's power index */
  end for

fcars_scheduler( )
  /* start the first scheduling */
  cars_scheduler (Ψ , B(t) );
  define_groups( );
  /* force the leading group to give up service */
  give_up_service ( );
  /* the compensation scheduling for lagging users */
  cars_scheduler ( g^g, G_lag);
  if (C_R > 0)
  /* if the lagging group can not absorb all given up service */
  /* return it to the leading and normal group */
     cars_scheduler ( C_R, G_lead + G_normal);
  end if
```

Figure A.3 Pseudo-code of FCARS algorithm: Part Two - Compensation

BIBLIOGRAPHY

[1] F. Adachi, M. Sawahashi, and H. Suda, "Wideband DS-CDMA for next-generation mobile communications systems," *IEEE Commun. Mag.*, vol. 36, pp. 56–69, September 1998.

[2] E. Dahlman, F. Gudmundson, M. Nilsson, and J. Skold, "UMTS/IMT-2000 based on wideband CDMA," *IEEE Commun. Mag.*, vol. 36, pp. 70–80, September 1998.

[3] T. Ojanpera and R. Prasad, "An overview of air interface access for IMT-2000/UMTS," *IEEE Commun. Mag.*, vol. 36, pp. 82–95, September 1998.

[4] N. Sollenberger, N. Seshadri, and R. Cox, "The evolution of IS-136 TDMA for Third-Generation wireless services," *IEEE Commun. Mag.*, vol. 6, pp. 8–18, June 1999.

[5] H. Zhang, "Service disciplines for guaranteed performance service in packet-switching networks," *Proc. IEEE*, vol. 83, pp. 1374–1396, Oct. 1995.

[6] S. Golestani, "A self-clocked fair queuing scheme for broadband applications," *Proc. IEEE INFOCOM*, pp. 636–646, April 1994.

[7] J. Bennett and H. Zhang, "WF2Q: Worst-case fair weighted fair queuing," *Proc. IEEE INFOCOM*, pp. 120–128, Mar. 1996.

[8] P. Goyal, H. Vin, and H. Chen, "Start-time fair queuing: A scheduling algorithm for integrated services," *Proc. ACM-SIGCOMM*, pp. 157–168, Augst 1996.

[9] D. Stidialis and A. Varma, "Efficient fair queueing algorithms for packet-switched networks," *IEEE/ACM Trans. On Networking*, vol. 6, pp. 175–185, April 1998.

[10] J. C. R. Bennett, D. C. Stephens, and H. Zhang, "Hierarchical packet fair queueing algorithms," *Proc. ACM-SIGCOMM*, pp. 143–156, 1996.

[11] D. C. Stephens, J. C. R. Bennett, and H. Zhang, "Implementing scheduling algorithms in high-speed networks," *IEEE J. Select Areas Commun.*, vol. 17, pp. 1145–1158, June 1999.

[12] A. Parekh and R. Gallager, "A generalized processor sharing approach to flow control – The single node case," *IEEE/ACM Trans. On Networking*, vol. 1, pp. 344–357, June 1993.

[13] X. Liu, E. Chong, and N. Shroff, "Transmission scheduling for efficient wireless utilization," *Proc. IEEE INFOCOM*, vol. 2, pp. 776–785, April 2001.

[14] P. Viswanath, D. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. on Inform. Theory*, vol. 48, pp. 1277–1294, June 2002.

[15] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana, and A. Viterbi, "CDMA/HDR: A bandwidth-efficient high-speed wireless data service for nomadic users," *IEEE Commun. Mag.*, pp. 70–77, July 2002.

[16] S. Oh and K. Wasserman, "Dynamic spreading gain control in multiservice CDMA networks," *IEEE J. Select Areas Commun.*, vol. 17, pp. 918–927, May 1999.

[17] O. Gurbuz and H. Owen, "Dynamic resource scheduling for variable QoS traffic in W-CDMA," *Proc. IEEE ICC*, 1999.

[18] T. E. Ng, I. Stoica, and H. Zhang, "Packet fair queueing algorithms for wireless networks with location-dependent errors," *Proc. IEEE INFOCOM*, pp. 1103–1111, March 1998.

[19] V. Bharghavan, S. lu, and T. Nandagopal, "Fair queuing in wireless networks: issues and approaches," *IEEE Personal Communications*, vol. 6, pp. 44–53, Feb. 1999.

[20] S. Lu, V. Bharghavan, and R. Srikant, "Fair scheduling in wireless packet networks," *IEEE/ACM Trans. On Networking*, vol. 7, no. 4, pp. 473–489, 1999.

[21] X. Liu, E. Chong, and N. Shroff, "A framework for opportunistic scheduling in wireless networks," *Computer Networks*, vol. 41, pp. 451–474, March 2003.

[22] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "CDMA data QoS scheduling on the forward link with variable channel conditions," *Bell Labs Technical Report*, April 2000.

[23] F. Berggren, S. Kim, R. Jantti, and J. Zander, "Joint power control and intracell scheduling of ds-cdma nonreal time data," *IEEE J. Select Areas Commun.*, vol. 19, pp. 1860–1870, Oct. 2001.

[24] R. Agrawal, A. Bedekar, R. La, R. Pazhyannur, and V. Subramanian, "Class and channel condition based scheduler for EDGE/GPRS," *Proceedings of SPIE*, vol. 4531, pp. 59–69, 2001.

[25] L. Xu, X. Shen, , and J. W. Mark, "Dynamic bandwidth allocation with fair scheduling for WCDMA systems," *IEEE Wireless Communication*, vol. 9, pp. 26–32, April 2002.

[26] S. Jafar and A. Goldsmith, "Adaptive multirate CDMA for uplink throughput maximization," *IEEE Trans. on Wireless Comm.*, vol. 2, pp. 218–228, Mar. 2002.

[27] I. Akyildiz, D. Levine, and I. Joe, "A slotted CDMA protocol with BER scheduling for wireless multimedia networks," *IEEE/ACM Trans. On Networking*, vol. 7, no. 2, pp. 146–158, 1999.

[28] A. Stamoulis and G. Gianakis, "Packet fair queueing scheduling based on multirate multipath-transparent CDMA for wireless networks," *Proc. IEEE INFOCOM*, vol. 3, pp. 1067–1076, 2000.

[29] M. Ard and A. Leon-Garcia, "A generalized processor sharing approach to time scheduling in hybrid CDMA/TDMA," *Proc. IEEE INFOCOM*, vol. 3, pp. 1164–1171, 1998.

[30] S. Choi and D. Cho, "Weighted fair queueing for data service in a multimedia CDMA system," *Proc. IEEE VTC*, vol. 1, pp. 286–291, Fall 2000.

[31] A. Sampath, P. S. Kumar, and J. M. Holtzman, "Power control and resource management for a multimedia CDMA wireless system," *Proc. IEEE PIMRC*, vol. 1, pp. 21–25, 1995.

[32] R. Guerin, H. Ahmadi, and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks," *IEEE J. Select Areas Commun.*, vol. 9, pp. 968–981, Sept. 1991.

[33] J. Ye, J. Hou, and S. Papavassiliou, "A comprehensive network management framework for next generation wireless networks," *IEEE Trans. on Mobile Computing*, vol. 1, pp. 249–264, Oct.-Dec. 2002.

[34] H. Wang and N. Moayeri, "Finite-state markov channel – a useful model for radio communication channels," *IEEE Trans. on Vehic. Tech.*, vol. 44, pp. 163–171, March 1995.

[35] T. Ottosson and A. Svensson, "On schemes for multirate support in DS-CDMA systems," *Wireless Personal communications*, pp. 265–287, March 1998.

[36] S. Borst and P. Whiting, "Dynamic rate control algorithms for HDR throughput optimization," *Proc. IEEE INFOCOM*, vol. 2, pp. 976–985, 2001.

[37] Y. Liu and E. Knightly, "Opportunistic fair scheduling over multiple wireless channels," *Proc. IEEE INFOCOM*, pp. 1106–1115, March 2003.

[38] D. Hochbaum, "A nonlinear knapsack problem," *Operations Research Letters*, vol. 17, pp. 103–110, April 1995.

[39] S. Martello and P. Toth, *Knapsack problems: Algorithms and computer implementations*. John Wily and Sons, 1990.

[40] R. Miller, *Optimization: Foundations and Applications*. John Wiley and Sons, 2000.

[41] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," *Proc. IEEE VTC*, vol. 3, Spring 2000.

[42] S. Kulkarni and C. Rosenberg, "Opportunistic scheduling policies for wireless systems with short term fairness constraints," *Proc. IEEE GLOBECOM*, vol. 1, pp. 533–537, Nov. 2003.

[43] B. Bensaou, K. Chan, and D. Tsang, "Credit-based fair queuing (CBFQ): A simple and feasible scheduling algorithm for packet networks," *Proc. IEEE ATM'97 Workshop*, pp. 589–594, May 1997.

[44] K. Gilhousen, I. Jacobs, R. Padovani, A. Viterbi, L. W. Jr., and C. W. III, "On the capacity of a cellular CDMA system," *IEEE Trans. on Vehic. Tech.*, vol. 40, pp. 303–312, May 1991.

[45] S. Kumar and S. Nanda, "High data rate packet communications for cellular networks using CDMA: Algorithms and performance," *IEEE J. Select Areas Commun.*, vol. 17, pp. 472–492, March 1999.

[46] S. Ramakrishna and J. M. Holtzman, "A scheme for throughput maximization in a dual-class CDMA system," *IEEE J. Select Areas Commun.*, vol. 16, pp. 830–844, Aug. 1998.