OPPORTUNISTIC SCHEDULING ALGORITHMS IN DOWNLINK CENTRALIZED WIRELESS NETWORKS

Rui Yin

Submitted in fulfillment of the academic requirements
for the degree of MScEng
in the School of Electrical, Electronic and Computer Engineering
at the University of KwaZulu-Natal, Durban, South Africa

October 28, 2005

To my parents, sister and 11. To my brother, landlord and landlady

As the candidate's supervisor and co-supervisor, we have approved this dissertation for submission.

Name: Prof. Dawoud Dawoud

Signed: _____

Date: _____

Name: Dr. Hongjun Xu

Signed: _____

Date: _____

Abstract

As wireless spectrum efficiency is becoming increasingly important with the growing demands for wideband wireless service scheduling algorithm plays an important role in the design of advanced wireless networks. Opportunistic scheduling algorithms for wireless communication networks under different QoS constraints have gained popularity in recent years since they have potentials of achieving higher system performance. In this dissertation firstly we formulate the framework of opportunistic scheduling algorithms. Then we propose three new opportunistic scheduling schemes under different QoS criteria and situations (single channel or multiple channel).

1. Temporal fairness opportunistic scheduling algorithm in the short term

   We replicate the temporal fairness opportunistic scheduling algorithm in the long term. From simulation results we find that this algorithm improves the system performance and complies with the temporal fairness constraint in the long term. However, the disadvantage of this algorithm is that it is unfair from the beginning of simulation to 10000 time slot on system resource (time slots) allocation - we say it is unfair in the short term. With such a scheme, it is possible that some users with bad channel conditions would starve for a long time (more than a few seconds), which is undesirable to certain users (say, real-time users). So we propose the new scheme called temporal fairness opportunistic scheduling algorithm in the short term to satisfy users' requirements of system resource in both short term and long term. Our simulation results show that the new scheme performs well with respect to both temporal fairness constraint and system performance improvement.

2. Delay-concerned opportunistic scheduling algorithm

   While most work has been done on opportunistic scheduling algorithm under fairness constraints on user level, we consider users' packet delay in opportunistic scheduling. Firstly we examine the packet delay performance under the long term temporal fairness opportunistic scheduling (TFOL) algorithm. We also simulate the earliest deadline-first (EDF) scheduling algorithm in the wireless environment. We find that

the disadvantage of opportunistic scheduling algorithm is that it is unfair in packet delay distribution because it results in a bias for users with good channel conditions in packet delay to improve system performance. Under EDF algorithm, packet delay of users with different channel conditions is almost the same but the problem is that it is worse than the opportunistic scheduling algorithm. So we propose another new scheme which considers both users' channel conditions and packet delay. Simulation results show that the new scheme works well with respect to both system performance improvement and the balance of packet delay distribution.

3. Utilitarian fairness scheduling algorithm in multiple wireless channel networks
   Existing studies have so far focused on the design of scheduling algorithm in the single wireless communication network under the fairness constraint. A common assumption of existing designs is that only a single user can access the channel at a given time slot. However, spread spectrum techniques are increasingly being deployed to allow multiple data users to transmit simultaneously on a relatively small number of separate high-rate channels. Not much work has been done on the scheduling algorithm in the multiple wireless channel networks. Furthermore in wire-line network, when a certain amount of resource is assigned to a user, it guarantees that the user gets some amount of performance, but in wireless network this point is different because channel conditions are different among users. Hence, in wireless channel the user's performance does not directly depend on its allocation of system resource. Finally the opportunistic scheduling mechanism for wireless communication networks is gaining popularity because it utilizes the "multi-user diversity" to maximize the system performance. So, considering these three points in the fourth section, we propose utilitarian fairness scheduling algorithm in multiple wireless channel networks. Utilitarian fairness is to guarantee that every user can get its performance requirement which is pre-defined. The proposed criterion fits in with wireless networks. We also use the opportunistic scheduling mechanism to maximize system performance under the utilitarian fairness constraint. Simulation results show that the new scheme works well in both utilitarian fairness and utilitarian efficiency of system resource in the multiple wireless channel situation.

Preface

The research work presented in this dissertation was performed by Mr Rui Yin, under the supervision of Prof. Dawoud Dawoud and Dr. Hongjun Xu, at the University of KwaZulu-Natal's school of Electrical, Electronic and Computer Engineering, in the Centre of Radio Access and Rural Technologies. This work was partially sponsored by Telkom South Africa Ltd and Alcatel Altech Telecoms as part of the Centre of Excellence programme.

The entire dissertation, unless otherwise indicated, is the author's original work and has not been submitted in part, or in whole, to any other universities for degree purposes.

Acknowledgments

I would like to express my sincere appreciation to my supervisor, Prof. Dawoud Dawoud and Dr HongJun Xu, for their excellent supervision and academic support during my research. Without their insightful advice and guidance, the accomplishment of this dissertation would not have been possible. Their brilliant ideas have directly contributed to the key chapters of this dissertation.

I would also like to thank Prof. F. Takawira for his helpful suggestions. Thanks to him for giving me a chance to study in this school, which is committed to academic excellence and innovation in research.

I am greatly indebted to my parents for their constant support and encouragement. They are the source of my inspiration and have been instrumental in my pursuit of higher education. Special thanks also go to my sister Xing Yin and brother Juke for their perennial love and care. 11-my fiancee, her presence is without a shadow of doubt one of the best things in my life.

I also want to thank my colleagues for the memorable time we have experienced as we work together. My communications with them will remain unforgettable. They all contributed to my development.

Finally, I am also grateful for the financial support received from Alcatel S.A. and Telcom during my MSc. study.

Contents

Contents vii

List of Figures x

List of Tables xiv

1 Introduction 1

1.1 Overview of resource allocation in wireless networks 2

1.2 Motivation 11

1.3 Dissertation overview 13

1.4 Original contribution in this dissertation 14

2 Scheduling algorithms in wirelss communication networks 16

2.1 Wireline Extension Wireless Scheduling Algorithms 16

2.2 The Framework of WEWS Algorithms [1] 18

2.2.1 Components of the Framework 18

2.2.2 Function for Each Components 19

2.2.3 Related Work 21

2.3 Opportunistic Scheduling Algorithms 25

2.3.1 Quality of Service 26

2.3.2 The Framework of Opportunistic Scheduling Algorithms 27

2.3.3 Related Work 32

2.4 Conclusion 35

3 Temporal fairness opportunistic scheduling algorithms 36

3.1 Temporal Fairness Opportunistic Scheduling Algorithm in Long Term (TFOL) 37

3.1.1 Problem Formulation 37

3.1.2 An Optimal Policy 38

3.1.3 Simulation Results 39

3.1.4 Conclusion 54

3.2 Temporal Fairness Opportunistic Scheduling Algorithm in Short Term (TFOS) 54

3.2.1 A New Policy 55

3.2.2 Simulation Results 57

3.2.3 Conclusion 60

4 Delay-concerned opportunistic scheduling algorithm 63

4.1 System Model 64

4.2 The Temporal Fairness Opportunistic Scheduling Algorithm of Long Term and EDF Scheduling Algorithm 65

4.2.1 Definition of algorithm 66

4.2.2 Simulation results 67

4.2.3 Discussion 69

4.3 Delay Concerned Opportunistic Scheduling 70

4.3.1 A New Scheme 70

4.3.2 Simulation Results 71

4.3.3 Discussion 76

4.4 Conclusion 77

5 Opportunistic scheduling algorithm under the utilitarian fairness constraint in multiple wireless channel system **79**

5.1 Utilitarian Fairness and System Model 80

5.1.1 Utilitarian Fairness 80

5.1.2 System Model 81

5.2 Problem Formulation 82

5.3 A New Scheduling Policy 83

5.4 Parameter Estimation 85

5.5 Simulation Results 87

5.5.1 Simulation Procedure 87

5.5.2 Simulation Results 89

5.6 Conclusion 95

6 Conclusion and Future work **97**

6.1 Conclusion 97

6.2 Future Work 100

List of Figures

1.1 Scheduling diagram in downlink system 8

1.2 Three users' time-varying SINR 11

2.1 Generic framework for WEWS algorithms 19

2.2 Generic framework of opportunistic scheduling algorithms 27

3.1 Two-tier cell structure 43

3.2 Users' performance values vs SINR 44

3.3 Time slots allocated to eight users 45

3.4 Users' performance under TFOL, round robin scheduling algorithm, greedy algorithm respectively 46

3.5 System's performance achieved by TFOL, round robin scheduling algorithm, greedy algorithm respectively 46

3.6 System performance when the number of users is 4, 6, and 8 respectively . . 47

3.7 Users' performance in TFOL, round robin, greedy scheduling algorithm respectively. 50

3.8 System performance when number of users is changed 51

3.9 Time slots allocated to users 52

3.10 Fairness deviation 53

3.11 Starvation time for Group 1 and 4 53

3.12 Implementaton procedures of the example in new algorithm 56

3.13 Fairness deviation in 1000000 time slots 59

3.14 Fairness deviation in 10000 time slots 59

3.15 Starvation time v.s. time slots 60

3.16 Users' performance in TFOL, round robin policy and the new scheme respectively. 61

3.17 Fairness deviation vs. time window size 61

3.18 The performance gain by the new scheme compared to the round robin policy. 62

4.1 Five State-Markov Chain Model 66

4.2 The packet delay violation probability of user 1 with the best channel condition and user 16 with the worst channel condition in TFOL when the deadline is 700 ms. 68

4.3 The packet delay violation probability of user 1 with best channel condition and user 16 with the worst channel condition in the EDF scheduling algorithm when the deadline is 700ms. 68

4.4 The packet drop ratio of user 1 and user 16 when the deadline changes from 0 to 1000ms in TFOL algorithm. 69

4.5 The packet delay violation probability of user one who has the best channel condition and user sixteen who has the worst channel condition in TFOL algorithm and new scheme when the deadline is 700, 800, 900, 1000 ms respectively. 72

4.6 The packet drop ratio of user one and user sixteen when the deadline changes from 0 to 1000 in TFOL algorithm and new scheme respectively. . 73

4.7 The number of time slots allocated to sixteen users in TFOL algorithm and the new scheme respectively when the deadline is 700 ms, the left bar is by TFOL algorithm and right bar is by the new scheme. 74

4.8 The fairness deviation factor η in TFOL algorithm and new scheme under the deadline from 100 to 1000. 75

4.9 User's performance (average throughput) by TFOL algorithm, EDF, the new scheme under the deadline 700ms, respectively, the middle bar is by EDF, the left bar is by TFOL and the right bar is by the new scheme. . . . 76

4.10 System throughput in TFOL algorithm, EDF, the new scheme respectively, x axis is packet deadline, y axis is system throughput. 77

5.1 Five State-Markov Chain Model 82

5.2 User average performance when $\sum_{i=1}^{N} = 0.72,\ a = 0.28$ 90

5.3 User average performance when $\sum_{i=1}^{N} = 0.76,\ a = 0.24$ 91

5.4 User average performance when $\sum_{i=1}^{N} = 0.84,\ a = 0.16$ 91

5.5 User average performance when $\sum_{i=1}^{N} = 0.84,\ a = 0.16$ 92

5.6 System average performance when tuning factor changes 93

5.7 System performance when user one's and users sixteen's performance requirements are increased, respectively. 94

5.8 Resource Utilization Efficiency (The bottom bar is the efficiency in Round Robin policy. The top bar is the efficiency improvement of our scheme compared to Round Robin policy). 95

List of Tables

3.1 The central point coordinates of the base station 42

3.2 Gaussian process parameters 49

4.1 Users' Mean Power Consumption per unit data rate (N=16) [2] 65

5.1 Users' Mean Power Consumption per unit data rate (N=16) [2] 81

5.2 User's performance requirement φ 90

5.3 User's performance requirement φ 92

5.4 User's performance requirement φ 93

5.5 User's performance requirement φ 94

List of Acronyms

1G	The First Generation Wireless Communication System
2G	The Second Generation Wireless Communication System
3G	The Third Generation Wireless Communication System
QoS	Quality of Service
FDMA	Frequency Division Multiple Access
TDMA	Time Division Multiple Access
CDMA	Code Division Multiple Access
T-CDMA	Time Code Division Multiple Access
GSM	Global Standard for Mobile Communication
PDC	Personal Digital Cellular
WCDMA	Wideband CDMA
CAC	Call Admission Control
SINR	Signal to Interference And Noise Ratio
DRC	Data Request Channel
PSMM	Pilot Strength Measurement Message
SCRM	Supplemental Channel Request Message
ARQ	Automatically Repeat Request
BER	Bit Error Rate
GPS	Generalized Processor Sharing
WEWS	Wireline Extension Wireless Scheduling
TFOL	Temporal Fairness Opportunistic Scheduling Algorithm of Long Term
TFOS	Temporal Fairness Opportunistic Scheduling Algorithm of Short Term
EDF	Earliest Deadline First WPS Wireless Packet Scheduling
CIF-Q	Channel Condition Independent Fair Scheduling Algorithm
SBFA	Server Based Fairness Approach
WFS	Wireless Fair Service Algorithm

HDR	High Data Rate
HOL	Head of Line
MLWD	Modified Largest Weighted Delay First
RR	Round Robin
WRR	Weighted Round Robin

Chapter 1

Introduction

Wireless communication is one of the most active areas of technology development of our time. This development is being driven primarily by the transformation of what has been largely a medium for supporting voice telephony into a medium for supporting other services, such as the transmission of video, images, text, and data. The demand for new wireless capacity is growing at a very rapid pace. Although there are, of course, still a great many technical problems to be solved in wireline communications, demands for additional wireline capacity can be fulfilled largely with the addition of new private infrastructure, such as additional optical fiber, routers, switches, and so on. On the other hand, the traditional resources that have been used to add capacity to wireless systems are radio bandwidth and transmitter power. Unfortunately, these two resources are among the most severely limited in the deployment of modern wireless networks: radio bandwidth because of the very tight situation with regard to useful radio spectrum, and transmitter power because mobile and other portable services require the use of battery power, which is limited. However, over the last two decades, there has been a tremendous growth in the number of users of wireless communication networks. These users are becoming increasingly sophisticated and require various services that can provide performance guarantees or Quality of Service. Hence, there is a substantial interest in providing high data rate services with heterogeneous QoS requirements. Compared with voice services, these services have complicated characteristics [3] [4]. This, coupled with a rapid growth in the demand

for wireless services, makes it difficult to estimate the required capacity accurately, due to the highly bursty characteristic of these services and occasional congestion within the network. Unfortunately, the system resource is simply not growing or growing at rates that can support anticipated demands of wireless capacity. Hence, appropriate resource allocation mechanisms are needed to design systems that are stable, efficient, and are able to provide users with QoS.

1.1 Overview of resource allocation in wireless networks

The first generation (1G) wireless networks are analog systems that used Frequency Division Multiple Access (FDMA) and accommodated only voice services. In the second generation (2G) wireless networks the digital technologies such as speech coding and bandwidth efficient modulation techniques are used. Both 1G and 2G wireless systems focus on voice services. The 3G and future wireless networks are designed to provide high data rate multimedia services. The services requiring high data rate in wireline networks such as video and email are expected to be supported in future generation networks as well. These services have diverse QoS requirements while most services in 1G and 2G wireless system are voice services that have the same QoS requirement. The voice service requires stringent delay bounds but relatively low data rate. File transmission service such as email do not require stringent delay bounds but stable data rate while video service needs high data rate and stringent delay bounds. Those services with different QoS requirements are expected to be provided in future wireless networks. Furthermore, services are highly asymmetric in the fourth and future generation wireless networks: downlink transmission is more important than uplink transmission. Hence, resource in downlink could become more precious. This implies that the efficient resource management of downlink is an important issue in future generation wireless networks. Also the various characteristics of services expected to be provided in the future wireless system make the resource management problem highly complex. Several key approaches have been proposed for resource management.

Multiple access

Multiple access techniques allow a communication medium to be shared among different users. In particular, three basic multiple access techniques, i.e., frequency-division multiple access (FDMA), time-division multiple access (TDMA), and code division multiple access (CDMA) [5], [6], are used in centralized networks. First generation systems are analog systems that use FDMA technique. TDMA and CDMA techniques are implemented in the second and third generation systems. In the second generation systems, the Global Standard for Mobile Communication (GSM) in Europe, the Personal Digital Cellular (PDC) in Japan, and the IS-136 in United States employ TDMA technique, while the IS-95 in United States uses CDMA technique. In the third generation systems, wideband CDMA (WCDMA) [7], [8], which evolves from GSM is specified in Europe and Japan while CDMA2000 [9], which evolves from IS-95, is specified in North America. The main goals of 3G systems are to provide universal access and global roaming, and to support high data rate multimedia services up to 2Mbps. In this dissertation we consider the TDMA system in chapter 3. In chapters 4 and 5, the T-CDMA is simulated.

Call admission control(CAC)

The challenges in the wireless networks are to guarantee quality of service requirements while taking into account the radio frequency spectrum limitations and radio propagation impairments. Call admission control is one method to manage radio resource in order to utilize radio resources efficiently and as many users as possible should be admitted into the system. However, if the number of admitted mobile users is too large, their requirement of QoS cannot be guaranteed. Hence, an objective of call admission control is to maintain a certain level of quality of service for existing users by admitting or rejecting requests of new arriving users, while admitting as many users as possible to maximize the system capacity.

In traditional FDMA and TDMA systems, CAC is simple, since there are a maximum number of channels that can be allocated to mobiles. When a new user arrives, the system only has to check the number of available channels. If there are a sufficient number of

available channels to accommodate the new call, the user is accepted. Otherwise, it is rejected. Empirical studies [10] have shown that a typical user is far more irritated when an ongoing call is dropped than a call blocked from the beginning. Thus, most of the effort of the CAC in FDMA and TDMA systems is devoted to handling handoff calls. Various handoff prioritizing schemes have been proposed in [11], [12]. These handoffs are closely related to CAC.

Contrary to FDMA and TDMA systems that have a "hard" capacity, CDMA systems have a "soft" capacity, since there is no concept of a maximum available number of channels in CDMA systems. The capacity in CDMA system is determined by the interference level. As more users are admitted, the interference to existing users is increased and the required QoS level of the calls may not be satisfactory. Thus, CAC in CDMA system can be treated as interference management. There has been a great deal of work done on CAC in CDMA systems [13], [14], [15].

Power control

Establishing and maintaining communication links is the most important function of radio resource management. In wireless systems, due to the propagation environment and the mobility of the users, the signal strength at the receiver level fluctuates significantly, and only by controlling transmitted power, communication links can be maintained at the desired QoS level. Power control is also used to decrease power consumption to increase battery life and decrease interference to other mobiles. Therefore, power control is an important resource management function in wireless networks, and other resource management schemes are closely related to it.

In FDMA [16] and TDMA based cellular networks, power control is used to manage co-channel interference improving resource reuse and increasing capacity. In CDMA cellular networks, power control is focused on balancing the Signal to Interference and Noise Ratio (SINR) of mobiles at the base station [17], [18], [19], [20]. Power control is an effective method to eliminate the near-far effect. The "near-far" effect is: if all mobiles transmit signal at the same level, mobiles that are closer to the base-station cause significant in-

terference to mobiles at the boundary of the cell and QoS requirements of mobiles at the boundary of the cell cannot be satisfied. In future generation wireless networks, services in the same network have diverse QoS requirements and, thus, power must be allocated to each mobile considering its own QoS requirements.

Handoff

Mobility is the most important feature of a wireless cellular communication system. Usually, continuous service is achieved by supporting handoff from one cell to another. Handoff is the process of changing the channel (FDMA-frequency, TDMA-time slot, CDMA-spreading code, or combination of them) associated with the current connection while a call is in progress. It is often initiated either by crossing a cell boundary or by a deterioration in quality of the signal in the current channel. Handoff is divided into two broad categories: hard and soft handoffs [21], [22], [23]. They are also characterized by "break before make" and "make before break". In hard handoffs, current resources are released before new resources are used; in soft handoffs, both existing and new resources are used during the handoff process. Poorly designed handoff schemes tend to generate very heavy signaling traffic and, thereby, a dramatic decrease in quality of service. The reason why handoffs are critical in cellular communication systems is that neighboring cells are always using a disjoint subset of frequency bands, so negotiations must take place between the mobile station (MS), the current serving base station (BS), and the next potential BS. Other related issues such as decision making and priority strategies during overloading might influence the overall performance [24], [25], [26], [27], [28].

Data rate adaptation

Voice-centric cellular systems are designed to provide good coverage for telephony services. In such a system, a minimum required signal to interference plus noise ratio (SINR) is guaranteed over at 90-95 percent of the coverage area. Contrary to voice services that require constant bit rates and stringent delay bounds, data services have diverse characteristics. Some of them are variable bit rate services and do not require stringent delay

bounds. Thus, such services can have variable bit rates for transmission. Generally, the required power to maintain the SINR at a fixed level increases with an increase in the data rate. For packet data service, larger SINR can be used to provide higher data rates by reducing coding or spreading and/or increasing the constellation density. Research shows that cellular spectral efficiency (in term of b/s/Hz/sector) can be increased by a factor of two or more if users with better links are served at higher data rates [29].

In TDMA wireless networks, data rate adaptation is done by time-slot aggregation, adaptive modulation, adaptive coding, and incremental redundancy. In a time-slot, by adjusting the modulation levels, the number of bits that can be transmitted in a symbol is adjusted. In general, the number of symbols that can be transmitted in a time-slot is fixed and, thus, the data rate can be adjusted by using adaptive modulation. Also, by adjusting the coding levels, the number of redundancy bits that are required to control the error can be adjusted, which implies that the number of information bits that are transmitted in a packet can be adjusted by using adaptive coding.

In CDMA rate adaptation is achieved through a combination of variable spreading, coding and code aggregation. Wideband CDMA (WCDMA) [5], [6] and CDMA2000 [9] systems achieve higher rates through a combination of variable spreading and coding. The WCDMA standards support data rates up to 2.048 Mbps in 5 MHz bands. In IS-95B code aggregation is used to support data rates up to 76.8 kbps. Variable spreading is achieved by adjusting the data rate while fixing the chip rate.

Research shows that fast rate adaptation is required to achieve high capacity on the fast fading channel [30]. In IS-856 [31], [32], the pilot bursts provide the mobile users with the means to estimate accurately and rapidly the channel conditions. Among other parameters, each mobile user estimates the received number of all resolvable multipath components and predicts the effective received SINR. This channel state information is then fed back to the base station via the reverse link data rate request channel (DRC) and updated as often as every 1.67ms.

Channel condition feedback is important to rate adaptation [30]. In CDMA systems, pilot strength measurements are used to estimate the SINR at the receiver. In IS-95B

and cdma2000, pilot strength measurements are provided to the base station through the pilot strength measurement message (PSMM) or included in the supplemental channel request message (SCRM). In both GPRS and EGPRS systems, measurement reports are included in supervisory automatically repeat request (ARQ) status messages. In TDMA systems, channel quality is estimated at the receiver and the information is provided to the transmitter through appropriately defined messages. The measurement report message in WCDMA additionally include block error rate, bit error rate (BER), received power, path loss, and downlink SINR measurements.

Scheduling

Scheduling controls the order of service for each individual user. Hence, it controls each individual service most directly. In wireline networks, scheduling schemes can be classified as work-conserving and nonwork-conserving schemes [33]. In a work-conserving scheme, if there is a packet to send in the queue, the server never idles [34], [35], [36], while in a nonwork-conserving scheme, even though there may be packets waiting to be transmitted in the queue, if there is no eligible packet to send, they are not served [37]. Hence, in general, a work-conserving scheme provides a higher average throughput and a lower average delay than a nonwork-conserving scheme. This is the reason why most efforts have so far been devoted to developing work-conserving scheduling schemes. Moreover, in many cases, the end-to-end delay bound is the more important QoS parameter than the average delay. Hence, there have recently been several proposals for the nonwork-conserving scheduling scheme [38]. The scheduling structure in downlink system of wireline network is shown in Fig. 1.1

In this dissertation we focus on the downlink system in cell-structured wireless networks. In Fig. 1.1.1 each user has a buffer in the base station. When users request the data from base station (n is the number of users), data/packets will be reserved in their corresponding buffers in the base station. Then the scheduler decides which user will be served (in TDMA system in one time slot only one user can be selected, in CDMA system several users can be served at the same time). We also call the buffer flow, queue or user.

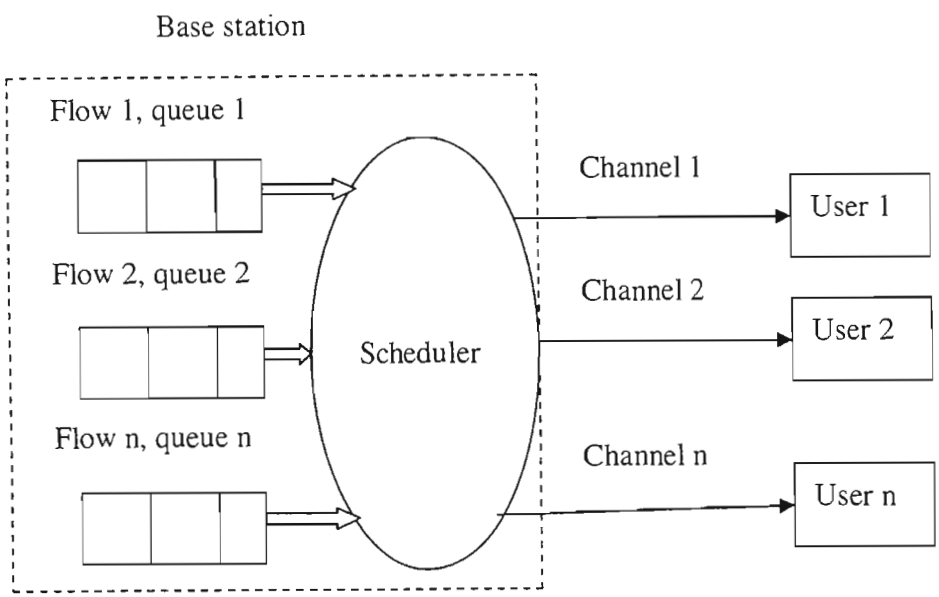

Figure 1.1: Scheduling diagram in downlink system

Scheduling algorithms which try to seek some fairness criterion are called fair scheduling, In the following, we are going to describe the fairness criterion in wireline networks briefly [34], [35].Consider a link that is being shared by a set F of data flows (users).Consider also that each flow $f \in F$ has a rate weight σ_f Each time instant t the rate allocated to a backlogged flow (nonempty flow) f is $\frac{\sigma_f \times C(t)}{\sum_{i \in B(t)} \sigma_i}$,where $B(t)$ is the set of nonempty queues and $C(t)$ is the link capacity at time t. Scheduler serves backlogged flows in proportion to their rate weights. Specifically, for any time interval $[t_1, t_2]$ during which there is no change in the set of backlogged flows $B(t_1, t_2)$, the channel capacity granted to each flow i, $W(t_1, t_2)$, satisfies the following property:

$$\forall i, j \in B(t_1, t_2), \left| \frac{W_i(t_1, t_2)}{\sigma_i} - \frac{W_j(t_1, t_2)}{\sigma_j} \right| = 0$$

The above definition of fair scheduling (fair queueing) is applicable to both channels with constant capacity and channels with time varying capacity. Since packet switched networks allocate channel access at the granularity of packets rather than bits, packetized scheduling algorithms must approximate the fluid model.The goal of a packetized scheduling algorithm is to minimize $\left| \frac{W_i(t_1,t_2)}{\sigma_i} - \frac{W_j(t_1,t_2)}{\sigma_j} \right|$ for any two backlogged flows i and j over an arbitrary time window $[t_1, t_2]$. This fairness criterion is also called Generalized Processor Sharing (GPS) fairness. In wireline there are three services that are sought to

satisfy:

1. **Fairness among backlogged flows.**
2. **Bounded delay channel access.**
3. **Throughput guarantee.**
4. **Provide full separation between flows: flows (users) are unaffected by the behaviour of other flows (users).**

In Fig. 1.1 users' channel conditions are the same and constant all the time because it is a wireline system. But in wireless system the channel condition is time varying. So the scheduling algorithms in wireline networks cannot be adapted to wireless systems directly because there are unique characteristics in the wireless system. Some of these characteristics are:

- Location-dependent and time-varying wireless link capacity

  In wireless networks, the channel conditions of mobile users are time varying and location-dependent. It is well known that radio signals propagate according to three mechanisms: reflection, diffraction, and scattering [10]. In free space, signal strength decays with the square of the path length. The signal received by a mobile user is a superposition of time-shifted and attenuated versions of transmitted signal. Radio propagation are related to three near independent phenomena: path-loss variation, slow log-normal shadowing, and fast multipath-fading. Path loss is caused by dissipation of the power radiated by the transmitter. Path loss models generally assume that path loss is the same at a given transmit-receive distance [39]. Shadowing is caused by obstacles between the transmitter and receiver that absorb power [40], [41]. When the obstacle absorbs all the power, the signal is blocked. Variation due to path loss occurs over very large distances (100-1000 meters), whereas variation due to shadowing occurs over distances proportional to the length of the obstructing object (10-100 meters in outdoor environments and less in indoor environments). Since variations due to path loss and shadowing occur over a relatively large distance, this

variation is sometimes referred to as large-scale propagation effects or local mean attenuation. Variation due to multipath occurs over very short distances, on the order of the signal wavelength, so these variations are sometimes referred to as small-scale propagation effects or multipath fading. Furthermore, a user receives interference from other transmissions, which is time-varying; and background noise is also constantly varying. Hence, users' channel conditions are location dependent and time varying.

- Multiuser diversity

  Mutiuser diversity is a form of diversity inherent in a wireless network, provided by independent time-varying channels across the different users. As we described in last section user's channel conditions are location-dependent and time-varying. Hence, if there are many users in the same cell, because of the mobility and different positions in the cell, users will experience different channel conditions. Signal to interference plus noise ratio (SINR) is a commonly used measure of channel conditions. Fig. 1.2 shows the time varying SINR of three mobile users.

- If the same resource is given to different users, the resultant network performance (e.g. throughput) could be different from user to user.

  In wireline networks when a certain amount of resource is assigned to a user, it is equivalent to granting the user a certain amount of throughput/performance value. However, the situation is different in wireless networks. For example, consider if the same amount of resource (power, time-slots, etc.) is allocated to user 2 and user 3 in Fig. 1.2 from 0 second to 10 second, it is likely that the throughput of user 2 will be much larger than that of 3. This is because during this time user 2 has a much better channel condition than user 3. So there is no direct relation between resource assignment and performance of users.

- channel errors are location-dependent and bursty in nature.

  Scheduling algorithms in wireless networks should take these factors into account.

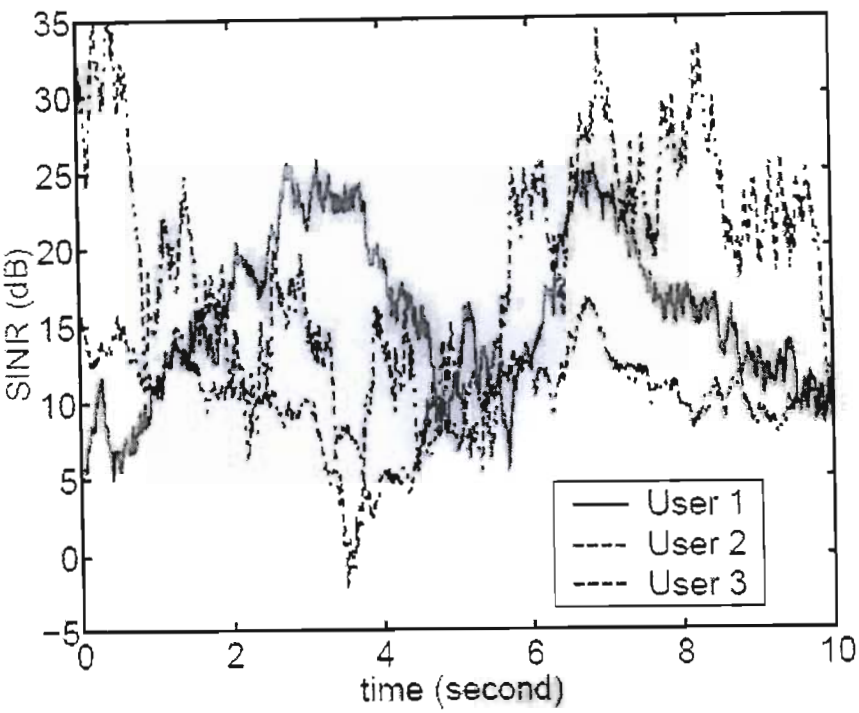

Figure 1.2: Three users' time-varying SINR

1.2 Motivation

Efficiently utilizing the scarce system resource to satisfy different QoS requirement of users is one of the most important issues in future generation wireless networks. Scheduling algorithm plays a critical role in resource allocation. Hence, in this dissertation we study scheduling algorithms in wireless communication networks. In the last section we showed that the scheduling schemes from the wireline domain cannot carry over to wireless systems because wireless channel has unique characteristics. These characteristics can be exploited to improve the efficiency of the scheduling algorithm.

Because channel conditions are time-varying, if users are served when they experience better channel conditions, they will have higher quality of service/performance. Research shows that cellular spectral efficiency (in terms of b/s/Hz/sector) can be increased by a factor of two or more if users with better links are picked up at a higher data rate [31]. Hence, a good scheduling algorithm in wireless should be able to exploit the variability of

channel conditions to achieve higher utilization of the resource.

Different users have independent time-varying channels in wireless networks, which is called multiuser diversity. To maximize the system performance/throughput, Knopp and Humblet [42] have shown that the optimal strategy is to schedule at any one time only the user with the best channel condition to transmit to the base station. This is based on the fact that in a wireless system with many users, whose channel conditions vary independently, there is likely to be a user whose channel condition is near its peak at any one time. Overall system throughput is maximized by allocating at any time the common channel resource to the user who can best utilize it. Hence a good scheduling algorithm should take advantage from multiuser diversity.

However, although the system performance/throughput is maximized by only scheduling the user with the best channel condition , it will incur unfairness. This is because some users always experience good channel conditions and some users always experience bad channel conditions, like user 1 and user 3 in Fig. 1.2. Hence, for maximizing system performance, only selecting the users with better channel conditions will starve users with bad channel conditions from resource access. For example, in Fig. 1.2 user 3 has worse channel condition than user 1 and user 2 all the time. If only scheduling users with good channel conditions, user 1 and 2 will be always selected. Hence, user 3's quality of service cannot be guaranteed. A good scheduling algorithm in wireless networks should be "fair" to all users. In our thesis two kinds of fairness criterion are introduced: temporal and utilitarian [43], [44]. Temporal fairness is that each user gets a fair share of system resource, and utilitarian fairness means that each user gets a certain share of overall system performance. We will discuss them in detail in chapter 3 and chapter 5.

Two classes of scheduling algorithms which take these characteristics into account have been proposed. One is named wireline extension wireless scheduling (WEWS) algorithm [45], [46], [47], [48], [49] and the other one is opportunistic scheduling algorithm [50], [51], [52], [53] , [54], [55](we will describe them in detail in chapter 2). In wireline extension wireless scheduling algorithms the characteristics of wireless channel are treated as a negative factor which should be subdued. On the other hand, the op-

portunistic scheduling algorithms utilize wireless characteristics to improve the system performance. So we choose the latter one as our topic. A great deal of work has been done on scheduling algorithms in wireless system, but none of them is perfect. Hence, based on the disadvantages of these algorithms, we propose a new scheme.

1.3 Dissertation overview

The outline for the remainder of this dissertation is as follows. In chapter 2, two classes of scheduling algorithms are described. We start describing the structure for each class of scheduling algorithm. Then we explain the function of each component in the structure. Following the description of each structure, a literature survey of corresponding scheduling algorithm is given. We analyze these algorithms according to the structures. The general formulation of opportunistic scheduling problems is also given.

In chapter 3, temporal fairness opportunistic scheduling algorithm of long term (TFOL) given in [43] is considered. We simulate this algorithm in two environments. The first one is an actual cell in which path loss and shadowing are taken into account to calculate users' channel conditions. In the second simulation environment, users' channel conditions are assumed to be time-correlated Gaussian process with different mean and variance. From the simulation results, we observe TFOL algorithm is not fair in the short term. So we present a new scheme called term temporal fairness algorithm of short term (TFOS). Some analysis of this algorithm is given. Then through the simulation we compare it to TFOL to examine how it works in the short term on temporal fairness.

In chapter 4, firstly the system model is given. Then the simulation results of packet delay distribution and packet drop ratio in temporal fairness opportunistic scheduling algorithm of long term (TFOL) are given. We also simulate earliest deadline first (EDF) scheduling algorithm works in wireless environment. Simulation shows that in TFOL users with good channel conditions have much better performance on packet delay than users who always experience relatively bad channel conditions. On the other hand, under the EDF algorithm, packet delay distribution of users is almost the same, but it is worse than

the opportunistic scheduling algorithm because it does not consider users' channel conditions. So we propose a new scheme called "channel concerned opportunistic scheduling algorithm" which considers both users channel conditions and packet delay. Then the simulation result on packet delay, temporal fairness and users' performance are given.

In chapter 5, we consider the opportunistic scheduling problem in multiple wireless channel networks. Firstly, we introduce our system model and formulate utilitarian fairness criterion mathematically. Then we present this scheduling problem mathematically. After this, an optimal algorithm is given and several properties of this algorithm are discussed. Then we explain how to update the fairness related parameter. Lastly, we present simulation results.

Finally, conclusions are drawn and topics for future work are discussed in Chapter 6.

1.4 Original contribution in this dissertation

The original contributions is this dissertation include:

1. In chapter 2, we present a general structure of opportunistic scheduling algorithm. We also formulate the general opportunistic scheduling problem mathematically in both the long and short term.

2. In chapter 3, we propose a new scheme named "temporal fairness opportunistic scheduling algorithm in short term (TFOS)". Some analysis is done on this scheme. Simulation shows that our new scheme satisfies the temporal fairness of short term, while exploiting wireless channel to improve the system performance.

3. In chapter 4, we examine the TFOL and earliest deadline delay first (EDF) algorithm on delay performance on packet level. Through the simulation results, we observe that TFOL does not take packet delay into account and EDF does not consider users' channel conditions. So a new scheme, which takes both users' channel conditions (in opportunistic way) and packet delay into account called "delay-concerned opportunistic scheduling algorithm", is proposed. By analysis, we know that in the new

scheme packet delay factor and channel conditions have the same power. Through the simulation results, we observe that the new scheme not only balances packet delay distribution of different users, but also exploits the characteristic of wireless channel to improve the system performance.

4. In chapter 5, utilitarian fairness criterion is considered. And an opportunistic scheduling algorithm under the utilitarian fairness constraint in multiple wireless channel system is proposed. Properties of this algorithm are given by analysis. Simulation results show that the new scheme fulfils utilitarian fairness criterion, while utilizing system resource efficiently to improve system performance.

Parts of this dissertation have been presented and submitted for the following conference and journals.

- R. Yin, D. Dawoud, Hj. Xu, "Opportunistic Scheduling Algorithms in Wireless Communication Networks", Proceeding of IEEE International Conference on Telecommunication (IEEE ICT) 2005, Cape town, South Africa, May, 2005.
- R. Yin, D. Dawoud, Hj. Xu, "Delay Concerned Opportunistic Scheduling Algorithm in Wireless Communication Networks", Poceeding of South African Telecommunications Networks and Applications Conference (SATNAC 2005), Champagne Sports, Drakensberg, South Africa, Sept 2005.
- R. Yin, D. Dawoud, Hj, Xu, "Utilitarian Fairness Constraint Opportunistic Scheduling Algorithm in Multiple Wireless Channel System", Submitted to IEEE 2005 Global Communications Conference (Globalcom 2005). St. Louis, Missouri, USA, November 2005, not published.
- R. Yin, D. Dawoud, Hj, Xu, "Delay Concerned Opportunistic Scheduling Algorithm in Wireless Communication Networks", Submitted to the SAIEE Transactions, under review.

Chapter 2

Scheduling algorithms in wirelss communication networks

In wire-line network, scheduling algorithm has long been a popular paradigm for achieving instantaneous fairness and bounded delays in channel access. However, adapting wireline scheduling algorithms to the wireless domain is nontrivial because of the unique characteristics of the wireless channel, such as location and time dependancy, channel contention and multiuser diversity. Consequently the scheduling algorithms for wireline networks do not apply directly to wireless networks.

2.1 Wireline Extension Wireless Scheduling Algorithms

Several wireline scheduling algorithms have been developed [45], [46], [47], [48], [49] for adapting the wireless domain, so we call them wireline extension wireless scheduling algorithms(WEWS). In this class of scheduling algorithms the wireless links between a base station and each of the mobile hosts are independent. Furthermore a two State-Markov channel model is used for the state of a wireless link, which is in either one of the two states: good (or error-free state or bad (or error) state. In a good state, the wireless link is assumed to be error-free and it works like a wireline link. If a link is in a bad

state, data cannot be transmitted on the link at all. The goal of this class of scheduling algorithms is to make short bursts of location-dependent channel error transparent to users by a dynamic reassignment of channel allocation over small time scales. The idea is to swap channel access between a backlogged flow (user) that perceives channel error and backlogged flows (users) that do not, with the intention of reclaiming the channel access for the former when it perceives a good channel.

Wireless scheduling seeks to provide the same service to flows in a wireless environment as traditional scheduling algorithm does in wireline environment. This implies providing bounded delay access to each flow and providing full separation between flows. However, in the presence of location-dependent channel error (due to different physical locations, some mobile hosts may enjoy error-free communication with the base station, while others may not be able to pick up communications with the base station at all, this is a so-called location-dependent error), the ability to provide short-term fairness will be violated. Channel utilization can be significantly improved by swapping channel access between error-prone and error-free flows at any time. This will provide long-term fairness but not instantaneous fairness. Since we need to compromise a complete separation (the degree to which the service of one flow is unaffected by the behaviour and channel conditions of another flow) between flows in order to improve efficiency, wireless scheduling necessarily provides a somewhat less stringent quality of service than wireline scheduling. In [1] the author defined the quality of service that this class of scheduling algorithms typically seeks to satisfy:

1. **Short-term fairness among flows that perceive a clean channel and long-term fairness for flow with bounded channel error (comparison with the error-free channel, the error-prone flows can catch up when their channels are error-free).**
2. **Delay bound for packets.**
3. **Short-term throughput bounds for flows with error-free channels and long-term throughput bounds for all flows with bounded channel error.**
4. **Support for both delay-sensitive and error-sensitive data flow.**

The short term fairness ensures that channel allocation is fair among backlogged flows that are able to transmit packets (good channel conditions). The long term fairness further specifies that even if a flow has received additional service in a previous time window, its degradation of service in any subsequent time window must be graceful, i.e. a flow that has received excess service in the past must not be starved of channel access at any time in the future. The delay bound is subject to the fact that channel error is bounded for any flow over some time period. Property four is very useful for handling both delay sensitive and error sensitive flows (users).

We define the error-free service of a flow as the service that it would have received at the same time if all channels had been error-free, under identical offered loads. A flow is said to be leading if it has received channel allocation in excess of its error-free service. A flow is said to be lagging if it has received channel allocation less than its error-free service. If a flow is neither leading nor lagging, it is said to be "in sync", since it's channel allocation is exactly the same as its error-free service.

2.2 The Framework of WEWS Algorithms [1]

In [1] the author presented a generic framework for WEWS algorithms and identified the key components of the framework.

2.2.1 Components of the Framework

As shown in Fig. 2.1, this class of wireless scheduling algorithms involves the following five components:

1. Error-free service model: defines an ideal service model assuming no channel errors (works like in wireline system). This is used as a reference model for channel allocation.

2. Lead and lag model: determines which flows are leading or lagging their error-free service, and by how much.

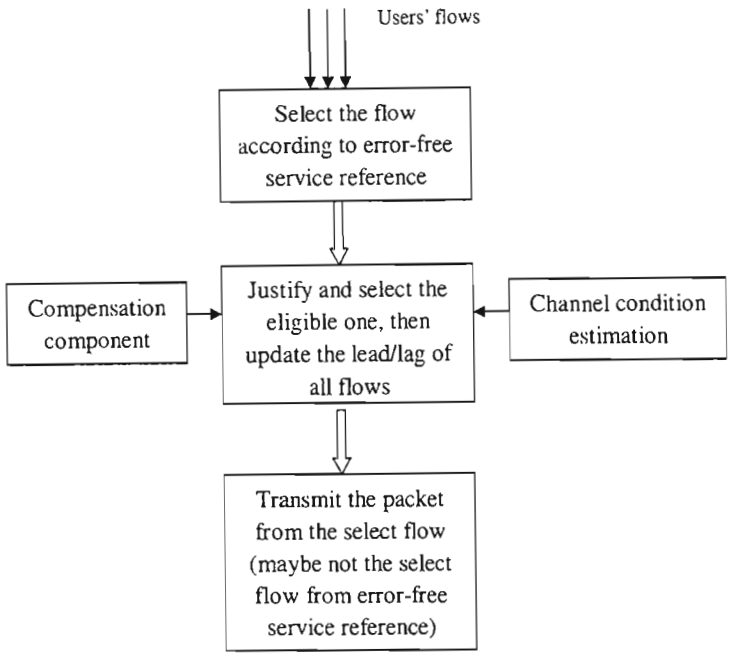

Figure 2.1: Generic framework for WEWS algorithms

3. Compensation model: compensates lagging flows that perceive an error-free channel at the expense of leading flows, and thus addresses the key issues of bursty and location-dependent channel error in wireless channel access.

4. Slot queue and packet queue decoupling: allows for the support of both delay-sensitive and error-sensitive flows in a single framework and also decouples connection-level packet management policies from link-level packet scheduling policies.

5. Channel monitoring and prediction: provides a (possibly inaccurate) measurement and estimation of the channel state at any time instant for each backlogged flow.

2.2.2 Function for Each Components

In this section we explain the function of each component in the framework, as given in [1].

- **Error-free service model**

  The error free service model provides a reference for how much service a flow should

receive in an ideal error-free channel environment. As mentioned before, the goal of WEWS algorithms is to approximate the error-free service model by making short channel errors transparent to a user, and only exposing prolonged channel errors to the flow.

- **Lead and lag model**

  The leading flows are the users with the actual received services more than idealized service (error-free service) and the lagging flows are the users with actual received service less than the idealized service. The author in [1] defines the lag of a lagging flow as the amount of additional service to which it is entitled in the future in order to compensate for lost service in the past, whereas the lead of a leading flow as the amount of additional service that the flow has to relinquish in future in order to compensate for additional service received in the past.

- **Compensation model**

  The compensation model is the key component of wireless scheduling algorithms. It determines how lagging flows make up their lag and how leading flows give up their lead. Leading flows are required to give up some of the slots that are allocated to them in error-free service so that lagging flows can use these slots to reduce their lag.

- **Slot queues and packet queues**

  Wireline scheduling algorithms assign tags to packets as soon as they arrive, which works well if we assume no channel error, i.e., a scheduled packet will always be transmitted and received successfully. However, in a wireless channel, packets may be corrupted due to channel error, and an unsuccessfully transmitted packet may need to be retransmitted for an error-sensitive flow. Retagging the packet will cause it to join the end of the flow queue and thus cause packets to be delivered out of order.

  Fundamentally, there needs to be a separation between "when to send the next packet," and "which packet to send next." The first question should be answered by the scheduler, whereas the second question is really a flow-specific decision and should be beyond the scope of the scheduler. In order to address these two questions,

one additional level of abstraction can be used in order to decouple "slots", the units of channel allocation, from "packets", the units of data transmission. When a packet arrives in the queue of a flow, a corresponding slot is generated in the slot queue of the flow and tagged according to the wireless scheduling algorithm. At each time, the scheduler determines which slot will get access to the channel, and the head of line packet in the corresponding flow queue is then transmitted. The number of slots in the slot queue at any time is exactly the same as the number of packets in the flow queue.

2.2.3 Related Work

In the previous section, we described the framework for the wireline extension wireless scheduling algorithms which is first discussed in [4]. In this section we will use the framework to analyze four representative algorithms. The four algorithms chosen are the wireless packet scheduling algorithm [56] , the channel-condition-independent fair scheduling algorithm (CIF-Q) [47], the server-based fairness approach (SBFA) [48] and the wireless fair service algorithm (WFS) [46]. These algorithms are also analyzed in [1].

1. Wireless Packet Scheduling (WPS) [46]

   The components of WPS scheduler are:

   - **Error-free Service Model:** the error-free model of WPS uses a variant of weighted round robin and WFQ [57], and is called WRR with spreading.
   - **Lead and Lag model:** in WPS, the lead and lag of a flow are used to adjust the weights of the flow in the WRR spreading allocation. The lead is treated as negative lag. Thus, WPS generates a "frame" of slot allocation from the WRR spreading algorithm. At the start of a frame, WPS computes the effective weight of a flow as the sum of its default weight and its lag, and resets the lag to 0. The frame is then generated based on the effective weights of flows. The lag and lead are bounded by a threshold.
   - **Compensation Model:** in WPS, in each slot of a frame, if the flow that is allocated to the slot is backlogged but perceived as error channel, then WPS

tries to swap the slot with a future slot allocation within the same frame. If this is not possible, then WPS increments the lag of the flow if another flow can transmit in its place, and the lead of this new alternate flow is incremented.

The lag/lead accounting mechanism described above maintains the difference between the real service and the error-free service across frames. By changing the effective weight in each frame depending on the result of the previous frames, WPS tries to provide additional service to lagging flows at the expense of leading flows. In the ideal case, in-sync flows are unaffected at the granularity of frames, though their slot allocations may change within the frame.

Disadvantage: although this algorithm does not disturb the in-sync flows, it also can starve the leading flows when the lagging flows begin to perceive a clean channel. So this algorithm does not provide a graceful linear degradation service for leading flows. Furthermore it provides poor short-term fairness guarantees.

2. Channel Independent Fair Scheduling Algorithm (CIF-Q) [47]
   The components of CIF-Q scheduler are stated as followings:

   - **Error-free Service Model:** in CIF-Q, the error-free service is simulated by STFQ (start time fair queueing) [58]. The lag or lead of a flow is maintained just as in IWFQ. In other words, the lag of a backlogged flow is incremented only when some other flow is able to transmit in its place. Lead is maintained as negative lag.
   - **Lead and Lag model:** when a lagging or in-sync flow i is allocated the channel, it transmits a packet if it perceives a clean channel. Otherwise, if there is a backlogged flow j that perceives a clean channel and transmits instead of i, then the lag of i is incremented and the lag of j is decremented.
   - **Compensation Model:** a leading flow i retains a fraction α of its service and relinquishes a fraction $1 - \alpha$ of its service, where α is a system parameter that governs the service degradation of leading flows. When a leading flow relinquishes a slot, it is allocated to the lagging flow with a clean channel and the largest normalized lag, where the normalization is done using the rate weight

of the flow. Thus, lagging flows receive additional service only when leading flows relinquish slots.

Disadvantages: although this algorithm overcomes one of the main drawbacks of the algorithms mentioned before (WPS) - the graceful degradation for leading flows. But sync flows may be disturbed during redistribution of channel allocations that cannot be used by lagging flows or the selected flow and a lagging flow may access the channel. The computational complexity is higher than WPS because it needs to compute the fraction of leading flows services to compensate the lagging flows in the time slot.

3. Server-Based Fairness Approach (SBFA) [48]

   The components of SBFA scheduler are stated as followings:

   - **Error-free Service Model:** SBFA provides a framework in which different wireline algorithms can be adapted to wireless domain. The error-free service in SBFA is the desired wireline scheduling algorithm that needs to be adapted to the wireless domain. For example, we can choose WFQ (weighted fair queueing) or WRR (weighted round robin) to be the error-free service.
   - **Lead and Lag model:** there is no concept of leading flows in SBFA and the lag of a flow is not explicitly bounded, and the order of compensation among lagging flows is according to the order in which their slots are queued in the LTFS.
   - **Compensation Model:** SBFA statically reserves a fraction of the channel bandwidth for compensating lagging flows. This reserved bandwidth is called a virtual compensation flow or a long-term fairness server (LTFS). When a backlogged flow is unable to transmit due to channel error, a slot request corresponding to that flow is queued in the LTFS. The LTFS is allocated a rate weight that reflects the bandwidth reserved for compensation. The scheduling algorithm treats LTFS the same as packet flows for channel allocation. When the LTFS flow is selected by the scheduler, the flow corresponding to the head-of-line slot in the LTFS is selected for transmission. Thus, in contrast with

other wireless fair scheduling algorithms, SBFA tries to compensate the lagging flows using the reserved bandwidth rather than swapping slots between leading and lagging flows. When the reserved bandwidth is not used, it is distributed among other flows according to the error-free scheduling policy. This excess service is essentially free since lead is not maintained.

Disadvantages: because this algorithm provides fairness guarantees as a function of the statically reserved LTFS bandwidth, the bounds are very sensitive to this reserved fraction. For example, a single flow could perceive many errors, thereby utilizing all the bandwidth of the LTFS flow. Other flows experiencing errors may not get enough compensation, resulting in unfair behaviour for the system. Another thing is since SBFA is designed, based on the reasoning that all flows whose wireless links are in a good state should always be served at its promised service rate and not a fraction of the promised rate, no restriction is imposed on flows receiving excessive service. Hence, a flow with a consistently good link may receive far more service than its promised share. In addition, when several flows share an LTFS, the rate at which these flows receive compensation is determined only by the service rate of LTFS and is independent of the flows' allocated service rates. The information of different service rate requirements of different flows is lost in the compensation process. Finally, the algorithm does not work well if the packet size of a flow is variable. To keep in-order transmission of a flow, a slot in LFTS is always associated with the HOL (head of line) packet of a flow. However, this HOL packet may not be the same packet with which the slot was originally associated.

4. Wireless Fair Service (WFS) [46]
   The components of WFS are:

   - **Error-free Service Model:** in WFS, the error free service is computed by the modified fair queueing algorithm described in error-free service model in order to achieve a delay-bandwidth decoupling in the scheduler. Unlike traditional fair queueing algorithms, WFS can support flows with high bandwidth and high delay requirements, as well as flows with low bandwidth and low delay requirements, due to the use of this modified scheduler.

- **Lead and Lag model:** the notion of lag and lead in WFS is the same as in CIF-Q. A flow can increase its lag only when another flow can transmit in its slot.
- **Compensation Model:** each flow i has a lead bound of l_i^{max} and a lag bound of b_i^{max}. A leading flow with a current lead of l_i relinquishes a fraction $\frac{l_i}{l_i^{max}}$ of its slots, whereas a lagging flow with a current lag of b_i receives a fraction $\frac{b_i}{\sum_{j \in S} b_j}$ of all the relinquished slots, where S is the set of backlogged flows. Effectively, leading flows relinquish their slots in proportion to their lead, and relinquished slots are fairly distributed among lagging flows.

Disadvantages: WFS achieves all the properties of the fair service model. It achieves both short-term and long-term fairness, as well as delay and throughput bounds, however high computational complexity and that the compensation for lagging flows takes longer than other algorithms are its problems.

2.3 Opportunistic Scheduling Algorithms

Channel fading is traditionally viewed as a source of unreliability that has to be mitigated. Information theory suggests an opposing view: channel fluctuations can instead be exploited by transmitting information opportunistically when and where the channel is strong. The theory has been translated into practice. A scheduling algorithm, which exploits the inherent multi-user diversity while maintaining fairness among users, has been implemented as the standard algorithm in Qualcomm's High Data Rate (HDR) system (1xEV-DO) [32] .The diversity benefit is exploited by tracking the channel fluctuations of the users and scheduling transmissions to users when their instantaneous channel quality is near the peak. In general, a user is served with better quality and/or a higher data rate when the channel condition is better. Hence, good scheduling schemes should be able to exploit the variability of channel conditions to achieve higher utilization of wireless resources. We call this class of scheduling algorithms opportunistic scheduling algorithms. By opportunistic, we mean the ability to exploit the variation of channel conditions. Consider a few users that share the same resource. The users have constantly varying channel

conditions, which imply constantly varying performance. The scheduling policy decides which user should transmit during a given time interval. Intuitively, we want to assign resource to users experiencing "good" channel conditions so that the resource can be used efficiently. At the same time, we also want to provide some form of fairness or QoS guarantees to all users. For example, allowing only users close to the base station to transmit with high transmission power may result in very high system throughput, but may starve other users, which incurs unfairness. So there is a tradeoff between system performance and the quality of service (QoS). In this section firstly we introduce some quality of service constraints. Then, the formulation of opportunistic scheduling algorithms will be presented. Finally the related work and the framework of opportunistic scheduling algorithms will be presented.

2.3.1 Quality of Service

There are two extra fairness constraints that the opportunistic scheduling algorithm seeks to satisfy, which are firstly presented in [43] [44].

1. Temporal Fairness (resource sharing fairness):we suppose that there are N users in a cell, each user i is assigned a fixed fraction of resource, denoted as ϕ_i, $0 \leq \phi_i \leq 1$ and $\sum_{i=1}^{N} \phi_i \leq 1$.

2. Utilitarian Fairness (system performance sharing fairness): we suppose that there are N users in our system, each user i is ensured to get at least a pre-allocation fraction φ_i of the system performance.

3. Generalized Processor Sharing Fairness (GPS) [34]: there are N users to request the service in the system and each user i has a weight ε_i. Let $S_i(\tau, t)$ be the amount of user i traffic served in the interval $[\tau, t]$. The GPS fairness is defined as $\frac{S_i(\tau,t)}{S_j(\tau,t)} \geq \frac{\varepsilon_i}{\varepsilon_j}$ for any user $i(i \neq j)$. GPS fairness constraint is same as that in (1.1).

In chapter 5 we will prove that GPS is a special case of utilitarian fairness constraint.

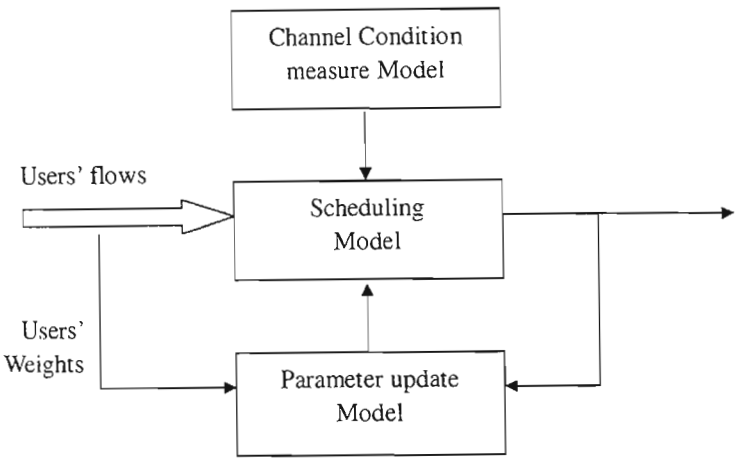

Figure 2.2: Generic framework of opportunistic scheduling algorithms

Delay bound of user packets is expressed as:

$$P_r(W_i > D_i) \leq \delta_i \tag{2.1}$$

where W_i is the steady state packet delay for user i, and parameters D_i and δ_i are the delay threshold and maximum probability of exceeding delay threshold, respectively. P_r is probability function.

2.3.2 The Framework of Opportunistic Scheduling Algorithms

In this section, we present generic structure for opportunistic scheduling algorithms introduced in [59](Fig. 2.2). In the following we identify the key components of the framework, and discuss the function of each component.

Components in the Framework

- Parameter update model: in this block users' parameter vector $\vec{v}(k) = [v_1(k), v_2(k), \ldots, v_N(k)]$ is updated according to users' weight and the users' parameter vector: $\vec{v}(k-1) = [v_1(k-1), v_2(k-1), \ldots, v_N(k-1)]$ in previous time slot, where N is the number of users. This parameter vector is related to the fairness constraint.

- Channel condition update model: in this model users' channel conditions are monitored and transmitted to the base station.
- Scheduling Model: this model decides which user/users should be served by the base station.

Function and Mechanism for Each Components

In this section we explain the function of each component in the framework and then introduce some mechanisms adopted for each component.

- Parameter update model:

  In opportunistic scheduling algorithms users' channel conditions are exploited to improve system performance. However, adapting the channel condition will easily lead to deviations from ideal fairness, memory of the scheduling decision history is required. Therefore, this model is to update the fairness related parameter vector $\vec{v}(k) = [v_1(k), v_2(k), \ldots, v_N(k)]$ in time slot k according to its previous value $\vec{v}(k-1) = [v_1(k-1), v_2(k-1), \ldots, v_N(k-1)]$ in time slot , users' performance distribution and the fairness constraint.

  The objective of opportunistic scheduling algorithms is to ensure fairness .while simultaneously exploiting users' channel conditions to increase the total system performance. So there are two conflicting goals (system performance optimization and fairness guarantees). It is an optimization problem. Many methods are employed to resolve this problem, e.g. adaptive method. Recently the stochastic algorithm is the most popular method [60]. In our dissertation we will use this algorithm to update the fairness related parameter.

- Channel condition measure model

  Channel conditions in wireless networks are time varying. The strategy of opportunistic scheduling is to improve the system performance by selecting user with high-quality channel when possible. So it is necessary to evaluate users' channel conditions before the scheduler makes a decision. In chapter 3 we use stochastic

model to capture user's channel conditions and in chapter 4 and 5 Markov chain is used to model the wireless channel. Normally there are three ways to measure users' channel conditions:

1. Signal to interference and noise ratio (SINR) can be used as a measure of channel conditions.
2. There is relationship between SINR and data rate (section 3.1, chapter 3) so in the second set of simulation in chapter 3 we employ users' data rate $\vec{r}(k) = [r_1(k), r_2(k), \ldots, r_N(k)]$, where N is number of users, to evaluate channel conditions. When a user experiences relatively good channel conditions, the data rate it has will be higher and vice versa.
3. The power requirement per unit data rate $\vec{c}(k) = [c_1(k), c_2(k), \ldots, c_N(k)]$, where N is the number of users, can also be used as an indication of a user's channel conditions. More power consumption per unit data rate means worse channel condition. In chapter 4 and 5 we measure channel conditions in this way.

- Scheduling model

  At the beginning of every time slot the scheduling model makes scheduling decisions based on the channel conditions and fairness related parameter. The output of this model is the decision of which user will be served in the following time slot.

  The objective of opportunistic scheduling is to optimize the system performance while providing some fairness constraint. Throughout the dissertation, we use throughput (in terms of bits/sec) as the system performance measure. The scheduling decision model has diversiform mechanism based on different QoS constraints. In the followings we consider, as examples, two mechanisms:

  1. For maximizing system performance while satisfying the temporal fairness constraint in long term [43], the problem is formulated mathematically as:

$$maximize \sum_{i=1}^{N} (E(r_i) \times I_{\{Q(\vec{r})=i\}}) \tag{2.2}$$

$$\text{Subject to} \quad E\{I_{\{Q(\vec{r})=i\}} \geq \phi_i \tag{2.3}$$

where $E()$ is and expectation function, $Q()$ is the scheduling policy, I is the indication function

$$I_b = \begin{cases} 1 & \text{if } b \text{ occurs,} \\ 0 & \text{otherwise} \end{cases}$$

$E(r_i) \times I_{\{Q(\vec{r})=i\}}$ is user $i's$ throughput in long term. We use throughput to stand for the system performance, so the summation of user throughput in 2.2 is the total system performance. This scheduling problem is subject to the temporal fairness constraint. So 2.3 means user i achieves at least ϕ_i fraction of system resource. In [43] the optimal policy $Q^*()$ is defined as:

$$Q^*(\vec{r}) = \underset{i}{argmax}(r_i + v_i) \tag{2.4}$$

Where $Q^*()$ stands for optimal scheduling policy, v_i is a parameter determined by the distribution of user's performance and fairness constraint.

2. For maximizing system throughput while satisfying the GPS (Generalized Processor Sharing) fairness constraint in the long term the problem is formulated mathematically as:

$$maximize \sum_{i=1}^{N} (E(r_i) \times I_{\{Q(\vec{r})=i\}}) \tag{2.5}$$

$$\text{Subject to} \quad \frac{E\{r_i \times I_{Q(\vec{r})=i}}{\sum_{i=1}^{N}(E(r_i) \times I_{\{Q(\vec{r})=i\}})} = \varepsilon_i \tag{2.6}$$

Similar to the first mechanism, equation 2.5 is to maximize the system performance (throughput) over the long term, which is the objective of the scheduling algorithm. However, it is subject to the GPS constraint: each user i achieving ε_i fraction of the system performance. In 2.6 the summation users' ε is equal to 1 . In [2] the optimal policy Q^* is defined as:

$$Q^*(\vec{r}) = \underset{i}{argmax}(r_i \times \nu_i) \tag{2.7}$$

Where the function of ν_i is same as v_i in 1.

In general the opportunistic scheduling problem can be stated as:

$$maximize \sum_{i=1}^{N} \sum_{k=1}^{M} (f(x_i(k)) \times I_{\{Q(\vec{r}(k))=i\}}) \tag{2.8}$$

$$\text{Subject to} \qquad F^j(i,k) \geq C_i^j \tag{2.9}$$

Where $f()$ is utility function. Utility is generally defined as a measure of satisfaction that a user derives from accessing the wireless resource. In 1 and 2 the utility function $f(x_i) = r_i$. $F^j()$ is the jth function to shape the quality of service constraints. C_i^j is the jth constraint factor for user i. M is the time window size. For long term (as $M \rightarrow \infty$) the equation (2.8) is converted to:

$$maximize E(\sum_{i=1}^{N} (f(x_i) \times I_{\{Q(\vec{r})=i\}})) \tag{2.10}$$

where the time index k is omitted. We use a stochastic model to capture the time-varying channel condition of each user. We assume that the stochastic process is stationary and ergodic. Hence we drop the time index k.

For the short term (2.8) can be written as:

$$maximize A(\sum_{i=1}^{N} (f(x_i(k)) \times I_{\{Q(\vec{r}(k))=i\}})) \tag{2.11}$$

Where $A()$ is an average function. It is a short term problem we cannot drop the time index k and we also cannot use the expectation function to calculate users' performance. So we use the average function $A()$ to replace expectation function. In every time window M the average system performance is calculated and it should be maximum.

There are several possible performance measures (utility function f). In our dissertation the performance measure is the throughput (in terms of bits/sec). Throughput is the number of information bits per time-slot successfully transmitted between the base station and the mobile user. Besides throughput, other issues could also be important to users like the "monetary value" of the throughput in terms of dollars/sec or power consumption (value of throughput-cost of power consumption). In summary, the performance measure is an abstraction

used to capture the time-varying and channel-condition-dependent "worth" of the system resource to a user.

2.3.3 Related Work

Opportunistic scheduling mechanisms for wireless communication networks are gaining popularity in recent years. In this subsection we will discuss some concurrent opportunistic scheduling algorithms.

(a) The greedy opportunistic scheme [43]

In the class of opportunistic scheduling algorithms the Greedy Opportunistic scheduling algorithm can gain the best system performance. So it is the up bound in system performance among opportunistic scheduling algorithms. The greedy opportunistic scheduling scheme can be stated as follows:

$$i = \underset{i}{argmax}(r_i(k)) \tag{2.12}$$

Where $r_i(k)$ is user $i's$ data rate in time slot k. In every time slot the scheduler chooses the user with the best channel condition to serve.

Advantages: in this algorithm the system performance is maximized.

Disadvantages: the greedy opportunistic scheduling algorithm is intrinsically unfair for users in system resource allocation. Users with continuously bad channel conditions may be starved in overall transmission time.

(b) The proportional fairness scheduling algorithm [50] [51]

Proportional fairness scheduler was suggested in [9] for the first time. Proportional fairness scheduler maximizes the product of the throughput delivered to all the users. The algorithm maintains a running average of each user's channel condition and attempts to deliver data at the requested peak rates, avoiding delivering data when the requested rates are at their lowest points. The scheduler is weighted to serve users that are improving their signal quality and weighted against users that are experiencing signal degradation. The disadvantaged users with worse channel conditions accumulate credits with the scheduler, increasing their priority in the system and their

throughput will start to improve. Suppose there are Nusers and $\overline{r}_i(k)$ is the estimate of average data rate for user i at time slot k, $i = 1, 2, \cdots, N$. Also, suppose that at time slot k, the current achievable data rate of user i is $r_i(k)$, $i = 1, 2, \cdots, N$. The algorithm works as followings:

- **Scheduling:** the user with the highest ratio of $\frac{r_i(k)}{\overline{r}(t)}$: $i = \underset{i}{argmax}(\frac{r_i(k)}{\overline{r}(t)})$ will be selected to serve at the beginning of each time slot.
- **Update average data rate using exponentially weighted low-pass filter:** for each user i, $\overline{r}_i(k+1) = (1 - \frac{1}{t_c}) \times \overline{r}_i + \frac{1}{t_c} \times r_i(k) \times I_i$, where I_i is an indication function, when user i is served at time slot $k+1$, I_i equals to 1, otherwise it is equal to 0. The value of parameter t_c used by the scheduling is related to the maximum amount of time for which an individual user can be starved.

Advantages: this algorithm takes the channel condition into account. Only is the user's request data rate higher than its average data rate, the user would be served by the base station. In this way, the system performance is improved.

Disadvantages: the system needs to update users' average data rate in every time slot, which increases the algorithm complexity. Furthermore although the algorithm mentioned the fairness, it did not describe how it works in the proportional fairness algorithm.

(c) Modified largest weighted delay first (MLWDF) [2] [52]

The author of [2] considers the problem of scheduling transmissions of multiple data users sharing the same wireless channel. Both delay and channel conditions are taken into account. The author also defined the throughput optimal: a scheduling algorithm is throughput optimal if it is able to keep all queues stable if this is at all feasible to do with any scheduling algorithm. $MLWDF$ can be described as :

$$i = \underset{i}{argmax}(a_i \times d_i(t) \times r_i(t)) \tag{2.13}$$

Where $d_i(t)$ is the head of the line packet delay for queue i, $r_i(t)$ is the channel capacity with respect to flow i , and a_i is arbitrary positive con-

stant. In this algorithm the choice of parameter a_i allows to control packet delay distributions for different users. Increasing the parameter a_i for user i, while keeping a_i of other users unchanged, reduces packets delays for this flow at the expense of a delay increase for other users (flows). Therefore, the delay distribution can be shaped.

Advantages: in this algorithm the author is concerned with both channel conditions and delay.

Disadvantages: in this algorithm the fairness is not considered and there is no discussion on how to obtain the value of a_i .

(d) The exponential rule [61] [62]

In [11], authors study an exponential rule:

$$i = \underset{i}{argmax}[b_i \times r_i \times exp(\frac{a_i W_i(t) - \overline{aW}}{1 + \sqrt{1 + \overline{aW}}})] \tag{2.14}$$

$$\overline{aW} = \frac{1}{N} \sum_i a_i W_i(k) \tag{2.15}$$

where $r_i(k)$ is the state of the channel of user i at time slot k, i.e., the actual data rate supported by the channel which is constant over one slot, $W_i(t)$ is the amount of time the HOL packet of user has spent at the base-station, $b_i \geq 0$ and $a_i \geq 0, i = 1, 2, \cdots, N$, are fixed constants. For "reasonable" value of b_i and a_i, this policy tries to equalize the weighted delays $aW_i(t)$ of all the queues when their differences are large. If one of the queues would have a larger (weighted) delay than the others by more than order $\sqrt{\overline{aW}}$, then the exponent term becomes very large and overrides channel considerations , hence leading to that queue getting priority. On the other hand, for small weighted delay differences (i.e., less than order $\sqrt{\overline{aw}}$), the exponential term is close to 1 and the policy becomes the proportionally fair rule. Hence, this policy gracefully adapts from a proportionally fair one to one which balances delays. The factor 1 in the denominator of the rule is present simply to prevent the exponent from blowing up when the weighted delays are small.

Advantages: the exponential rule takes both users' delay and users' channel conditions into account. It is throughput optimal and in co-oporation with a token queue mechanism allows the algorithm to support a mixture of real-time and non-real-time data over HDR with high efficiency.
Disadvantages: with the exponential factor the computational complexity is higher in this algorithm.

2.4 Conclusion

In this chapter we introduce two classes of scheduling algorithm in wireless communication networks: one is wireline extension wireless scheduling algorithms and the other is opportunistic scheduling algorithms. We present their structures and the related work also described. Compared to the WEWS, we observe there are three merits in opportunistic scheduling algorithms:

1. The opportunistic scheduling algorithms exploit and utilize the wireless channel conditions to improve the system performance. However, in WEWS the character of fluctuation in the wireless channel is a negative factor.

2. In WEWS, the wireless channel is modeled as two state-Markov Chain (either "good" or "bad"), which is too simple to characterize the realistic wireless channel. In opportunistic scheduling algorithms the wireless channel has continuous states.

3. The framework of opportunistic scheduling algorithms is simpler than that of WEWS, because the opportunistic scheduling algorithms do not take the compensation which is the most complex part in WEWS into account. So the computational complex in opportunistic scheduling algorithms is lower than that of WEWS.

Considering these three advantages we select opportunistic scheduling algorithms as our main research topic.

Chapter 3

Temporal fairness opportunistic scheduling algorithms

From chapter two we know a good scheduling algorithm should be able to exploit the variability of the users' channel conditions to achieve higher utilization of wireless resource. But allowing only users with good channel conditions to transmit may result in very high system performance, but may starve other users with poor channel conditions, which causes unfairness. In this chapter we study opportunistic scheduling problems under the temporal fairness constraints in the long and short term based on the system model presented in Section 3.1.3.2 and 3.1.3.3. In section 3.1 we replicate the temporal fairness scheduling algorithm in the long term (TFOL) [43] based on two sets of system models. Through simulation results, we show that this algorithm actually does not satisfy the temporal fairness constraint in the short term. So in Section 3.2 we propose a new scheduling policy which fits the short term temporal fairness criterion and is also able to exploit users' channel conditions to improve system performance (opportunistic).

3.1 Temporal Fairness Opportunistic Scheduling Algorithm in Long Term (TFOL)

The temporal fairness constraint is that every user in the system is assigned a fixed fraction of system resource and in this chapter we focus on the Time Division Multiple Access (TDMA) system. So the resource is time slots and on average (long term) each user should be allocated a fixed portion of time slots in TFOL.

3.1.1 Problem Formulation

As we discussed in Section 2.2.2 system performance will be measured by the throughput. The system throughput is equal to the summation of users' average throughput, so in the following user's channel condition in time slot k will be represented by its data rate $r_i(k)$ which is a random variable with respect to user i . So in time slot the users' performance vector is $\vec{r}(k) = \{r_1(k), r_2(k), \cdots, r_N(k)\}$, where N is the number of users. The scheduling problem is stated as follows: given $\vec{r}(k)$, determine which user should be scheduled in time-slot k. We determine a policy Q to be a mapping from performance-vector space to index set $1, 2, \cdots, N$. The objective of TFOL is to exploit users' time varying channel conditions to maximize the system performance in the long term under the temporal fairness constraint. Accordingly this scheduling problem can be stated formally as follows:

$$\underset{Q\in\theta}{maximize} \sum_{i=1}^{N} (E(r_i) \times I_{\{Q(\vec{r})=i\}}) \tag{3.1}$$

$$\text{Subject to} \qquad E\{I_{\{Q(\vec{r})=i\}} \geq \phi_i \tag{3.2}$$

Where I is an indication function:

$$I_b = \begin{cases} 1 & \text{if } b \text{ occurs,} \\ 0 & \text{otherwise} \end{cases}$$

We use θ to denote the set of all feasible policies. It is a long term scheduling problem, so $E()$ is an expectation function in 3.1 and 3.2. Accordingly $\sum_{i=1}^{N} E(r_i \times I_{\{Q(\vec{r})=i\}})$ represents the average system performance. Equation 3.1 stands for the goal of the scheme which is to maximize the system average performance. The time slot index k of r_i is dropped in 3.1 because we assume the long term users' data rate (performance) $\vec{r} = \{r_1(k), r_2(k), \cdots, r_N(k)\}$ is stationary and ergodic. Equation 3.2 means that user i is selected at least $\lfloor \phi_i \times$ number of time slots$\rfloor$ times (in our system time slots are system resource) to be served by the scheduler.

The summation of user i resource allocation ϕ_i does not have to be one. More generally each user i is assigned ϕ_i system resource, where $0 \leq \phi_i \leq 1$ and $\sum_{i}^{N} \phi_i \leq 1$. This allows more flexibility in resource allocation. We call the extra resource: $\sigma = 1 - \sum_{i=1}^{N} \phi_i$ tuning factor. The larger the tuning factor σ, the more easily the temporal fairness will be satisfied (less restrictive of the fairness constraint), and the greater the opportunity to improve the system performance because more extra system resources can be allocated to users with relatively good channel conditions. If $\sigma = 0$, the temporal fairness would be the most restrictive and the scheme has minimum chance to improve the system performance. On the other hand if $\sigma = 1$, there is no fairness constraint and this causes the scheduler to have the most amount of freedom to improve system performance. Under this situation the optimal scheduling policy tends to a greedy algorithm (this point is firstly discussed in [43]).

3.1.2 An Optimal Policy

The optimal policy has been proposed in [43] . It is defined as follows:

$$Q(\vec{r}) = \underset{i}{argmax}(r_i + v_i) \tag{3.3}$$

where the v_i is chosen for user i such that:

- $min(v_i) = 0$;
- $E\{I_{\{Q(\vec{r})=i\}}\} \geq \phi_i$ for all i,
- For all i, if $E\{I_{\{Q(\vec{r}=i)\}}\} \geq \phi_i$, then $v_i = 0$.

According to [43] the parameter v_i in 3.3 is an "offset" factor which is adjusted in every time slot to satisfy the fairness requirement. We know that without the fairness constraint the scheduler would always select the "best" user to serve in every time slot so as a result the optimal policy would be $Q(\vec{r}) = argmax_i(r_i)$. But if there is a fairness requirement, the scheduling policy would schedule the "relatively-best" user to serve. In 3.3 user i is "relatively-best" if $r_i + v_i \geq r_j + v_j$ for all j, $i \neq j$. If $v_i \geq 0$, user $i's$ resource requirement is not satisfied (e.g. because of poor channel condition) and it has to take advantage of some other users. Thus the basic idea of this policy is to give users who experience relatively poor channel conditions the amount of resource equivalent to their minimum requirements. So when $E\{I_{\{Q(\vec{r})=i\}}\} \geq \phi_i$ for all users, the policy changes to greedy algorithm because every user's resource requirement is satisfied:$v_i = 0$ for user i. The extra system resource would be allocated to users with the best channel conditions in the next time slot.
In [43] the stochastic approximation method is used to estimate user's parameter v, which is stated as below:

$$v_i^{k+1} = v_i^k - a^k \times (I_{\{Q(\vec{r})=i\}} - \phi_i) \tag{3.4}$$

where a^k is the step size and converges to 0 as k increases.

3.1.3 Simulation Results

The distinctive feature of TFOL is to exploit wireless varying channel conditions: the policy dynamically schedules a user to transmit in a given time slot based on the users' current channel conditions. At the same time it guarantees that every user gets its system resource requirement. For a comparison, similar to [43], we simulate three scheduling algorithms: round robin (non-opportunistic scheduling algorithm), TFOL scheme and greedy opportunistic scheduling algorithm. The reason we choose round robin and greedy algorithms is because the round robin algorithm is the benchmark of fairness (total resource is evenly divided amongst all users) and the greedy algorithm has the up-bound of system performance in the class of opportunistic scheduling algorithms.

In simulations we consider performance and temporal fairness as primary measures. Firstly

we compare the time slots allocated in TFOL scheme to that of round robin scheduler to show how our policy performs in temporal fairness. Secondly, for evaluating the system performance of TFOL algorithm, we compare the system performance obtained in TFOL to that of the non-opportunistic policy - round robin scheme and the greedy scheduling scheme. We will also investigate the relationship between the number of users and system performance. Finally the shortage of TFOL policy on temporal fairness in the short term will be shown.

We have two simulation environments. The first one is to simulate the TFOL algorithm in an actual cell. The aim of this set of simulation results is to show the users' performances and the allocation of time slots to users in those three algorithms. In the second set of simulation results the relationship between the number of users and the system performance and how the users' 'elasticity' influences their performance are examined. The simulation is also used to show the fairness deviation and starvation time different users experience in TFOL scheme.

Implementation Procedure

In this thesis we focus on the downlink in wireless communication networks. There are four steps in our simulation:

Step 1. If user i is active, it will measure the receiving power level from the central base station and the interference power received from neighboring cells. Then based on these measurements, it will calculate SINR (signal to noise plus interference ratio). According to the relation between SINR and performance, it will estimate its own data rate under the current channel condition.

Step 2. Active users transmit their data rate information to the base station.

Step 3. Base station chooses the user to serve according to the temporal fairness opportunistic policy, round robin policy and greedy algorithm policy respectively.

Step 4. The base station updates users' fairness parameter vector $\vec{v}$ of temporal fairness

opportunistic scheduler by the stochastic approximation algorithm which is

$$v_i^{k+1} = v_i^k - a^k(I_{\{Q(\vec{\tau})=i\}} - \phi_i)$$

The system performs the above steps mentioned in every time slot. Let N be the number of active users. Active user i has a temporal requirement of system resource, $\phi_i = \frac{1}{(number\ of\ active\ users)}$. When the number of active users number N changes, the base station will update the fairness factor accordingly. According to [43] in our simulation we set initial value of vector $\vec{v}$ to zero and $a^k = 0.01$ is a constant in every time slot.

Simulation of an Actual Cell

System model: Our simulation environment is the same as that in [63]. We consider a multi-cell system consisting of a central cell surrounded by hexagonal cells of the same size. The base station is at the center of each cell and simple omni-directional antennas are used by mobiles and base stations. The base station transmission power is $10W$. We focus on the performance of the downlink of the central cell because downlink communication is more important for data services. The frequency reuse factor is 3 and co-channel interference from the first and second tier (Fig. 3.1) neighboring cells is taken into account. The cell radius is 1000m. The cells' central coordinates are shown in Table 3.1 and the cell we considered is cell 0. We assume that each cell has a fixed number of frequency bands. Usually there are tens of users in each cell sharing different frequency bands. We focus on one frequency band, which is shared by 8 users in the central cell. The scheduling policy decides which user should transmit in this frequency band in each time-slot. All users have exponentially distributed "on" and "off" periods. The velocities of mobile users are independent random variables uniformly distributed between the minimum (2km/h) and the maximum velocity (100km/h). The direction of mobile users are independent random variables uniformly distributed between 0 and 2π. A mobile user chooses its velocity when it becomes active and the velocity changes during that on-period. The direction of a mobile user changes periodically. When a user becomes active, its location is uniformly distributed in the cell. If a user moves out of the border, we assume that it reappears at a

Table 3.1: The central point coordinates of the base station

Base Station No.	$X(m)$	$Y(m)$	Base Station No.	$X(m)$	$Y(m)$
1	1500	$500 \times \sqrt{3}$	10	1500	$-1500 \times \sqrt{3}$
2	1500	$-500 \times \sqrt{3}$	11	0	$-2000 \times \sqrt{3}$
3	0	$-1000 \times \sqrt{3}$	12	−1500	$-1500 \times \sqrt{3}$
4	−1500	$-500 \times \sqrt{3}$	13	−3000	$-1000 \times \sqrt{3}$
5	−1500	$500 \times \sqrt{3}$	14	−3000	0
6	0	$1000 \times \sqrt{3}$	15	−3000	$1000 \times \sqrt{3}$
7	3000	$1000 \times \sqrt{3}$	16	−1500	$1500 \times \sqrt{3}$
8	3000	0	17	0	$2000 \times \sqrt{3}$
9	3000	$-1000 \times \sqrt{3}$	18	1500	$1500 \times \sqrt{3}$

point that is symmetric to the exiting point about the central base station. In every time slot users' data rate depends on their SINR and is calculated according to Fig. 3.2. We calculate SINR by following equation:

$$SINR_i = \frac{P_i G_i}{I_i + \sum_{k=1}^{k=17} P_k G_k} \tag{3.5}$$

where P_i is the total signal power transmitted to user i, G_i is the path gain from the base station to user i, and I_i is the noise for user i, $\sum_{k=1}^{k=17} P_k G_k$ is the external interference (from other base stations). Since in this chapter we consider at one time slot there is only one user served by base station, there is no internal inference. The average system capacity is $9.5kbps$.

As a point of discussion in [43] we adopt the path-loss model (Lee's model [39]) and the slow log-normal shadowing model. Specifically, the channel gain $g(k)$(in dB) in time-slot k between an arbitrary user at a distance d from a base station is given as:

$$g(k) = l_p(k) + s(k) \qquad (dB) \tag{3.6}$$

where $l_p(k)$ and $s(k)$ are terms representing path-loss and shadowing, respectively. The path loss $l_p(k)(dB)$ is given as:

$$l_p(k) = K - 38.4log_{10}(d(k)) - \alpha_0 \tag{3.7}$$

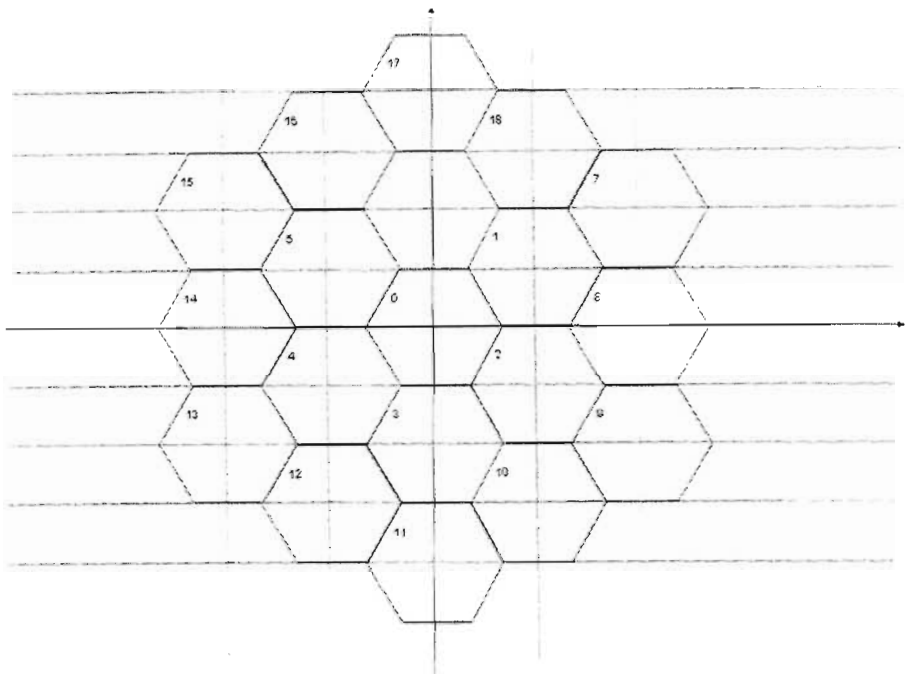

Figure 3.1: Two-tier cell structure

where α_0 is a correction factor used to account for different base station and mobile station (MS) antenna heights, transmitting powers, and antenna gains, and $K = 103.41$ is a constant in the simulation assuming that the transmission power of a base station is fixed at $10W$.

Shadowing is the result of the transmitted signal passing through or reflecting off some random number of objects such as buildings [10], hills, and trees. The shadowing term $s(k)$ (in dB) is usually modeled as a zero-mean stationary Gaussian process with autocorrelation function given as:

$$E(s(k)s(k+m)) = \sigma_0^2 \varepsilon^{vT/D} \tag{3.8}$$

where ε is the correlation between two points separated by a spatial distance D(meters), and v is velocity of the mobile user. T is the length of the time slot. In our simulation, we use a value of $\sigma_0 = 4.3dB$, corresponding to a correlation of 0.3 at a distance of 10 meters.

In this simulation set users' performances are functions of SINR. After calculating SINR we can get the users' performances according to their functions which are different for users as shown in Fig. 3.2. In this set of simulations initially we set user one and two

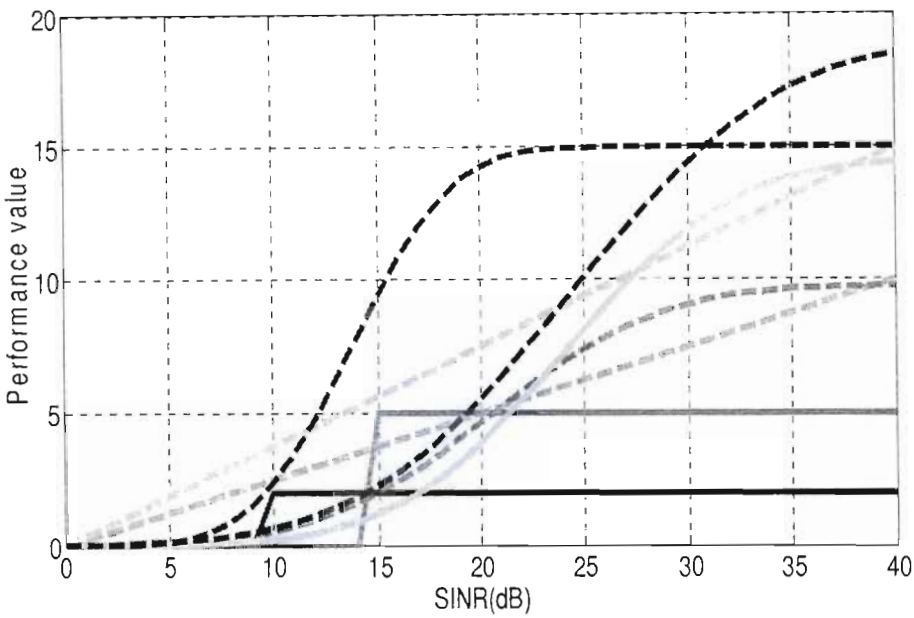

Figure 3.2: Users' performance values vs SINR

are far away from a base station (bad channel conditions because more path loss from equation 3.7), user 3, user 4, user 5 and user 6 are in between and user 7 and 8 are near to the base station (good channel conditions because of less path loss by equation 3.7). Each user has exponentially distributed "on" (5000ms) and "off" (2500ms) periods. We use data rate (kbps) to stand for users' performance in Fig. 3.2 and we run 1000000 time slots and one time slot is 10ms.

Results discussion: In this experiment we evaluate the user fairness satisfaction and the performance improvement in TFOL. The relationship between system performance and the number of users is also displayed.

- The fairness result for the first set simulation is shown in Fig. 3.3. Y axis is the number of time slots (system resource) allocated to users. This gives the number of times slots out of 1000000. The number of time slots allocated to users by TFOL policy is almost virtually the same as those allocated to users by the round robin policy. This means that TFOL policy works as well as the round robin policy in allocating resources to users in a fair manner. On the other hand the greedy algorithm biases the users who experience the better channel condition, so allocating

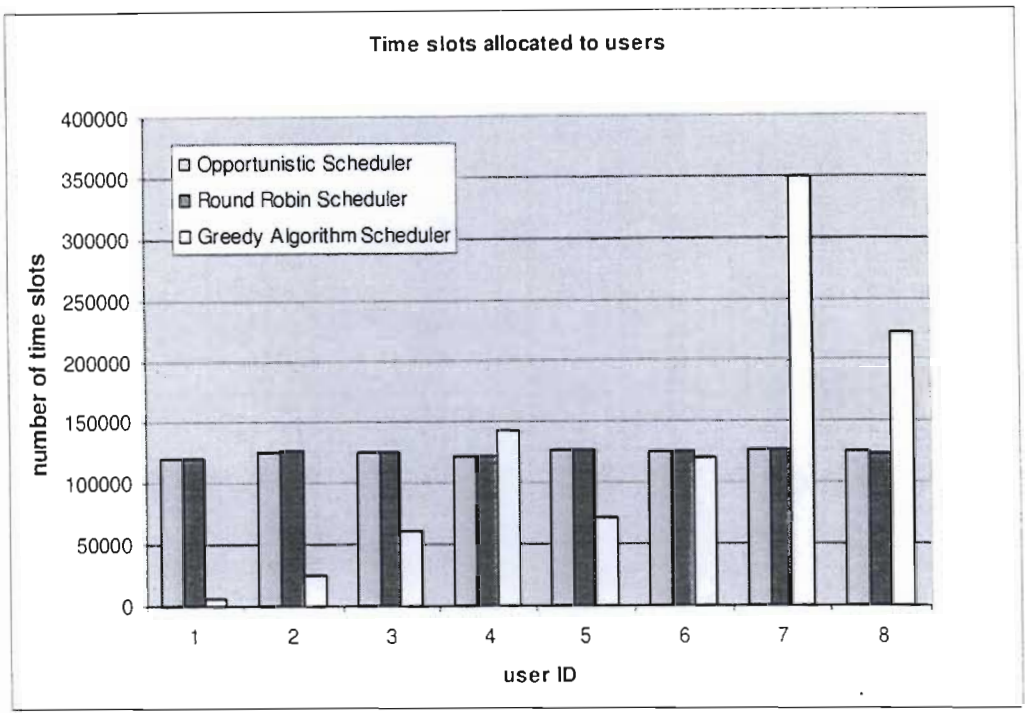

Figure 3.3: Time slots allocated to eight users

most system resource (time slots) to them, which is intrinsically unfair.

- The users' performance result is shown in Fig. 3.4. Users' performance of the opportunistic scheme is much higher than that of round robin - non opportunistic scheduling algorithm. User one's performance is increased by 16%, user eight is increased by 135% compared to that of the round robin. Because the greedy algorithm is intrinsically unfair - allocating most resource (here is time slots) to users who experience better channel conditions, the users who have the best chance to experience the better channel condition, like user 7 and user 8 will get much better performances than with the other policy, but the other users' performances are even worse than that of the round robin scheduler.

- The system performance result for the first set simulation is shown in Fig. 3.5. The system achieved the highest performance - benchmark in performance by the greedy algorithm. The system performance achieved by TFOL scheme is much higher than the non-opportunistic round robin scheme, which is the result we expected.

- How the number of users influences the system performance achieved by TFOL is

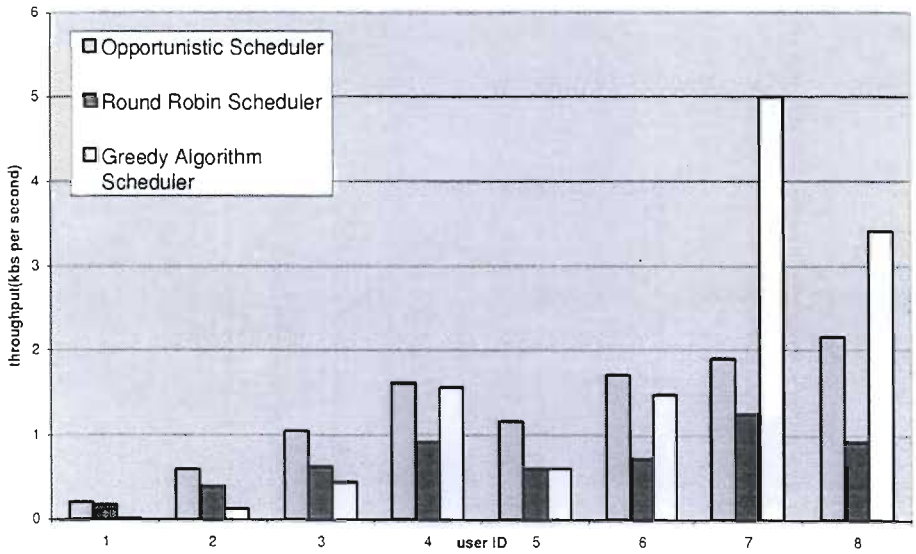

Figure 3.4: Users' performance under TFOL, round robin scheduling algorithm, greedy algorithm respectively

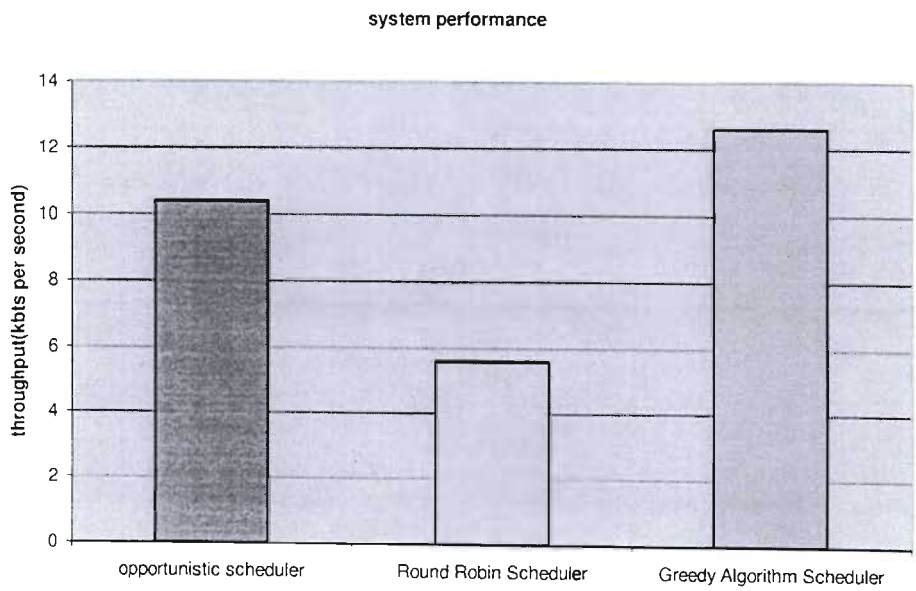

Figure 3.5: System's performance achieved by TFOL, round robin scheduling algorithm, greedy algorithm respectively

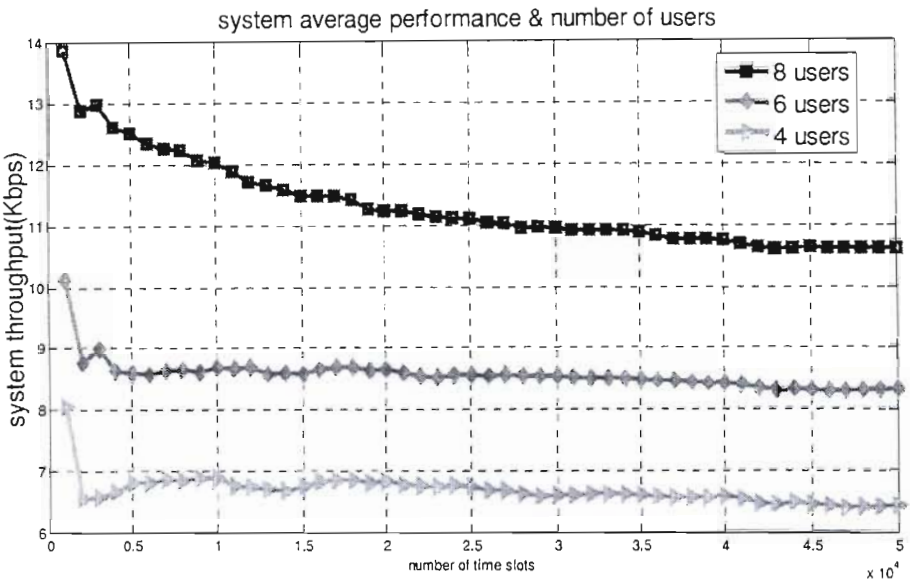

Figure 3.6: System performance when the number of users is 4, 6, and 8 respectively

shown in Fig. 3.6. Here we only run 50000 time slots because the users average performance converges to a constant value after these time slots. We simulated three groups of users. The first group has 8 users, the second group has 6 users and the third has 4 users. Fig. 3.6 shows the average system performance increases as the number of users increase. But in the next section we observe this is not always a fact. The average performance of every group is higher at the beginning but tends to become constant later. This is because at the beginning of the simulation the parameter vector $\vec{v}$ (for controlling the fairness) does not converge to the optimal vector $\vec{v^*}$, in equation $Q(\vec{r}) = argmax_i(r_i + v_i)$ the temporal fairness factor v_i has less power then user performance r_i, so the value of r_i dominates the equation. It works like the greedy algorithm which can gain the highest performance.

Another Set of Simulation

In this subsection we show how user's performance is influenced by the fluctuation of its channel condition and we also show the relationship between the user number and the system performance. Furthermore we examine how TFOL works in the short term from two scenarios: one is the fairness deviation in the short term. The other is the starvation

time for different groups of users in the short term. In our experiment, one time slot is 10ms and the duration of the simulation is 1000000 time slots.

System model: Here we focus on the downlink system in TDMA system. We assume:

- The users' channels are independent of one another.
- There are 16 users in the system with exponential on with a mean=5000ms and off period with a mean=2500ms and we also assume that when the user is on, he always has data to transmit.
- The user's data rate (channel condition) is time-correlated Gaussian process with different mean and variance.
- Let β_i be the auto-regression correlation factor of user i. In each time slot the users' data rate is updated as

$$r_i^{k+1} = \beta_i \times r_i^k + (1 - \beta_i) \times H_i^k \tag{3.9}$$

  where $\{H_i^k\}$ is a sequence of Gaussian random variables. The mean of $\{H_i^k\}$ is the same as the users' performance distribution, but deviation γ_n^i is different. The mean and deviation of both users' Gaussian process and sequence $\{H_i^k\}'s$ Gaussian process are tabulated on Table 3.2. In order to simulate user experiencing different channel conditions we assume different users' Gaussian process is with different means. Table 3.2 shows that the Gaussian process mean of user 1, 2, 3 and 4 is only 4. Those users which have poor average channel conditions, we call group 1. In 3.2 user 5, 6, 7 and 8 have better average channel conditions than the users in group 1, we call it group 2. In the same way we call user 9, 10, 11, 12 group 3. User 13, 14, 15, 16 have the best channel conditions, we call them group four. There are 4 groups and each group has 4 users with different Gaussian process deviations. We notice that in equation 3.6 in each time slot user' channel condition are independent. Here in equation 3.9 user' channel condition in time slot $[k]$ is correlated to it's channel condition in time slot $[k-1]$.

Results and discussion:

Table 3.2: Gaussian process parameters

User ID.	$mean$	$Autoreg.coefficient(\beta_i)$	User dev.	$Sequencedev.(\gamma_n^i)$
1	4	0.3	10.8	20
2	4	0.4	6.9	16
3	4	0.5	4.0	12
4	4	0.6	2.5	8
5	8	0.3	10.8	20
6	8	0.4	6.9	16
7	8	0.5	4.0	12
8	8	0.6	2.5	8
9	12	0.3	10.8	20
10	12	0.4	6.9	16
11	12	0.5	4.0	12
12	12	0.6	2.5	8
13	16	0.3	10.8	20
14	16	0.4	6.9	16
15	16	0.5	4.0	12
16	16	0.6	2.5	8

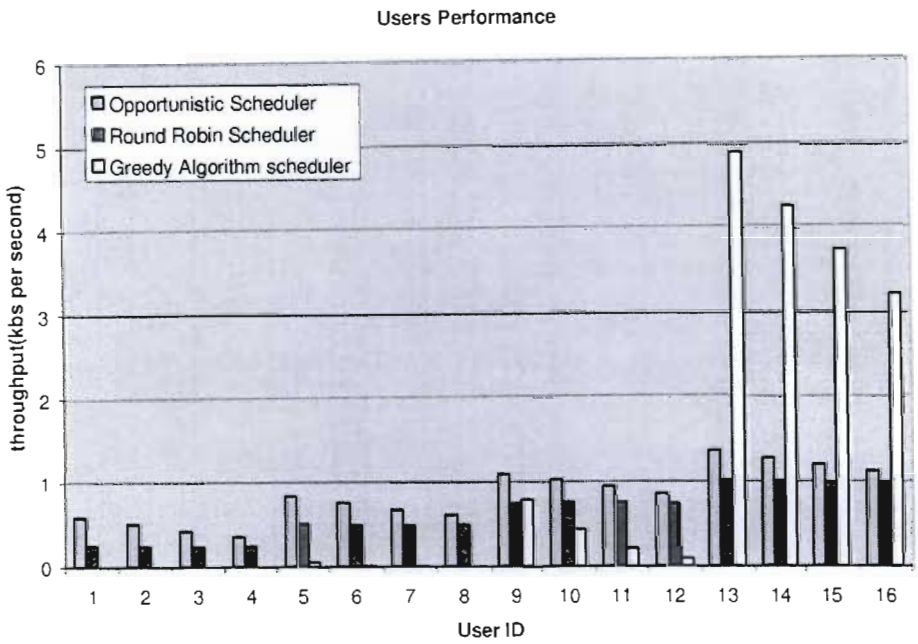

Figure 3.7: Users' performance in TFOL, round robin, greedy scheduling algorithm respectively.

- Fig. 3.7 shows that the performances of users with better average channel conditions (with higher Gaussian process mean) are always greater than those with worse channel conditions (with lower Gaussian process mean). Furthermore, even in the same group in which users have the same average channel conditions (same Gaussian process means), the users' performances are different because they have different fluctuations in channel conditions (different Gaussian process deviations). The user who has the largest fluctuations in channel conditions (higher Gaussian process deviation) in each group always has the highest performance, which means users with more 'fluctuating/elastic' channel conditions will achieve more performance than users who are 'stable'.

- How the number of users influences the system performance is shown in Fig. 3.8. It shows that as the number of users increases the system performance (throughput) increases. But the system performance is slightly decreased as the number of users is increased from 24 to 32. In general, the scheme would have more chance to select the user with a relative good channel condition to serve when the number of users

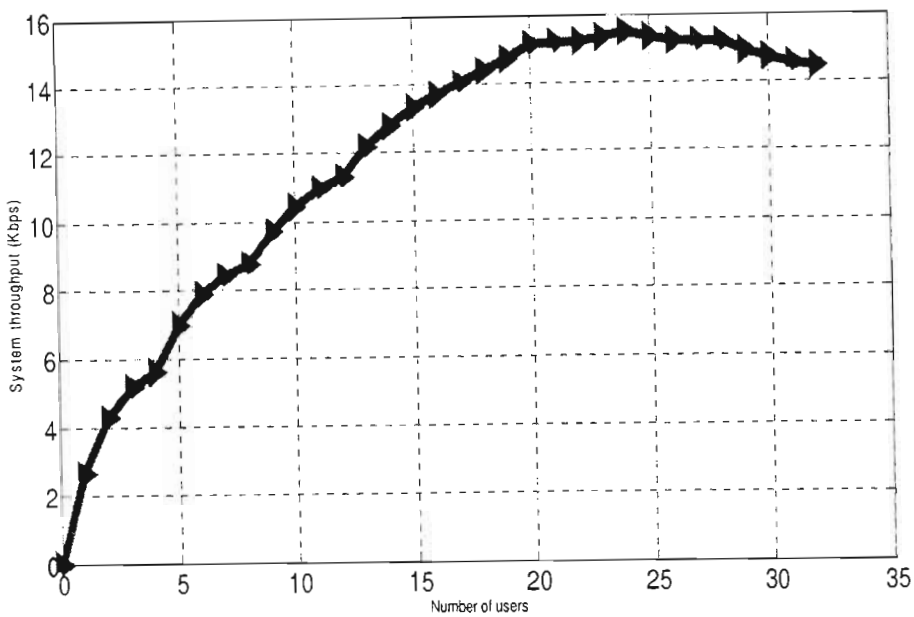

Figure 3.8: System performance when number of users is changed

increases. But if users with relatively bad channel conditions join to the system, the system performance would be compromised. This is because part of system resource (time slots) has to be assigned to those users to gain the temporal fairness. The tuning factor is decreased, so less extra system resource can be allocated to users with good channel conditions in TFOL. Hence it is a question of practical importance to decide the number of users sharing the same channel, which is the admission control issue.

- Time slots (system resource) allocated to 16 users are shown in Fig. 3.3. As we expected time slots assigned to each user in TFOL are equal to that in the round robin scheme. In order to test how the TFOL scheduler works on fairness in the short term we define the fairness deviation factor (FDF):

$$EDF = \sum_{i=1}^{N} \frac{(Timeslotstouseriinopportunisticalgorithm - timeslotstouseriinroundrobin)}{timeslotstouseriinRoundRobin}$$

We check the fairness deviation factor in every 5000 time slots. Fig. 3.10 shows how the fairness deviation factor changes with the time slots. It is observed that the fairness deviation factor decreases exponentially with the increase of the time slots. Furthermore at the beginning (from 5000 to 20000 time slots) the fairness deviation

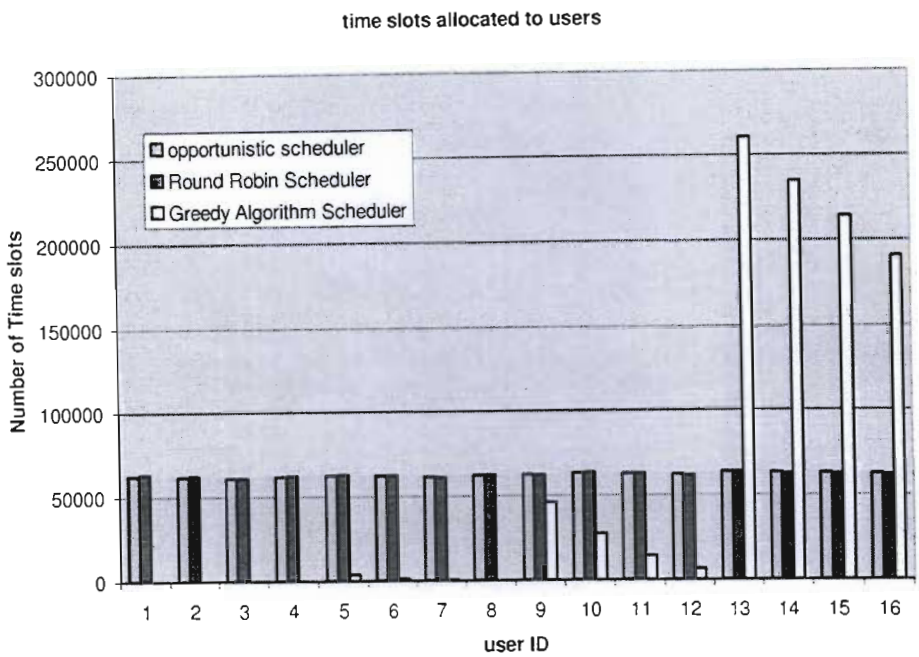

Figure 3.9: Time slots allocated to users

factor is quite high, which is greater than 0.5. After 1000000 time slots the fairness deviation factor decreases to almost zero. This means that TFOL scheduler works well in the long term but not in the short-term where fairness is concerned.

- To further explain how unfair TFOL scheduler works in the short-term we examine the average starvation time (ms) for different groups. Although there are four groups in our system we only examine the average starvation time slots of the group 4 (with best average channel conditions) and the group 1 (with worst average channel conditions). The results are shown in Fig. 3.11. It is observed that the average starvation time experienced by the first group and the fourth group is extremely different during the first 20000 time slots. The group with poor channel conditions experiences heavier resource (time slots) starvation than the group with good channel conditions during the first 20000 time slots. This means that in the short term TFOL scheme does not work well in resource allocation. But as the time slots (running time) increase the average starvation time for these two groups reaches nearly the same value.

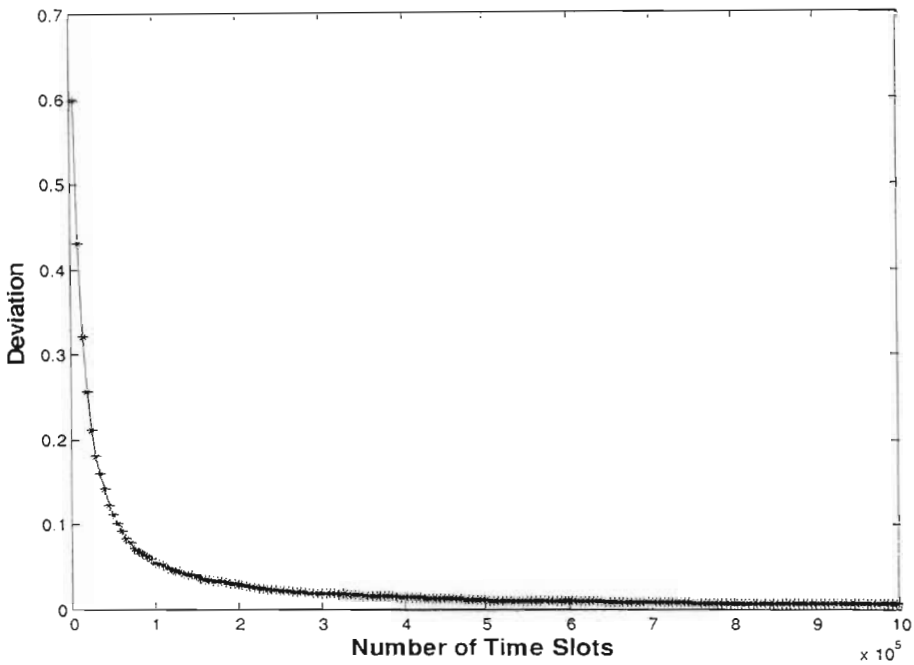

Figure 3.10: Fairness deviation

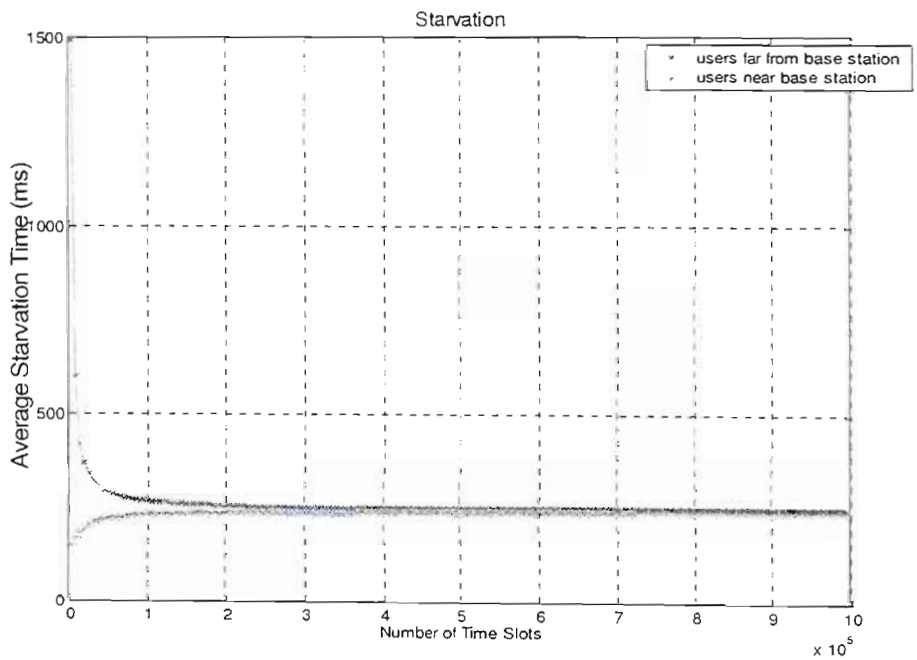

Figure 3.11: Starvation time for Group 1 and 4

3.1.4 Conclusion

In this section we investigated how TFOL works in temporal fairness and system performance both in the long and short term. In the first set of simulation results we observe that TFOL policy maximizes the average system performance while assigning each user the system resource it requires. We use the second model to investigate the influence of users' channel conditions fluctuation on their performance. In this model we assume each user's performance is time-correlated Gaussian process with different mean and variance, reflecting users random-varying channel conditions, Simulation results illustrated the performance of the TFOL policy, showing the significant gains over the non-opportunistic scheduling algorithm-round robin. The disadvantage of this policy comes from the fact that it is unfair in the short term. The reason which causes the fairness deviation in the short term is that the parameter vector $\vec{v}$ in TFOL does not converge to the optimal value in the short term by the stochastic approximation method. It is important to find a new scheme that takes the short term fairness into consideration. The proposed algorithm given in the next section is a possible solution.

3.2 Temporal Fairness Opportunistic Scheduling Algorithm in Short Term (TFOS)

Motivated by the remark in the last section, a new scheme is proposed under temporal fairness constraint in the short term in this section. Based on the definition of long term fairness criterion in [43] the short term fairness criterion is defined as follows.

Short-term fairness criterion: if user i has a predetermined weight ϕ_i and $\sum_{i=1}^{N} \phi_i \leq 1$ the short term fairness criterion is defined as being that each user is allocated at least ϕ_i fraction of time slots in a fixed time window size M, $\sum_{k=1}^{M} I_{\{Q(\vec{r}(k))=i\}} \geq M \times \phi_i$.

3.2.1 A New Policy

The proposed new scheme is stated as:

$$Q(\vec{r}) = \underset{i \in B(k)}{argmax}(r_i(k)) \tag{3.10}$$

- At the beginning of time window, $B(0)$ is a set of all users i.e., $B(0) \Leftarrow \{1, \cdots, N\}$ for N users;
- if $\sum_{k'=1}^{k} I_{\{Q(\vec{r}(k'))=i\}} = M \times \phi_i$, $B(k+1) \Leftarrow B(k)/i$, where / is the subtraction of the set;
- If for all users $\sum_{k'=1}^{k} I_{\{Q(\vec{r}(k'))=i\}} = M \times \phi_i$, then $B(k+1) \Leftarrow B(0)$ and the set remains $B(0)$ until the start of new time window.

$B(k)$ is the set of users at time slot k. The policy is used to meet short term fairness criterion by the 'opportunistic' way. In this scheme in every time window only after every user is selected by the base station $\lfloor M \times \phi_i \rfloor$, $i = 1, 2, \cdots . N$, times then the remainder time slots in this time window can be allocated to those users with good channel conditions.For example, there are two users, user one and user two in the system and $\phi_1 = 0.2$, $\phi_2 = 0.3$. We assume the time window size $M = 10$. Thus, according to their system resource requirements, user one should be picked up by base station to transmit $M \times \phi_1 = 2$ times and user two should be selected to transmit by base station $M \times \phi_2 = 3$ times in every time window. According to 3.10, if in the first and second time slot of the new time window user one is picked up by base station, in the third time slot the base station can only select other users with the exception of user one to transmit. This is because according to the new rule if one user's system resource requirement is satisfied and there are still some other users whose requirements of the system resource are not satisfied, in the following time slot this user would not be selected to transmit by base station until all users in the system attain their share of system resource. Thus, in the following three time slots user two will be chosen to transmit (there are only two users in the system). After user one and user two have been allocated the time slots (system

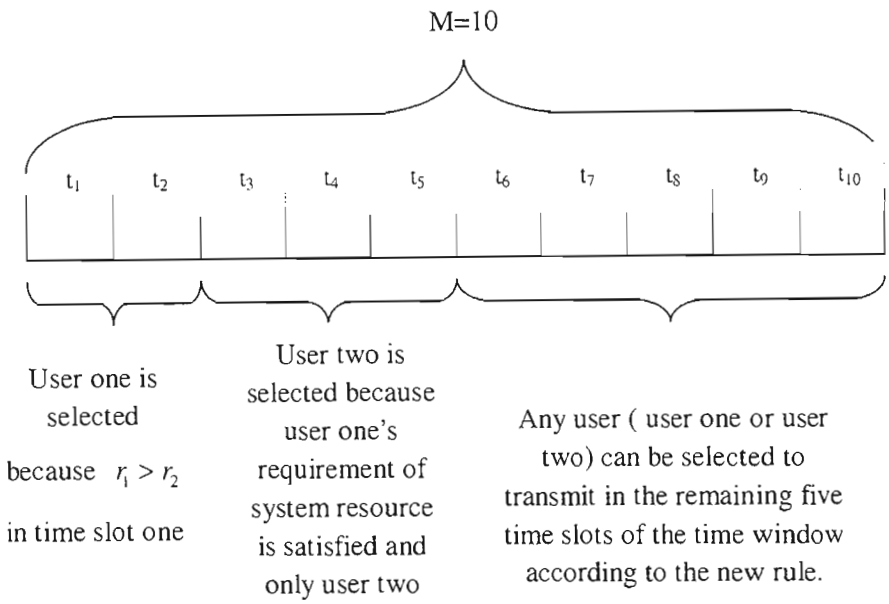

Figure 3.12: Implementaton procedures of the example in new algorithm

resource) they.desired, the remaining five time slots the base station can select any one of them to serve according to 3.10. The whole procedure is shown in Fig. 3.12. In the new scheme the longest starvation in our system is no longer than $2 \times M - 2$. The larger the time window M, the more system performance would be gained in the short-term, but the worse the algorithm in temporal fairness. If time window M increases to infinite, the new scheme becomes greedy algorithm which is the benchmark in the system performance gain in the class opportunistic scheduling algorithms. If the window M is the same as the user number and all users have the same weight ϕ_i, the new scheme becomes 'Round Robin' which is the benchmark in temporal fairness. So our scheme actually is a "bridge" between fairness and system performance gain in the short term.

Lemma 1 *In round robin, $\phi_i = \phi_j$ and $\sum_{i=1}^{N} \phi_i = 1$. We declare that the performance gained in this new policy is not less than that of round robin under the same temporal fairness constraint condition.*

Proof: In the new scheme the utility function is $f(x_i) = r_i$ which monotonously non-decrease with user's data rate r_i. From the definition of the new scheme we have $r_Q \geq r_{RR}$

(Q and RR stand for our new scheme and round robin policy, respectively) in each time slot of time window, we further get $\sum_M f(r_Q) \geq \sum_M f(r_{RR})$ in time window M.

The performance gain of our new scheme is less than that of TFOL scheme because the short term fairness means more restriction, which makes the scheme have less 'opportunity' to improve its performance. We will see this point in our simulation results.

3.2.2 Simulation Results

The distinctive feature of TFOS policy is that it can improve system performance while guaranteeing the temporal fairness in the short term. So in the simulation we will consider the performance and fairness as two main measures and for comparison we also simulate round robin algorithm, TFOL scheme and greedy scheduling policy for comparison. Similar to the last section, the starvation time experienced by different users and fairness deviation are also examined to show how the new policy works in the short term.

Implementation Procedure

Here we focus on downlink in wireless communication networks. For user i we set an initial counter C_i zero. Our simulation system implements according to the steps below :

Step1. The first step is to set time window size M which is variable in our simulation. We set the counter $C_i = 0$ for each user.

Step2. The second step is to evaluate the users' channel conditions and then transfer them to the base station.

Step3. The base station selects user i according to our policy: $Q(\vec{r}) = argmax_{i \in B(k)}(r_i)$.

Step4. The base station updates users' counter vector $C_i(k)$: $C_i(k) = C_i(k-1) + I_{\{Q(k)=i\}}$. If $C_i(k) = M \times \phi_i$ then the $B(k+1) = B(k)/i$. If in time slot k for every user the counter reaches its share $C_i(k) = M \times \phi_i$ the base station resets the users' counter to zero and $B(k+1)$ to the set of all users in time slot $k+1$.

The system performs the above steps in every time window. When the new time window comes, the system will restart and repeat the above steps. For simulation simplicity, we assume that as long as users are active, they will always have packets to send.

System model: For comparing with TFOL algorithm in Section 3.1, we simulate the same system given in section 3.1.3, using the time-correlated Gaussian process with different mean and variance in Table 3.2 to stand for users' performances. But the difference in this system is we assume that users are always on. We also set users' system resource requirements, $\phi_i = \phi_j$ and $\sum_{i=1}^{N} \phi_i = 1$.

Results and discussion:

- Fig. 3.13 shows the fairness deviation results of temporal fairness schedulers in both the long and short term, where the time window size is set to the number of users and fairness deviation factor is sampled after every 5000 time slots. It is observed that at any time slot (no matter whether it is long term or short term) TFOS scheduler equally allocates the resource (time slots) to every user, which means that our new scheme complies strictly with the temporal fairness criterion. But TFOL policy in section 3.1 is not fair in resource (time slots) allocating in the short term at all. To make this point clear, we ran the simulation 10000 time slots (short term) and sampled fairness deviation factor in every 100 time slots. The fairness deviation results are shown in Fig. 3.14. The results show that our new scheme works very well in temporal fairness in the short term.

- The starvation results are shown in Fig. 3.15. The average starvation time for group 1 and 4 is almost the same in the short term. But Fig. 3.11 on page 19 shows that TFOL scheduler biases users who are near the base station too often, which makes users who experience relatively poor channel conditions starve for a longer period in the short term. So the new scheme also works well with respect to starvation time.

- Fig. 3.16 shows that the users' performance gain of our new scheme is greater than that of the round robin policy and less than that of TFOL policy. There is also a tradeoff between fairness and system performance by adjusting the size of the time window in the short term. We sampled the fairness deviation factor in every 100

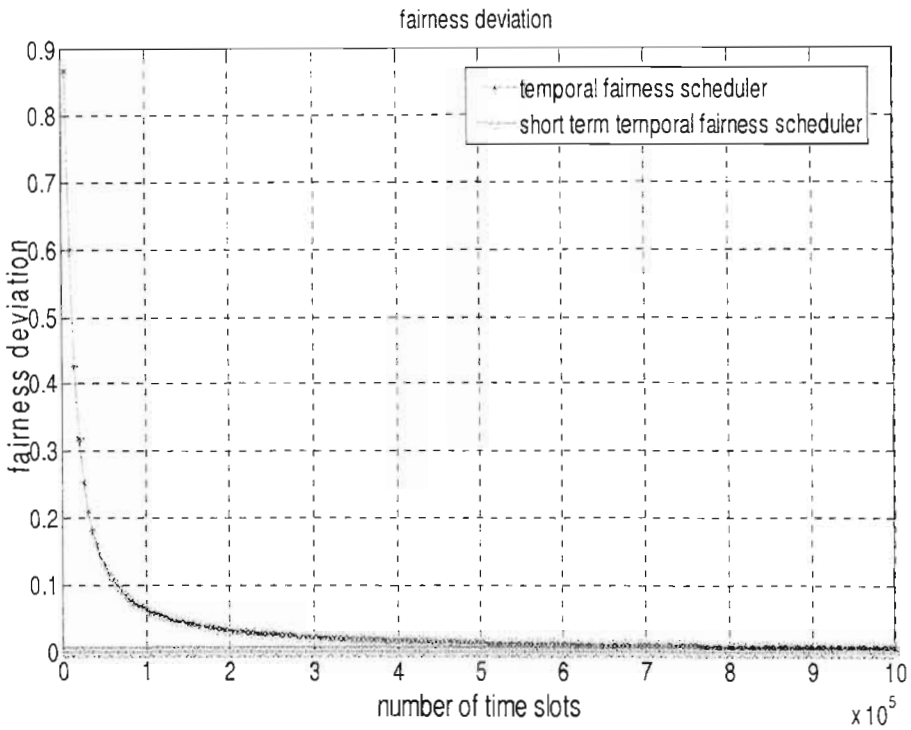

Figure 3.13: Fairness deviation in 1000000 time slots

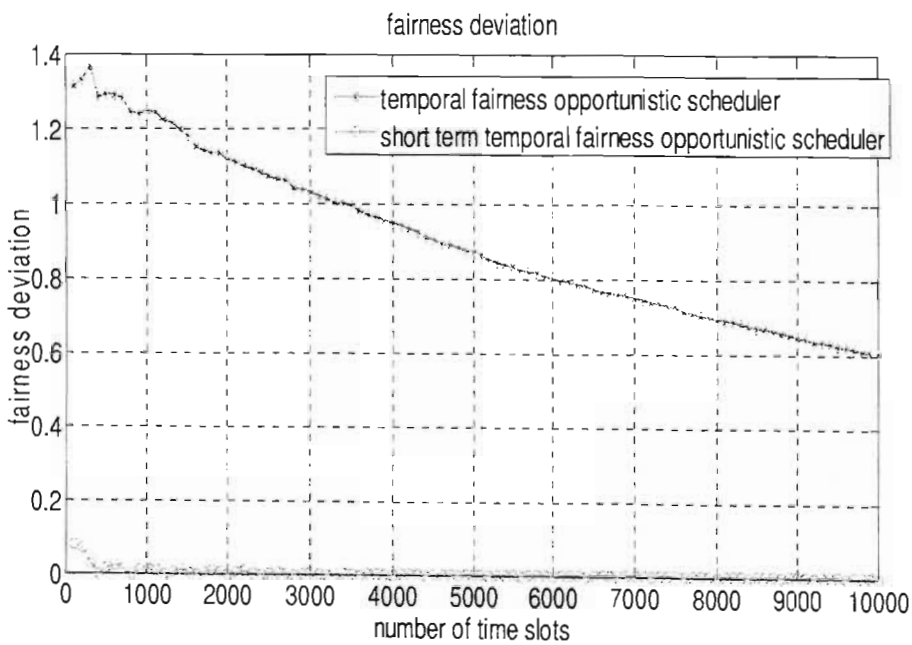

Figure 3.14: Fairness deviation in 10000 time slots

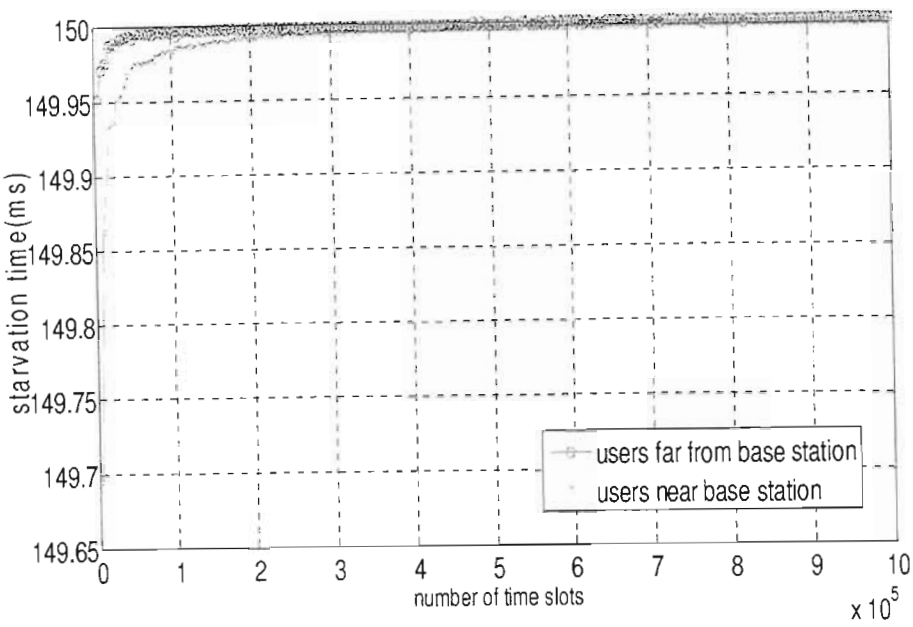

Figure 3.15: Starvation time v.s. time slots

time slots and the simulation run 20000 time slots with different window size M. These results are shown in Fig. 3.17 which shows that the fairness deviation increases with the size of time window in the short term. For testing how the running time and time window size impact on the performance gain, we sample performance gain in every 5000 time slots for different time window size. Fig. 3.18 shows that the performance gain compared to the round robin policy ($\frac{f(Q)-f(RR)}{f(RR)}$)increases as well in the short term. As the simulation running time increases the performance gain compared to the round robin in different time window size tends to be equal.

3.2.3 Conclusion

In this section we proposed a new opportunistic scheduling algorithm under the temporal fairness constraint in the short term. The new policy improves the average system performance while satisfying user's requirement for the system resource in the short term. To compare the proposed policy to the opportunistic scheduling algorithm under the temporal fairness constraint in the long term, we used the same system model as section 3.1.3.3. Simulation results showed that users' requirements for system resource are fulfilled in the

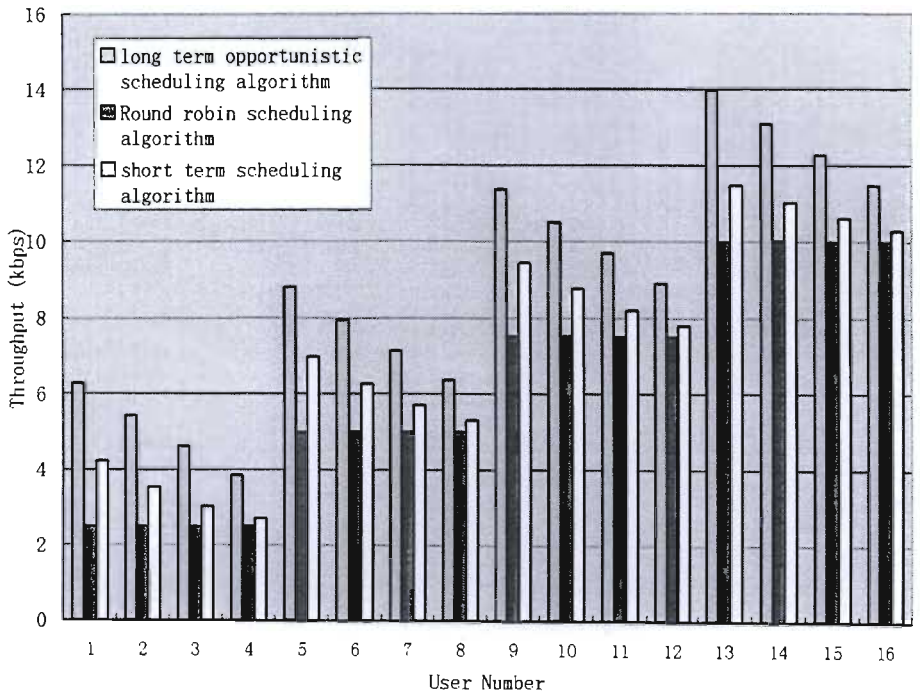

Figure 3.16: Users' performance in TFOL, round robin policy and the new scheme respectively.

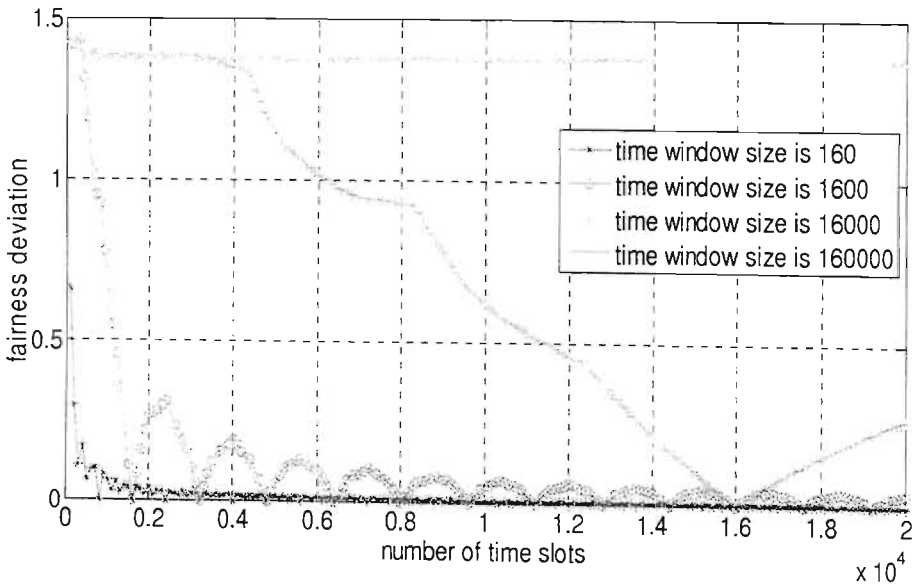

Figure 3.17: Fairness deviation vs. time window size

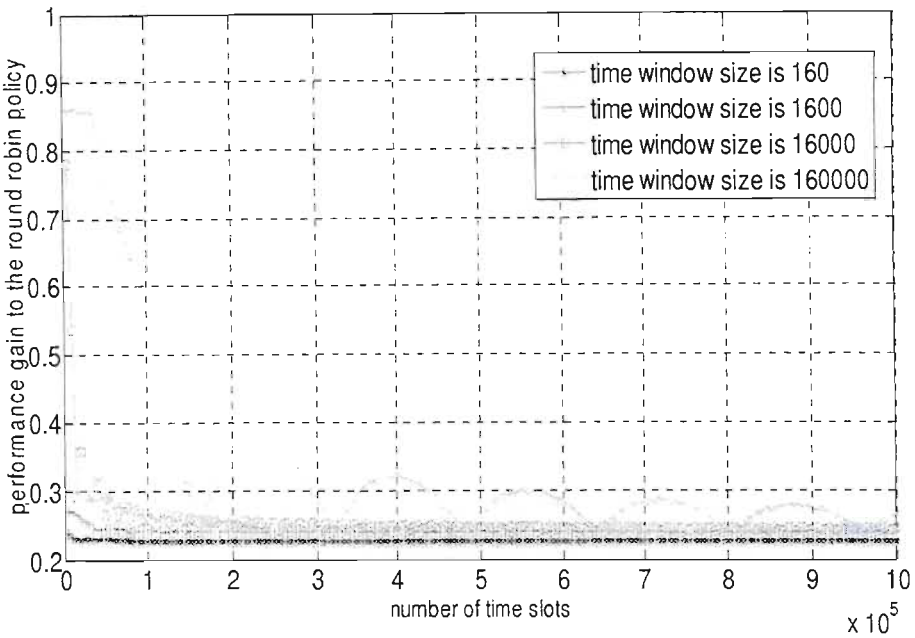

Figure 3.18: The performance gain by the new scheme compared to the round robin policy.

short term and the system performance is improved compared to the non-opportunistic scheduling algorithm. But the performance achieved in TFOS is less than that in TFOL because the short term constraint is stricter than the long term constraint. We also discussed how the size of the time window influences system performance in the short term and there is a tradeoff between fairness and system performance. In the next chapter we will analyze TFOL policy in packet level to examine how it works in the packet delay.

Chapter 4

Delay-concerned opportunistic scheduling algorithm

In this chapter we are investigating the case when the performance is measured by the packet delay. We start as in chapter 3, by considering the delay when using TFOL (temporal fairness opportunistic scheduling algorithm in long term). Also we consider the case of using the earliest deadline first (EDF) scheduling policy in wireless environment and the effect of that on the packet delay performance. The simulation results show that TFOL bias users with relatively good channel condition have improved system performance. The results also show that users with bad channel condition have a worse performance in packet delay distribution. The EDF which works well in wire-line networks does not perform well in wireless networks on packet delay performance. To overcome the shortage in TFOL algorithm we propose a new scheme which considers both channel condition of user and also the packet delay. The simulation results show that the new scheme works well with respect to both system performance improvement and the balance of packet delay distribution.

4.1 System Model

In this section we consider the forward link of time slotted code division multiple access (TDMA) cell that supports real time data users [2]. The scheduler in the base station dominates the downlink transmissions in this system. In our system the total transmission power is treated as a system resource constraint and user's power consumption per unit data rate is the indication of its channel condition. The total transmission power is limited to 1 in any cell.

In our model we consider the following:

- The wireless channel is a time-varying process driven by user mobility and channel shadowing. A *Markov* chain is used to model the fading process of users' channel.
- We normalize the transmission power and generate the mean of users' power consumption per unit data rate which is the same as [2].
- We also assume that all base stations transmit signals at maximum power at all times.
- I_{ext} is the relative out-of-cell interference. The interference I_{ext} that a user experiences is a random variable, which is distributed over all positions in the cell and over log-normal fading.
- We used the assumption made in [2]. The assumption is that each base station has a maximum transmit power of 2, but only half of this power is dedicated to data. The in-cell interference is also simply assumed to be equal to the in-cell received power. The target signal-to-interference ratio $\frac{E_b}{I_o}$ is assumed to be $7dB$ and system bandwidth is $4MH_z$.
- We use the equation:

$$\frac{E_b}{I_o} = \frac{Bandwidth}{TransmitRate}[\text{Signal-to-Noise Ratio}] \tag{4.1}$$

to generate the mean of users' power consumption per unit data rate. The mean power consumption per unit data rate is tabulated in Table 4.1.

Table 4.1: Users' Mean Power Consumption per unit data rate (N=16) [2]

UserID	$c_i(W/bps)$	UserID	$c_i(W/bps)$
1	2.508×10^{-6}	9	4.307×10^{-6}
2	2.518×10^{-6}	10	4.533×10^{-6}
3	2.518×10^{-6}	11	5.229×10^{-6}
4	2.598×10^{-6}	12	6.482×10^{-6}
5	2.771×10^{-6}	13	6.635×10^{-6}
6	2.924×10^{-6}	14	7.257×10^{-6}
7	3.623×10^{-6}	15	7.395×10^{-6}
8	4.142×10^{-6}	16	7.470×10^{-6}

In this chapter we are mainly interested in scheduling users with time sensitive traffic under the wireless channel condition, so we assume:

1. Traffic model: we assume that there are $N = 16$ users in the system and packets for each mobile user are generated by a *Possion* process. The flow of packets is generated at the rate $28 packets/s$ and the packet size is constant, 1024 bits, which corresponds to the typical rate required for the real time users like audio [31].

2. Wireless channel model: because the wireless channel is a time-varying process driven by user mobility and channel shadowing we model the fading process of users' channel by a five states Markov chain. The state-transition diagram for the channel model is given in 4.1. The average system capacity is $219 kbps$.

4.2 The Temporal Fairness Opportunistic Scheduling Algorithm of Long Term and EDF Scheduling Algorithm

In this section we firstly define two scheduling algorithms - temporal fairness opportunistic scheduling algorithm of long term (TFOL) and the earliest deadline delay first (EDF). Then we simulate those two algorithms in the system we described in section 4.1 to examine the packet delay distribution and packet drop ration by these two algorithms.

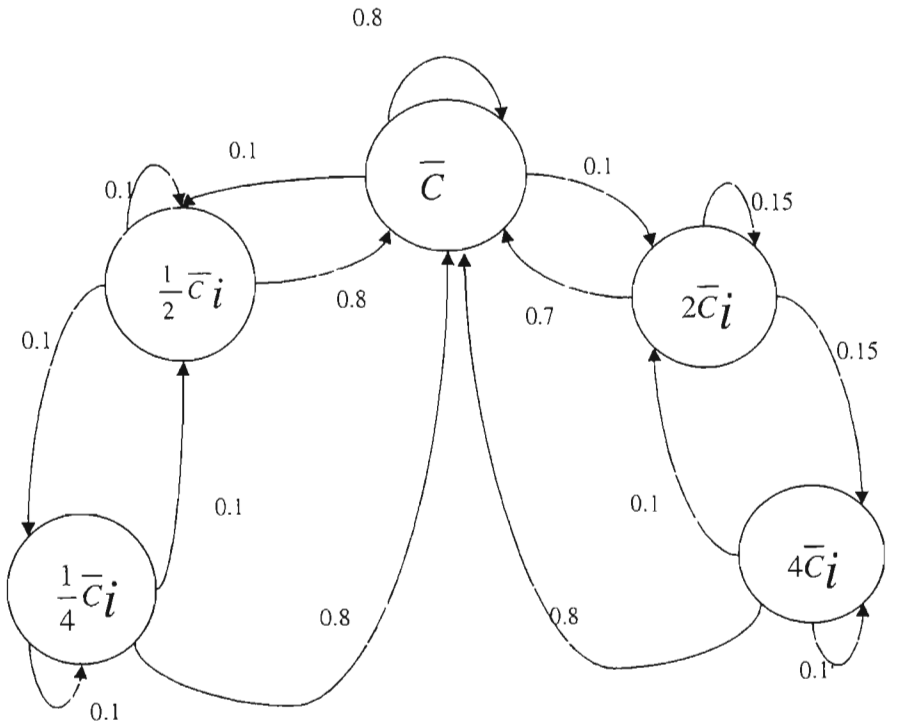

Figure 4.1: Five State-Markov Chain Model

4.2.1 Definition of algorithm

The temporal fairness opportunistic scheduling algorithm in the long term (TFOL): From chapter 3 we know that the objective of TFOL algorithm is to maximize the system performance under the temporal fairness constraint. The optimal policy is:

$$Q(\vec{r}) = \underset{i}{argmax}(r_i + v_i) \tag{4.2}$$

The parameter v_i is updated in each time slot using the stochastic approximation method, which is stated in Section 3.1.2. So in every time slot the base station estimates each user $i's$ parameter v_i and then selects one user to transmit according to the equation 4.2.

The earliest deadline first (EDF) scheduling algorithm: intuitively in every time slot the base station chooses to serve the user for which the deadline of the packet at the top of the queue is closest:

$$Q(\vec{d}) = \underset{i}{argmin}(D_i - d_i) \tag{4.3}$$

In equation 4.3 the parameter d_i is the delay experienced by the head of packet since its entrance to the user $i's$ queue in the base station, $\vec{d}$ is users' delay vector, D_i is the deadline or delay threshold of user i.

4.2.2 Simulation results

The main purposes of our experiments are to examine packet delay distribution and packet loss ratio in TFOL algorithm and EDF policy. In TFOL algorithm we consider all users to have the same system resource requirements, (i.e. $\phi_i = \phi_j$ for any $i \neq j$).

- **Packet delay distribution:** in TFOL algorithm all users would be allocated the same portion of system resource (time slot). In Table 4.1 with $ID = 1$ user has the smallest mean power consumption per unit data rate (good channel condition) and with $ID = 16$ user has the largest mean power consumption per unit data rate (bad channel condition). Hence, providing the same number of time slots, user one would have better performance in packet delay than user sixteen. Fig. 4.2 shows the packet delay distribution for user one and user sixteen. The result shows that our analysis of user one having a much better performance in packet delay than user sixteen is correct. In the EDF, the user's delay is the only parameter to affect the performance in packet delay. Hence packet delay distributions for users who have different channel conditions would not deviate too much. Fig. 4.3 shows the packet delay distribution of user one and user sixteen in the EDF. User one (best channel condition) and user sixteen (worst channel condition) have almost the same packet delay distribution but both users perform worse when compared to TFOL scheme.

- **Packet's drop ratio:** the packet drop ratios under different deadlines in TFOL scheduling algorithm are shown in Fig. 4.4. As the deadline increases the packet drop ratio of both users (one and sixteen) decreases. User one still performs better than user sixteen in terms of packet drop ratio because if they are given the same system resource (time slots) user one with good channel condition would transmit more data than user sixteen with relatively bad channel condition.

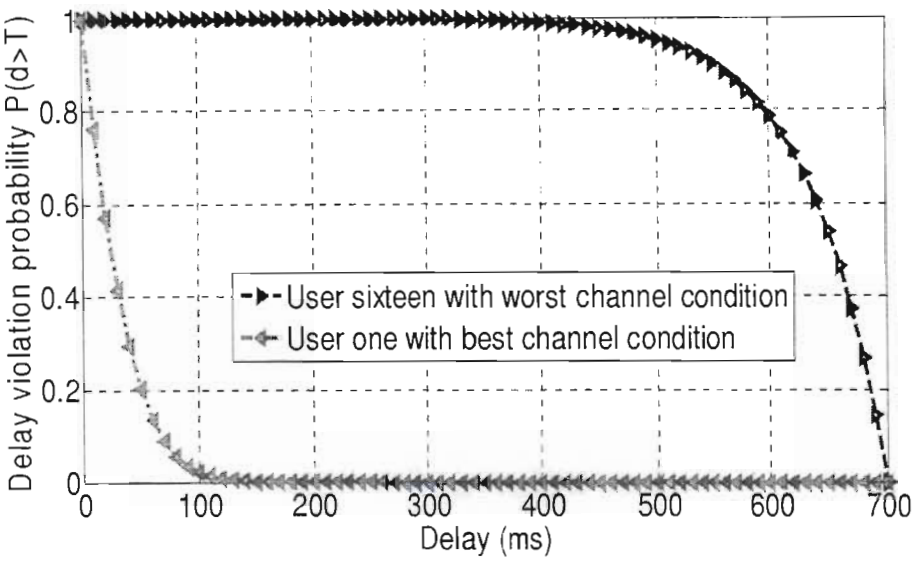

Figure 4.2: The packet delay violation probability of user 1 with the best channel condition and user 16 with the worst channel condition in TFOL when the deadline is 700 ms.

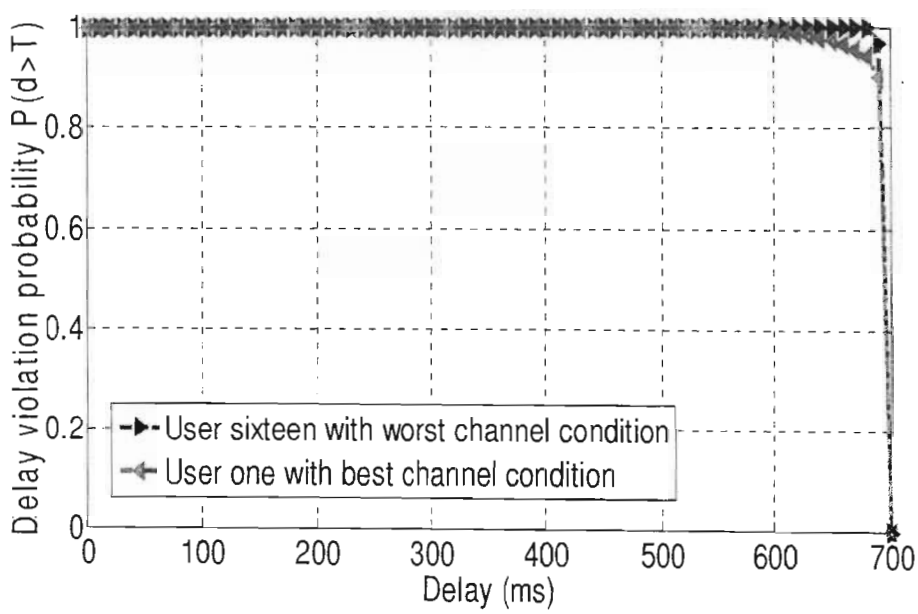

Figure 4.3: The packet delay violation probability of user 1 with best channel condition and user 16 with the worst channel condition in the EDF scheduling algorithm when the deadline is 700ms.

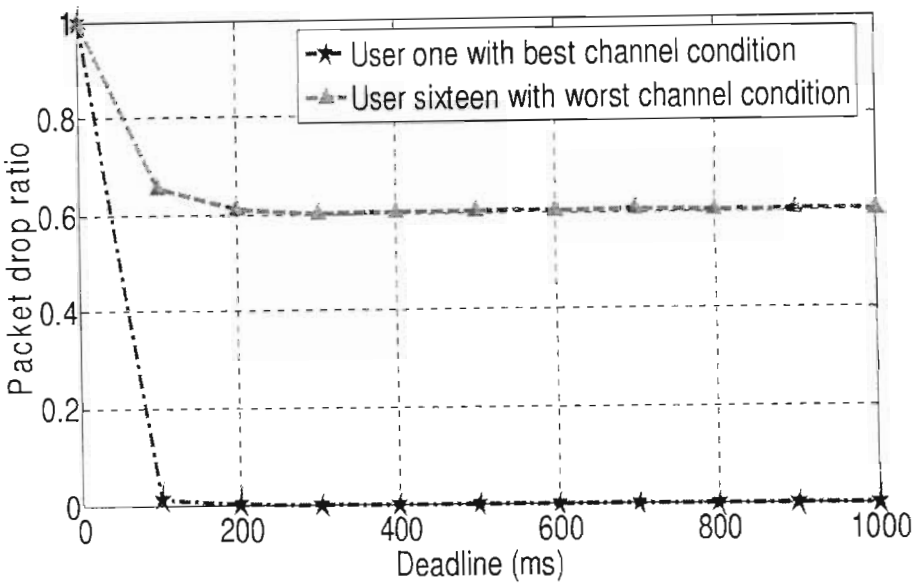

Figure 4.4: The packet drop ratio of user 1 and user 16 when the deadline changes from 0 to 1000ms in TFOL algorithm.

4.2.3 Discussion

Through the simulation results we found that:

- TFOL algorithm is unfair in terms of packet delay. The user who always experiences relatively good channel condition performs much better than the user who always experiences bad channel condition in packet delay and packet drop ratio. The reason is that the temporal fairness opportunistic scheduling algorithm only considers the user's channel condition and the fairness constraint when the base station allocates the time slot.

- Although the EDF ensures that the packet delay distribution for each user is comparable, the overall packet delay distribution that each user experiences is worse than that of TFOL algorithm. So it does not work well in packet delay under wireless environment. This is because the EDF does not take user's channel condition into account.

Based on these results, we are introducing a new scheduling policy that takes into consideration both the user's channel condition and packet delay .

4.3 Delay Concerned Opportunistic Scheduling

In this section we propose a new scheduling scheme to balance the user's packet delay performances while increasing the system performance. In the new scheduling scheme we take both the user's channel conditions and packet delay into account.

4.3.1 A New Scheme

The new scheme is expressed mathematically by the equation:

$$Q(\vec{d},\vec{c}) = \underset{i}{argmax}(\frac{d_i}{D_i} \times \frac{\bar{C}_i}{c_i}) \tag{4.4}$$

Where:

D_i :The deadline of packets in user queue in base station.

d_i :Delay experienced by the head of packet in user $i's$ packet's queue.

c_i : The power requirement per unit data rate of user i.

$\bar{C}_i$:The mean power consumption per unit data rate of user i.

The user $i's$ mean power consumption per unit data rate $\bar{C}_i$ is shown in Table 4.1, which is constant. The packet delay and user channel condition have the same power in the equation 4.4. When a certain queue has its HoL (head of line) packet waiting in the base station for a relatively long period, the weight of delay $\frac{d_i}{D_i}$ in the equation would grow significantly due to the contribution of d_i until it overcomes the other factor $\frac{\bar{C}_i}{c_i}$ in 4.4. This factor is to balance the user's packet delay difference and decrease packet drop ratio. On the other hand, when all users' HoL packet delays are almost the same, i.e. their waiting times to deadline is close, the factor $\frac{d_i}{D_i}$ will be common to all users, another factor $\frac{\bar{C}_i}{c_i}$dominates the equation. So the policy reduces to a proportional fairness algorithm which exploits the diversity of user's channel condition to improve the system performance.

4.3.2 Simulation Results

In this section we use the new scheme to calculate the packet delay distribution, packet drop ratio, temporal fairness, user performance and system performance respectively. The simulation results of the proposed algorithm are then compared with those when using TFOL algorithm and the EDF scheduling algorithm.

- **Packet delay distribution:** in Fig. 4.5 we draw the packet delay distribution using TFOL algorithm and the new scheme. We calculate the delay distribution for deadline value of 700ms, 800ms, 900ms and 1000ms respectively. Fig. 4.5 shows that the difference between user one (best channel condition) and user sixteen (worst channel condition) in packet delay distribution decreases tremendously when compared to TFOL algorithm. This is because our new algorithm not only takes the user's channel condition into account but also considers user's packet delay. In TFOL algorithm the base station has to allocate the system resource (time slots) to the user with good channel condition to meet the fairness constraint even though it already has better performance in packet delay. As a result a user with bad channel condition cannot have extra system resource to compensate for the poor performance in packet delay. But in the new scheme because the packet delay is considered, user's channel condition and packet delay have the same power. Extra system resource is allocated to users with bad channel condition to increase its packet delay performance. In the new scheme the fairness in resource allocation will be compromised (we will explain this point later).

- **Packet drop ratio:** As expected, the packet drop ratio of user one (best channel condition) and sixteen (worst channel condition) tends to zero as the deadline increases as shown in Fig. 4.6 . The packet drop rate due to the deadline violation has a direct relationship to the packet delay distribution. In the new scheme the packet delay and user's channel condition are considered equally. Though, if the delay characteristics of all users are about the same, then the algorithm will be reduced to the proportional fairness scheduling algorithm. The fairness in this algorithm is still the second consideration while in TFOL algorithm the fairness in resource (time slots)

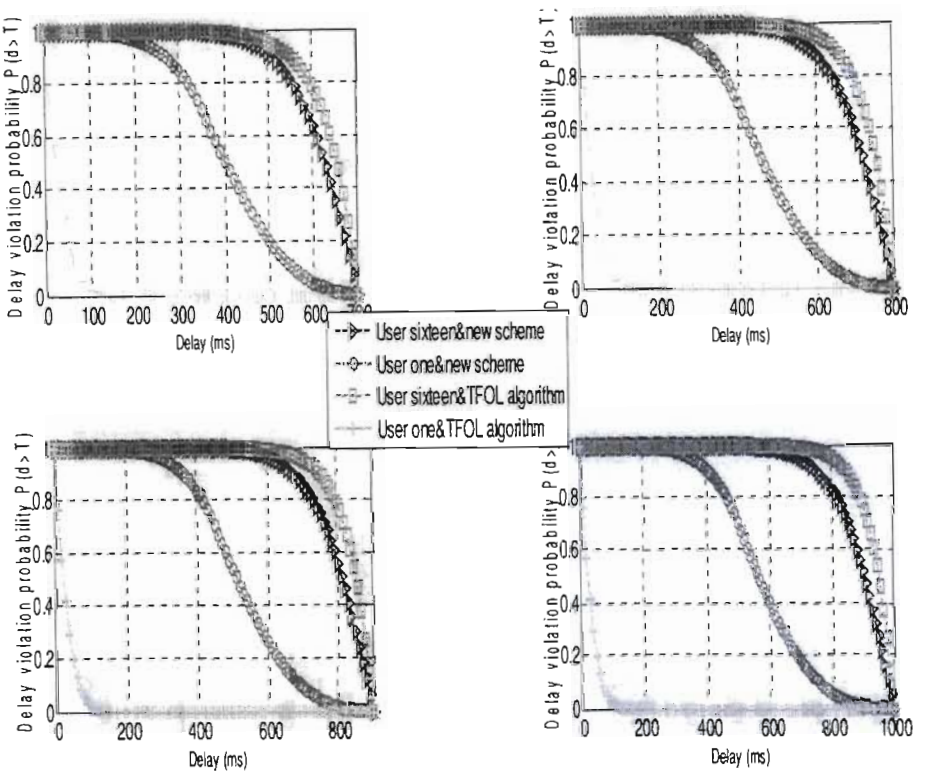

Figure 4.5: The packet delay violation probability of user one who has the best channel condition and user sixteen who has the worst channel condition in TFOL algorithm and new scheme when the deadline is 700, 800, 900, 1000 ms respectively.

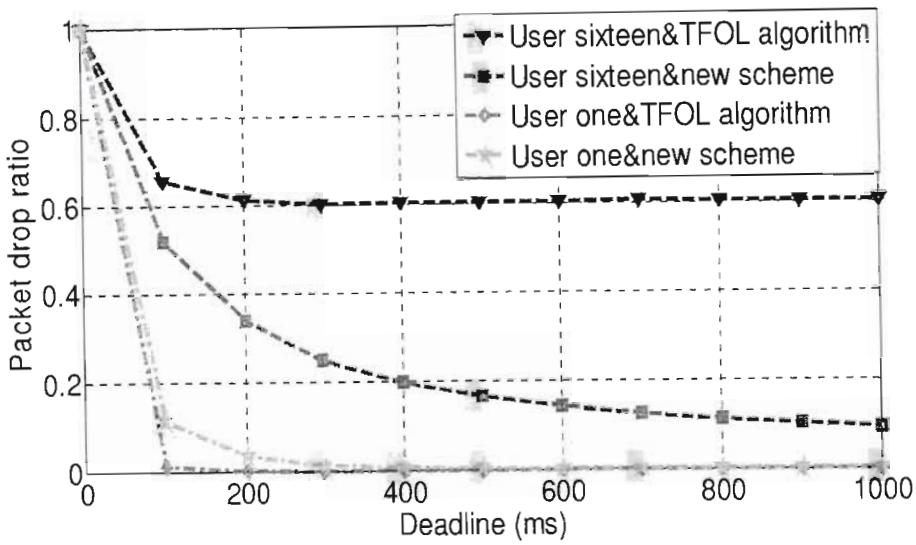

Figure 4.6: The packet drop ratio of user one and user sixteen when the deadline changes from 0 to 1000 in TFOL algorithm and new scheme respectively.

allocation and user's channel condition remain the first concern. So in TFOL, in order to meet the fairness, the base station has to allocate time slots to users with good channel conditions though its packet drop ratio is low.

- **Fairness:** TFOL algorithm is designed for the fairness of system resource allocation. The new scheme is more concerned with the packet delay than the fairness. Time slot allocation with 200ms deadline is displayed in Fig. 4.7. In TFOL algorithm all users' system resource requirements are the same. So, in the graph the time slots allocated to different users in TFOL are more even than those in the new scheme. If users are assumed always on and the simulation is run on user level, the number of time slots allocated to different users is almost equal [64] using TFOL algorithm. In Fig. 4.7 there are deviations from the ideal one on the time slot allocation in TFOL algorithm because if there are no packets in the queue of users with good channel conditions the time slots have to be allocated to users with bad channel conditions, which causes the fairness violation. On the other hand, packet delay and user's channel condition are the main influencing factors in the new scheme. Accordingly, packets allocated to users are quite different. Users with bad channel conditions (user 16, 15, 14, 13) receive more time slots to guarantee that their packet

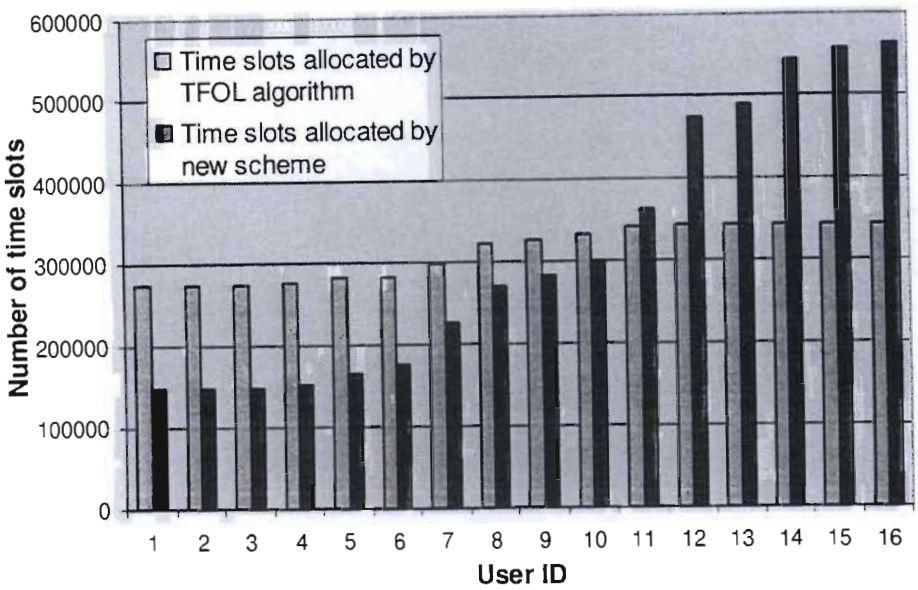

Figure 4.7: The number of time slots allocated to sixteen users in TFOL algorithm and the new scheme respectively when the deadline is 700 ms, the left bar is by TFOL algorithm and right bar is by the new scheme.

delay distribution does not vary too much from users with good channel conditions. This means that the new scheme sacrifices the fairness to gain the balance in the users' packet delay distribution. To examine how TFOL and the new scheme work in temporal fairness criterion, we define the fairness deviation factor:

$$\eta = \sum_{i=1}^{N} \left| 1 - \frac{\text{Time slots allocated in algorithm}}{\phi_i \times \text{Total time slots}} \right| \tag{4.5}$$

The fairness deviation under different deadline is shown in Fig. 4.8. The deviation factor by using TFOL algorithm is always lower than 0.1, when using our new scheme it is approximately 0.5. This is because when the difference of users' packet delay performances is large, the time slots would be allocated to users with poor packet delay performances as compensation.

- **System performance & user performance:** In the new scheme the users' channel conditions are considered to improve the system performance. Accordingly the system performance and user's performance should be greater than those non-opportunistic scheduling algorithms (EDF, Round Robin, Short remained time first

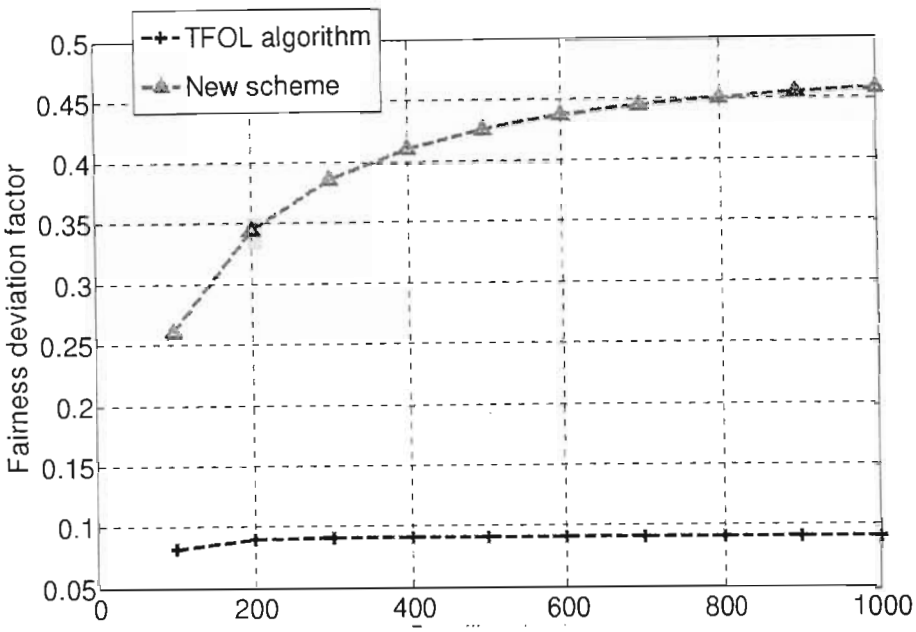

Figure 4.8: The fairness deviation factor η in TFOL algorithm and new scheme under the deadline from 100 to 1000.

and so on). In our simulations, throughput is used to stand for performance. Fig. 4.9 and Fig. 4.10 show the user's performance and system performance in TFOL algorithm, EDF and the new scheme, respectively. As expected both the system performance and user's performance achieved in the new scheme are much higher than the EDF (non-opportunistic scheduling), with gains of 25% to 146%. But user's performance and system performance achieved in TFOL algorithm are better than in the new scheme. This is because more time slots are allocated to users with poor channel conditions (like user sixteen) to improve their performance in packet delay. This means that the packet delay is stricter than the fairness constraint so the new scheme has less chance of exploiting the users' channel diversity than TFOL scheme. Another observation is that although more time slots are allocated to user 12, 13, 14, 15, 16 which is shown in Fig. 4.7, but in 4.9 these users' performances are not better than that of other users. So in the wireless communication environment more resource allocated to users does not mean that these users are guaranteed to get better performance. This is because the users' channel conditions are not stable (constant), if resource is given to users when their channel conditions are poor, they

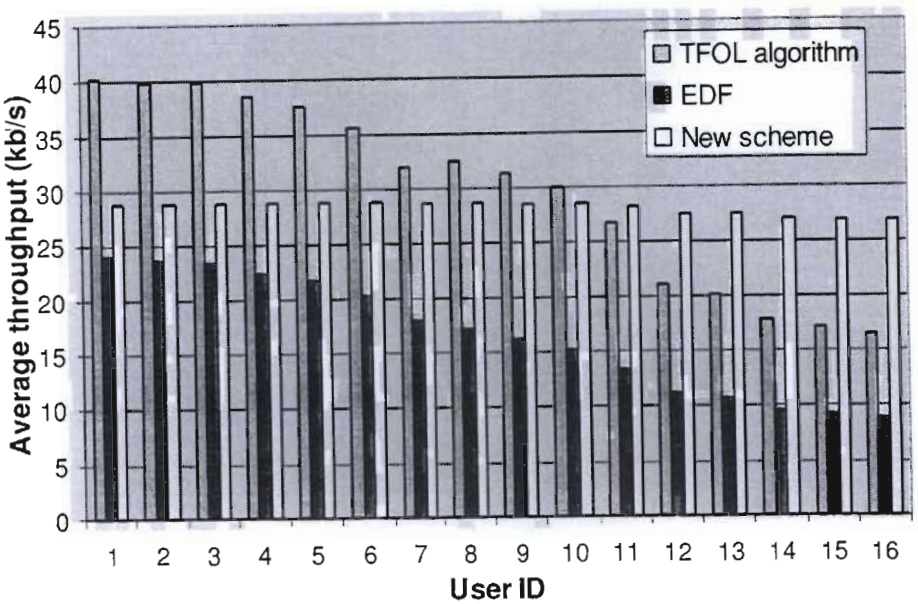

Figure 4.9: User's performance (average throughput) by TFOL algorithm, EDF, the new scheme under the deadline 700ms, respectively, the middle bar is by EDF, the left bar is by TFOL and the right bar is by the new scheme.

can not gain higher performance and the resource is wasted. So the temporal fairness constraint does not really work with wireless environment.

4.3.3 Discussion

In this section we introduce a new scheduling scheme which considers both user's channel condition and packet delay. From the simulation results we observed:

- The new scheme takes both user's channel condition and user's packet delay into account. Hence the delay of users with different channel condition is balanced. Meanwhile user's performance is increased by exploiting users' channel diversity compared to the non-opportunistic scheduling algorithm (EDF).
- Although the new scheme works well in balancing users' packet delay, the temporal fairness criterion is compromised. The system performance is not maximized compared to TFOL algorithm because the delay constraint is stricter than the fairness

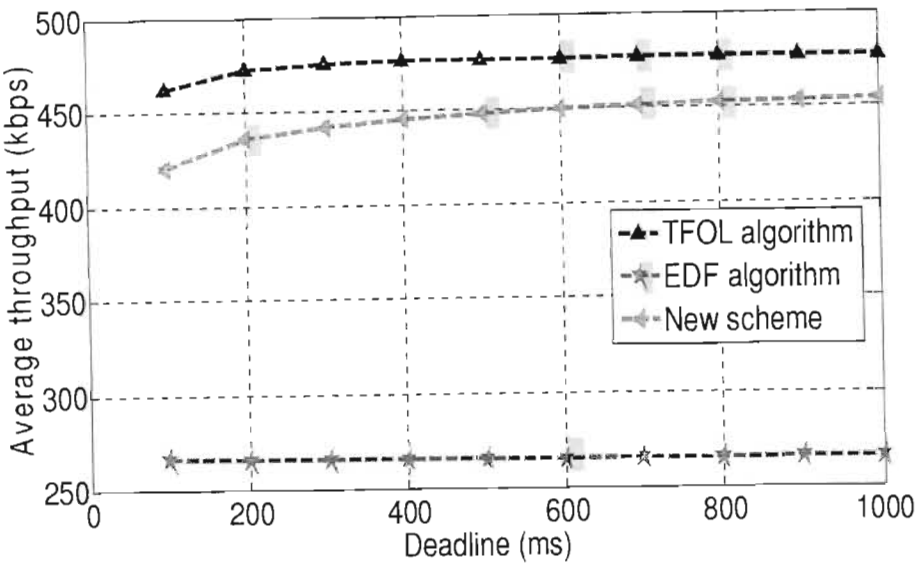

Figure 4.10: System throughput in TFOL algorithm, EDF, the new scheme respectively, x axis is packet deadline, y axis is system throughput.

constraint.

- In the wireless environment, users receiving more resource does not guarantee that they would achieve a better performance.

4.4 Conclusion

In this chapter firstly the performance of temporal fairness opportunistic scheduling algorithm in the long term and the EDF scheduling algorithm in packet delay and packet drop ratio are given. TFOL algorithm biases users with good channel condition in packet delay and packet drop ratio. The EDF scheduling algorithm does not work well in delay under the wireless channel circumstance. So a new scheme which considers both user's channel condition and packet delay is proposed. Through the analysis and simulation we prove that the new scheme works well in both packet delay and packet drop ratio. But there are also some shortages to this algorithm. Firstly, to guarantee the balance of packet delay distribution between users with good channel condition and users with bad channel condition the new algorithm has to sacrifice the fairness. Secondly, because the delay con-

straint is stricter than the fairness constraint, the system performance gain achieved by new scheme is lower than TFOL algorithm. Finally, we observe that the temporal fairness constraint does not fit wireless environment because it cannot guarantee that every user would obtain a certain performance although under this constraint every user's system resource (time slots in our system) requirement is fulfilled. So, in the next chapter we will introduce the new fairness constraint which suits the wireless environment.

Chapter 5

Opportunistic scheduling algorithm under the utilitarian fairness constraint in multiple wireless channel system

In wire-line network, when a certain amount of resource is assigned to a user, it guarantees that the user gets some amount of performance, but in wireless network this point is different because channel conditions are different among users. So in wireless channel the user's performance is not related directly to the system resource allocated to it as we discussed in chapter 4. Furthermore, providing service differentiation in wireless networks has attracted much attention in recent research. Existing studies so far have focused on the design of scheduling algorithm in the wireless network in which only a single user can access the channel at a given time slot, i.e., time division multiple access (TDMA). However, emerging spread spectrum high-speed data networks utilize multiple channel via orthogonal codes [5] or frequency-hopping patterns such that multiple users can transmit concurrently. There has not been much work done about the scheduling algorithm in the multiple wireless channel networks. Finally the opportunistic scheduling mechanism for

wireless communication networks is gaining popularity because it utilizes the "multi-user diversity" to maximize the system performance.

So, considering the above mentioned three points we propose the utilitarian fairness scheduling algorithm in this chapter. Utilitarian fairness is to guarantee that every user can get at least a fixed fraction of system performance which is pre-defined. Hence the proposed algorithm is suitable in wireless networks. We also use the opportunistic scheduling mechanism to maximize system performance under the utilitarian fairness constraint. Simulation results show that the new scheme works very well in both utilitarian fairness and utilitarian efficiency of system resource.

5.1 Utilitarian Fairness and System Model

This section starts by introducing the utilitarian fairness criterion and then the system model will be described. The utilitarian fairness is more general than the GPS fairness. We will prove that the GPS fairness criterion is a special case of the utilitarian fairness criterion. In this chapter, as in chapter 4, we will consider the T-CDMA system. However, in chapter 4 the whole system power is allocated to one user in one time slot, so only a single user is selected by the base station in one time slot. In this chapter we consider the situation where, on any one time slot, the base station can select multiple users.

5.1.1 Utilitarian Fairness

The utilitarian fairness criterion is defined in 2.2.1. In our system we use throughput to evaluate user and system performance. The utilitarian fairness constraint is used to ensure that every user gets at least a pre-allocated fraction of the system performance. This can be formulated by the following equation:

$$E(r_i \times I_{\{Q(\vec{r})=i\}}) \geq \varphi_i \times E(r_{Q(\vec{r})}) \tag{5.1}$$

where $E()$ is the expectation function. $E(r_i \times I_{\{Q(\vec{r})=i\}})$ is user $i's$ performance and $E(r_{Q(\vec{r})})$ is system performance in long term under the scheduling policy $Q(\vec{r})$.

Table 5.1: Users' Mean Power Consumption per unit data rate (N=16) [2]

UserID	$c_i(W/bps)$	UserID	$c_i(W/bps)$	UserID	$c_i(W/bps)$	UserID	$c_i(W/bps)$
1	7.470×10^{-6}	5	6.482×10^{-6}	9	4.142×10^{-6}	13	2.598×10^{-6}
2	7.395×10^{-6}	6	5.229×10^{-6}	10	3.623×10^{-6}	14	2.518×10^{-6}
3	7.257×10^{-6}	7	4.533×10^{-6}	11	2.924×10^{-6}	15	2.518×10^{-6}
4	6.635×10^{-6}	8	4.307×10^{-6}	12	2.771×10^{-6}	16	2.508×10^{-6}

5.1.2 System Model

We consider scheduling problem for a wireless T-CDMA system accessed by multiple users in which a centralized scheduler at the base station controls downlink scheduling. In data CDMA system, a number of higher-rate orthogonal channels are available for data transmission (typically fewer than the number of users). In this chapter, total transmission power is considered to be the system resource constraint while the power requirement per unit data rate is used as an indication of a user's channel condition. Consider N users accessing the system such that user i has a set of possible transmitting rates in time slot k given by $r_i(k) \in \{0, r_i^1(k), \cdots, r_i^M(k)\}$, where $(M+1)$ denotes the number of the possible rates for user i, and rate 0 indicates that the user is not scheduled at that time slot. In time slot k, user i experiences a certain wireless channel condition $c_i(k)$ abstracted as a power consumption per unit data rate in order to guarantee a certain SINR. In this chapter we use the same way as section 4.1 to generate $c_i(k)$ and we also normalize the total power to 1 and assume there are 16 users in the system(this system model is firstly used in [2]). The users mean power consumption per unit data rate $c_i(k)$ is calculated and listed in Table 5.1. We assume that the user data rate is equal to $\frac{Transmission\ power}{power\ consumption\ per\ uit\ data\ rate}$ Because $c_i(k)$ is a random process reflecting the user's channel condition as driven by user mobility and channel shadowing, we model $c_i(k)$ using a five states *Markov* chain. The state-transition diagram for the channel model is given in Fig. 5.1.

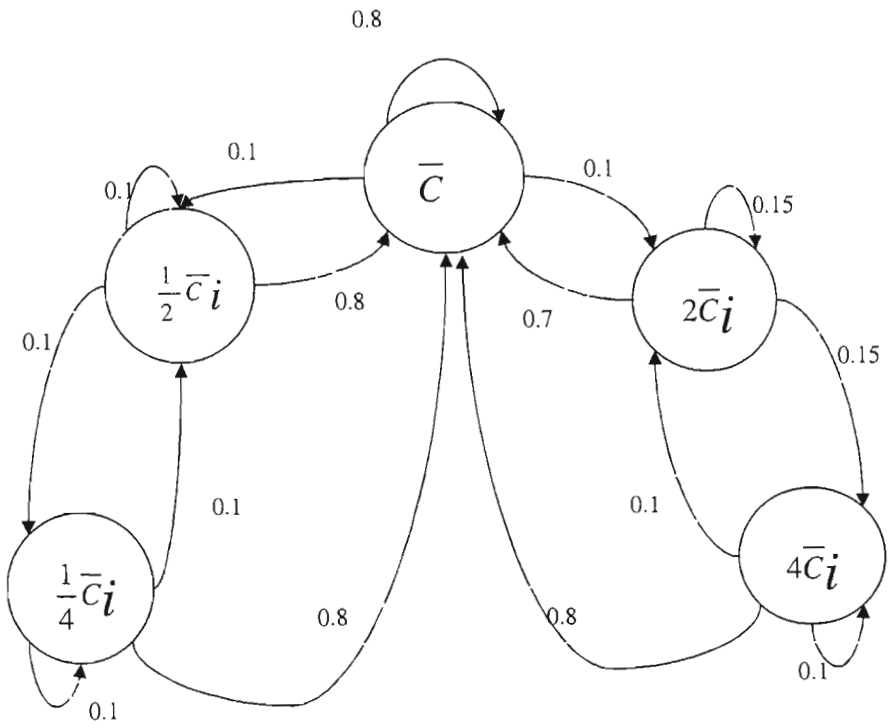

Figure 5.1: Five State-Markov Chain Model

5.2 Problem Formulation

The multi-channel scheduling problem is to select the time slot, channels, and rates for the transmission of queued users. The objective of our scheduling policy is to maximize the system performance subject to the utilitarian fairness constraint and system resource constraint (in our system the power and time slots are the system resource). Because we use the power consumption per unit data rate $c_i(k)$ as the indication of channel condition, the problem can be stated as:

$$\underset{Q\in\theta}{maxmize} E(r_{Q(\vec{r})}) \tag{5.2}$$

$$\text{Subject to } 1.\ E(r_i \times I_{\{Q(\vec{r})=i\}}) \geq \varphi \times E(r_{Q(\vec{r})}),\ i \in N \tag{5.3}$$

$$2.\ \sum_{i=1}^{N} c_i(k) \times r_i(k) \leq p \text{ in every time slot } k \tag{5.4}$$

where θ is the set of feasible policies, $E(r_{Q(\vec{r})}) = \sum_{i=1}^{N} E(r_i \times I_{\{Q(\vec{r})=i\}})$ is system performance, p is the system power (in our system $p = 1$), and $c_i(k) \times r_i(k)$ is user $i's$ power consumption in time slot k when it is scheduled in data rate $r_i(k)$. Equation 5.3 represents the mathematic form of the utilitarian fairness constraint and Equation 5.4 represents the system power constraint. If the system has other constraints, they can be added into this scheduling optimization problem.

5.3 A New Scheduling Policy

We define an optimal policy as:

At the beginning of the time slot the base station selects users to transmit based on the following equation:

$$Q(\vec{r}) = \underset{i \in S}{argmax}(\frac{v_i(k)}{c_i(k)}) = argmax_{i \in S}(r_i(k) \times v_i(k)) \tag{5.5}$$

$$\text{Then} \quad S = S/i \tag{5.6}$$

$$\text{Then select}: \quad r_i = max(0, r_i^1, \cdots, r_i^M) \text{ as user } i's \text{ data rate} \tag{5.7}$$

$$\text{if} \quad \sum_i c_i(k) \times r_i(k) < 1, \ i \in \text{user has been selected, then repeat 5.5, 5.6, 5.7} \tag{5.8}$$

In equation 5.5 v_i is a parameter related to user i and is decided by user $i's$ fairness parameter φ_i and user $i's$ channel condition. We use the stochastic approximation method to estimate it in every time slot. This method will be introduced in the next section. In this chapter we consider the case when in one time slot several users can be selected by the base station to be served. At beginning of every time slot the base station sorts users according to the value of $r_i(k) \times v_i(k), r_1(k) \times v_1(k) > r_2(k) > \cdots > r_N(k) \times v_N(k)$. Then scheduler chooses data rate $r_i = max(0, r_i^1, \cdots, r_i^M)$ beginning with ordered user 1 and proceeding sequentially until user j so that the maximum power limit is reached.

In the algorithm, the function of parameter v_i is to control the resource allocation to satisfy every user's performance requirement. If $v_i > min_j(v_j)$, then the user $i's$ channel condition is worse and its performance does not reach its required value $\varphi_i \times E(r_{Q(\vec{r})})$.

So the scheduler has to allocate another user's resource to it. On the other hand, if $E(r_i \times I_{\{Q(\vec{r})=i\}}) > \varphi_i \times E(r_{Q(\vec{r})})$, then user i gets more performance than its minimum requirement. So the user cannot take advantage of other users. In this algorithm the scheduler only allocates the required resource to the users with bad channel conditions to guarantee that they get their required system performance. The extra resource is given to the users with relatively good channel conditions. This is the meaning of "opportunistic".

We set $a = 1 - \sum_{i=1}^{N} \varphi_i$. We call a tuning factor. Under the optimal policy, higher value a which is brought on by decreasing the performance requirements of users with bad channel conditions will gain higher system performance. When $a = 1$, our scheduling algorithm is changed to the *Greedy Scheduling Algorithm*. The smaller its value, it is harder to satisfy all users' performance requirements, less extra resource allocated to the users with relatively good channel conditions to increase the whole system performance. We will prove this point in the simulation.

According to [59], several propositions can be derived:

Proposition 1 *If $a = 0$, the utilitarian fairness opportunistic scheduling satisfies the* GPS *fairness constraint as well.*

Proof: Because of $a = 0$, we have $\sum_{i=1}^{N} \varphi_i = 1$. Under our scheduling scheme for any user i its performance should be $E(r_i \times I_{\{Q(\vec{r})=i\}}) = \varphi_i \times E(r_{Q(\vec{r})})$, otherwise it is infeasible. So the *GPS* fairness constraint holds:

$$\frac{E(r_i \times I_{\{Q(\vec{r})=i\}})}{\varphi_i} = \frac{E(r_j \times I_{\{Q(\vec{r})=j\}})}{\varphi_j} \tag{5.9}$$

to user i, j, $i \neq j$

Proposition 2 *The difference between two users in performance in the utilitarian fairness opportunistic scheduling scheme has the high and low bound, which can be stated below:*

$$\frac{\varphi_i}{(\varphi_i + a)} \leq \frac{E(r_i \times I_{\{Q(\vec{r})=i\}})}{E(r_j \times I_{\{Q(\vec{r})=j\}})} \leq \frac{(\varphi_i + a)}{\varphi_j} \tag{5.10}$$

Proof: The largest fraction of system performance that user i can get is $(\varphi_i + a)$ if other users get exactly the fraction of system performance they are pre-allocated. The mini-

mum fraction of system performance that user i can achieve is φ_i which is pre-allocated, otherwise the utilitarian fairness constraint would be violated. So there is a high and low bound in the difference of performance between two users, which is stated in 5.10.

Proposition 3 *Increasing the performance requirements of users who always experience bad channel conditions will impair the system performance.*

Proof: The system performance can be expressed as:

$$E(r_{\{Q(\vec{r})\}}) \leq \frac{E(r_i \times I_{\{Q(\vec{r})=i\}})}{\varphi_i} \tag{5.11}$$

So if user i with bad channel condition (small $E(r_i \times I_{\{Q(\vec{r})=i\}}$) has a large performance requirement (large φ_i) it will compromise the whole system performance significantly.

5.4 Parameter Estimation

In the scheduling process the base station needs to estimate and update users' parameters at every time slot as mentioned above. The parameter vector $\vec{v}$ is related to the fairness constraint and user's performance distribution. In practice this distribution is not priorly known and we therefore need to estimate it. We still use the stochastic approximation method to do this job. To estimate and update users' parameters $\vec{v} = \{v_1, v_2, \cdots, v_N\}$ via stochastic approximation method, we firstly need to define a function: $f(\vec{v}) = \{f_1(v_1), f_2(v_2), \cdots, f_N(v_N)\} = \{0, 0, \cdots, 0\}$, where the function f is:

$$f(v_i) = (\varphi_i - \frac{E(r_i)}{E(r)}) = (\varphi_i - \frac{E(r_i)}{\sum_i^N E(r_j)}) \tag{5.12}$$

where $E(r_i) = E(r_i \times I_{\{Q(\vec{r})=i\}})$ and $E(r) = E(r_{Q(\vec{r})})$. Now updating the users' parameters is converted to find the root of function f. The stochastic approximation is an effective technique for finding the root of function $f(\cdot) = 0$. Suppose we try to solve the root of the function $f(x) = 0$, where f is a continuous function and a vector with one root vector $\vec{x}$. Via the stochastic approximation method the root of $f(x)$ can be estimated recursively by the equation:

$$x^{k+1} = x^k - a^k \times y(x^k) \tag{5.13}$$

where a^k is the step size which is 0.01 in our simulation. If we can not obtain the value $f(x^k)$ directly, we can use a noise measurement of $f(x^k)$, i.e. $y^k = f(x^k) + e^k$ where e^k is noise and $E(e^k) = 0$ (e^k is white noise). Then the algorithm:

$$x^{k+1} = x^k - a^k \times y^k \tag{5.14}$$

converges to the root of function $f(x)$ with probability 1 (for detail please read [60]).

In equation 5.12 we can not measure the $\frac{E(r_i \times I_{\{Q(\vec{r})=i\}})}{\sum_{j=1}^{N} E(r_j \times I_{Q(\vec{r})=j})}$ directly because it includes the expected value. But we have the observed value (noisy value):

$$y^k = (\varphi_i - \frac{T_i(k)}{\sum_{j=1}^{N} T_j(k)}) \tag{5.15}$$

where $T_i(k)$ is the estimated value of user expected throughput (performance) at time slot k. We update $T_i(k)$ by an exponentially weighted low-pass filter at the beginning of every time slot.

$$T_i(k) = (1 - \psi) \times T_i(k) + \psi \times r_i(k) \tag{5.16}$$

Where ψ is a filter parameter, in our simulation $\psi = 0.001$. So users' parameters vector $\vec{v}$ can be updated by the equation:

$$v^{k+1} = v^k - a^k \times (\varphi_i - \frac{T_i(k)}{\sum_{j=1} NT_j(k)}) \tag{5.17}$$

When $v_i^k = min(v_j^k)$, $j = 1, 2, \cdots, N$, we also need to guarantee that the user $i's$ performance: $E(r_i) \geq \varphi_i \times E(r)$. Otherwise the fairness parameter vector $\vec{v^k}$ is an unfeasible parameter set because the fairness constraint is violated for user i. To project the parameter vector $\vec{v^k}$ to the feasible set, we need another way to update user $i's$ parameter. Intuitively we can see that $E(r_i)$ is an increasing function of its fairness parameter v_i. So if $v_i^k = min_j(v_j^k)$ and $E(r_i) < \varphi_i \times E(r)$, we increase the value of v_i to increase the value of $E(r_i)$. We use the equation:

$$v^{k+1} = v^k + b \tag{5.18}$$

to update user $i's$ fairness parameter. In our simulation $b = 0.02$. This method is first used in [59] under the single channel situation.

5.5 Simulation Results

To demonstrate the new scheme in multiple wireless channel networks two scenarios are simulated:

1. Discrete data rate set: in one time slot the base station selects as many users as it can under the condition that the total transmission power that users consume does not exceed $p = 1$. Let $\bar{r}$ denote the finite set of data rates which the base station can use for a transmission. The maximum rate r_i at which we can transmit signal to user i is given by $r_i = max\{r_i \in \bar{r}, c_i \times r_i \leq 1\}$. If we transmit signals to user i, we typically incur a waste in power of $1 - r_i(k) \times c_i(k) \geq 0$. So this leads us to adopt the utilitarian fairness scheduling algorithm to the multiple wireless channel networks under the discrete set of data rates. The transmission rate $r_i(k)$ at time slot k is assigned as follows. Firstly we generate the sorted list: $\frac{v_1^k}{c_1(k)} < \frac{v_2^k}{c_2 k} < \cdots < \frac{v_N^k}{c_N(k)}$. For $p > 0$ and $c > 0$, we define the function $r(p; c) = max\{r_i(k) \in \bar{r} : c_i^k \times r_i(k) \leq p\}$ .Then the rate $r_i(k)$ assigned to mobile i is computed iteratively by use of $r_1^k = r(1; c_1^k)$and $r_i^k = r(1 - \sum_{j=1}^{i-1} r_j^k \times c_j^k; c_i^k)$ The set of data rate $\bar{r}$ for our simulation is $\{614.4, 307.2, 153.6, 76.8, 38.4, 19.2, 9.6, 0\}$,where all rates are in *kbps*.

2. Continuous rate set: the base station supports continuous data rate set $[0, \infty)$, then in every time slot the whole transmission power is only given to one user. In the system the data rate to the selected user in one time slot is equal to transmission power (in our system it is equal to 1) divided by its power consumption per unit data rate, which can be expressed as: $r_i(k) = 1/c_i(k)$. Other users' data rates are 0.

5.5.1 Simulation Procedure

In our system, in order to make the decision, the base station needs to obtain information of each data rate (channel condition) at the beginning of the time slot. The performance value of a user can be estimated either by the user or by the base station, based on the channel condition from the pilot signal. Here we focus on the downlink system in wireless communication networks, so the users will measure the received pilot signal from

base-station and the inference, then users measure their own data rates from their SINR (signal to noise plus interference ratio). After these steps, users transmit their data rate information to the base station via the feedback channel. In data *CDMA* systems, a number of higher-rate orthogonal channels are allocated to users. In this chapter, the scheduling algorithm in multiple wireless channel system is to decide which users should be selected and in which data rate as well as how much power should be allocated to them. Our simulation system works according to the followings steps:

1. User will measure the received power level from the central base station via the pilot signal, and the interference power received from neighbouring cells. Then, based on these measurements, the user calculates its own power per unit data bit consumption.

2. Users transmit their power per unit data rate consumption information to the base station. Then by $\frac{v_i(k)}{c_i(k)}$, the base station generates the sorted list:

$$\frac{v_1(k)}{c_1(k)} > \frac{v_2(k)}{c_2(k)} > \cdots > \frac{v_N(k)}{c_N(k)}$$

3. Base station selects the data rate in the data rate set which we have defined in the last part to serve the user from the beginning of the sorted list until the total power limit (in our system it is 1) is reached.

4. Base station updates users' fairness parameter vector $\vec{v}$ by the stochastic approximation algorithm.

In the second step after the base station selects user i , the base station will try to allocate the maximum data rate to the user,$r_i(k) = Max(\bar{r})$. If this causes the total power consumption exceed 1,$\sum_j c_j \times r_j + c_i \times r_i > 1$(j is the user who has been selected by base station). Then the base station will try the second largest data rate and so on. If no value in the data rate set fits the current user, its data rate will be zero.

The system performs the steps above at the beginning of every time slot. We consider traffic in which all flows are continuously backlogged such that the achieved fairness and system performance is totally related to the scheduling process and the channel condition without any other factors due to traffic fluctuation.

5.5.2 Simulation Results

In our simulation we consider performance and fairness as our main measures. The simulation consists of the following:

1. How our new scheme works on utilitarian fairness constraint.
2. User's performance and system performance.
3. The impact of tuning factor a , $a = 1 - \sum_{i=1}^{N} \varphi_i$, on the system performance.
4. How the performance requirements of users who experience poor channel conditions impacts on the system performance.
5. The efficiency of the system resource utilization in the new scheme.

- **Fairness:** In this subsection we examine the fairness performance for our scheduling algorithm. We have four sets of users' performance requirements which are tabulated in Table 5.2. We separate users into four groups. And, in each group user 1 4, user 5 8, user 9 12, user 13 16 have the same performance requirements. We run simulation 1000000 time slots. From Fig. 5.2, Fig. 5.3 and Fig. 5.4, we observe that for each user their performance requirements are fulfilled. Meanwhile user 16 gets much more performance than its performance requirement. The reason is that after each user's performance requirement is reached, the scheduler would give the extra resource to users who experience good channel conditions (low power per unit data rate consumption). User 16 has more chance to experience good channel condition (from Table 5.2). In Fig. 5.2, when $\sum_{i=1} N\varphi_i = 1$ in which the utilitarian fairness is converted to *GPS* fairness criterion, we observe our new scheduling scheme satisfies *GPS* fairness constraint.

- **System performance & tuning factor** a**:** To examine how the tuning factor a influences the system performance we use the same set of user performance requirements as in Table 5.3. Fig 5.6 displays that as the tuning factor a increases the system performance increases accordingly. When $a = 0$, the system performance

Table 5.2: User's performance requirement φ

$\varphi_1 \sim \varphi_4(User1 \sim 4)$	0.01	0.02	0.03	0.05
$\varphi_5 \sim \varphi_8(User5 \sim 8)$	0.02	0.02	0.03	0.05
$\varphi_9 \sim \varphi_{12}(User9 \sim 12)$	0.05	0.05	0.05	0.05
$\varphi_{13} \sim \varphi_{16}(User13 \sim 16)$	0.1	0.1	0.1	0.1
Summation	0.72	0.76	0.84	1
Tuning factor a	0.28	0.24	0.16	0

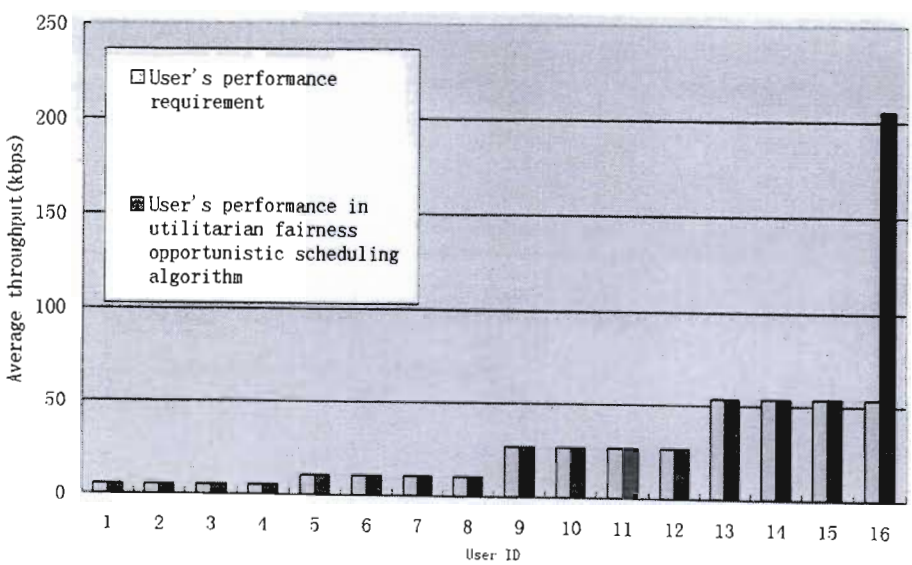

Figure 5.2: User average performance when $\sum_{i=1}^{N} = 0.72,\ a = 0.28$

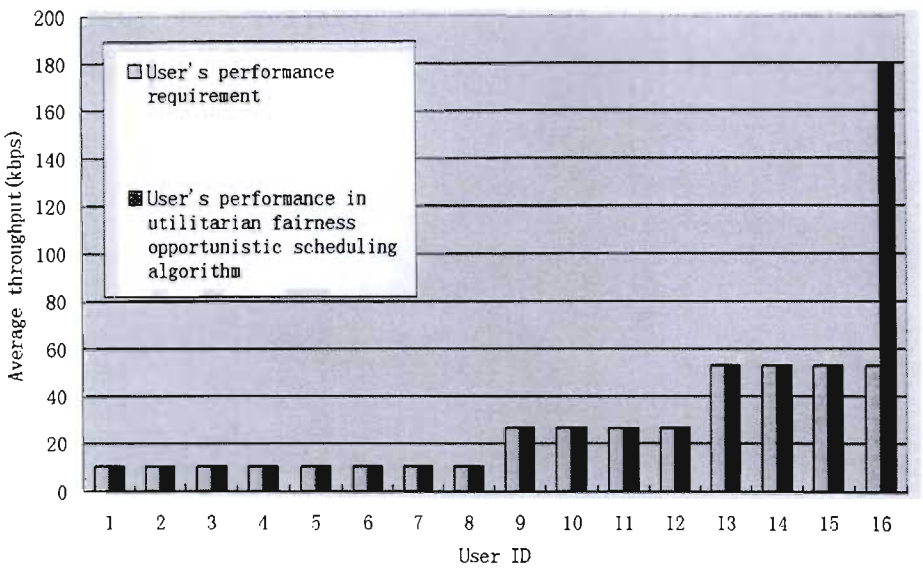

Figure 5.3: User average performance when $\sum_{i=1}^{N} = 0.76,\ a = 0.24$

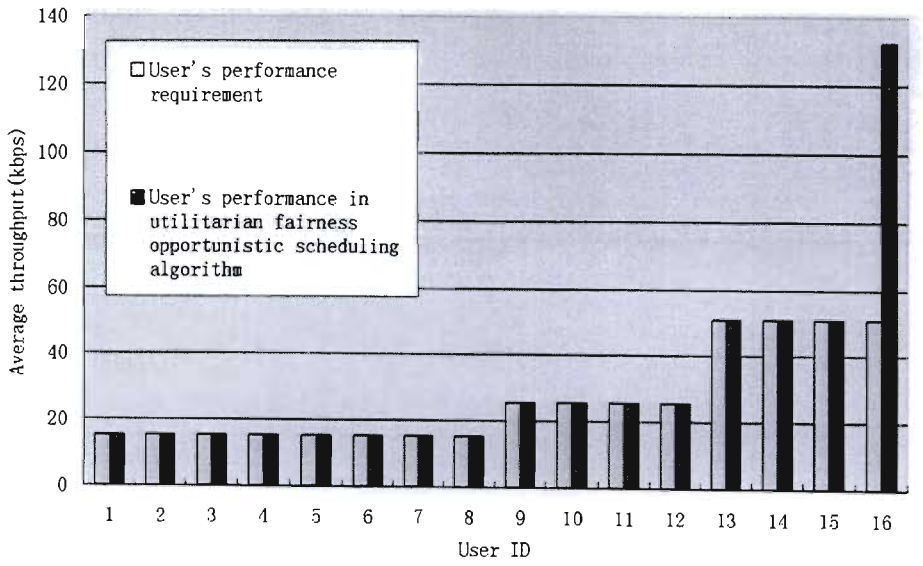

Figure 5.4: User average performance when $\sum_{i=1}^{N} = 0.84,\ a = 0.16$

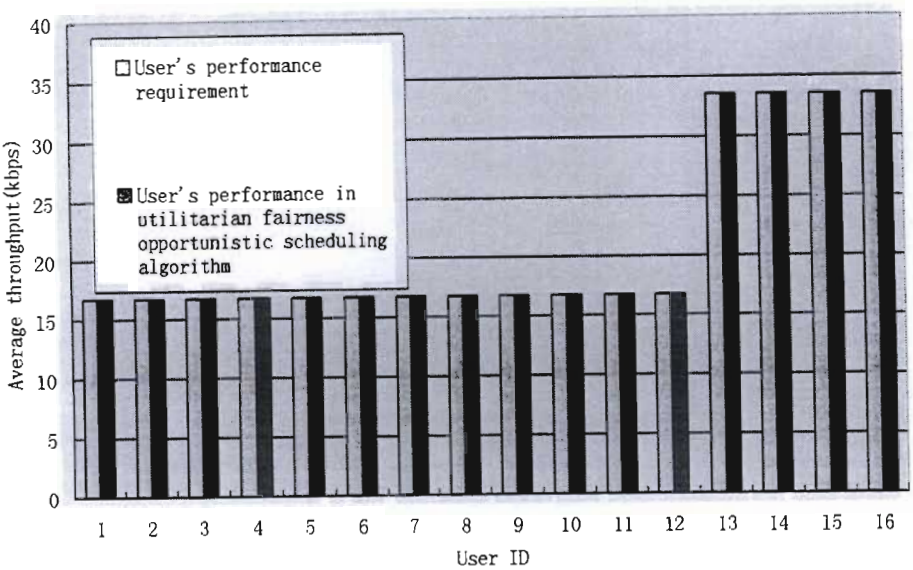

Figure 5.5: User average performance when $\sum_{i=1}^{N} = 0.84,\ a = 0.16$

Table 5.3: User's performance requirement φ

$\varphi_1 \sim \varphi_4(User1 \sim 4)$	0.05	0.03	0.02	0.01
$\varphi_5 \sim \varphi_8(User5 \sim 8)$	0.05	0.04	0.03	0.02
$\varphi_9 \sim \varphi_{12}(User9 \sim 12)$	0.05	0.05	0.04	0.03
$\varphi_{13} \sim \varphi_{16}(User13 \sim 16)$	0.1	0.1	0.1	0.1
Summation	1	0.88	0.76	0.64
Tuning factor a	0	0.12	0.24	0.36

reaches the minimum value. As we show in the last section when $a = 0$ our scheduling scheme agrees with the *GPS* fairness constraint, so there is a trade-off between the improvement of the system performance and *GPS* fairness constraint.

- **System performance & user performance requirement:** In Table 5.1 user 1 has the worst channel condition while user 16 has the best channel condition. To examine how the user's requirement of performance influences the system performance, we change the set in Table 5.2 to those sets in Table 5.4 and Table 5.5 and then compare the system performance. The results shown in Fig. 5.7 indicate that, under the same tuning factor , as the performance requirement for the user who experiences poor channel condition (user one) increases, the system performance drops

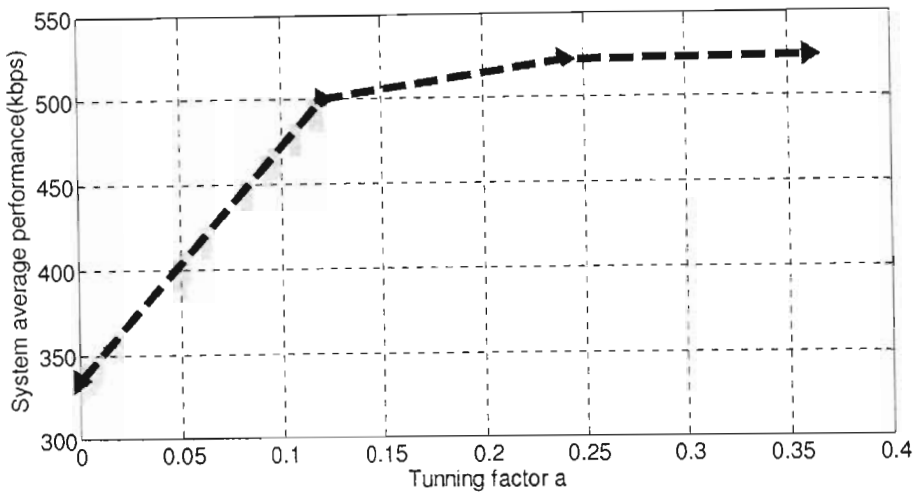

Figure 5.6: System average performance when tuning factor changes

Table 5.4: User's performance requirement φ

$\varphi_1 \sim \varphi_4(User1 \sim 4)$	0.01	0.02	0.03	0.05
$\varphi_5 \sim \varphi_8(User5 \sim 8)$	0.02	0.02	0.02	0.02
$\varphi_9 \sim \varphi_{12}(User9 \sim 12)$	0.05	0.05	0.05	0.05
$\varphi_{13} \sim \varphi_{16}(User13 \sim 16)$	0.1	0.1	0.1	0.1
Summation	0.72	0.76	0.8	0.88
Tuning factor a	0.28	0.24	0.2	0.12

drastically. The results also show that as the performance requirement for the user who experiences relatively good channel condition (user sixteen) increases the system performance does not change much. This is because although the performance requirements for those users that always experience good channel conditions increases, there is not a great increase of the system performance since the tuning factor decreases. In other words, no matter whether we increase performance requirements for those users that experience good channel conditions the extra resource would be allocated to them.

- **Resource Utilization Efficiency:** To test how efficiently our algorithm allocates the resource to users we implement the second set of simulation. We assume the system supports continuous data rate. So in every time slot only one user is picked

Table 5.5: User's performance requirement φ

$\varphi_1 \sim \varphi_4(User1 \sim 4)$	0.01	0.01	0.01	0.01
$\varphi_5 \sim \varphi_8(User5 \sim 8)$	0.02	0.02	0.02	0.02
$\varphi_9 \sim \varphi_{12}(User9 \sim 12)$	0.05	0.05	0.05	0.05
$\varphi_{13} \sim \varphi_{16}(User13 \sim 16)$	0.1	0.11	0.12	0.14
Summation	0.72	0.76	0.8	0.88
Tuning factor a	0.28	0.24	0.2	0.12

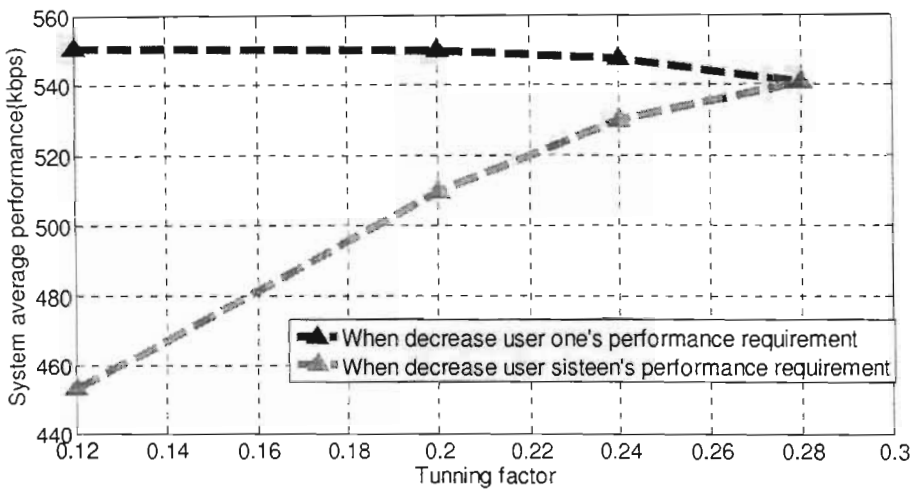

Figure 5.7: System performance when user one's and users sixteen's performance requirements are increased, respectively.

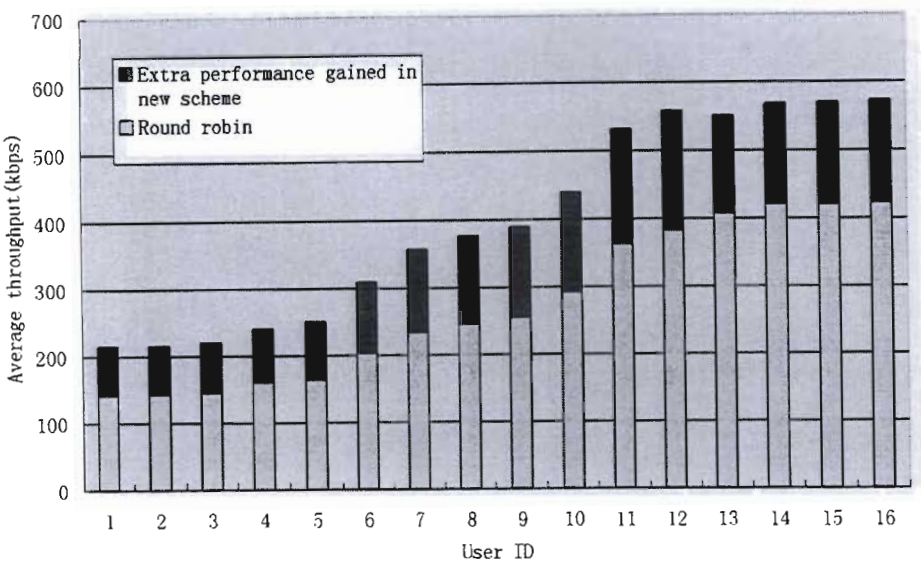

Figure 5.8: Resource Utilization Efficiency (The bottom bar is the efficiency in Round Robin policy. The top bar is the efficiency improvement of our scheme compared to Round Robin policy).

up by the base station, which can show us how the algorithm clearly allocates the system resource-time slot. The initial requirement of the system performance is the same as in Table 5.2. To examine the utilization efficiency of the resource (time slots) we define: (efficiency $= \frac{\text{User Performance(throughput)}}{\text{User Resource Consumption(time slots)}}$) and we also simulate the non-opportunistic scheduling algorithm Round Robin for comparison. In this section user's resource consumption is the allocated time slots not the power. Fig. 5.8 shows that the resource utilization of our new scheme is much higher than no opportunistic scheduling algorithm.

5.6 Conclusion

In this chapter we formulated the opportunistic scheduling problem in multiple wireless channel communication networks under the utilitarian fairness constraint. The optimal algorithm for this problem is proposed and analysed. By considering the power consumption over different channels by different users, the algorithm maximizes system perfor-

mance while satisfying the utilitarian fairness constraint. Simulation results show the new scheduling policy works well in both maximizing system performance and maintaining the utilitarian fairness constraint.

Chapter 6

Conclusion and Future work

With the increase in demand for quality of service, more attention has been paid to the efficient utilization of the limited resource in wireless systems. Different techniques have been developed on the resource allocation in wireless networks, such as admission control, power control and handoff. Scheduling is an important technique among them because it controls the order of service for each individual user. However, the scheduling techniques employed in wireline networks are not applicable to the wireless networks because of the unique characteristics in wireless channels such as bursty errors and location-dependent, multiuser diversity and time-varying channel conditions.

6.1 Conclusion

Two classes of scheduling algorithms have been proposed in wireless systems [1], [59] . The first one is to adapt the wireline scheduling algorithms to the wireless environment, we call it WEWS, and the other one is opportunistic scheduling algorithms. In chapter 2, we explain their structures. Compared to WEWS, opportunistic scheduling algorithms can take advantage of characteristics of wireless channel. So we choose opportunistic scheduling algorithm as our main research topic.

The opportunistic scheduling algorithms can exploit the time-varying channel conditions

to improve the system performance. For example, the base station selects the user with the best channel condition every time as it will maximize the system performance. However, this will make some users, with relatively bad channel conditions, starve from the resource access, which is unfair. So the opportunistic scheduling algorithms should maintain some form of fairness while exploiting the wireless channel conditions to improve the system performance. In chapter 3 the opportunistic scheduling problem under the temporal fairness constraint is studied. We observe that the temporal fairness scheduling algorithm of long term (TFOL) which is first proposed in [43] is actually unfair in the short term. So the new scheme under the temporal fairness constraint of short term is proposed.

In chapter 4, we further simulate TFOL in packet level to study how it works on packet delay distribution for different users. Earliest Deadline First (EDF) which is the benchmark on delay performance in wireline networks is also simulated in wireless system. Simulation results show that in TFOL there is a huge gap between users with good channel condition and users with bad channel condition on packet delay distribution. In EDF, both users with good channel condition and users with bad channel condition have almost the same performance on packet delay distribution, but both of them are worse than that of TFOL. In order to balance the packet delay distribution among different users and improve the system performance in an opportunistic way, we propose a new scheme which takes both channel condition and packet delay into consideration, called delay-concerned opportunistic scheduling algorithm.

In wireline networks, user's performance is directly related to the system resource allocated to them. However, in wireless systems there is no direct relationship between the resource and performance because of the unique characteristics in wireless channel. The temporal fairness we introduced in chapter 3 and chapter 4 is to guarantee a fixed fraction of system resource allocated to each user in the system. Hence, it does not fit wireless systems. The utilitarian fairness criterion is introduced in chapter 5. Under the utilitarian fairness criterion, each user gets at least a pre-defined fraction of system performance. A common assumption in chapter 3 and chapter 4 is that only a single user can access the channel at a given time. However, spread spectrum techniques are increasingly being deployed to allow multiple data users to transmit simultaneously on a relatively small

number of separate high-rate channels. In particular, multiple logical channels can be created via different frequency hopping pattern or via orthogonal code in Code Division Multiple Access (CDMA). Hence in chapter 5 we consider the opportunistic scheduling in multiple wireless channels system (T-CDMA). We formulate the opportunistic scheduling problem in multiple wireless channels system and propose the optimal scheme. Several properties are given on this scheme. We define the tuning factor a . As a increasing, the system performance will increase. Furthermore, if performance requirements of users with bad channel condition are increased, the total system performance will be impaired tremendously. By considering resource consumption over different channels, the algorithm allows system operators to jointly optimize the transmission through multiple channels for total throughput maximization while maintaining the utilitarian constraints.

Opportunistic scheduling also has its own shortages and limitations.

1. The signaling costs involved in all opportunistic scheduling schemes are high because scheduling decisions inherently depend on channel conditions. Users or the base station need constantly to estimate the channel condition.

2. In chapter 3 we know there is a tradeoff between the short term fairness and short term system performance. In general the greater the improvement in the short term performance, the less the short term fairness.

3. In chapter 3 we observe the opportunistic scheduling algorithms exploit the fluctuation of channel conditions, the greater the fluctuation of channel conditions, the larger the number of users, the better the system performance. On the other hand the fluctuation of channels should be slow so that users or the base station can estimate it in time. So this is another issue in opportunistic scheduling algorithm.

4. 4. From chapter 4 we know the opportunistic scheduling algorithms cannot provide the delay bound to the real time users because it does not take packet delay into account. If the algorithm considers both packet delay and channel conditions, the system performance will be compromised. Hence, there is a tradeoff between packet delay and system performance.

6.2 Future Work

Many interesting problems are yet to be resolved in opportunistic scheduling. We know scheduling algorithm is an important part of resource allocation to provide high-rate data and quality of service in wireless networks. The opportunistic scheduling scheme in its current form is a network-layer problem. However, its performance is closely related to physical layer designs. Estimation errors occur in all opportunistic scheduling schemes. On the one hand, we need better understanding of the effect of channel estimation errors on scheduling schemes. On the other hand, it calls for better channel estimation techniques and smart coding schemes (e.g., incremental redundancy transmission schemes with turbo codes). Further, it is also important to study the performance of opportunistic scheduling in multiple antenna systems. In summary, a better understanding of physical-layer technologies or even cross-layer designs can be potentially beneficial.

The opportunistic scheduling problems studied here can increase the overall effective capacity of the wireless network. This means that the network can now accommodate more users or higher-data-rate users. Thus, we know that keeping all else fixed, the admissable region of the wireless network will increase by using opportunistic scheduling schemes. A challenging problem that still remains is how to make intelligent admission control decisions on whether or not to allow a new user into a cell. Although admission control is a difficult problem in wireless systems whether or not opportunistic scheduling is used, it is more challenging in the context of opportunistic scheduling because opportunistic scheduling increases the system dynamics.

Bibliography

[1] T.Nandagopal, S.Lu, and V.Bharghavan, "A unified architecture for the design and evaluation of wireless fair queueing algorithms," (Seattle, WA), pp. 132–142, In Proceedings of the ACM/IEEE International Conference on Mobile Computing and Networking, Aug. 1999.

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

[3] Y. Cao and V. Li, "Scheduling algorithms in broadband wireless networks," *Proceedings of the IEEE*, vol. 89, no. 1, pp. 76–87, 2001.

[4] I. Stojmenovic, *Handbook of wireless networks and mobile computing.* Jon Wiley and Sons Ltd, Feb. 2002.

[5] A. Viterbi, *CDMA: Principles of Spread Spectrum Communication.* Addison-Wesley, 1995.

[6] S. Glisic and B. Vucetic, *Spread Spectrum CDMA Systems for Wireless Communications.* Artech House Publishers, 1995.

[7] E. Dahlman, B. Gudmundson, M. Nilsson, and J. Skold, "Umts/imt-2000 based on wideband cdma," *IEEE Communications Magazine*, vol. 5, pp. 70–80, Sep. 1998.

[8] R. Prasad and T. Ojanpera, "An overview of cdma evolution toward wideband cdma," *IEEE Communications Survey & Tutorial*, vol. 1, pp. 2–29, 4^{th} Quarter 1998.

[9] D. N. Knisely, S. Kumar, S. Laha, and S. Nanda, "Evolution of wireless data services:is-95 to cdma2000," *IEEE Communications Magazine*, vol. 5, pp. 140–149, Oct. 1998.

[10] A. Goldsmith, *Wireless Communications*. Stanford University, 2004.

[11] N. D. Tripathi, J. H. Reed, and H. F. Banlandingham, "Handoff in cellular systems," *IEEE Personal Communications*, vol. 5, pp. 26–37, Dec. 1998.

[12] A. E. Leu and B. L. Mark, "A discrete-time approach to analyze hard handoff performance in cellular networks," *IEEE Transactions on Wireless Communications*, vol. 3, pp. 26–37, Sep. 2004.

[13] Y. Ishikawa and N. Umeda, "Capacity design and performance of call admission control in cellular cdma systems," *IEEE Journal on Selected Areas in Communications*, vol. 15, pp. 1627–1635, Oct. 1997.

[14] Q. Gao and A. Acampora, "Performance comparisons of admission control strategies for future wireless networks," vol. 3, pp. 284–288, WCNC 2002-IEEE Wireless Communications and Networking Conference, Mar. 2002.

[15] D. Shen and C. Ji, "Admission control of multimedia traffic for third generation cdma network," No. 1, pp. 1077–1086, IEEE INFOCOM 2000 - The Conference on Computer Communications, Mar. 2000.

[16] J. Zander, "Performance of optimum transmitter power control in cellular radio systems," *IEEE Transactions on Vehicular Technology*, vol. 41, pp. 51–62, Feb. 1992.

[17] D. M. Novakovic and M. L. Dukic, "Evolution of the power control techniques for ds-cdma toward 3g wireless communication systems," *IEEE Communications Survey & Tutorials*, vol. 3, pp. 2–15, 4^{th} Quarter 2000.

[18] R. Yates, "A framework for uplink power control in cellular radio systems," *IEEE Journal on Selected Areas in Communications*, vol. 43, pp. 1341–1347, Sep. 1995.

[19] M. Xiao, N. Shroff, and E. Chong, "Utility-based power control in cellular wireless systems," No. 1, pp. 412–421, In Proceedings of IEEE INFOCOM 2001, Apr. 2001.

[20] J.-F. Chamberland and V. V. Veeravalli, "Decentralized dynamic power control for cellular cdma systems," *IEEE Transactions on Wireless Communications*, vol. 2, pp. 549–559, May 2003.

[21] A. E. Leu and B. L. Mark, "An efficient timer-based hard handoff algorithm for cellular networks," vol. 4, pp. 1207–1212, WCNC 2003 - IEEE Wireless Communications and Networking Conference, Mar. 2003.

[22] C. W. Sung, "Analysis of fade margins for soft and hard handoffs in cellular cdma systems," *IEEE Transactions on Wireless Communications*, vol. 2, pp. 431–435, May 2003.

[23] D. Wong and T. J. Lim, "Soft handoffs in cdma mobile systems," *IEEE Personal Communications*, vol. 4, pp. 6–17, Dec. 1997.

[24] Y. Pan, M. Lee, J. B. Kim, and T. Suda, "An end-to-end multipath smooth handoff scheme for stream media," *IEEE Journal on Selected Areas in Communications*, vol. 22, pp. 653–663, May 2004.

[25] S. Sharma, N. Zhu, and T. cker Chiueh, "Low-latency mobile ip handoff for infrastructure-mode wireless lans," *IEEE Journal on Selected Areas in Communications*, vol. 22, pp. 643–652, May 2004.

[26] B. Hamdaoui and P. Ramanathan, "A network-layer soft handoff approach for mobile wireless ip-based systems," *IEEE Journal on Selected Areas in Communications*, vol. 22, pp. 630–642, May 2004.

[27] Y. Fang and I. Chlamtac, "Analytical generalized results for handoff probability in wireless networks," *IEEE Transactions on Communications*, vol. 50, pp. 396–399, Mar. 2002.

[28] D.-J. Lee and D.-H. Cho, "Channel-borrowing handoff scheme based on user mobility in cdma cellular systems," No. 1, pp. 685–689, ICC 2000-IEEE International Conference on Communications, Jun. 2000.

[29] S. Nanda, K. Balachandran, and S. Kumar, "Adaptation techniques in wireless packet data services," *IEEE Communications Magazine*, vol. 38, pp. 54–64, Jan. 2000.

[30] A. Goldsmith and S. G. Chua, "Adaptive coded modulation for fading channels," *IEEE Transactions on Communications*, vol. 46, pp. 595–602, May 1998.

[31] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushyana, and A. Viterbi, "Cdma/hdr: a bandwidth-efficient high-speed wireless data service for nomadic users," *IEEE Communications Magazine*, vol. 38, pp. 70–77, Jul. 2000.

[32] "1xev: 1x evolution, is-856 tia/eia standard," *QUALCOMM Inc.*

[33] M. Markaki, E. Nikolouzou, and I. Venieris, "Performance evaluation of scheduling algorithms for the internet," *in ATM and IP 2000, IFIP Workshop*, Jul. 2000.

[34] A. K. Parekh and R. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: the multiple node case," *IEEE/ACM Transaction on Networking*, vol. 2, pp. 137–150, Apr. 1994.

[35] A. K. Parekh and R. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: the single-node case," *IEEE/ACM Transactions on Networking*, vol. 1, pp. 344–357, Jun. 1993.

[36] A. Demers, S. Keshav, and S. Shenker, "Analysis and simulation of a fair queueing algorithm," pp. 3–12, In Proceeding ACM SIGCOMM'89, 1989.

[37] C.Kalmanek, H.Kanakia, and S.Keshav, "Rate controlled server for very high-speed networks," vol. 1, IEEE Global Telecommunications Conference, Dec. 1990.

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

[39] W. C. Y. Lee, *Mobile Communications Engineering*, vol. 1. New York: McGraw-Hill, 1982.

[40] V. Erceg, L. J. Greenstein, S. Y. Tjandra, S. R. Parkoff, A. Gupta, B. Kulic, A. A. Julius, and R. Bianchi, "An empirically based path loss model for wireless channels in suburban environments," *IEEE Journal on Selected Areas in Communication*, pp. 1205–1211, Jul. 1999.

[41] S. S. Ghassemzadeh, L. J. Greenstein, A. Kavcic, T. Sveinsson, and V. Tarokh, "Indoor path loss model for resisdential and commercial buildings," pp. 3115–3119, Proceeding Vehicle Technology Conference, Oct. 2003.

[42] R. Knopp and P. Humblet, "Information capacity and power control in single cell multiuser communications," in Proceeding IEEE International Communication Conference (ICC'95), Jun. 1995.

[43] X. Liu, E. Chong, and N. Shroff, "Opportunistic transmission scheduling with resource-sharing constraints in wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 19, Oct. 2001.

[44] X. Liu, E. Chong, and N. Shroff, "Transmission scheduling for efficient wireless resource utilization with minimum utility guarantee," (Atlantic City, America), Proceedings of the IEEE VTC Fall 2001, Oct. 2001.

[45] P. Bhagwat, A. Krishna, and S. Tripathi, "Enhancing throughput over wireless lan's using channel state dependent packet scheduling," pp. 1133–1140, In Proceeding INFOCOM'96, Mar. 1996.

[46] S. Lu, V. Bharghavan, and R. Sirkant, "Fair scheduling in wireless packet networks," *IEEE/ACM Trans. Networking*, vol. 7, pp. 473–489, Aug. 1999.

[47] T. S. E. Ng, I. Stoica, and H. Zhang, "Packet fair queueing algorithms for wireless networks with location-dependent errors," vol. 7, pp. 1103–1111, In INFOCOM (3), 1998.

[48] P.Ramanathan and P.Agrawal, "Adapting packet fair queueing algorithms to wireless networks," (Dallas, TX), pp. 1–9, In Proceedings of the ACM/IEEE International Conference on Mobile Computing and Networking, Oct. 1998.

[49] D.Eckhardt and P.Steenkiste, "Effort-limited fair(elf) scheduling for wireless networks," (Tel Aviv, Isreal), pp. 1097–1106, In Proceedings of IEEE INFOCOM, Mar. 2000.

[50] J. Holtzman, "Cdma forward link waterfilling power control," vol. 3, pp. 1663–1667, In Proceedings of IEEE Vehicular Technology Conference 2000-Spring, Oct. 2000.

[51] J. Holtzman, "Asymptotic analysis of proportional fair algorithm," vol. 2, In Proceedings of 12^{th} IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 2001.

[52] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Communications Magazine*, vol. 39, pp. 150–153, Feb. 2001.

[53] S. Shakkottai and R. Srikant, "Scheduling real-time traffic with deadlines over a wireless channel," (Seattle, WA), In Proceedings of ACM Workshop on Wireless and Mobile Multimedia, Aug. 1999.

[54] S. Shakkottai and A. Stolyar, *Scheduling algorithms for a mixture of real-time and non-real-time data in HDR.* Bell Laboratories Technical Report, 2000.

[55] N. Joshi, S. Kadaba, S. Patel, and G. Sundaram, "Downlink scheduling in cdma data networks," In Proceedings of ACM Mobicom 2000, 2000.

[56] S. Lu, V. Bharghavan, and R. Sirkant, "Fair queueing in wireless packet networks," (Cannes, France), pp. 63–74, In Proceedings of the ACM SIGCOMM'97, Sep. 1997.

[57] J. C. R. Bennett and H. Zhang, "wf^2q: Worst-case fair weighted fair queueing," vol. 7, (San Francisco, America), pp. 120–128, In Proceedings of IEEE INFOCOM, Mar. 1996.

[58] P. Goyal, H. Vin, and H. Chen, "Start-time fair queueing: A scheduling algorithm for integrated service access," (Palo Alto, CA), pp. 157–168, In Proceedings of ACM SIGCOMM'96, Aug. 1996.

[59] X. Liu, E. Chong, and N. Shroff, "A framework for opportunistic scheduling in wireless networks," *Computer Networks Journal (Elsevier)*, vol. 41, no. 10, pp. 451–474, 2003.

[60] H. J. Kushner and G. Yin, *Stochastic Approximation and Recursive Algorithms and Application.* Springer, 2003.

[61] S. Shakkottai and A. Stolyar, "Scheduling for multiple flows sharing a time-varying channel: The exponential rule," *Transaction of the AMS*, vol. A volume in memory of F. Karpelevich, 2001.

[62] S. Shakkottai and A. Stolyar, "Scheduling of a shared a time-varying channel: The exponential rule stability," (New York, America), In INFORMS Applied Probability Conference, Jul. 2001.

[63] X. Liu, "Opportunistic scheduling in wirless communication networks," PhD Dissertation, 2002.

[64] R. Yin, D. Dawoud, and H. Xu, "Opportunistic scheduling algorithms in wireless communication networks," (Cape town, South Africa), Proceeding of IEEE International Conference on Telecommunication (IEEE ICT) 2005, May 2005.