Modeling and Resource Management in Wireless Multimedia WCDMA Systems

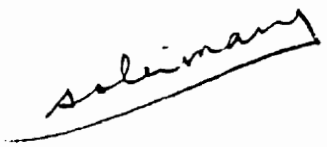
by

Majid Soleimanipour

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical Engineering

Waterloo, Ontario, Canada, 1999

The University of Waterloo requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.



Abstract

The focus of this thesis is on modeling and optimal resource management in wireless multimedia wideband code-division multiple access (WCDMA) systems. Resource management in CDMA systems has two major roles: to increase spectral efficiency by controlling cochannel interference, and to accommodate multimedia services by proper adjustment of the allocated resources. In this thesis, the system model supports a new class of services with limited delay tolerance in addition to real time and delay insensitive services which have been already considered in the literature. The basic control variables in CDMA resource management are transmit powers, data rates, and base station assignments (handoff decision). Previous works have included a subset of these variables in their resource management algorithms. In this thesis, these control variables are combined in a mathematical programming problem which maximizes the profit gained by a wireless multimedia service provider subject to satisfying the service and quality of service (QoS) requirement for each user. The profit is obtained as the difference between the network revenue and cost. A pricing scheme has been developed to map the network throughput onto the network revenue and the handoff switching overhead onto a certain cost. In this pricing scheme, every user pays proportionally to its instantaneous data rate and quality of service. The mathematical programming problem is analyzed, restructured, and solved for single- and multi-cell systems. The single-cell solution has the advantage of low complexity and global convergence in comparison with the previous work. Maximum achievable throughput (capacity) of a single cell is mathematically evaluated and has been used as the benchmark for performance measure of single- and multi-cell systems. For multi-cell systems, different solution approaches lead to different results. The best result is generated by the improved mixed-integer nonlinear programming (I-MINLP) algorithm which achieves up to

94% of the capacity in a network with 9 base stations, equivalent to a reuse efficiency of 0.94. The sensitivity of the resource management solution to erroneous input data (path gains) is examined. It has been shown that the higher the capacity utilization, the higher the vulnerability to estimation error. As a practical issue, resource management at different operation levels (centralized, partially decentralized, decentralized, and fully decentralized) is discussed. To facilitate the centralized implementation, a more practical version for the I-MINLP algorithm has been developed. The centralized and partially decentralized schemes are preferred for their high performance and flexibility. The contribution of this thesis is confined to the reverse link of WCDMA systems. We have assumed that necessary mechanisms to perform handoff procedures, when a user is assigned to a new base station, exist and have not addressed soft handoff. A resource management algorithm that combines the result of this research with the well known closed-loop power control is suggested to enhance reverse-link FDD-mode resource management in IMT-2000, the global standard for third-generation wireless systems. A part of this research has been reported in [1, 2].

Acknowledgements

All praise is due to God, the Lord of the Worlds, the Beneficent, the Merciful, the One, on Whom all depend, and none is like Him.

Many thanks to my parents for their support and encouragement throughout my life. I can never thank my wife and sons enough for their patience and understanding during my studies.

I acknowledge Professor George H. Freeman's supervision of this research and thank him for his valuable comments as well as his friendliness and moral support. To my other supervisor Professor Weihua Zhuang, I would like to express my appreciation for the insight and precision with which she co-supervised my research. Special thanks to her for reviewing the thesis during her maternity leave.

I am grateful to Professor J. P. Black for his constructive and effective comments.

I further acknowledge the financial support of the Ministry of Culture and Higher Education of Iran (Imam Hossein Division) for this research as well as the contribution of the Natural Sciences and Engineering Research Council of Canada (NSERC).

To my noble people
who rose up for independence, freedom, justice,
and revival of their dignity and great civilization;

to the memory of the great father
and eminent leader of my people;

and to the memory of the beloved ones
who are and will be missed for ever.

Contents

List of Tables

List of Figures

Abbreviations and Symbols

ARQ	Automatic repeat request
BER	Bit error rate
BPF	Band pass filter
B_{cb}	Coherent bandwidth
C	Network capacity
C_c	Cell capacity
C_n	Normalized capacity
CBR	Constant bit rate
CDMA	Code division multiple access
CIR	Carrier-to-Interference ratio
CLPC	Closed-loop power control
D	File size vector
DICOPT	Discrete continuous optimizer
DS-CDMA	Direct sequence CDMA
ER	Equality relaxation
E_b	Energy per bit
FDD	Frequency division duplex
FDMA	Frequency division multiple access
FTP	File transfer protocol
GAMS	General algebraic modeling system
I	Interference at a base station

I_0	Interference density
I-MINLP	Improved MINLP
IMT-2000	International mobile telecommunications for year 2000
IP	Integer programming
LCD	Long constrained delay data bearer service
LDD	Low delay data bearer service
LFP	Linear fractional programming
LP	Linear programming
LSA	Least signal attenuation
M	Number of base stations
MILP	Mixed-integer linear programming
MINLP	Mixed-integer nonlinear programming
MPA	Minimum power assignment
N	Number of wireless users
NBS	Nearest base station
NLP	Nonlinear programming
OA	Outer approximation
OLPC	Open-loop power control
PN	Pseudo-random noise
P_{inf}	Infeasibility probability vector
$P_{\max}$	Maximum user output power vector
$Q(.)$	Q-function
QPSK	Quadrature phase shifting key
$Q_{\max}$	Maximum Received signal power vector
QoS	Quality of service
RMINLP	Relaxed MINLP

R	Network throughput
R_c	Cell Throughput
R_n	Normalized Throughput
$R_{\max}$	Maximum rate vector
$R_{\min}$	Minimum rate vector
S	Set of all feasible assignments
S-MINLP	Simplified MINLP
SIR	Signal-to-Interference (plus noise) ratio
SNR	Signal-to-noise ratio
TC	Transmission cycle
TDD	Time division duplex
TDMA	Time division multiple access
T_b	Data bit duration
T_f	Time frame duration
T_m	Delay spread
T_s	Data symbol duration
UDD	Unconstrained delay data bearer service
VBR	Variable bit rate
VSG	Variable spreading gain
W	bandwidth
WCDMA	Wideband CDMA
a	Assignment vector
b	Binary assignment matrix
c	Continuous version of b
$c(t)$	Spreading PN sequence
e	Error estimation matrix

f	Interference ratio
$f_{d,\max}$	Maximum doppler frequency
g	Path gain matrix (vector in a single cell)
$\hat{g}$	Estimation of path gain matrix
h	Handoff variable
$h_{\max}$	Maximum number of handoffs per time frame
i	Global index referring to a wireless user
k	Global index referring to a base station
l	Global index referring to an assignment out of M^N possible assignments
p	User output power vector
q	Received signal power vector
r	Data rate matrix (vector in a single cell)
r_s	Source rate vector
r_{cod}	Coding rate
w	W/γ
λ	Vector of price per unit transmitted data (kbps)
λ_h	Cost per handoff switching
γ	Target SIR per bit vector
η	Background noise
τ	Delay bound vector
$\bar{\tau}$	Time average delay
$\hat{\tau}$	Average delay bound vector
τ_r	Residual delay bound vector

Chapter 1

Introduction

Wireless personal communications is the fastest growing segment of telecommunications [3]. This is driven by the desire for the change from wired fixed place-to-place communications to wireless mobile person-to-person communications, and to be free from physical connections to communication networks. On the other hand, increasing demand for new wireless services, such as internet connections, is pushing the technological barriers towards implementation of wireless multimedia[1] communications. Typical multimedia applications include teleconferencing, medical imaging, entertainment and educational video, and advertizing [4].

1.1 Motivation

A typical wireless multimedia system supports different services with a wide range of transmission rates and various error and delay performances. The network structure

[1]The term multimedia refers to the representation, storage, retrieval, and distribution of information expressed in multiple media such as: text, audio, video, graphics, and image [4].

consists of mobile handsets and terminals (for voice, data, image, and low-rate video), base stations, and a broadband network connecting base stations and control entities in the system. Figure 1.1 illustrates a general view of a broadband network with wired and wireless access points. In the wireless subnet, each communication

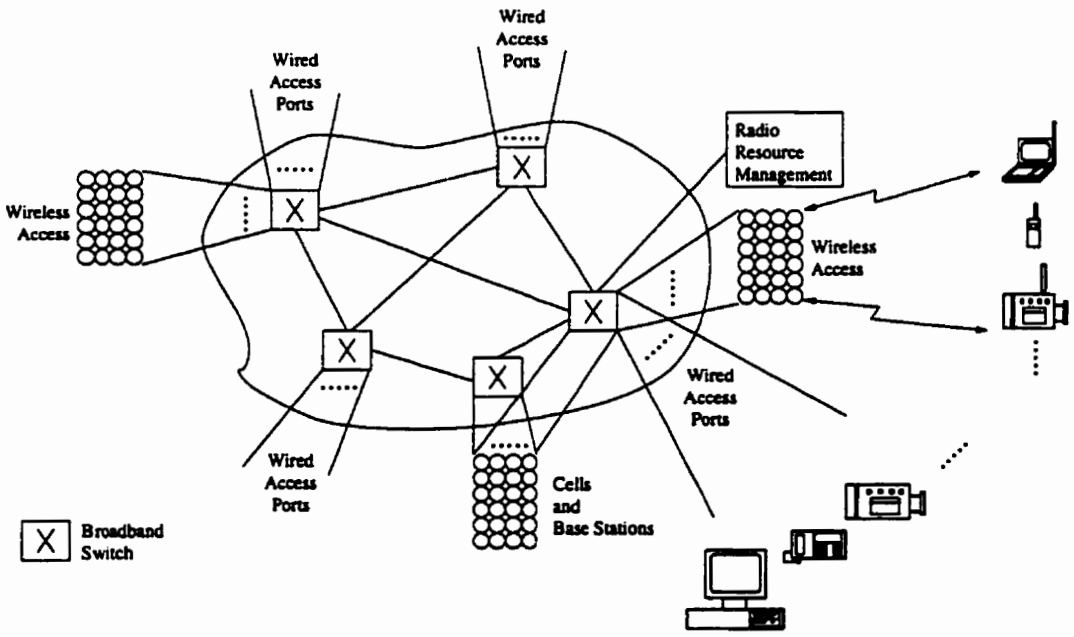


Figure 1.1: A typical wired/wireless multimedia network.

link is initiated by, or terminated at, a mobile/fixed user through base stations and the broadband network. Transmissions from mobiles to base stations (uplinks) and from base stations to mobiles (downlinks) take place in different radio channels (frequency division duplex, FDD) or the same radio channels but different time slots (time division duplex, TDD). Table 1.1 [5] illustrates the required transmission rates (bandwidth) for typical multimedia services in the uplink and downlink of a wireless network.

A signal transmitted over a mobile channel undergoes attenuation due to path

Table 1.1: Multimedia services and the uplink and downlink transmission rates [5].

Multimedia Services	Downlink Bandwidth	Uplink Bandwidth
Broadcast Video Broadcast TV Enhanced Pay per View	1.5 - 6 Mbps/channel	None
Interactive Video Video-on-Demand Interactive TV Interactive Games Information Retrieval Services	64 kbps - 6 Mbps	9.6 - 64 kbps
Internet (www, ftp, telnet) Voice	14.4 kbps - >10 Mbps	14.4 - 128 kbps
Symmetric Data Desktop Multimedia Work-at-Home Video Conferencing Video Telephony Fax	9.6 kbps - 2 Mbps	9.6 kbps - 2 Mbps
Small Business/Home Internet Homepage Internet Information Server	9.6 - 384 kbps	64 kbps - 1.5 Mbps

loss, shadowing due to uneven terrain, buildings, and other obstacles much larger than the wavelength of the radio channel frequency, and fading as a result of multiple paths and a time-variant channel. Thus, the received signal power varies and fluctuates randomly. If all mobiles transmit at the same power level, the signal from a near mobile usually will be received at a strong level, whereas that of the farthest one will be received at a very weak level; this is known as the near-far effect. The difference between these two received signal power levels can be in the range of 80 $\sim$ 100 dB [6, 7] which will cause an excessive amount of cochannel interference[2] and saturation of the weaker signal's receiver.

The frequency spectrum utilization of current mobile systems is very low. Optimistic estimates, as reported in [8], show that the highest achievable capacity among the second generation mobile systems falls under 20 percent of the Shannon channel capacity. This low utilization is due to the fact that the capacity of a Rayleigh fading channel, on the average, approaches that of a Gaussian channel only if the channel bandwidth or the number of diversity branches approaches infinity [9]. The implication is that in a real system, the capacity of a wideband mobile system using many diversity paths can be close to the channel capacity. Thus, a wireless CDMA[3] system is a potential candidate for high-capacity mobile communications due to the common wideband radio channel being shared among all the users. Further advantages, as will be discussed in Chapter 2, add to the CDMA capability in accommodating multimedia services and reliable communications over

[2]Cochannel interference refers to the interfering signals from the same radio channel. In conventional narrowband systems, cochannel interference implies inter-cell interference since a channel can be used once in a cell. In CDMA, however, it includes both inter- and intra-cell interferences since the same radio channel is used for users in every cell.

[3]CDMA (code division multiple access) technique versus other multiple access techniques is explained in Chapter 2. It refers to direct-sequence CDMA throughout this thesis.

fading channels. Hence, wireless communication standard bodies in Asia, Europe, and North America have proposed CDMA as the major multiple access technique for the third generation of wireless systems [10, 11, 12].

Power control algorithms have been developed to alleviate the near-far problem by compensating for the variations and fluctuations of the mobile signal power at the base station receiver. This power adjustment results in an increase in the system capacity [6, 13, 14] in terms of the number of simultaneous calls. In CDMA systems, where the radio channels are neither separated in frequency nor separated in time, the total interference level is a limiting factor for the system capacity [15]. Thus, excessive use of power by a single user could inhibit the communication of other users and reduce the capacity significantly. Therefore, power control and resource management are more important and essential in CDMA systems. CDMA uplink power control has been the center of focus in many investigations, because each user's signal in a cell follows a different propagation path and experiences different channel characteristics. On the other hand, the intra-cell interference and the desired signals on a downlink path undergo the same channel impairments and the relative power levels are preserved. In fact, on a single-cell basis, there is no need for downlink power control if the background noise is negligible. In a multi-cell environment, downlink power control should be employed but uplink power control is more complicated.

Besides combating the near-far problem, it has been shown, e.g. in [16], that power control can be engaged to accommodate different qualities of service (QoS's) in terms of bit error rate (BER). In the context of multimedia CDMA, the variable processing gain, as defined on page 13, can be employed as another means to control the BER. Thus, in a poor channel condition, where even the maximum available power fails to maintain the target BER, an increase in the processing gain, or

equivalently a decrease in the transmission rate, can further improve the average signal-to-interference ratio (SIR) and the BER. On the other hand, as proposed in this research, the data rate can be regulated such that the delay requirement of each user is met. Thus, the error and delay performance of a multimedia CDMA system can be controlled as a function of the resource (power and rate) budgets. It is highly desirable to manage the network resources efficiently so that the QoS requirement for each user is satisfied and the network resources are utilized maximally .

An important factor in allocating power and data rate to a mobile is the base station assignment, i.e., determination of the access point of the mobile to the wireless network. In other words, the amount of the allocated resources may vary significantly with different assignments. Conventionally, a mobile user is connected to the closest base station or to the one whose broadcast pilot signal is sensed as the strongest, implying that the path gain from the mobile to that base station is the highest with respect to all other base stations. We call the former assignment the nearest base station (NBS) assignment and the latter one the least signal attenuation (LSA) assignment. NBS assignment is valid in the absence of shadowing and fading. If the traffic is evenly distributed over the whole network and, consequently, each base station sees the same total interference at its receiving antenna, LSA assignment provides maximum SIR, thus the best performance. In unevenly-distributed traffic, however, a base station with high local traffic, despite being the choice of LSA, may receive a mobile signal with a lower SIR level than a nearby base station with a lighter local traffic. Therefore, an assignment decision based on the global traffic, at least in a cluster of nearby base stations, outperforms the LSA. We are interested in examining the performance of a globally-decided assignment as compared to the LSA, in other words, combining the base station assignment with the resource allocation process.

Optimization techniques are commonly employed in the literature for power control and resource management of cellular systems, as reviewed in the next chapter. Many power control algorithms are proposed for narrowband mobile systems to minimize the total transmitted power and increase the system capacity. Most of these algorithms are designed for conventional voice communications and do not address the problems involved in a multi-service environment such as the accommodation of different rates and service qualities for different users. As for CDMA systems, related literature has focused on variable-rate transmission and variable BER requirements for real-time services. However, many non-real-time services with various delay requirements have not been addressed. Moreover, the proposed models are limited to single-cell environments where roaming problems and associated challenges such as base station assignments are disregarded.

1.2 Objective

The objective of this thesis is to develop algorithms for optimal resource management of the reverse link of a wireless multimedia wideband CDMA (WCDMA) system in order to maximize utilization of network resources and capacity while guaranteeing service qualities.

The following network resources and related control variables are considered: the mobile transmitted powers, transmission rates, number of handoffs, and base station assignments[4]. The role of power control is to accommodate different BER requirements as well as to solve the near-far problem. Transmission rate control

[4]Base station assignment is equivalent to handoff decision. In this thesis only hard handoff decisions are considered and it has been assumed that necessary mechanisms to perform handoff procedures, when a user is assigned to a new base station, exist in the network.

is employed to satisfy both BER and delay performance requirements. The delay performance for non-real-time services is controlled by adjusting their service time within the range of their tolerable delays. The number of handoffs is controlled to reduce the signaling overhead due to handoff switchings. Among all feasible assignments, an assignment which yields the best value of an objective function is selected. The objective function is expressed in terms of a weighted sum of the network transmission rates and the number of handoffs. It reflects the network efficiency and is formulated in Chapter 4.

The development of the algorithms involves the following steps:

- *Introducing a model to support wireless multimedia services and accommodate different QoS's based on the CDMA technique.*
- *Translating the objective into an optimization problem based on the presented model.*
- *Developing efficient algorithms to solve and perform optimal resource management in single- and multi-cell environments.*
- *Investigating implementation issues including sensitivity analysis to measurement errors, centralized and decentralized implementations, and compatibility with the proposals for the third generation wireless systems.*

By fulfilling the above steps, the following contributions to the resource management problem in multimedia WCDMA have been achieved. The system model supports a new class of services with limited delay tolerance as well as real time and delay insensitive services which have been already considered in the literature. All major control variables are included in the resource management problem and are optimally obtained and allocated. Previous work has considered a subset of the

variables. A novel pricing strategy and a new throughput measure for multimedia services are developed. The single-cell solution has the advantage of low complexity and global convergence in comparison with the previous work. Resource management in a multi-cell system, despite its very high complexity, has achieved a high throughput performance within a reasonable and practicable time. The results of multi-cell solution for multimedia applications are new and can be used to enhance the reverse-link FDD-mode resource management in IMT-2000.

1.3 Overview of the Thesis

Chapter 2 includes a brief background on CDMA principles and characteristics followed by a discussion of the performance of CDMA systems in a mobile environment. A literature review of optimal power control and resource management in different systems (narrowband and wideband), applications (single and multiple service), and network environments (single and multiple cells) in addition to a detailed description of previous work relevant to this thesis is presented in this chapter. Chapter 3 is devoted to laying out the physical and mathematical description of our system model including service classes and QoS's. In Chapter 4, a mathematical programming problem is built to model optimal resource management in the multimedia CDMA system. Mathematical expressions for the objective function and constraints including a pricing scheme for multimedia services are developed and presented in this chapter. Moreover, the structure of the problem is analyzed in terms of the type and complexity. Resource management algorithms as the solutions to the problem are developed in Chapter 5 and validated by numerical results and simulations. Chapter 6 includes discussions related to implementation issues. Sensitivity of the resource management algorithms to the estimation errors

in path gain measurements, centralized versus decentralized schemes, and compatibility with the recent proposals for the third generation of wireless systems are discussed in this chapter. The conclusion and suggested future work are presented in Chapter 7. Some relevant theories and techniques in optimization theory are provided in the Appendix.

Chapter 2

Background and Literature Review

The focus of this thesis is on the problem of optimal power control and resource management in a CDMA context and multimedia application. This chapter provides a brief background on wireless CDMA systems and their characteristics, and a review of the relevant previous work. More details of the literature closely related to this work are included as well. Background information on optimization theory and techniques including relevant definitions, theorems, and algorithms are provided in Appendix A.

2.1 Mobile CDMA Systems

The introduction of the spread spectrum technique for mobile communications goes back to 1978 when a frequency-hopping system for high-capacity mobile communications was proposed in [17]. This proposal, which was the first spread-spectrum

approach to mobile communications, stimulated considerable interest and controversy. In particular, the application of direct-sequence code-division multiple access (DS-CDMA) in cellular systems was introduced at the beginning of the 1990s [6, 18], years after its application in military and satellite communications. For simplicity, throughout this proposal the term CDMA stands for DS-CDMA.

Unlike FDMA (frequency division multiple access) and TDMA (time division multiple access), which divide the available frequency spectrum into narrowband channels and assign each radio channel to one or more calls, CDMA is a wideband spread-spectrum scheme that spreads multiple calls across a wide segment of the allocated spectrum. Each individual call is assigned a unique code that permits it to be distinguished from the multitude of calls transmitted simultaneously over the same band. As long as the receiving end has the right code, it can pick its own signal out from all the others.

The principle of direct-sequence spread spectrum is as follows. Consider the transmitted signal $As(t)c(t)\cos(\omega_c t + \phi)$ in Figure 2.1, where $s(t)$ is the encoded and interleaved signal, $c(t)$ is the pseudo-random spreading signal comprising a bipolar signal (± 1) and $A\cos(\omega_c t + \phi)$ is the carrier. Coding techniques are employed in error-prone mobile channels to enhance the BER performance or reduce the required power. Interleaving techniques randomize bursts of error that occur in a fading environment to increase coding gain for codes designed designed for memoryless channels. The band-pass filter (BPF) is assumed to be ideal and to pass the modulated and spread signal without distortion. The signal is impaired by the channel noise, interference, shadowing, and multipath propagation. The received signal is coherently multiplied by the spreading sequence $c(t)$ and detected at the receiver. It is a common method to use a particular spreading signal with different time shifts for multiple access signals. Thus, if the autocorrelation of $c(t)$ satisfies

the condition $E[c(t)c(t+\tau)] \approx 0$ for all $|\tau| > T_{chip}$, where T_{chip} is the chip duration of $c(t)$, then only received signals with the same time-shifted code are detected and all other signals with any time shift $|\tau| > T_{chip}$ act as wideband interferences at the receiver.

This condition prescribes that $c(t)$ must have the properties of white noise. It is a common practice to compromise between a long pseudo-random code with low autocorrelation and a shorter sequence with a faster acquisition and tracking time [19]. An important parameter of a CDMA system is the spreading factor

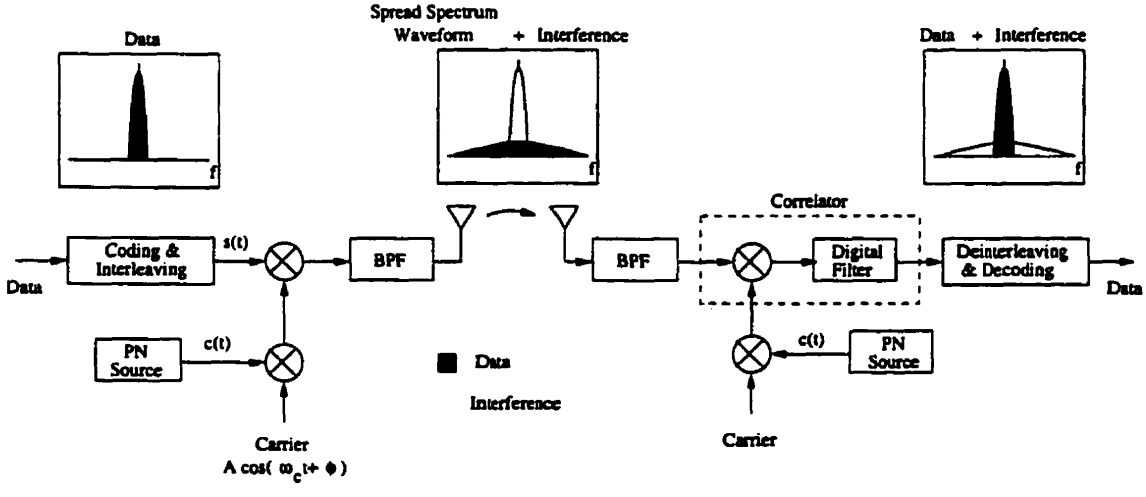


Figure 2.1: A view of the direct sequence spread spectrum concept.

or processing gain which is defined as the ratio of the data bit duration (T_b) to the spreading signal chip duration (T_{chip}), or the ratio of the spread spectrum bandwidth (W) to the data rate for user i (r_i), that is

$$\text{Processing gain} = \frac{T_b}{T_{chip}} = \frac{W}{r_i}. \tag{2.1}$$

The processing gain is an indication of the excessive use of the frequency band for spreading signals and most of the CDMA advantages, as explained below, are direct

results of this excessive use of the spectrum. In general, it is desirable to have a high processing gain to take advantage of CDMA characteristics. However, since the radio channel bandwidth is limited, increasing the processing gain imposes a certain restriction on the data rate. We recognize this inherent limit in CDMA and will set an upper bound on the transmission rate to preserve a high processing gain.

The performance of a CDMA system in terms of the spectral efficiency and BER is highly affected by the amount of the multiple access (cochannel) interference. Therefore, it is essential to use a number of techniques for interference reduction such as power control, sectorization and voice activity monitoring to achieve a high performance [6]. Interference from the adjacent channel is another degrading factor. In CDMA, interference from the adjacent channel refers to two different type of interference interchangeably. One is from the adjacent wideband radio channel which can be alleviated by a guard band. The other one is the interference caused as a result of non-zero cross-correlation between CDMA codes. The latter is also referred to as the multiple-access interference in some literature. Some of the important characteristics and advantages of CDMA are explained in the following.

Frequency Reuse - In multi-cell narrowband systems such as FDMA and TDMA, a radio channel can be used so that the cochannel interference for mobiles sharing this channel is less than a certain level. In practice, only a fraction of the frequency band can be used in a sector or cell. For example, in a hexagonal cell structure, FDMA systems normally use one-seventh of the available radio channels, a frequency reuse factor of $1/7$. To increase the reuse efficiency, narrowband systems have to employ sophisticated frequency planning schemes and dynamic channel allocation. CDMA shares the same spectrum in all cells (universal frequency reuse) and achieves a frequency reuse factor of approximately 1, thus gaining several times higher spectrum efficiency and getting rid of the frequency-planning complexity.

Power Control - Due to an inherently high cochannel interference in CDMA systems, power control is an absolute requirement which affects the overall system performance. Extensive research, e.g. [20, 75, 22], has shown that power control error has a significant effect on both reverse- and forward-link capacity. Therefore, power control has to be dealt with carefully in the design of CDMA cellular systems.

Soft Handoff - Handoff or handover is the mechanism for the transition of a mobile station from one cell or sector to another cell or sector. If this transition is done abruptly, i.e., connecting to the new cell after disconnecting from the old cell, it is called hard handoff. If the transition is gradual, i.e., establishing the connection to the new cell before disconnecting from the old cell, it is called soft handoff. Hard handoff can fail if there is no radio link in the new cell, or the mobile is switched to a wrong cell, or the mobile fails to hear the command to switch the channel. By simultaneous communications via two or more base stations, soft handoff greatly reduces link drops during transitions and enhances the capacity by exploiting the link diversity. Obviously, CDMA is able to perform soft handoff very easily because the same radio channel is used in the adjacent cells. However, each handoff is subject to a signaling overhead to establish the new connection and to update the corresponding user information.

Diversity - In addition to common diversity techniques used in narrowband and wideband systems, e.g., time diversity by interleaving, there are certain types of diversity which inherently exist or are easily employed in CDMA systems such as:

- link diversity during soft handoff, and
- inherent frequency diversity due to the wide bandwidth which resolves the multipath components of the received signal and provides the receiver with several

independent paths.

To combine the received signals from different diversity paths, a simple and common method is selection combining [23]. In this method, received signals from diversity paths are compared and, usually, the one with the highest carrier-to-noise ratio is selected. A popular and optimum method of combining is maximal ratio combining in which diversity branches are co-phased and then weighted proportionally to their signal level before summing [23]. However, due to its need for accurate channel estimates, maximal ratio combining is more appropriate for slow fading channels. This technique can be implemented by RAKE receivers in CDMA systems. Because CDMA uses a large bandwidth, the time resolution for resolving the signals from diversity branches is high, therefore, maximal ratio combining can be implemented efficiently in CDMA systems.

Error Performance of CDMA in Mobile Channels

In a Gaussian channel, the probability of error for a single direct-sequence spread-spectrum signal is shown to be uniquely identified by the signal-to-noise ratio (SNR) per bit [24]. If the number of users and the processing gain is high enough, it is quite common to approximate the sum of ambient noise and multiple-access interference in a CDMA system by an uncorrelated Gaussian signal with a flat spectrum. Thus, a one-to-one relation between the BER and signal-to-interference plus noise ratio (SIR) per bit does still exist. The SIR per bit for user i is calculated as

$$\left(\frac{E_b}{I_0}\right)_i = \frac{W}{r_i}\frac{q_i}{I} \tag{2.2}$$

where E_b is the energy per bit, I_0 is the interference plus noise density, W is the channel bandwidth, r_i is the data rate, q_i is the received power, and I is the total noise plus interference seen by user i. For example, for an uncoded signal using

QPSK (quadrature phase shift keying) modulation, an SIR per bit of 9 dB will guarantee a BER of less than 10^{-5} [23]. In [25], however, it has been argued that the Gaussian approximation will over-estimate the error performance if the mobile system is low-populated or there are dominant multiple access components (no power control).

In a flat (frequency non-selective) fading channel, whose coherence bandwidth[1] B_{cb} is larger than the signal bandwidth, the signal spectrum remains flat but the received power fluctuates randomly. In this case, the average SIR per bit, $\overline{E_b/I_0}$, represents an average BER performance [23]. A flat fading condition can be applied if $W < B_{cb} = 1/T_m$ or $T_m < 1/W$ where T_m is the delay spread.

In a frequency selective fading channel, $T_m > 1/W$, multipath signals are resolvable and can be exploited by a RAKE receiver for higher spectrum efficiency. This channel is modeled by a tapped delay line (transversal filter) with $T_m W$ taps [23]. Typical values for T_m range from .1 to 3 and 3 to 10 μsec in an indoor and outdoor environment, respectively [7]. Thus, indoor and outdoor CDMA channels with bandwidths of greater than 1 MHz fall in the category of frequency selective fading and the multipath effect can be mitigated at a cost of higher receiver complexity. The error performance of this channel is also characterized by the average SIR per bit.

2.2 Optimal Power Control

Since the introduction of cellular mobile systems, power control has been an important part of the resource management. This problem has been among many

[1]Coherence bandwidth is a part of the spectrum which is faded simultaneously.

problems in numerous applications which have been approached by optimization techniques. Many power-control algorithms are proposed in the literature for narrowband systems to minimize the mobiles' transmitted powers and increase the system capacity. Despite most of these algorithms being designed for single-service communications, the ideas can be applied to multi-service applications, as well. Optimal resource management for CDMA systems in multi-service communications has been considered in the literature. Researchers have focused on variable-rate transmission and variable BER requirements for real-time services. However, non-real-time services with various delay requirements have not been addressed. Moreover, the previously proposed models are limited to single-cell systems and the problems associated with the multi-cell environment have not been considered.

Many power-control algorithms have been proposed to combat cochannel interference and the near-far problem by regulating the mobiles' transmitted powers such that each user communicates with a certain target SIR while spending the minimum required power. This objective has been formulated as an optimization problem in the literature. Power control for cochannel interference reduction in satellite systems has been investigated in [30] where the concept of carrier-to-interference ratio (CIR) balancing is introduced for voice communications with the same BER requirements. CIR balancing is intended to balance interference such that each user experiences the same CIR level. In [31], this concept is extended to spread-spectrum systems. In this case, the effect of cochannel interference has been taken into account and shown to be dominant (compared to the background noise). It is also shown that CIR balancing improves the system capacity significantly. Another trend in power control algorithms is to keep the received signal power at a constant level [6]. Although this is very efficient in controlling intra-cell interference in CDMA systems, it has been shown in [32] that it has a limited ef-

fect on inter-cell interference. A good mathematical model and a centralized power control algorithm are presented in [13]. In this algorithm, the transmitted power vector is optimized in a centralized and synchronous fashion to balance CIR. By a synchronous algorithm, it is meant that all the mobiles' transmitted powers are updated synchronously, otherwise, the algorithm is asynchronous.

In the above algorithms, the problem of optimizing the transmitted power vector is identified as an eigenvalue problem and the optimum power vector is found by inversion of a nonnegative matrix which is composed of channel gains for each individual user. The main problem with these centralized algorithms is computational complexity and knowledge of path gains at the control center. Hence, for practical purposes, iterative and distributed versions of these algorithms have been developed. These algorithms use locally-measured path gains [14],[33]-[36]. The other problem is that they require knowledge of one or more of the parameters such as path gains (in general path gains from each user to a number of base stations), CIR, received interference power, or bit error rate (BER). Stochastic power control for CDMA systems in [37] employs an iterative and distributed algorithm to find the optimal power vector. The advantage of this algorithm is that a mobile only needs to know the matched filter output at its corresponding base station and its path gain.

2.3 Combined Power and Base Station Assignment

An important fact in optimizing power control is that the previously proposed algorithms often find the optimal power vector assuming that the assignment of a

mobile to a base station is determined and controlled by other criteria.

However, it is likely that there exists a better solution under another assignment. This has been the motivation for developing optimal solutions for combined power control and base station assignment. In [38], an iterative algorithm, namely minimum power assignment (MPA), for both synchronous and asynchronous cases, is proposed, and convergence problems are discussed. In each iteration, every mobile finds a base station to which the required power is minimum, then, based on the updated assignments, the new power vector is derived.

A very similar MPA algorithm is developed independently by Hanly in [39]. In this work, assuming a deterministic log-linear propagation model, the set of potential base stations to be connected to a particular user is confined to nearby base stations. The combined optimum power and base station assignment, as a result, expands or contracts cell-site radii with respect to their traffic load. The proposed algorithm increases the system capacity and is implemented decentrally. Each mobile must have knowledge of its path gains to all base stations (or at least a subset of nearby base stations) and their corresponding interference levels to select the one which requires the least amount of power. The higher capacity, however, is achieved at the cost of higher complexity of mobile terminals and communication overhead. In [40], a dynamic load-sharing algorithm is proposed to eliminate the communication overhead in Hanly's algorithm. The idea is to push the mobiles in the boundary regions to an early handoff when a cell is overloaded and the target SIR per bit cannot be maintained. This is done by a stepwise decrease in the pilot signal power of the overloaded base station. A low pilot strength results in cell size reduction as well as an overall increase in the transmitted power in the reverse link inversely proportional to the pilot signal power. It can be argued that, first, the extra power has an adverse effect on the system capacity, and second, a

higher power for all mobiles has almost no effect on the SIR because the interference would also increase proportionally (background noise is very small compared to the cochannel interference). Furthermore, when a mobile is pushed to an early handoff to another base station, the highest path gain is still to the crowded base. This means that under similar interference levels at both base stations, the user needs more power to communicate with the new base station. If the interference level is lower at the new base station, then the required power would be less. How much less depends on the difference in the path gains and the interference levels. So there is no guarantee of less power in the new cell and, consequently, no guarantee for improvement in the old cell.

A combined channel allocation, base station assignment, and power control algorithm is also proposed in [41]. This algorithm considers the combined problem for a system of 2 base stations and computes a maximum matching in a graph that captures the topological characteristics of the mobile locations. A very similar work is reported in [42]. The solution approach is different in the sense that a Lagrangian relaxation method is employed and a multi-cell model is considered. This work does not provide a real-time solution because of the high computational complexity. The details of the MPA algorithm are described below.

Minimum Power Assignment Algorithm

We are interested in the MPA algorithm [38] for its formulation of the optimal power control and base station assignment in one model. Although this algorithm is designed primarily for narrowband systems, where dynamic channel allocation is a challenging part of the network operation (to reduce cochannel interference and increase capacity), the idea can be applied to CDMA systems to manage the handoff process very efficiently and to distribute the traffic load in order to mitigate

local network congestion.

The network consists of M base stations and N users. All time-dependent parameters are defined at discrete-time n, when the algorithm updates power and base station assignment vectors. Each user is connected, or assigned, to one (hard handoff) base station. If user i is assigned to base station k, we denote this assignment, following [38], as $a_i = k$ where $i \in \{1, \cdots, N\}$ and $k \in \{1, \cdots, M\}$. With a combination of M base stations and N users, there are in total M^N distinct assignments. All possible assignments are numbered from 1 to M^N and each particular assignment is specified by a superscript. Thus, the l^{th} assignment is represented by vector a^l with length N.

The path gain $g_{ik}^l(n)$ denotes the attenuation of the signal transmitted from user i to the base station k under the specific assignment vector $a^l = [a_1^l, \cdots, a_N^l]$. If user i transmits at a power level $p_i(n)$ (assumed to be unlimited), the received signal from user i at the base station k is equal to $p_i(n)g_{ik}^l(n)$ at time n. The path gain as a function of a particular assignment a_j^l is also useful and is represented as $g_{ia_j}^l(n)$. If the assignment $a_j^l = k$, then, $g_{ik}^l(n) = g_{ia_j}^l(n)$. The optimization problem is modeled as

$$\underset{a^l, p(n)}{\text{Minimize}} \quad \sum_{i=1}^{N} p_i(n) \tag{2.3}$$

subject to

$$\left(\frac{C}{I}(n)\right)_i = \frac{g_{ia_i}^l(n)p_i(n)}{\sum_{j=1, j\neq i}^{N} g_{ja_i}^l(n)p_j(n) + \eta_{a_i}^l} \geq \gamma_i \tag{2.4}$$

$$p_i(n) \geq 0 \tag{2.5}$$

where $a_i^l \in \{1, \cdots, M\}$, $i \in \{1, \cdots, N\}$, and $l \in \{1, \cdots, M^N\}$, $p(n)$ is the transmit power vector, and $\eta_{a_i}^l$ is the receiver noise at the assigned base station. The CIR

constraint in (2.4) guarantees the error performance for user i specified by γ_i and can be written as

$$p_i(n) \geq G_i^l(n)p(n) + \sigma_i^l(n) \tag{2.6}$$

where $\sigma_i^l(n) = \gamma_i \eta_{a_i}^l / g_{ia_i}^l(n)$ and $G_i^l(n)$ is the i^{th} row of the $N \times N$ path gain matrix $G^l(n)$ defined as

$$G_{ik}^l(n) = \begin{cases} \gamma_i g_{ja_i}^l(n)/g_{ia_i}^l(n) & \text{for } i \neq j \\ 0 & \text{for } i = j. \end{cases} \tag{2.7}$$

The set of feasible solutions under assignment l is defined as

$$P^l(n) = \left\{ p(n) \geq 0 | p(n) \geq G^l(n)p(n) + \sigma^l(n) \right\} \tag{2.8}$$

where $\sigma^l(n)$ is an $N \times 1$ vector whose element i is $\sigma_i^l(n)$. The set $P^l(n)$ describes a cone of feasible power vectors in which, if $p(n) \in P^l(n)$ then $\alpha p(n) \in P^l(n)$ for $\alpha \geq 1$. Figure 2.2 depicts the set of feasible power vectors for a system of 2 users

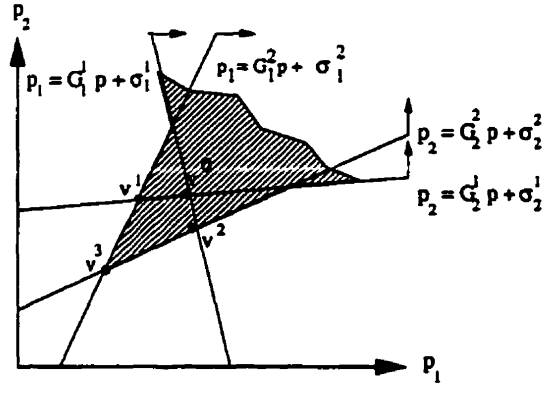


Figure 2.2: The feasible region and optimal solution for a system of 2 base stations and 2 mobiles [38].

and 2 base stations. In this figure the line $p_i(n) = G_i^l(n)p(n) + \sigma_i^l(n)$ describes the minimum power that user i needs to communicate with the assigned base (1

or 2) as a function of the transmitted power of the other user. There are four possible assignments. For each assignment, the feasible region is the area of the upper triangle formed by the intersection of the lines $p_1(n) = G_1^j(n)p(n) + \sigma_1^j(n)$ and $p_2(n) = G_2^j(n)p(n) + \sigma_2^j(n)$ where $j = 1, 2$. The minimum power required for each assignment is the intersection point or the vertex v^l. In Figure 2.2 the optimal solution is $v^3 = v^*$ and MPA algorithm should search for this global minimum. The union of all feasible regions is a nonconvex set $\bigcup_l P^l$, the shaded region in Figure 2.2. Given a power vector $p(n)$ and an assignment vector $a^l(n)$, the MPA algorithm iteratively searches for the optimal point v^* using the iteration

$$p(n+1) = G^l(n)p(n) + \sigma(n) \tag{2.9}$$

where n is the current iteration number and $G^l(n)$ is the path gain matrix for the assignment a^l at iteration n. In this algorithm, each user evaluates the minimum power required to transmit to all base stations at iteration $n+1$ based on the transmitted power of other users at iteration n. Then it switches to the base station corresponding to the lowest minimum required power, and sets its output power to the lowest required power level. Thus, a new assignment vector and a new power vector are obtained. In the iteration equation (2.9), if $G^l(n)$ is a nonnegative matrix and $p(n)$ is a feasible solution, from the Perron-Frobenius theorem [43], the above iteration converges to a unique fixed point if and only if the largest eigenvalue of $G^l(n)$ is less than unity.

The MPA algorithm has several problems:

- In the framework of the optimization problem, the new assignment vector should be obtained based on the current value of the power vector, whereas in MPA, each user's assignment is obtained based on the previous values of other users' output power. This may not be a major problem only if the channel condition

and users' locations do not change during two consecutive time intervals for updating transmitted powers.

- The initial power vector in (2.9) must be a feasible power vector. It is not clear how the algorithm finds this initial feasible value.

2.4 Handoff Trade-offs

A major problem with the combined power control and base station assignment optimization is that the number of new assignments (handoffs) can be untolerably high, as each handoff requires additional signalings for establishing the new connection as well as updating users' information in the network. There is also a high probability of "ping-pong" effect (constant switching between base stations). The trade-off between the best connection quality and handoff switching cost has been investigated independently from power control optimization. An approach based on statistical control theory is taken to obtain an optimal strategy for the handoff problem in [44]. Using a Markov decision process formulation and dynamic programming, the optimum decision is characterized by a threshold on the difference between the measured powers that the mobile receives from the base stations. A similar approach is taken in [45] where the cost and reward functions for deciding on a mobile-assisted handoff are formulated in a more general form. It includes the model in [44] as a special case.

2.5 Resource Management for Multimedia

Most power control algorithms mentioned so far are intended to optimize certain variables for conventional mobile systems designed for voice communications. In recent research publications, interest has been shown in the CDMA technique, particularly for multimedia applications. High flexibility and multipath resistance are some of the desirable characteristics of CDMA for future mobile systems [46]. Power control algorithms for multimedia CDMA have been proposed as a means to control the communication quality in terms of BER as well as solving the near-far problem.

To accommodate mixed traffic with different rates and QoS requirements, the concept of variable spreading gain (VSG) is proposed for CDMA systems in [47]. In this method, users with different transmission rates and the same BER requirements are allocated power levels proportional to their rates and, for users with the same rates, power levels are proportional to their E_b/I_0 requirements. In [16], multimedia sources are separated into multiple substreams with each substream's delay and loss characteristics individually negotiated between the source and network. In this work, only delay-sensitive substreams with different E_b/I_0 requirements are considered. It is suggested that power control be employed to achieve different QoS's for different substreams. The proposed power control algorithm is similar to MPA [14] and is intended to minimize total transmitted power for the downlink in a single-cell network. Moreover, the feasibility condition is analytically derived and used for the call admission purposes.

Single-Cell Multimedia CDMA System

In the following, we present a detailed review of [48] as it is considered relevant to our research. In this paper, a multimedia CDMA system is considered for a

single cell. It is intended to minimize total transmitted power in the cell to reduce total interference to other cells, and maximize total data rates in two separate optimization problems for the uplink transmissions. The first optimization is not directly related to our proposed research as we will consider a multi-cell system and control the system resources on a network-based configuration. Thus, we are not concerned about the effect of the interference outside of the network. In the following, we will describe the details of the second optimization problem, that is, maximizing the total transmission rate.

The objective is that the system wishes to give each user the best throughput under specified constraints. For example, if all constant bit-rate (CBR) services are taken care of, the system assigns the remaining rate such that the cell throughput is maximized. In the proposed model, there are N users in a single cell. The received powers at the base station and mobile transmission rates, represented by the vectors $q(n) = [q_1(n), \cdots, q_N(n)]$ and $r(n) = [r_1(n), \cdots, r_N(n)]$, respectively, constitute the system variables and are assumed to be subject to fixed maximum received power vector $Q_{\max} = [Q_{1,\max}, \cdots, Q_{N,\max}]$ and constant minimum rate vector $R_{\min} = [R_{1,\min}, \cdots, R_{N,\min}]$, respectively. The inequality

$$\left(\frac{E_b}{I_0}(n)\right)_i = \frac{W}{r_i(n)} \frac{q_i(n)}{\sum_{j=1, j\neq i} q_j(n) + \eta} \geq \gamma_i, \quad i \in \{1, \cdots, N\} \tag{2.10}$$

where W is the channel bandwidth and η is the background noise at the base station receiver, guarantees the error performance if γ_i is the target SIR per bit for user i. It is stated that, at the optimal solution, the constraint (2.10) is met with equality. In other words, it is not desirable to allocate extra power to achieve better quality than what is expected from γ_i.

The problem is formulated as

$$\underset{q(n), r(n)}{\text{Maximize}} \quad \sum_{i=1}^{N} r_i(n) \tag{2.11}$$

subject to

$$\frac{W}{r_i(n)} \frac{q_i(n)}{\sum_{j=1, j\neq i} q_j(n) + \eta} = \gamma_i \quad (2.12)$$

$$0 < q_i(n) \leq Q_{i,\max} \quad (2.13)$$

$$r_i(n) \geq R_{i,\min}. \quad (2.14)$$

This problem is reformulated by using the value of $r_i(n)$ evaluated from (2.12) and substituting in (2.11) and (2.14), thus, the rate vector is omitted and the problem changes to

$$\underset{q(n)}{\text{Maximize}} \quad \sum_{i=1}^{N} \frac{W}{\gamma_i} \frac{q_i(n)}{\sum_{j=1, j\neq i} q_j(n) + \eta} \quad (2.15)$$

subject to

$$0 < q_i(n) \leq Q_{i,\max} \quad (2.16)$$

$$\frac{W}{\gamma_i} \frac{q_i(n)}{\sum_{j=1, j\neq i} q_j(n) + \eta} \geq R_{i,\min}. \quad (2.17)$$

The above substitution reduces the number of variables and transfers the nonlinearity in the constraints to the objective function. The new formulation has the advantage of having a convex search space restricted to the surface of a polyhedron defined by the set of linear constraints. The gradient projection method [49] is used for solving the problem. Any initial vector is first checked for feasibility, i.e., satisfying all constraints, then is used as the first iteration for the gradient projection method. It is found that the method converges quickly to the solution. However, for larger values of $R_{i,\min}$, the algorithm converges to local minima, implying non-convexity of the objective function. To avoid the local minima, different initial guesses have been used. The work in [48] has several problems and limitations:

- In a multimedia environment, the sum $\sum_{i=1}^{N} r_i(n)$ does not represent the network throughput, or spectral efficiency, as each user may transmit data with a

different quality. Thus, the value of the sum may be increased by reducing the transmission quality without affecting the network performance.

- The multi-cell environment and related problems are not addressed. For example, in a single-cell system, the set of linear constraints has a very simple form and analytical expressions for the feasibility condition can be derived easily. This follows because (2.17) has the algebraic form

$$A_i q(n) \geq \eta \tag{2.18}$$

where A_i is the i^{th} row of the symmetric matrix A defined as

$$A_{ij} = \begin{cases} W_i & \text{if } i = j \\ -1 & \text{if } i \neq j \end{cases} \tag{2.19}$$

with $W_i = W/(\gamma_i R_{i,\min})$. Matrix A has a simple form and an analytical solution for (2.17) with equality exists. Solving for $q_1(n)$, we get [48]

$$(W_1 + 1)q_1(n)\left[1 - \sum_{j=1}^{N} \frac{1}{W_j + 1}\right] = \eta. \tag{2.20}$$

Positivity of the bracketed term implies

$$\sum_{j=1}^{N} \frac{1}{W_j + 1} < 1 \tag{2.21}$$

which provides the condition for not having an empty feasible set. This result is also reported in [16] using a different approach. The structure of matrix A does not change if we consider the transmitted power instead of the received power and the condition (2.21) remains valid. When dealing with multi-cell systems while working with the transmitted power p_i, the matrix A is no longer symmetric and derivation of an analytical solution is more challenging. Moreover, the mobility of users is restricted to a limited area and there is no

handoff problem and corresponding complexities. In general, the analysis of a multi-cell system is much more challenging in all analytical, numerical, and operational aspects.

- The proposed multimedia model does not include a wide range of multimedia services such as non-real-time delay-sensitive and delay-tolerant services, as categorized in class II and III in our system model in Chapter 3. The model in [48] considers only real-time services and, in fact, is very similar to the current voice-communication mobile systems except capturing variable rates and BER requirements. Consequently, the QoS is limited to error performance, and other important quality measures such as delay bounds have not been taken into account.

2.6 Summary

Although extensive academic efforts have been devoted to optimal resource management, the ground for further research and significant contributions potentially exists in this area. As far as a true wireless multimedia system is considered, there is no proposed model that comprehensively supports a wide range of multimedia applications with corresponding QoS requirements. Development of such a model is of great importance. For practical indoor and outdoor applications, the resource management problem should be defined based on a multi-cell system, thus, related challenging problems such as base station assignments should be contained in the resource management process. Algorithms should be developed to solve for the optimal values of the system resources to achieve certain objectives such as maximum throughput subject to satisfactory QoS's. The results of these developments would be of great interest to wireless service providers.

Chapter 3

WCDMA System Model

A wireless network including N users and M base stations is considered with no pre-specified cell borders. The task of assigning a user to a base station is combined with resource management. In this chapter, a detailed physical and mathematical description of the proposed wireless multimedia CDMA system model, including service classes and QoS's, is presented.

3.1 Supported Services

3.1.1 Service Classes

We accommodate three classes of wireless services similar to those in [50]. Class I consists of highly delay-sensitive real-time connections with zero delay-tolerance such as voice and low-rate video. This class has the highest service priority over the other classes.

Class II includes non-real-time delay-sensitive services with a small delay bound

(in the range of seconds) such as remote log-in, FTP (file transfer protocol) and similar applications associated with transport control protocols (TCP). If a channel is unavailable or impaired by deep fading, this type of connection can be queued for a short time. This class has service priority over class III.

Class III includes delay-tolerant services such as paging, electronic mail, voice mail, fax and data file transfer, and can be conveyed at the earliest possible time.

Both CBR and VBR (variable bit rate) services are supported in each class. A user priority can also be defined to satisfy urgent needs of mobiles. This priority shall be negotiated at the call initiation phase and dominates service priority.

In recent IMT-2000 proposals, e.g. [10]-[12], similarly, three classes of data services are considered: low delay data bearer services (LDD), long constrained delay data bearer services (LCD), and unconstrained delay data bearer services (UDD). In contrast to LCD, our Class II services include the whole range of short and long delay tolerant services.

3.1.2 Quality of Service

The model supports a class-based QoS. In the scope of this research, BER and delay bound identify the QoS.

Bit Error Rate

Bit error rate is a major indicator of the quality of service and, as described in 2.1.1, related to the average E_b/I_0 at the receiver. The relation between BER and average SIR per bit is one-to-one and is specified by the type of channel coding, modulation, and structure of the receiver. For any communication system, it is highly desirable

to employ different techniques such as forward error correction (FEC), bandwidth-efficient modulation, and diversity schemes to reduce the required SIR per bit and use the available resources more efficiently. The target BER for voice and data are typically 10^{-3} and 10^{-6}, respectively [10]-[12]. Lower BER's are achievable by using ARQ for delay-tolerant services. For example, it is shown in [51] that ARQ lowers the required SIR per bit by more than 3 dB for error-sensitive data requiring a BER of 10^{-9}, using a BCH code with a rate 6/7 and 4 independent diversity paths. Under the assumption of the availability of accurate path gain estimates, allocated resources satisfy QoS requirements and a transmission does not take place unless the performance is guaranteed. Therefore, ARQ is not applicable in this case. In practice, however, a perfect path gain estimate is not available and the number of erroneous received data bits may be higher than that could be corrected by FEC. In this case, we assume that error-sensitive data of classes II and III can be retransmitted.

Delay Bound

Each mobile data terminal of class II and class III services is assumed to have a sufficiently large data buffer for queueing purposes. These storage elements are necessary to buffer outputs of the traffic sources during their off-periods or when the allocated rates[1] are less than the source rates. Off-periods happen during situations such as degraded channel conditions, handoff blocking, and local or network congestion, and are controlled by the resource management algorithm. These services are queued in favor of allocating more resources to the higher-priority services.

[1]In this thesis, the allocated rate refers to the symbol rate in physical layer. Given a coding and modulation scheme, and the amount of overhead data, there is a one-to-one mapping between symbol rate and data rate.

We assume that the queueing delay at a mobile terminal is the only type of delay in the network and is controlled to not exceed a target delay bound for each user. There are, however, other types of delay such as processing, propagation, transmission and retransmission delays that are assumed to be relatively low and negligible. Blocking delay is not considered withing the scope of this work. Delay bound is the maximum tolerable delay for a particular service and varies with the type and class of that service. We define the *residual delay bound* $\tau_r(n)$ at time n as the difference between the delay bound (τ) and the total accumulated queueing delay and service time. For example, if user i starts data transmission at time n_0 with a delay bound τ_i, at time n the residual delay bound will be $\tau_{r,i}(n) = \tau_i - (n - n_0)$. We map $\tau_{r,i}(n)$ into a minimum-rate limit, $R_{i,\min}$, which guarantees the delay performance for each user i and is an important parameter of the proposed model. The minimum-rate limit is evaluated for different applications in Section 3.5.2.

3.2 Medium Access

A synchronized reverse-link WCDMA model is considered. Synchronization is assumed at the frame level. The chip rate is fixed and the total bandwidth W is used by all users. Information is collected from each traffic source in the form of fixed-length packets that are transmitted in synchronized, fixed-length time frames or transmission cycles (TC) with a period of T_f. In each packet, the number of information symbols is determined by the allocated data rate to user i times the packet time duration, $r_i T_f$. For example, if $T_f = 10$ ms, the number of symbols per packet will be 1280 and 80 for allocated rates of 128 ksps and 8 ksps, respectively. In practice, a number of symbols are reserved for labeling, synchronization, and packet-quality-check for retransmission purposes.

Packet Configurations

The packets can be configured in different ways as follows.

Variable Symbol Duration

In this configuration, the symbol duration is adapted to the optimum allocated rate, given the packet length is constant. Thus, the existing traffic sources are multiplexed perfectly by the optimal resource management algorithm (Figure 3.1-a). The cost, however, is a higher receiver complexity due to the need for a variable-impulse-

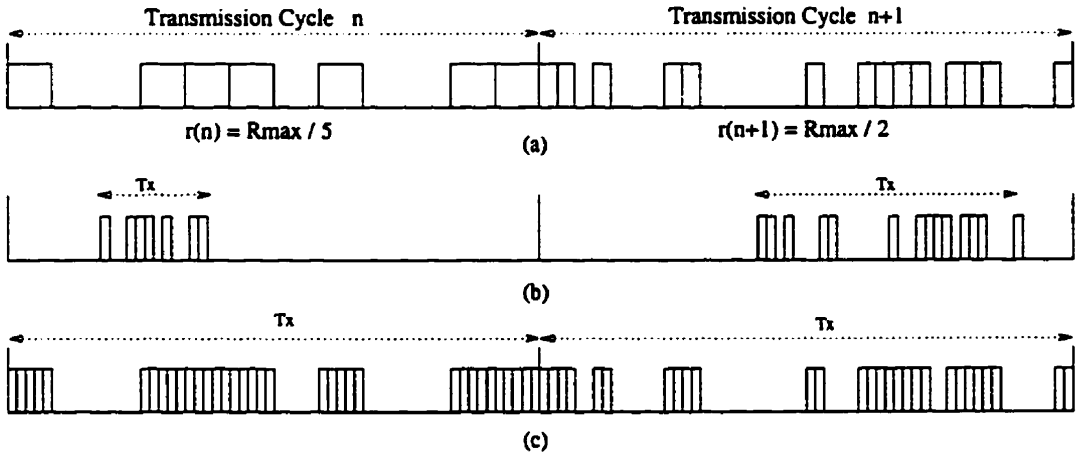


Figure 3.1: Packet configurations: (a) variable symbol duration, b) fixed symbol duration, (c) repetition code. Rectangular pulses represent 1's and zero levels represent 0's.

response matched filter for optimum detection of the data in each time-frame. Furthermore, data-rate information must be provided at the receivers through separate control channels.

Fixed symbol duration

The specified number of symbols in a packet are transmitted at the maximum rate, thus, the duty cycle of transmission varies in consecutive frames, as suggested in [52]. This scheme allows a less complex receiver but multiplexing of the data will not be perfect. The best statistical multiplexing can be achieved by randomly locating the bursts of data within the frames (time-hopping) as shown in Figure 3.1-b.

Repetition code

To preserve a perfect multiplexing while using a fixed symbol duration, repetition codes can be employed, as proposed in [53]. In this method, each symbol is transmitted at the channel peak rate and repeated such that a 100% duty cycle is acquired. For example, if the allocated rate is $R_{max}/7$, each symbol will be repeated 7 times consecutively. Thus, the advantages of the first and second methods are achieved at the same time. This method requires the data rates to be allocated in discrete values (integer multiples) which introduces integer programming complexities to the optimization problem (see Figure 3.1-c).

Our system model supports all three packet structures, however, the third configuration with 100% duty cycle, perfect multiplexing and extra coding gain is our preference.

3.3 Physical Channel

We assume that channel variations with respect to T_f are small and effectively constant during one frame. For medium speed mobiles in an urban area ($v < 50$ km/hr) and a carrier frequency of 1.8 GHz, the maximum Doppler frequency is

$f_{d,\max} = 90$ Hz. This requires a transmission cycle less than the coherence time of the channel, i.e., $T_f < 1/90$ s. In an indoor wireless network, however, real-time implementation of the algorithm is more feasible because $f_{d,\max}$ is smaller (in the range of 10-20 Hz) and T_f can be increased so that a larger time is available for the algorithm to compute and update new optimal resource values. With a bandwidth of $W = 5$ MHz and a typical delay spread of $T_m = 3$ μs in an outdoor microcellular environment [7], the channel is a frequency-selective fading channel because $W > 1/T_m$. It can be modeled by a tapped delay line with $T_m W = 15$ taps. Furthermore, since the data bit duration $T_b(\geq 1/128)$ ms, assuming a maximum data rate of 128 ksps, is much larger than T_m, the effect of intersymbol interference in this channel is negligible.

A mobile transmitted signal is impaired by this fading channel and received at the base station with a certain path gain. The path gains are stochastic in nature and have time-varying values (stochastic processes). The statistical properties of the path gains depend on the channel characteristics. We assume that path gains from the users to the base stations are given at any time. In practice, estimates of the path gains are provided by certain measurements in the network, for example, by measuring the level of a reference pilot signal from a base station in the forward link and using it as the estimate of the reverse-link gain [53]. For our analytical and simulation purposes, we will consider path loss and shadowing effects which are modeled as an attenuation of fourth-order and a log-normal distribution [54].

3.4 Control Variables

Having knowledge of each user's type and QoS requirement, and path gains, the resource management algorithm controls the transmitted powers, $p(n) = [p_1(n), \cdots, p_N(n)]$,

and transmission rates, $r(n) = [r_1(n), \cdots, r_N(n)]$, over the best base station assignment, $a^l(n) = [a_1^l(n), \cdots, a_N^l(n)]$, such that each user meets its required error and delay performance, while the network throughput is maximized, and the number of handoffs is controlled at a desired level.

We define a time-dependent variable $h(n)$ as the number of handoffs. A handoff process is initiated only when the new assignment for a user i is different from the previous assignment, that is, $a_i^l(n) \neq a_i^j(n-1)$. Therefore, the number of handoffs at time n can be evaluated as

$$h(n) = \sum_{i=1}^{N} f_{h,i}(n) \tag{3.1}$$

where $f_{h,i}(n)$ is defined as

$$f_{h,i}(n) = \begin{cases} 1 & \text{if } a_i^l(n) \neq a_i^j(n-1) \\ 0 & \text{if } a_i^l(n) = a_i^j(n-1), \end{cases} \tag{3.2}$$

with l and $j \in \{1, \cdots, M^N\}$. Using a binary assignment variable b_{ik}, defined as

$$b_{ik} = \begin{cases} 1 & \text{if user } i \text{ is assigned to base station } k \\ 0 & \text{otherwise,} \end{cases} \tag{3.3}$$

the handoff variable h can be denoted as

$$h(n) = \sum_{i=1}^{N} \sum_{k=1}^{M} [b_{ik}(n) \oplus b_{ik}(n-1)]. \tag{3.4}$$

The algorithm updates these control variables at the beginning of each frame. The rate vector $r(n)$ represents the data rate of the signals after coding and modulation. If channel coding with rate r_{cod} is employed with a modulation scheme with spectral efficiency of 1 bit/s/Hz, the actual data rate for user i is $r_{cod} \times r(n)$. We assume that enough data is available at the mobile terminal queue to load a complete time frame based on the allocated rate. Otherwise, a small proportion

of the network resources will be wasted. As described in Chapter 4, this problem is addressed from a wireless service provider perspective and formulated as an optimization problem.

3.5 Control Strategy

A major task of the resource management algorithm is to satisfy QoS requirements for all users by optimal allocation of the available resources. QoS's are controlled as follows.

3.5.1 BER control

The VSG concept, as described in Section 2.5, is applicable in our model because the chip rate is constant and packet loads vary according to the allocated data rate. Thus, increasing power and/or decreasing data rate reduces the BER. The BER requirement is reflected in the target SIR per bit and given by the vector $\gamma = [\gamma_1, \cdots, \gamma_N]$. To attain the target SIR per bit, different streams of data can use various coding schemes. A convolutional code with rate 1/3 and constraint length 9 is well known for voice services requiring a BER of 10^{-3} and used in IS-95 and proposed for the next generation wireless systems. For data applications, however, more interest is shown in turbo codes with a typical rate of 1/3 [10, 12]. Our resource management algorithms use γ as the target SIR per bit. The coding schemes and receiver structure are transparent to the algorithms.

Another means to control BER in variable-rate transmission is the use of different coding rates at each frame. A lower coding rate reduces the required average SIR per bit for a specific BER but, at the same time, adds more redundancy into

the data and increases the service time which may affect the delay performance. We will control the BER by varying the power and data rate for each user. Employing variable coding rate needs further investigations and can be part of the future work.

As explained previously in this chapter, with accurate knowledge of path gains, BER performance is guaranteed with the allocated power and rate. In a practical situation, where path gains are not accurate, a received time-frame may have a BER higher than the target value. In this case, the packet will be retransmitted (ARQ) for class II and III services with very low BER requirements.

3.5.2 Delay control

Delay performance requirements are controlled by setting specific limits on the minimum rate, $R_{\min}(n) = [R_{1,\min}(n), \cdots, R_{N,\min}(n)]$, allocated to each user. This minimum-rate limit is time-varying for class I and II services and captures the delay requirement of the service. In the following, we derive a mathematical expression for the minimum-rate limit in certain applications.

Minimum-Rate Limits

One of the main features of our multimedia system model is to accommodate different delay performance requirements by appropriate control of the transmission rates. The control policy is to set a dynamic lower bound on the data rate and map the delay bound onto this lower limit. The advantage is that the network is able to span the service time of a user and exploit the unused resources more efficiently.

The delay bound varies from zero to infinity. All class I services have zero delay tolerance. In this case the minimum rate cannot be less than the source rate r_s,

otherwise, the call must be dropped. Therefore,

$$r_i = r_{s,i} = R_{i,\min} = R_{i,\max}, \qquad \text{for user } i \in \text{class I.} \tag{3.5}$$

For class III services, the delay tolerance is unlimited. That is, the network can postpone any class III services until there are enough resources available and all class I and II are taken care of. Therefore, for this category

$$R_{i,\min} = 0, \quad \text{for user } i \in \text{class III.} \tag{3.6}$$

The problem of calculating $R_{i,\min}$ rises when $0 < r_i < \infty$, i.e., for class II services. We discuss this problem for two different cases. The first case covers the situations where a user intends to transfer a specific amount of data, as in FTP and transmission of stored images or video files and can tolerate a certain amount of delay. The second case includes applications such as electronic talk, remote login, and other on-line services where the total amount of data to be transmitted is not readily known. Thus, the network should set a limit on the average delay to maintain a certain delay performance. In what follows, mathematical expressions for the minimum-rate limit are derived for the above applications.

File transfer applications - Assuming that user i wants to transmit a total amount of data D_i with a maximum tolerable delay τ_i, the data rate at time n should satisfy

$$R_{i,\min}(n) \le r_i(n) \le R_{i,\max} \tag{3.7}$$

where $R_{i,\max}$ is the maximum channel rate. Assuming the service to user i started at time 0, the minimum required time to transmit D_i amount of data is $T_i(0) = D_i/R_{i,\max}$. With a delay bound τ_i, the transmission time can be spanned from $T_i(0)$ to τ_i, where we have assumed $\tau_i \ge T_i(0)$. At time $n+1$, the residual delay bound

is given as

$$\tau_{r,i}(n+1) = \tau_i - (n+1)T_f. \tag{3.8}$$

The minimum required time to transmit the residual amount of data is equal to

$$\begin{aligned} T_i(n+1) &= \frac{D_i - \sum_{j=0}^{n} r_i(j)T_f}{R_{i,\max}} \\ &= \frac{D_i - \sum_{j=0}^{n-1} r_i(j)T_f - r_i(n)T_f}{R_{i,\max}}, \quad \text{for } n > 0. \end{aligned} \tag{3.9}$$

The necessary condition for the delay performance to be met is

$$T_i(n+1) \leq \tau_{r,i}(n+1). \tag{3.10}$$

Substituting (3.8) and (3.9) in (3.10) and solving for $r_i(n)$, we get

$$r_i(n) \geq \frac{D_i}{T_f} - \sum_{j=0}^{n-1} r_i(j) - (\frac{\tau_i}{T_f} - n - 1)R_{i,\max}, \quad \text{for } n > 0. \tag{3.11}$$

Comparing (3.7) and (3.11), the desired lower bound is derived as

$$R_{i,\min}(n) = \frac{D_i}{T_f} - \sum_{j=0}^{n-1} r_i(j) - (\frac{\tau_i}{T_f} - n - 1)R_{i,\max}, \quad \text{for } n > 0. \tag{3.12}$$

□

The delay bound for this category of applications can also be maintained by setting a constant minimum-rate limit as

$$R_{i,\min} = \frac{D_i}{\tau_i}. \tag{3.13}$$

By this approach, the service is practically restated as a Class I CBR services. If the network has sufficient available capacity, the constant minimum-rate limit will guarantee the delay bound. Otherwise, it will be more likely to have an infeasible solution and a higher dropping rate.

On-line applications - Let $r_{s,i}$, $\tau_{f,i}$, $\bar{\tau}_i$, and $\hat{\tau}_i$ be the source rate, delay incurred in a frame, average delay, and average delay bound, respectively. During the n^{th} time-slot, the lower bound on the delay in a frame is

$$\tau_{t,i}(n) \geq \frac{r_{s,i}(n) - r_i(n)}{R_{i,\max}} T_f. \tag{3.14}$$

By definition we have

$$\bar{\tau}_i(n) = \frac{1}{n} \sum_{j=1}^{n} \tau_{f,i}(j). \tag{3.15}$$

Using (3.14) and the objective that $\bar{\tau}_i(n) \leq \hat{\tau}_i$, we can conclude

$$r_i(n) \geq \sum_{j=1}^{n} r_{s,i}(j) - \sum_{j=1}^{n-1} r_i(j) - \frac{n}{T_f} R_{i,\max} \hat{\tau}_i. \tag{3.16}$$

The right hand side of (3.16) gives the minimum-rate limit for on-line applications. □

In both cases, the initial value of $R_{i,\min}$ should be set depending on the delay bound from 0 to $R_{i,\max}$. For example, if τ_i or $\hat{\tau}_i$ is very low, $R_{i,\min}(0)$ must be set at a large value, otherwise, it can be set at 0 or a small value.

The right hand sides of (3.12) and (3.16) can be a negative value which implies that the residual delay is sufficiently large and the service can be delayed at time n. To avoid a negative rate, we set the minimum-rate limit at $\max[0, R_{i,\min}(n)]$ in our mathematical programming problem.

Hence, the required lower bound on the data rate can be updated regularly in terms of the known parameters D_i, τ_i, $\hat{\tau}_i$, T_f, and $R_{i,\max}$. It is also required to store the sum of previous allocated rates for this purpose which is quite practical.

3.6 Physical Constraints

3.6.1 Maximum Rate

The transmission rates are subject to an upper limit, $R_{\max} = [R_{1,\max}, \cdots, R_{N,\max}]$, to preserve an acceptable processing gain. In our model, a typical value of the radio channel bandwidth is 5 MHz and typical values of the maximum rate are 128 and 256 ksps per PN code. The latter rate is the equivalent physical rate for the LCD service in [10] and can be used for comparison purposes.

High data-rate requirements are supported by using additional PN codes, as suggested by the multi-code technique [55, 56], and accomplished by employing parallel channels (codes). Thus, total capacity of the network can be allocated to one user if there are no other service requests and the user has sufficient data to transmit. These parallel channels undergo exactly the same physical paths and experience the same channel impairments. As a result, the received signals can be detected by a uniform receiver structure, e.g., a RAKE receiver with exactly the same number of taps and equal coefficients. The synchronization and sequence of data in parallel channels are not distorted because the network resource controller allocates equal resources to each channel due to the same channel characteristics.

3.6.2 Maximum Power

In a real system, mobile transmitted power is subject to a certain maximum level due to its physical characteristics. For example, a vehicular mobile terminal is able to boost more power than a mobile handset with strictly limited size and weight. To simplify the analysis and avoid the complexity of integer programming, we assume that the mobile transmitter output can vary continuously and provide any power

level prescribed by the power control algorithm within a specified limit. In practice, however, the mobile transmitter power output is set to a number of discrete levels. In the next chapter, we will study the effect of this assumption on the resource allocation performance. The minimum power is set to zero for all users. This assumption is necessary for non-real-time services because they might be delayed at a particular time without any transmission. This requires that their powers and rates to be set to zero at certain times. But for a real-time service, zero power is equivalent to dropping the call. In Chapter 4, we shall see that the rate constraints do not allow a zero power for a real-time user.

3.7 Wireless System Throughput and Capacity

As pointed out in Chapter 2.5 on page 28, the sum $\sum_{i=1}^{N} r_i$ does not properly represent the network throughput and spectral efficiency in multimedia applications. For further elaboration, consider two bit rates r_1 and r_2 with E_b/I_0 requirements γ_1 and γ_2, respectively. Assuming $r_1 < r_2$ and $\gamma_1 \gg \gamma_2$, it is obvious that more network resources are required to accommodate r_1 than r_2 while r_1 has less contribution in the sum $\sum_{i=1}^{N} r_i$. A quantity such as the transmission rate weighted by the SIR per bit requirement is capable of removing such an ambiguity in the sum $\sum_{i=1}^{N} r_i$ and representing the throughput more precisely. Thus, $\gamma_1 r_1$ and $\gamma_2 r_2$ are measures of both data rates and transmission qualities. In comparison with the conventional measure $\sum_{i=1}^{N} r_i$ for single service applications, the sum $\sum_{i=1}^{N} \gamma_i r_i$ quantitatively includes a scaling factor and is not compatible with the conventional sum. A suitable measure of the throughput for multimedia must be equivalent to $\sum_{i=1}^{N} r_i$ when applied to single-medium. One way to account for the scaling factor in the product is dividing the product-sum by the average SIR per bit of the users

in the network. Thus, the network throughput will be equal to $\sum_{i=1}^{N} r_i$ if SIR per bit requirements of all users are the same.

Definition 1 The *network throughput* R, in symbol per second, is defined as the sum of the transmission rates weighted by the E_b/I_0 requirements normalized over the average target E_b/I_0's as

$$R = \frac{\sum_{i=1}^{N} \gamma_i r_i}{\frac{1}{N}\sum_{i=1}^{N} \gamma_i} \tag{3.17}$$

$$= \frac{1}{\bar{\gamma}} \sum_{i=1}^{N} \gamma_i r_i \tag{3.18}$$

where $\bar{\gamma} = \frac{1}{N}\sum_{i=1}^{N} \gamma_i$. The *cell throughput* R_c is defined as R/M and the *normalized throughput* refers to the throughput per unit of bandwidth per cell ($R/W/M$ or R_c/W). Consequently, the *network capacity* C, *cell capacity* C_c, and *normalized capacity* C_n refer to the maximum achievable network throughput, cell throughput, and normalized throughput, respectively.

3.8 Summary

In this chapter, the physical and mathematical description of our model is laid out in as much detail as required within the scope of our research. The resource management algorithms are developed based on this model. Further elaborations on the system model will be provided as necessary.

Chapter 4

Optimal Resource Management

In this chapter, a mathematical programming model based on our research objective and the proposed system model is developed. A service-provider perspective is selected to formulate the objective function. It relates the network revenue and cost to the system resources and control variables. The QoS requirements and physical limitations constitute the set of constraints. The structure of the problem is analyzed and, to some extent, simplified. This analysis is followed by a discussion of the type of variables as continuous or integer. This is essential in the mathematical programming problem and its computational complexity.

4.1 Mathematical Programming Model

A wireless multimedia network has limited resources which must be shared among a large number of users with different sorts of demands. This resource sharing is to be managed efficiently so that the network resources are utilized maximally and every user's need is satisfied. This goal may be translated into mathemati-

cal programming problems with different formulations. For example, an objective function derived for a non-profitable wireless computing network in a campus may be different from that of a commercial public wireless service provider, although both providers seek maximum efficiency. Thus, the definition of an objective function is commonly considered a subjective issue and depends on the application and the nature of the problem. A brief background on optimization is provided in Appendix A.

4.1.1 Objective Function

We are interested in a network which provides multimedia services such as voice, data, low-rate video, internet connections, and on-line services to the public. Our research is an attempt to formulate the problem of resource management in a CDMA-based system from a public-service-provider perspective. It only includes the wireless link and does not capture the related issues in an end-to-end basis. In this perspective, the overall wireless network efficiency should somehow be related to the profit made. Therefore, we identify the main sources of revenue and different costs involved in service provisioning, and express our objective in the form of maximizing the profit, that is, the difference between the revenue and costs.

Revenue - Without going into details of commercial issues, we consider that the network revenue is directly proportional to the total data rate weighted by the QoS's provided in the network. In practice, however, there are different pricing policies under different plans which are negotiated at the time of subscription of each mobile user. These plans usually consider a combination of the air-time, fixed monthly fees, and other parameters which are beyond the scope of this research. Our pricing scheme is based on the following propositions:

1. The network revenue from each user is directly proportional to its transmission rate.

2. Lower BER services utilize more resources (power) and are proportionally more expensive.

3. Services with higher delay tolerances are less expensive because of their flexibility in time which helps the resource management center (RMC) to span their service times and exploit the unused resources more efficiently. We realize that the tolerable delay for each service may vary with time, therefore, the revenue is a function of time in general.

The above propositions imply that the cost for a user to transmit a bit of data over the wireless network is a function of the required BER quality, transmission speed, and its delay-tolerance level. Selection of a function to relate these factors is a subjective issue and can be decided by different service providers differently. This issue has not been addressed in the literature so far. In what follows, we derive a typical pricing function to formulate the resource management problem. The second proposition should be captured by a monotonically-increasing function of the required SIR per bit or a monotonically-decreasing function of the target BER. Such a function in linear form can be equal to γ_i, thus, transmission of a bit with SIR requirement of 10 will cost twice that of one requiring an SIR of 5. The function to capture delay tolerances in the range of 0 to ∞ should be a monotonically-decreasing function of the residual delay bound and cover the whole range of non-negative numbers. An exponential function satisfies these requirements. It also has the flexibility to be shaped as desired by the service provider. This function can be expressed as $[A + B\exp(-\tau_{r,i}(n)/D)]$ and varies from $A + B$ for zero delay tolerant services to A for unlimited delay tolerant services. A, B and D are positive real

constants. A is used to prevent the function from being 0, and B and D can be used to shape the function as desired. The value of these parameters are set equally to 1 in our simulations in Chapter 5. Thus, a real-time service is more expensive than a delay-insensitive service by a factor of $(A + B)/A$ for the same BER. The rate of change in price is controlled by D. For small values of D, the price of delay-insensitive services drops faster and vice versa. The above functions can be combined in many ways. We incorporate these functions in the pricing function in the form of coefficients λ_i for the allocated rates. These coefficients represent the price of a transmitted bit as a function of its QoS. Thus,

$$\text{Total Revenue } = \sum_{i=1}^{N} \lambda_i(n) r_i(n) \tag{4.1}$$

where

$$\lambda_i(n) = \left[A + B \exp\left(\frac{-\tau_{r,i}(n)}{D} \right) \right] \gamma_i. \tag{4.2}$$

Costs - Many expenses are incurred in a wireless network operation. Within the scope of this work, only resource-related costs which are affected by and expressed as a function of the existing variables in our problem are taken into account. A major cost for the network operation is the number of handoffs. Each handoff is associated with additional signalings for establishing a new connection and transferring the user's information (such as the new base station assignment) to the network data base. We map the network signaling overhead to a certain cost and define λ_h as the cost per handoff, therefore, the handoff cost at time n is $\lambda_h h(n)$, where $h(n)$ is defined in (3.1) and (3.4). λ_h is a constant and is assumed to be the same for all services and equal to 1 in our simulations in the next chapter. It is adjustable and affected by many factors in the network. We assume λ_h is pre-determined and known.

Profit - The overall profit is

$$\text{Total profit } = \sum_{i=1}^{N} \lambda_i(n) r_i(n) - \lambda_h h(n) \tag{4.3}$$

and our objective is to maximize this profit subject to satisfying the service requirements for all users.

4.1.2 Constraints

The set of constraints includes the lower and upper bounds for the power and rate vectors, upper bounds for the number of handoffs, and the SIR requirements. Among these bounds, only $R_{\min}(n)$ is time-variant and must be updated per time frame.

As described in Chapter 2, E_b/I_0 in a CDMA system must satisfy the condition

$$\left(\frac{E_b}{I_0}(n)\right)_i = \frac{W}{r_i(n)} \frac{g^l_{ia_i}(n) p_i(n)}{\sum_{j=1, j\neq i}^{N} g^l_{ja_i}(n) p_j(n) + \eta} \geq \gamma_i, \quad i \in \{1, \cdots, N\}. \tag{4.4}$$

Similar to [48], we are interested in avoiding over-spending network resources beyond the required level for each mobile; thus, the equality in (4.4) would provide sufficient conditions to achieve the target BER. Rearranging (4.4) around the equality sign and omitting the time indices to simplify the notations, we get

$$\frac{W}{\gamma_i} g^l_{ia_i} p_i - \sum_{j=1, j\neq i}^{N} g^l_{ja_i} p_j r_i - \eta r_i = 0, \quad i \in \{1, \cdots, N\}. \tag{4.5}$$

Equation (4.5) forms a system of N non-linear quadratic equations with $2N$ variables. In the following, we re-arrange this equation in the algebraic quadratic form. Let

$$w_i = W/\gamma_i \tag{4.6}$$

$$q^l_i = w_i g^l_{ia_i} p_i - \sum_{j=1, j\neq i}^{N} g^l_{ja_i} p_j r_i - \eta r_i \tag{4.7}$$

$$= q^l_{i,ln} + q^l_{i,qu} \tag{4.8}$$

where

$$q^l_{i,ln} = \frac{W}{\gamma_i} g^l_{ia_i} p_i - \eta r_i \tag{4.9}$$

$$q^l_{i,qu} = -\sum_{j=1, j\neq i}^{N} g^l_{ja_i} p_j r_i \tag{4.10}$$

are linear and quadratic terms of q^l_i, respectively. Thus, q^l_i can be written in an algebraic form as

$$q^l_i = (c^l_i)' x + \frac{1}{2} x' C^l_i x, \tag{4.11}$$

where

$$x = [p_1, \cdots, p_N, r_1, \cdots, r_N]' \tag{4.12}$$

$$c^l_i = \nabla q^l_{i,ln} \tag{4.13}$$

$$= [0, \cdots, \underbrace{w_i g^l_{ia_i}}_{\text{element } i}, \cdots, \underbrace{\eta}_{\text{element } N+i}, \cdots, 0]' \tag{4.14}$$

$$C^l_i = \nabla^2 q^l_{i,qu} \tag{4.15}$$

$$= \begin{bmatrix} \emptyset & \hat{C}^l_i \\ (\hat{C}^l_i)' & \emptyset \end{bmatrix}, \tag{4.16}$$

and $'$ is the transpose sign. In (4.16), the partitions $\emptyset$ and $\hat{C}^l_i$ are $N \times N$ matrices. The $\emptyset$ partition is an all-zero matrix. In $\hat{C}^l_i$, the only nonzero column is column i with elements

$$\text{Column } i \text{ of } \hat{C}^l_i = [-g^l_{1a_i}, \ldots, -g^l_{(i-1)a_i}, 0, -g^l_{(i+1)a_i}, \ldots, -g^l_{Na_i}]' \tag{4.17}$$

We notice that c_i and C_i are functions of the path gains and are specified by the type of assignments and propagation conditions.

Now we can summarize the defined optimization problem as in Figure 4.1 where h and λ_i are defined in (3.1,3.4) and (4.2), respectively. The objective function

$$\underset{a^l}{\text{Maximize}} \quad \left\{ \max_{p,r,h} \sum_{i=1}^{N} \lambda_i r_i - \lambda_h h \right\}$$

subject to

$$\begin{aligned} 0 &\leq p_i \leq P_{i,\max} \\ R_{i,\min} &\leq r_i \leq R_{i,\max} \\ h &\leq h_{\max} \\ [c_i^l]' x &+ \frac{1}{2} x' C_i^l x = 0 \end{aligned}$$

Figure 4.1: The original mathematical programming problem

maximizes the total profit at time n in terms of the variables p, r, and h over all assignment vectors $a^l \in S$ where S is the set of all feasible assignments. In other words, for every assignment vector a nonlinear subproblem should be solved. It should be noted that in an exhaustive solution process, h is constant because at any time the current and previous assignment vectors are known. Having this knowledge, h is known and does not enter into the optimization process. The cardinality of S satisfies $|S| \leq M^N$ and can be astronomically large. In other words, the problem has a max-max structure consisting of a nonlinear programming (NLP) subproblem and an assignment problem. Therefore, our problem should be categorized as an *NLP large scale optimization problem.*

The NLP subproblem has a linear objective function and a set of quadratic constraints and can be categorized as a quadratic problem. Theorems 2 and 3 in Appendix A.1 state that if the quadratic form is positive semidefinite, it is convex

and converges to a global optimum. That is, if C_i^l in Figure 4.1 is positive semidefinite, a global optimum is guaranteed and the problem can be solved efficiently, that is, with a fast convergence to the global optimum. The following corollary, however, proves that C_i^l is indefinite.

Corollary 1 The quadratic constraint

$$[c_i^l]'x + \frac{1}{2}x'C_i^l x = 0 \tag{4.18}$$

is indefinite.

Proof: Using Definition 3 in Appendix A.1, it is enough to show that C_i^l is indefinite. For any non-zero vector $y \in \Re^{2N}$ we have

$$y'C_i^l y = -2y_{N+i} \sum_{j=1, j\neq i}^{N} g_{ja_i}^l y_j, \tag{4.19}$$

where y_{N+i} and y_j are the $(N+i)^{th}$ and j^{th} element of vector y, respectively. The value of (4.19) can be either positive or negative. For example, it is positive if y_{N+i} and y_j have different signs and negative if both have equal signs for all j's. Therefore, $y'C_i^l y$ is neither positive semidefinite nor negative semidefinite, but indefinite. □

An indefinite quadratic problem is non-convex, and global optimization of a linear objective function subject to indefinite quadratic constraints is an NP-complete problem [58]. Problems with this type of structure are difficult to solve both for global and local optima [77, 58]. Using the concept of equivalent mathematical programming, as defined in Appendix A.1 Definition 5, we attempt to find equivalent problems with certain preferences in their structures. These preferences are: (1)

linearity, (2) convexity or quasi-convexity (see Definition 3), (3) convexity in feasible space, (4) diagonality and/or sparsity of the Hessian matrix (see Definition 2) of the nonlinear functions, respectively [28, 77, 58]. The first two types of structures guarantee that a local optimum is the global optimum and are the most desirable structures for our problem. The third type is preferred due to the possibility of checking for the existence of a feasible solution, and availability of many algorithms which efficiently find at least a local optimum over a convex feasible search space. One important feature of the third type is that we can readily check for the existence of a feasible solution (not optimal), e.g. by running *the first phase of the simplex method* [1]. This property is very useful especially when dealing with a large scale problem such as ours. The fourth type of structure can be solved with less computational complexity by many algorithms that use the second-order partial derivatives to find the optimum solution.

As an alternative mathematical programming problem, similar to [48], we substitute the value of r_i from (4.5) to the objective function and rate constraint in Figure 4.1 which results in another mathematical model as shown in Figure 4.2, where $i \in \{1, \cdots, N\}$. In the mathematical programming of Figure 4.2, the number of variables is reduced and the constraints are linearized. On the other hand, the objective function alters from a linear form to a complex nonlinear fractional structure. In the next chapter, we will take necessary steps to reach the most efficient solution for the problem in Figure 4.2 for single- and multi-cell systems.

[1] The simplex method is a practical way to solve linear optimization problems. The first phase of this method checks for the feasibility condition [29] and is able to find out whether the feasible set is an empty set (see Appendix A.2).

$$\underset{a^l}{\text{Maximize}} \quad \left\{ \max_{p,h} \sum_{i=1}^{N} \lambda_i \frac{w_i g^l_{ia_i} p_i}{\sum_{j=1, j\neq i}^{N} g^l_{ja_i} p_j + \eta} - \lambda_h h \right\}$$

subject to

$$\begin{aligned} 0 \leq p_i &\leq P_{i,\max} \\ R_{i,\min} \leq \frac{w_i g^l_{ia_i} p_i}{\sum_{j=1, j\neq i}^{N} g^l_{ja_i} p_j + \eta} &\leq R_{i,\max} \\ h &\leq h_{\max} \end{aligned}$$

Figure 4.2: An equivalent mathematical programming with linear feasible space

4.2 Model Variables: Integer or Continuous?

In order to solve the developed mathematical programming problem in Figure 4.2, a fundamental question to be addressed is whether the model variables are integral or continuous. Technologically and practically, the allocated power and data rate have a discrete nature and, in principle, they should be regarded as integer variables and solved accordingly. But it is well known that integer problems are inherently much harder to solve than the corresponding continuous problems. Mathematically, integer models involve many times more calculation in their solution. While, for example, a linear problem with thousands of constraints and variables can almost certainly be solved in a reasonable amount of time, a similar situation does not hold for an integer linear problem. A great deal of effort is exerted to avoid integer programming in model building [59]. An immediate suggestion to solve an integer or mixed-integer problem, therefore, is to solve the relaxed problem and round off

any non-integer variable to the nearest integer. Obviously, this approach may result in other problems such as infeasibility of the rounded values and being quite far from the optimal integer solution. In this section, we will examine the effect of relaxing power and rate variables in our problem. If we are able to show or develop some condition so that the rounded solution is

1. feasible, and
2. close enough to the optimum,

then, we can deal with the power and rate variables as being continuous and, thus, reduce the computational complexity in a large scale. We conduct this investigation for power and rate variables independently using available data from the solution for a single-cell, as derived in the next chapter.

4.2.1 Data Rate Variables

For the feasibility of the rounded solution, the following corollary defines the direction of rounding to guarantee the feasibility condition.

Corollary 2 Any allocated rate r is feasible if

$$r_{\min} \leq r \leq r_{opt}, \tag{4.20}$$

where r_{opt} is the optimal rate.

Proof: The SIR constraint in Equation (4.4) implies that

$$r_i \leq \frac{w_i g^l_{ia_i} p_i}{\sum_{j=1, j\neq i}^{N} g^l_{ja_i} p_j + \eta}. \tag{4.21}$$

r_{opt} is the closest value of r to the hard limit $r_{\max}$ that satisfies the SIR constraint. Any $r \leq r_{opt}$ that does not violate the SIR constraint, that is $r \geq r_{\min}$, is feasible. □

Therefore, taking the floor of any calculated continuous r gives a set of feasible discrete rates that can be allocated to users. The floor can be taken with respect to a set of defined rate levels.

The next step is to find out how the rounding procedure affects the network throughput and how many rate levels are needed to not lose too much throughput. For this purpose, we use the results of a large number of runs for a single cell with 40 users requiring a target E_b/I_0 of 3.3 dB, under the condition as specified in Table 5.2 on page 91. The throughput versus the number of rate levels is depicted in Figure 4.3. This figure shows that the throughput loss decreases as the number of rate levels increases. It can be seen that with more than 2^8 levels, the loss is fairly close to zero and for 2^7 levels, it is less than 3 percent. Thus, a larger number of rate levels favors higher spectral utilization as well offering high flexibility.

4.2.2 Power Variables

Unlike the rate variables, there is no guarantee that the rounded power levels satisfy the feasibility condition. On the one hand, any increase in the power level of a user increases the interference level in the network, thus degrading the SIR per bit for some other users and increasing the probability of not maintaining the target SIR per bit (infeasible solution). On the other hand, decreasing power for a user, while mitigating the communication quality for other users, degrades its own communication condition. Therefore, statistical methods should be employed to express the effect of the rounding off process. One way is to round off continuous

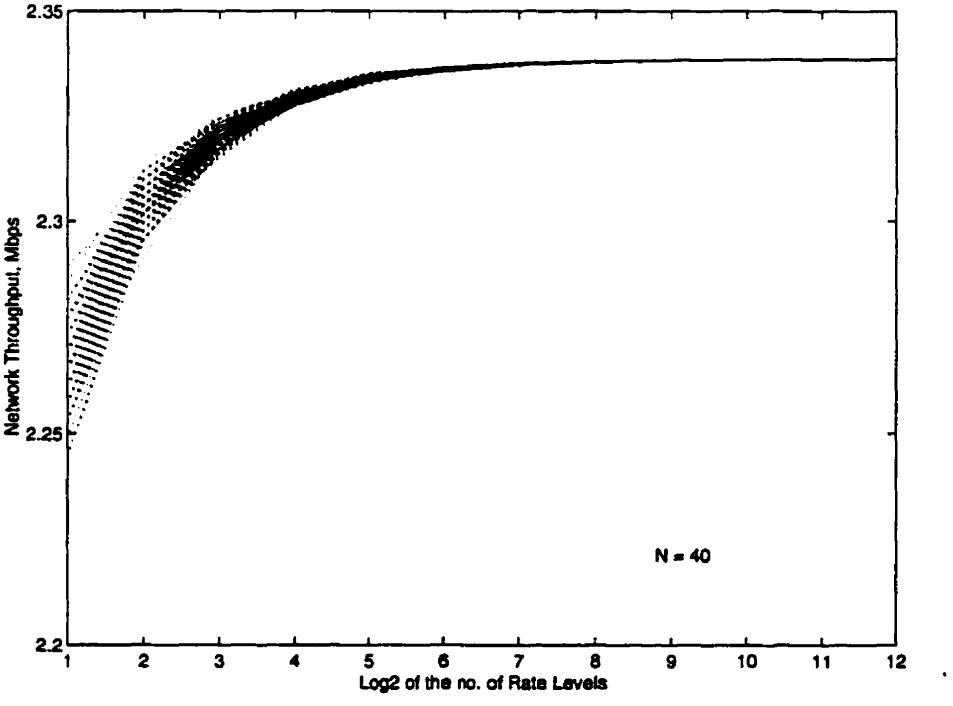


Figure 4.3: Throughput versus discretized rate.

solutions in a large number of runs with respect to different sets of discrete power levels and find the percentile of infeasible cases. Such results are derived using the same data as in the case of data rates and sketched in Figure 4.4. According to this figure, if the number of power levels exceeds 2^6, the corresponding discretized power levels will be valid with a probability close to one. Regarding the required number of power levels in CDMA mobile terminals, the current American standard IS-95 prescribes 128 levels, which lies within the desirable range.

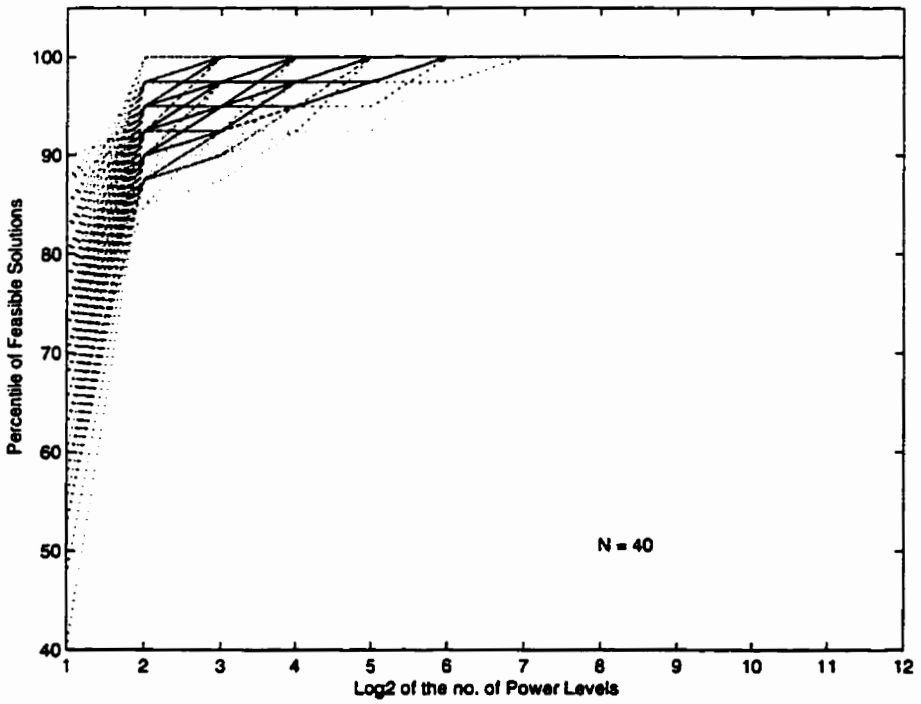


Figure 4.4: Feasibility versus discretized power.

4.3 Summary

In this chapter, a mathematical programming problem is developed to maximize the network profit while satisfying the service qualities and physical constraints. A proposed pricing scheme is used to map the weighted throughput onto the network revenue. We also observed that if the number of rate levels is more than 128, at a cost of less than 3 percent of the throughput, we can treat rate variables as continuous and free our mathematical programming problem from integer programming complexities. Similarly, the same approach can be taken with respect to power variables if the number of discretized power levels is more than 64. Different approaches to solving the mathematical programming problem will be discussed and examined in the next chapter.

Chapter 5

Resource Allocation Algorithms

In this chapter, our attempt is to efficiently solve the mathematical programming problem, as derived in the previous chapter and presented in Figure 4.2, in single- and multi-cell systems. A desirable solution must be efficient in the sense that it is globally optimum or very close to it and, needs low computational complexity. We have tried different approaches to achieve such a solution. In the following, a description of the solution process is presented in detail.

5.1 Single-Cell Solution

Optimal resource management for a single-cell system, supporting only Class I services, is addressed in [48] as reported in Chapter 2. Their formulation of the problem is a special case of our mathematical model described in Figure 4.2. This special case is solved in [48] employing the gradient projection method for nonlinear problems. It is reported that the algorithm converges to local minima in certain cases. These local minima imply that the problem is non-convex. No specific solu-

tion is proposed in the literature to overcome this problem except trying different initial values.

5.1.1 Mathematical Model for a Single Cell

In the following lemma and theorem, we prove the existence of an efficient and convex solution for the mathematical programming problem in Figure 4.2 for a single cell ($M = 1$) if the number of users is large enough ($N \gg 1$). In other words, our optimization problem can be translated into a linear programming problem for a single-cell environment.

Lemma 1 For $N \gg 1$ and $M = 1$, there exists an equivalent linear program (LP) for the optimization problem in Figure 4.2.

Proof: When $M = 1$, all users are assigned to a single cell and the large set of M^N assignments reduces to one. Thus, there is only one path gain for user i which can be expressed as g_i. In addition, the value of h is zero because no handoff takes place. Thus, the mathematical program in Figure 4.2 reduces to its single-cell version as in Figure 5.1. For a large N, the approximation

$$\sum_{j=1, j\neq i}^{N} g_j p_j \approx \sum_{j=1}^{N} g_j p_j \tag{5.1}$$

is valid. By this approximation, the summation in the objective function becomes

$$\sum_{i=1}^{N} \frac{\lambda_i w_i g_i p_i}{\sum_{j=1, j\neq i}^{N} g_j p_j + \eta} \approx \frac{\sum_{i=1}^{N} \lambda_i w_i g_i p_i}{\sum_{j=1}^{N} g_j p_j + \eta} \tag{5.2}$$

$$= \frac{m'p}{g'p + \eta} \tag{5.3}$$

$$\underset{p}{\text{Maximize}} \quad \left\{ \sum_{i=1}^{N} \frac{\lambda_i w_i g_i p_i}{\sum_{j=1, j\neq i}^{N} g_j p_j + \eta} \right\}$$

subject to

$$0 \leq p_i \leq P_{i,\max}$$

$$R_{i,\min} \leq \frac{w_i g_i p_i}{\sum_{j=1, j\neq i}^{N} g_j p_j + \eta} \leq R_{i,\max}$$

Figure 5.1: Mathematical programming for single-cell systems

where

$$m = [\lambda_1 w_1 g_1, \cdots, \lambda_N w_N g_N]' \tag{5.4}$$

$$g = [g_1, \cdots, g_N]' \tag{5.5}$$

Thus, the problem in Figure 5.1 can be written in a new structure as shown in Figure 5.2. According to Definition 6 in Appendix A.1, the new structure is a linear fractional programming (LFP) problem.

Theorem 4 in Appendix A proves that for any LFP problem, there exists an equivalent LP problem. Using this theorem, the LFP problem in Figure 5.2 is equivalent to the LP problem in Figure 5.3, where u and y are defined as

$$u = \frac{1}{g'p + \eta} \tag{5.6}$$

$$y = up. \tag{5.7}$$

Therefore, resource management in a single cell is modeled as an LP problem which can be solved efficiently by LP methods. □

$$\underset{p}{\text{Maximize}} \quad \left\{ \frac{m'p}{g'p + \eta} \right\}$$

subject to

$$\begin{aligned} 0 \;\leq\; p_i \;&\leq\; P_{i,\max} \\ R_{i,\min} \leq \frac{1}{\lambda_i} \frac{m_i p_i}{g'p + \eta} \;&\leq\; R_{i,\max} \end{aligned}$$

Figure 5.2: Linear fractional programming for single-cell systems

$$\underset{y,u}{\text{Maximize}} \quad \{m'y\}$$

subject to

$$\begin{aligned} 0 \;\leq\; y_i \;&\leq\; P_{i,\max}\, u \\ \frac{R_{i,\min}}{w_i g_i} \leq y_i \;&\leq\; \frac{R_{i,\max}}{w_i g_i} \\ g'y + \eta u \;&=\; 1 \end{aligned}$$

Figure 5.3: Linear programming model for single-cell systems

5.1.2 Single-Cell Capacity

The following corollary gives a mathematical expression for the cell capacity.

Corollary 3 The capacity of a single cell is $W/\bar{\gamma}$.

Proof: The throughput of a single cell, as defined in Definition 1 on page 46, is

$$R_c = \frac{1}{\bar{\gamma}} \sum_{i=1}^{N} \gamma_i r_i \leq \frac{1}{\bar{\gamma}} \sum_{i=1}^{N} \gamma_i \frac{W}{\gamma_i} \frac{g_i p_i}{\sum_{j=1, j \neq i} g_j p_j + \eta} \tag{5.8}$$

$$\leq \frac{1}{\bar{\gamma}} \frac{\sum_{i=1}^{N} W g_i p_i}{\sum_{j=1} g_j p_j + \eta} \approx \frac{W}{\bar{\gamma}}. \tag{5.9}$$

In writing (5.8), the SIR constraint (4.4) on page 51 has been used. Equation (5.9) is derived using the approximation (5.2) on page 62 for a large N and the assumption that η is negligible with respect to the interference term. Therefore, the cell capacity is

$$C = C_c \approx \frac{W}{\bar{\gamma}}. \tag{5.10}$$

□

Thus, in a populated and interference-limited cell the capacity is independent of the system parameters such as $P_{\max}$, $R_{\max}$, N and g. It is affected by the bandwidth and error performance requirement. In a single service case, where γ_i is unique for all users, the above capacity is equivalent to what has been given in [6] in terms of the number of users in the cell.

5.1.3 Numerical Results

The optimal resource management algorithm for a single-cell system has been implemented as a linear program in MATLAB. A preliminary assumption is that call

admission control independently takes care of the number of users in the system. It is required that at any time the inequality

$$\sum_{i=1}^{N} \gamma_i R_{i,\min} \leq C_c \tag{5.11}$$

be valid. The call admission control should keep controlling this requirement when admitting new users. When there is no feasible solution, part of the network resources should be released to serve current users. One way is to remove a user and rerun the algorithm to seek for a feasible solution. A priority management can be applied for this purpose. The priority level of each user can be negotiated at call initiation. Another way is to remove the user who is using a relatively high amount of resources in the network, e.g., a user who is in a deep fading condition. The removal process can be delayed for a short time if the call can tolerate small interruptions equal to a few frame intervals.

The algorithm works as shown in Figure 5.4. The numerical results for two different scenarios including Class III services are presented in this section.

First Simulation Scenario

In the first scenario, 50 users are admitted into the cell and no new user will be added during their service times. The users are randomly located in a 4-km-wide square cell with uniform distribution. The base station is located at the center of the cell. All users are assumed to be in the standstill state during their communications, thus, the time frame can be set at a relatively larger period such as $T_f = 1$ sec. The system parameters are set at: $W = 5$ MHz, $\gamma_i = 10$ dB, $R_{i,\max} = 128$ kbps, and $P_{i,\max} = 1$. Transmissions are assumed to follow a fourth-order log-linear propagation law (no shadowing). Each user has a stored data or image file to transmit. The size of these files, after encoding and using a modulation

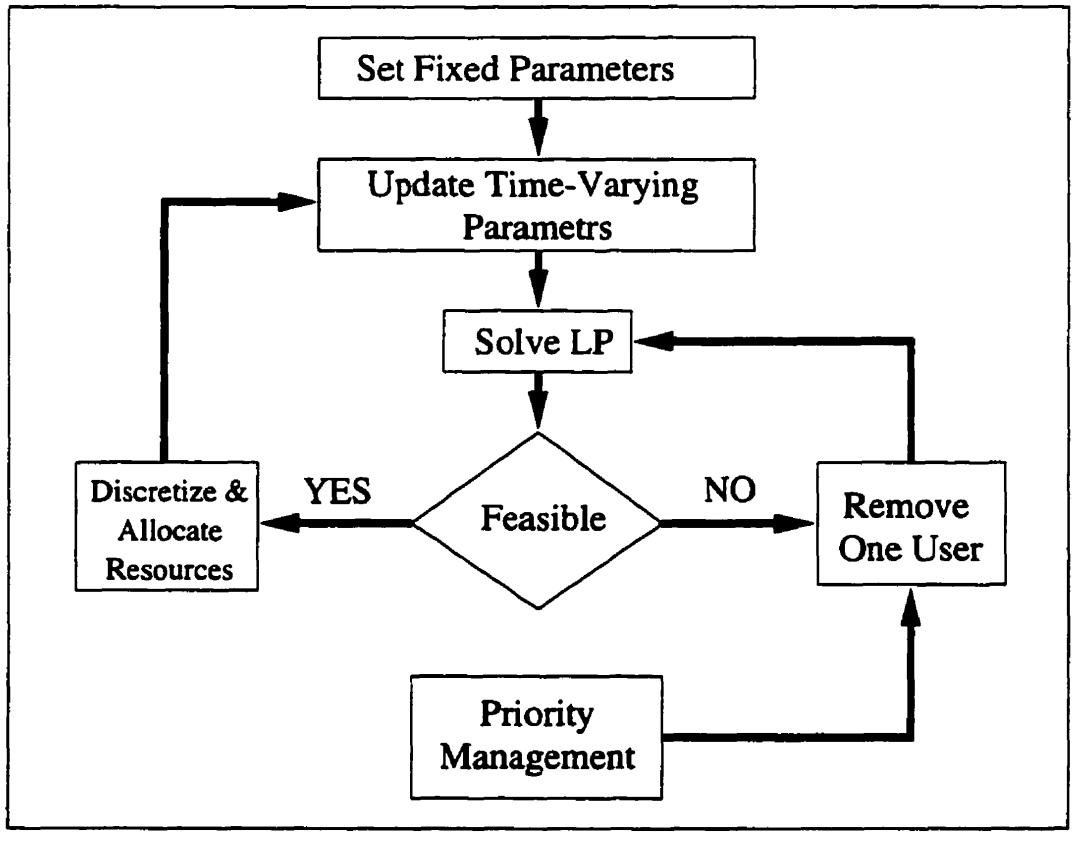


Figure 5.4: Single-cell solution flow chart.

with the spectral efficiency of 1 bit/sec/Hz, are uniformly distributed with mean 500 kbits and standard deviation 96 kbits. The transmission starts at time $n = 0$ and continues until every user in the network has sent its file. Figure 5.5 illustrates the allocated resources (power and rate), the residual amount of data, and the interference seen at each frame by three selected users with path gains $g_2 = .11$, $g_{13} = .03$, and $g_{29} = 90.3$. Users number 13 and 29 have the lowest and highest path gains, respectively. It is found that the required time to complete the transmission for all users is 57 frames. User 29 is serviced in a short time (3 frames) due to a high allocated rate. User 13 is the last user being serviced in the network. It can

also be observed that:

- if a mobile is able to transmit at the maximum rate, it is allocated the minimum power that satisfies the target BER;
- if a mobile is not able to transmit at the maximum rate, it is allocated maximum power to achieve the highest rate that satisfies the target BER;
- all users have a share in the network resources and communicate reliably although their throughput may be very small. For example, in Figure 5.5-a, the allocated rates to users #2 and #13 at the time of starting their services (n=8 and 15) are in the range of a few kbps.

To illustrate the behavior of the algorithm when Class I and II services exist, we change the service status of users 2 and 3 to Class II and I, respectively. Under the very same conditions, we assume that user 2 has a delay bound of 20 sec. That is, user 2 must be serviced within 20 time frames rather than 46 as in the previous case. Figures 5.6-a and -b depict the allocated rates and residual data for this user in the new condition. It can be seen that by the end of the 20^{th} time frame, the data is transferred successfully. The effect of this change on user 13 is a slight delay equal to one frame. User 29 has not been affected. In the other test, user 13 is assumed to be served as a Class I CBR service with a rate of 64 kbps. Figures 5.6-c and -d show the same data in this case. While satisfactory service has been offered to user 13, i.e., a constant 64 kbps rate during the service time, user 29 is not affected but the service time for user 2 has increased slightly. In both cases, when additional resources are allocated to specific users, other users in the network will experience more service delay due to the decrease in the available system resources.

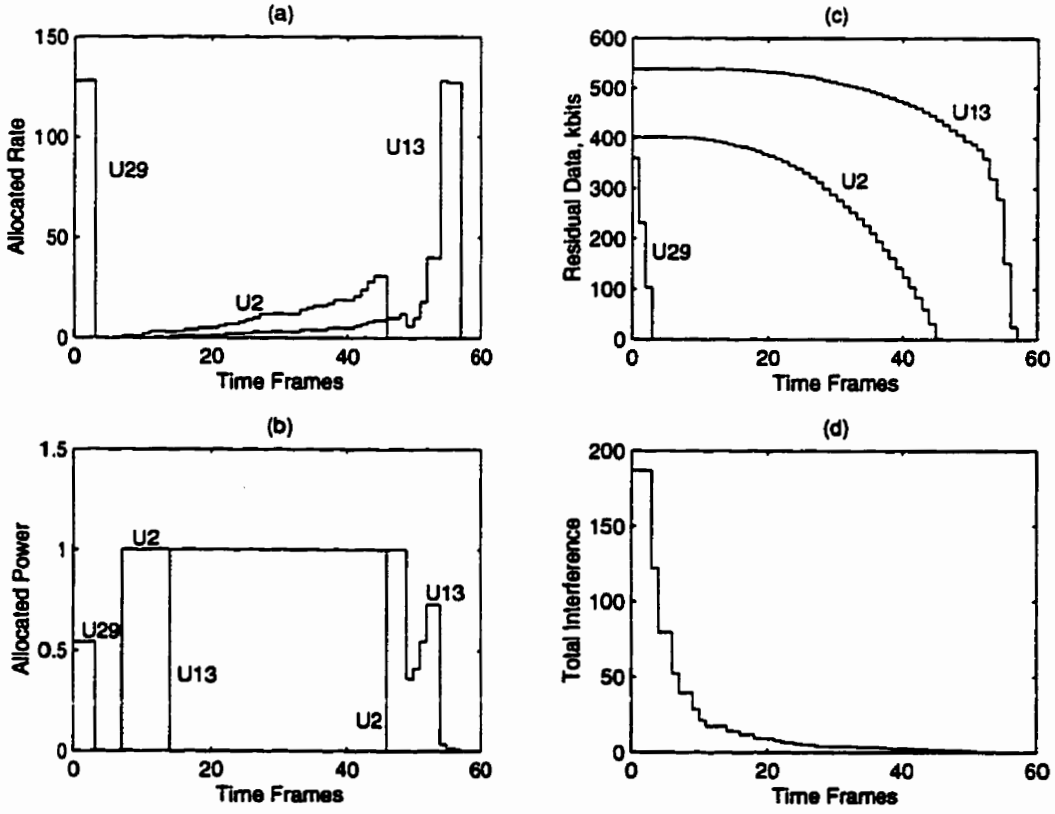


Figure 5.5: Allocated resources, residual data, and the total interference seen by three selected users out of 50 Class III users.

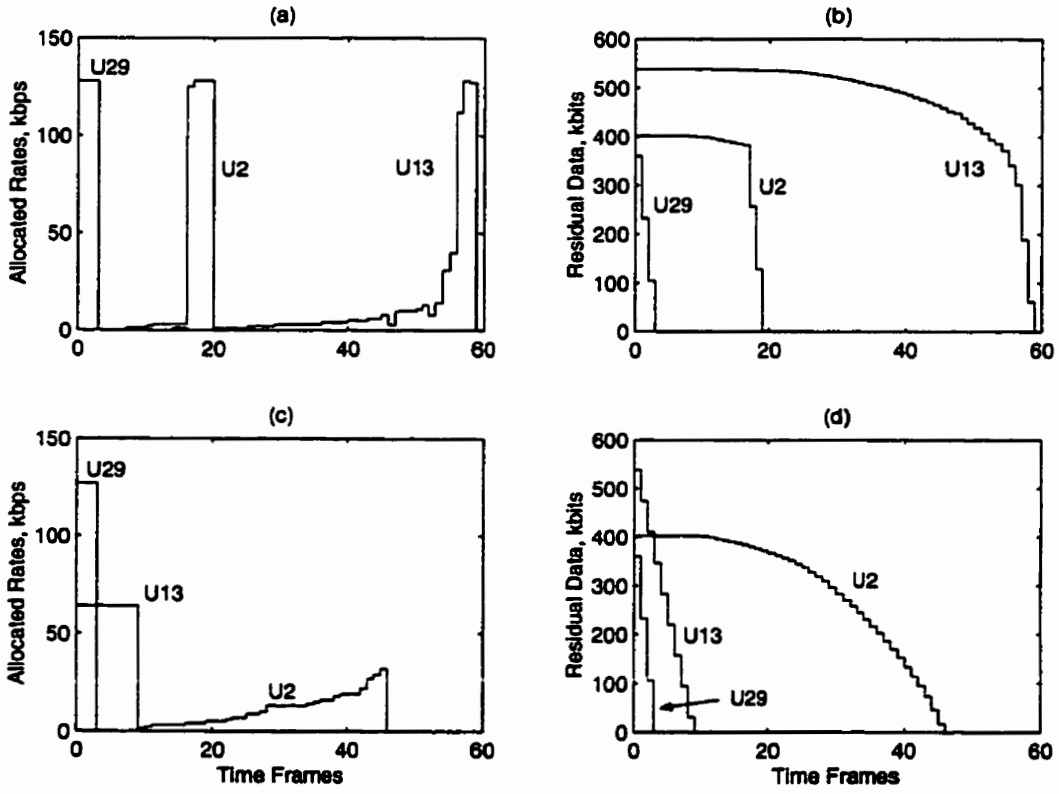


Figure 5.6: Allocated rates and residual data in the presence of one Class II (a and b) and one Class I user (c and d).

Second Simulation Scenario

In the second scenario, the number of admitted calls increases from a single call to 50 calls. The network throughput and sum of the allocated powers versus the number of users, with and without discretization, are sketched in Figure 5.7. For discrete allocation, since the throughput and power sum are case dependent, the aggregate results for 50 independent runs are illustrated. Each dotted line presents the sum of the allocated rates and powers for 50 uniformly-distributed users in the cell.

The observations can be summarized as:

1. The network throughput saturates at a certain level. This level for continuous resource allocation is equal to the capacity since

$$\frac{W}{\bar{\gamma}} = \frac{5 \text{ MHz}}{10} = 500 \text{ kbps} \tag{5.12}$$

   For discrete resource allocation with 128 rate levels, the cell throughput is less than the capacity. This reduction on average is 3.4 percent for 50 users in the cell.

2. There are no dropped calls.

3. The number of admitted Class III users can be fairly large, assuming that the cross correlations between each user's PN code and other users' are negligible. However, admitting too many calls into the network has several drawbacks. One is the increase in the computation time which grows with N in polynomial time. Another drawback is that the allocated rates become very small due to the limited system capacity.

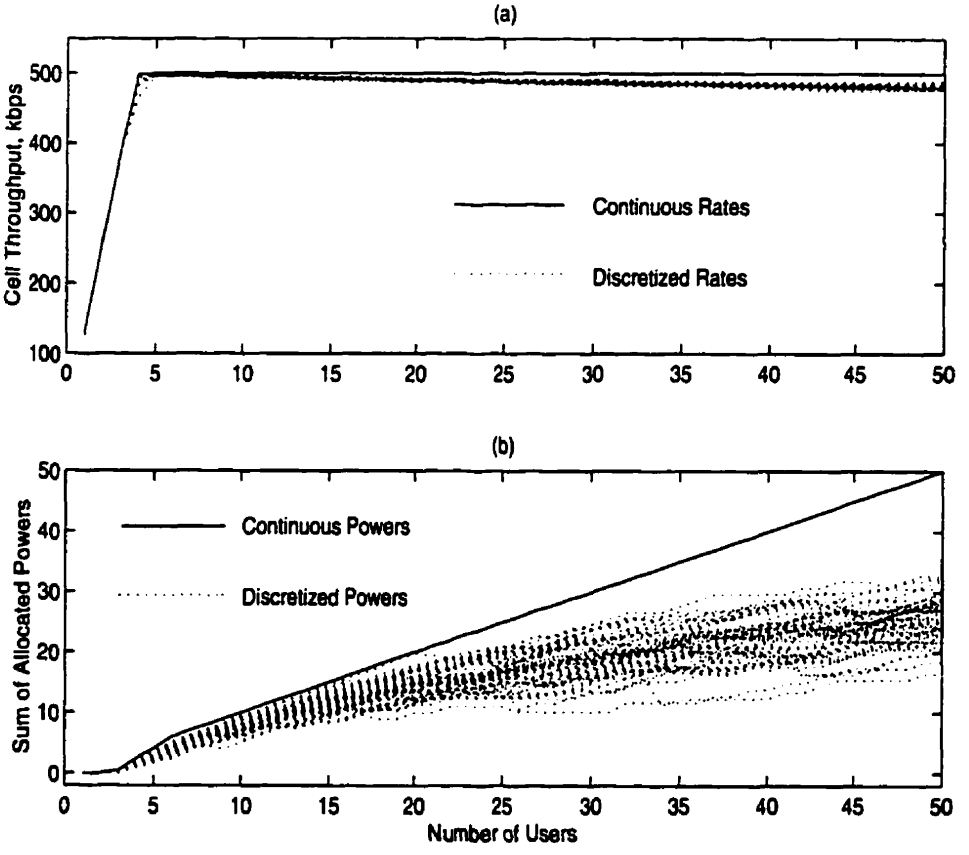


Figure 5.7: Throughput and power sum variations with the number of users for continuous and discrete resource allocation.

4. A lower maximum rate results in a lower capacity utilization because for very low traffic, despite availability of resources and capacity, users are unable to take on as many resources as they need. This problem will not show up if multi-code transmission is used and the number of PN codes is replaced by the number of users in Figure 5.7. The issue of different bounds on the cell throughput will be elaborated more in the next section.

5. Regarding the solution when power and rate variables are continuous, as N

increases, the interference seen by each user increases. To keep a target SIR per bit for a single call, the remedy is to increase the power and/or decrease the rate. Since our resource management algorithm tends to maximize the network throughput, it tries to compensate for the additional interference with extra power. Once the output power has reached the maximum level, the algorithm starts to decrease the rate. When the network throughput saturates, i.e., the capacity is reached, all active users are allocated the maximum power; thus, the total allocated power grows linearly with the number of users. In this case, the SIR per bit can be written as

$$\left(\frac{E_b}{I_0}\right)_i = \frac{W}{r_i}\frac{g_i P_{\max}}{\sum_{j=1}^{N} g_j P_{\max} + \eta} \approx \frac{W}{r_i}\frac{g_i}{\sum_{j=1}^{N} g_j} = \gamma_i, \qquad N \gg 1 \quad (5.13)$$

$$r_i = \frac{W}{\gamma_i}\frac{g_i}{\sum_{j=1}^{N} g_j} \quad (5.14)$$

where it is assumed that the maximum power is equal for all users. It is interesting to note that in a populated single-cell system, a user's throughput is almost independent of the output powers implying that the same throughput can be achieved by lower power levels as well if η is negligible.

6. When the rate and power values are discretized, a slight decrease in the cell throughput is observed. This is the outcome of the rounding procedure which grows with the number of users. Overall, the reduction has not exceeded 4% of the throughput for less than 50 users in the cell. The effect on the power sum, however, is favorable and results in a lower transmitted power in the cell. It happens due to avoiding resource allocation to those users which have very small share in the cell throughput (small rates below the first discretization level) but transmit at their maximum power. Therefore, discretization results in a more power-efficient algorithm. In other words,

continuous resource allocation is less power-efficient as many users transmit at maximum rate while transmitting small data rates.

Comparison with Previous Work

We described the previous related work [48] in Chapter 2. For further performance evaluation, it is helpful to compare our results from the LP algorithm with what has been reported in [48]. We adapt our system parameters to the simulation condition of [48] to maximize $\sum_i r_i$. In their simulation, a bandwidth of 1.25 MHz, a number of voice users with a minimum rate of 8 kbps and SIR per bit of 5 and data users requiring a minimum rate of 4 kbps with $\gamma = 8$ are considered. The maximum received powers at the receivers are 1 and 0.5 watts, respectively. Indices v and d are used to refer to the resource values for voice and data users. For example, N_v, q_d, r_v represent number of voice users, received power of data users, and data rate of voice users, respectively. Table 5.1 presents the results given in [48] together with ours. The unit of data rate is kbps. Our single-cell solution (the LP algorithm) can

Table 5.1: Comparison results.

No. of Users		Results from [48]					Our Results				
N_v	N_d	q_v	q_d	r_v	r_d	$\sum_i r_i$	q_v	q_d	r_v	r_d	$\sum_i r_i$
10	1	1	0.3	22.1	4	225	1	0.3	21.6	4	220
25	1	0.7	0.5	9	4	229	0.7	0.5	9.1	4	231.5
1	5	1	0.1	102	4	122	1	0.5	52.6	16.4	134.6
1	20	1	0.5	20.8	6.25	145.8	1	0.5	20.4	6.3	146.4

be compared with that in [48] as follows:

Supported services: The LP algorithm supports Class II services and covers a wider range of multimedia applications.

Global convergence: In [48], a nonlinear programming problem is solved by the gradient projection method which finds a local optimum and there is no guarantee to find the global optimum without additional precautions and complexity. Our single-cell solution is unique and globally optimum as a result of solving an LP problem.

Computational complexity: An exact comparison in terms of computational complexity is not feasible without knowledge of the method and implementation details. In general, it is well known that an NLP problem is more complex than an LP problem of the same size. In particular, gradient projection method is a feasible direction method to project the gradient into the feasible space. In [48], it is said that 40 to 100 iterations are needed for their NLP algorithm to converge to a local maximum. To project the gradient onto the feasible space, a number of matrix multiplications and inversions are required. If the linear feasible space is defined by $Ax = b$, $(AA')^{-1}$ is one of the necessary computations in each iteration [77]. Using MATLAB, this operation for $N_v = 25$ and $N_d = 1$ needs at least 83 kflop (floating point operations). Regarding the number of iterations, 3.3 to 8.3 Mflop computation is needed in total. The same problem is solved by our LP algorithm with 766.1 kflop in MATLAB. When we consider other necessary operations for the gradient projection method, the advantage of our single-cell solution in terms of low computational complexity will be clearer.

Performance: In general, the global solution of our LP algorithm must provide a higher performance in terms of the sum of allocated rates ($\sum_i r_i$). This has

been validated in Table 5.1, except for the first result. The exception can be explained based on the approximation made in linearizing the problem in Figure 5.1. The approximation ($N \gg 1$) affects the results for smaller numbers of users in the cell.

5.2 Multi-Cell Solution

In a multi-cell environment (general case), the mathematical programming problem in Figure 4.2 must be solved for any feasible assignment. That is, solving an NLP subproblem a very large number of times. The complexity of this problem is extremely high, and derivation of an efficient and accurate solution is very challenging and looks far from realistic. In what follows, we examine different approaches to the problem. In the first approach, we try to solve the problem as is, that is with a max-max structure. In the second approach, a reformulation to change the max-max form into a single problem is examined.

5.2.1 Solution of Max-Max Problem

A straightforward approach to solve the max-max problem in Figure 4.2 requires

1. an efficient and accurate solution for the NLP subproblem, and
2. some criteria to reduce the feasible assignment set dramatically.

NLP subproblem

We have already shown in the previous chapter on page 54 that the NLP subproblem is nonconvex. It is well-known that in nonconvex programming, there does not

exist a universally best method for all kinds of problems; thus different approaches may prove to be best fitted to different problems. Even methods considered to be ineffective may some day become viable as a result of successful research [60].

One way to solve the NLP subproblem is to use available solvers directly. One of the most popular NLP solvers is MINOS (modular in-core nonlinear optimization system) [61]. It has been developed at the Systems Optimization Laboratory at Stanford University and its development is still continuing. Linearly constrained models are solved with a very efficient and reliable reduced gradient technique utilizing the sparsity of the model. Models with nonlinear constraints are solved with a method that iteratively solves subproblems with linearized constraints and an augmented Lagrangian objective function. CONOPT [62, 63, 64], developed by ARKI Consulting and Development, Denmark, is another option that is well suited for models with very nonlinear constraints. It has a fast method for finding a primary feasible solution that particularly fits models with few degrees of freedom, i.e., the number of model variables is approximately the same as the number of constraints. The optimization toolbox in MATLAB provides another NLP solver that is based on the sequential quadratic programming (SQP) method [77]. Since our problem has linear constraints, the MINOS solver should give a better result than other solvers. It is noteworthy that MINOS and CONOPT are used as solvers in the General Algebraic Modeling System (GAMS), a high-level modeling language (code) for mathematical programming problems consisting of a language compiler and a set of integrated high-performance solvers [65].

Many optimization algorithms, including those in the above packages, have been developed to find at least one local optimum for nonlinear problems and are available. None of the existing algorithms, however, guarantees a global optimal solution unless the problem is convex or quasi-convex. Since we are interested in a global

solution and our problem is a nonlinear non-convex problem, we have the following two options to approach the problem. The first approach is problem-oriented and attempts to linearize or convexify the problem, at least in an approximate sense [78]. The second approach is methodology-oriented and has been developed mostly in the last decade. This approach involves methods such as: globalization of local optima with grid and random search, sequential improvement of local optima, enumeration of all optima, and branch and bound [58]. We will go into the details of the latter methods only if the first approach fails. The first approach has effectively solved the problem for a single-cell system. For the general case, one possible way is to exploit the fractional structure of the objective functions in the alternative models using a similar approach as in Theorem 4 in Appendix A.1 on page 142.

As an alternative to using NLP solvers, we have generalized Theorem 4 to linearize the NLP subproblem and use LP methods to solve it. This approach is favorable since the solution is unique and globally optimum. Given an assignment vector a, the optimization problem converts to the form shown in Figure 5.8. Here, without loss of generality, we have assumed $\lambda_h = 0$ for simplicity of the analysis. As explained at the end of this section, using certain criteria, the assignment set can be significantly reduced and the assignments that violate the constraint on h in Figure 4.2 can be removed. The generalized version of Theorem 4 is developed as follows.

Theorem 1 The linear multi-fractional programming problem in Figure 5.8 has an equivalent LP problem.

Proof: Define

$$u_{a_i} = \frac{1}{\sum_{j=1, j\neq i}^{N} g_{ja_i} p_j + \eta} \tag{5.15}$$

$$\underset{p}{\text{Maximize}} \quad \sum_{i=1}^{N} \frac{\lambda_i w_i g_{ia_i} p_i}{\sum_{j=1, j\neq i}^{N} g_{ja_i} p_j + \eta}$$

subject to

$$0 \leq p_i \leq P_{i,\max}$$

$$R_{i,\min} \leq \frac{w_i g_{ia_i} p_i}{\sum_{j=1, j\neq i}^{N} g_{ja_i} p_j + \eta} \leq R_{i,\max}$$

Figure 5.8: NLP subproblem for a typical assignment

$$y_{a_i} = p\, u_{a_i} \tag{5.16}$$

where the variable u_{a_i} is positive and the vector y_{a_i} is non-negative. Accordingly,

$$y_{ia_i} = p_i u_{a_i} \leq P_{i,\max}\, u_{a_i} \tag{5.17}$$

$$w_i g_{ia_i} y_{ia_i} = w_i g_{ia_i} p_i u_{a_i} = \frac{w_i g_{ia_i} p_i}{\sum_{j=1, j\neq i}^{N} g_{ja_i} p_j + \eta} \tag{5.18}$$

$$\sum_{j=1, j\neq i}^{N} g_{ja_i} y_{ja_i} + \eta u_{a_i} = u_{a_i}\Big(\sum_{j=1, j\neq i}^{N} g_{ja_i} + \eta\Big) = 1 \tag{5.19}$$

where $i \in \{1, \cdots, N\}$ and $a_i \in \{1, \cdots, M\}$. Thus, the point (y_{a_i}, u_{a_i}) is feasible. Conversely, if (y_{a_i}, u_{a_i}) is feasible, and the point $(y_{a_i}, 0)$ is infeasible, then, $u_{a_i} > 0$, and $p = y_{a_i}/u_{a_i}$ satisfies the constraints. Therefore, (5.15) and (5.16) map the optimization problem one-by-one onto the equivalent problem as presented in Figure 5.9. The first constraint in this figure is a combination of the following constraints.

$$0 \leq y_{ia_i} \leq P_{i,\max} u_{a_i} \tag{5.20}$$

$$\frac{R_{i,\min}}{w_i g_{ia_i}} \leq y_{ia_i} \leq \frac{R_{i,\max}}{w_i g_{ia_i}}. \tag{5.21}$$

The result is an LP problem. □

$$\underset{y,u}{\text{Maximize}} \quad \sum_{i=1}^{N} \lambda_i w_i g_{ia_i} y_{ia_i}$$

subject to

$$\frac{R_{i,\min}}{w_i g_{ia_i}} \leq y_{ia_i} \leq \min\left[P_{i,\max} u_{a_i}, \frac{R_{i,\max}}{w_i g_{ia_i}}\right]$$

$$\sum_{j=1, j\neq i}^{N} g_{ja_i} y_{ja_i} + \eta u_{a_i} = 1$$

Figure 5.9: Equivalent LP problem for the NLP subproblem

Using the equivalent LP problem, a simulation is carried out for a small scale network with 2 base stations and 3 users in [2]. To find the optimal throughput, the LP problem is solved for all 8 possible assignments and the maximum throughput over all assignments is selected in every time frame. The results are plotted for more than 50 frames and compared with the case of nearest base station assignment. On average, optimal assignments result in 11 percent higher throughput.

Assignment Problem

The NLP subproblem or its equivalent LP problem in Figures 5.8 and 5.9 should be solved for different base station assignments a^l where $l \in S$ and S, the set of

feasible assignments, is a subset of M^N possible assignments. In this case, the cardinality of S, $|S|$, has a significant impact on the complexity of the solution. Thus, it is very important to eliminate infeasible and invalid assignments and avoid unnecessary computations.

The SIR and handoff constraints can have a significant role in reducing $|S|$. In what follows, the effects of these constraints are studied.

Using SIR Constraint - The SIR constraint limits $|S|$ due to the fact that reliable communications can usually take place only within a certain range and through a number of nearby base stations.

Corollary 4 There exists a lower bound on the path gain g_{ik} beyond which reliable communications from user i to base station k is not possible.

Proof: The SIR constraint for user i connected to base station k is given in Figure 5.8. The lower bound on the path gain can be evaluated based on the best possible traffic condition in the network. This condition occurs when there are no interfering users in the network and user i transmits at its maximum power and minimum rate. Substituting these values in the constraint and evaluating g_{ik}, we get

$$g_{ik} \geq \frac{\eta \gamma_i R_{i,\min}}{W P_{i,\max}} \tag{5.22}$$

which gives the desired lower bound. When the path gain is smaller than this bound, under no circumstance can the SIR per bit at the receiver satisfy the target BER. □

It is also desirable to develop an analytical expression for the feasibility condition when the nonlinear problem has linear constraints. Having such an expression

derived, it is possible to find out whether an assignment is feasible by performing the first phase of the simplex method.

The following corollary provides an analytical feasibility condition for a system of two base stations and two users.

Corollary 5 Let $M = 2$ and $N = 2$. The assignment a^l, where $l = 1, \cdots, 4$, is feasible if

$$\frac{g_{1a_1}g_{2a_2}}{g_{1a_2}g_{2a_1}} > \frac{\gamma_1\gamma_2 R_{1,\max}R_{2,\max}}{W^2} \tag{5.23}$$

Proof: See Appendix A.3.

This condition relates the locations and propagation media of the users to their service qualities and is in agreement with the condition (2.21) on page 29. For $N = 2$, (2.21) is a special case of (5.23).

Having the lower bound in Corollary 4, all assignments to base stations with a path gain below the lower bound or invalid in the feasibility condition can be removed. Similarly, extending condition (5.23) to other values of N and M, we can perform a feasibility test for each assignment before going through the optimization process.

Using Handoff Constraint - The other factor in reducing $|S|$ is the limited number of handoffs, $h_{\max}$. With this constraint, as in (4.2) on page 56, the cardinality of S drops significantly. If we let at most $h_{\max}$ users switch to a new base station, we have

$$|S| \leq \sum_{j=0}^{h_{\max}} \binom{N}{j} (M-1)^j. \tag{5.24}$$

This value is derived based on the fact that there are $(M-1)^j$ different assignment vectors with j different elements. Obviously, for $h_{\max} = N$, $|S|$ is equal to M^N. As an example, let $N = 20$ and $M = 5$. If $h_{\max} = 4$, the number of assignments reduces from $5^{20} = 9.54 \times 10^{13}$ to 425×10^3. This number will further be reduced to less than 6.2×10^3 if each mobile finds its best assignment from the 2 nearby base stations.

The solution of the max-max problem suffers from the limit on the number of users and base stations as the computational complexity increases exponentially with N and M, no matter how efficiently the NLP subproblem is solved. A completely different approach in solving our problem is to reformulate it to a less complex problem, preferably changing the structure from a max-max form to a single problem. In the next section, we examine a reformulation process and look into different possible solutions and their complexity and closeness to the optimal solution.

5.2.2 Reformulation

The main problem with our mathematical programming model in Figure 4.2 arises from the size of the assignment set and the fact that we have to solve an NLP or its equivalent LP problem for a large number of times. The reformulation will help only if it deals with this problem effectively. Let b_{ik} be a binary variable as defined in Section 3.3. At any time, user assignments in a certain configuration can be represented by a matrix whose rows include one nonzero element equal to one. Using binary assignment variables b_{ik}'s, we can reformulate the problem as in Figure 5.10.

The new formulation maximizes the network profit while satisfying the necessary

$$\underset{b,p}{\text{Maximize}} \quad \sum_{i=1}^{N}\sum_{k=1}^{M} \frac{\lambda_i w_i g_{ik} b_{ik} p_i}{\sum_{j=1, j\neq i}^{N} g_{jk} p_j + \eta} - \lambda_h h$$

subject to

$$\begin{aligned}
0 &\leq p_i \leq P_{i,\max} \\
R_{i,\min} &\leq \frac{w_i g_{ik} p_i}{\sum_{j=1, j\neq i}^{N} g_{jk} p_j + \eta} \leq R_{i,\max} \\
\sum_{k=1}^{M} b_{ik} &= 1 \\
h &= \sum_{i=1}^{N}\sum_{k=1}^{M} \left[b_{ik} \oplus \bar{b_{ik}} \right] \\
h &\leq h_{\max}
\end{aligned}$$

Figure 5.10: Reformulated problem for multi-cell systems

conditions in the same way as the original problem. The characteristics of the new formulation follow

1. Reformulation leads to a mixed integer nonlinear programming problem (MINLP).
2. Assignments and the NLP subproblem are combined in one problem.
3. The handoff decision variable is defined as a function of the binary assignment variables.

As discussed on page 56, the MINLP problem is much harder to solve than the previous NLP subproblem, however, instead of solving a less complex problem for a large number times, only a single complex problem needs to be solved. A detailed

description and latest developments in solving MINLP problems are provided in Appendix A.5.

We have adopted different methods and algorithms to solve our MINLP problem, namely, relaxed MINLP (RMINLP), LSA, MINLP, and improved MINLP.

RMINLP Algorithm

An immediate solution to any integer programming problem, as explained in Chapter 4, is to relax the integer variables and solve as an NLP problem. We define a continuous version of the variable b_{ik} as c_{ik} where

$$0 \leq c_{ik} \leq 1. \tag{5.25}$$

The relaxed solution will provide values between 0 and 1 for the assignment variables. Since a user is connected to only one base station, it is reasonable to assign the mobile to the base station k where $c_{ik} > c_{ij}$ for all $j \neq k$. It should be noted that the programming problem in Figure 5.10 has been formulated such that the data rate r_{ik} is feasible for all i and k. Besides, rounding c_{ik}'s off to the closest integer does not always work because it is possible that all assignment values lie under 0.5 where rounded values all become zero. With the new continuous assignment variable, the variable h is represented in terms of c_{ik}. Due to the fact that

$$h = \sum_{i=1}^{N}\sum_{k=1}^{M}[b_{ik} \oplus \bar{b_{ik}}] \tag{5.26}$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{M}\left|b_{ik} - \bar{b_{ik}}\right|, \tag{5.27}$$

we consider

$$h = \sum_{i=1}^{N}\sum_{k=1}^{M}\left|c_{ik} - \bar{b_{ik}}\right| \tag{5.28}$$

in the RMINLP problem. Figure 5.11 shows the RMINLP mathematical programming problem. Thus, user i is assigned to base station k and allocated r_{ik} and p_i. The RMINLP algorithm can be summarized as in Figure 5.12.

LSA Algorithm

An important case to study is the conventional assignment method where the assignments are decided independently based on the least signal attenuation (LSA). When the assignment is fixed and known, the linear algorithm in Figure 5.9 finds a unique optimum solution. Thus, the LSA algorithm solves the resource management problem for the fixed LSA assignment. This solution will be useful in comparison with the results of the optimal assignment.

MINLP Algorithm

MINLP problems include the complexities of both NLP and integer programming (IP) problems and have proved to be very expensive and difficult to solve. A number of developments, since the mid 1980's, have paved the way to solve MINLP problems. Advances in NLP and MILP (mixed integer linear programming) solvers, in particular, the development of the outer approximation (OA) algorithm [66] and its extension with the equality relaxation (OA/ER) strategy [67] have played a significant role in this process. An implementation of the OA/ER algorithm in a program called DICOPT (DIscrete Continuous OPTimizer) [68, 69] has provided a tool to solve a family of MINLP problems. DICOPT is now available as a solver with the GAMS modeling language. An interesting feature of GAMS/DICOPT is the use of existing optimizers to solve its subproblems (NLP subproblem and MILP master-problems). Thus, any new development and enhancements in the NLP and

$$\underset{c,p}{\text{Maximize}} \quad \sum_{i=1}^{N}\sum_{k=1}^{M} \frac{\lambda_i w_i g_{ik} c_{ik} p_i}{\sum_{j=1, j\neq i}^{N} g_{jk} p_j + \eta} - \lambda_h h$$

subject to

$$\begin{aligned}
0 &\leq p_i \leq P_{i,\max} \\
R_{i,\min} &\leq \frac{w_i g_{ik} p_i}{\sum_{j=1, j\neq i}^{N} g_{jk} p_j + \eta} \leq R_{i,\max} \\
\sum_{k=1}^{M} c_{ik} &= 1 \\
h &= \sum_{i=1}^{N}\sum_{k=1}^{M} \left| c_{ik} - \bar{b_{ik}} \right| \\
h &\leq h_{\max}
\end{aligned}$$

Figure 5.11: RMINLP problem

Solve Figure 5.11 problem as RMINLP using NLP solvers,

For every user i

Find k **such that** $c_{ik} \geq c_{ij}\ \forall\ j \neq k$

Assign user i **to base station** k

Allocate floor(r_{ik}) **and floor**(p_i)

End

Figure 5.12: RMINLP algorithm

MILP solvers become automatically available to DICOPT. The DICOPT algorithm and its employed techniques are explained in Appendix A.5.

The class of MINLP problems that can be solved by DICOPT has a certain structure. First, integer variables must be binary. Second, binary variables should appear linearly so nonlinearities are involved in only the continuous variables. In our MINLP problem, however, the second requirement is not valid as the binary variables are involved in nonlinearities in the objective function.

Using an intuitive method, we adapt the structure of our problem such that the DICOPT solver can be applied to solve it. For this purpose, we use the continuous version of the assignment variable as an auxiliary variable and replace b_{ik} by c_{ik} in the objective function. We add a new constraint to equate c_{ik} with b_{ik}, as well. This new version of our MINLP problem is compatible to the acceptable structure for DICOPT, as shown in Figure 5.13. The resulting assignments are binary integers, if the problem has an integer solution. Otherwise, DICOPT outputs the relaxed solution which can be dealt with in the same way as in the RMINLP algorithm. In this case, also, we use h as defined in (5.28). If the integer solution exists, c_{ik} will be equivalent to (5.27). Figure 5.14 illustrates the algorithm for the solution of the MINLP problem.

Improved MINLP

While any initial point in convex problems eventually converges to a unique optimum solution, the starting points in nonconvex problems result in different local optima. A local optimum can vary from very close to far from the global optimum. Therefore, a "good" starting point can lead to a local optimum sufficiently close to the global solution. Such a point in regard to the assignment variables is the

$$\underset{b,c,p}{\text{Maximize}} \quad \sum_{i=1}^{N}\sum_{k=1}^{M} \frac{\lambda_i w_i g_{ik} c_{ik} p_i}{\sum_{j=1, j\neq i}^{N} g_{jk} p_j + \eta} - \lambda_h h$$

subject to

$$\begin{aligned}
0 &\leq p_i \leq P_{i,\max} \\
R_{i,\min} &\leq \frac{w_i g_{ik} p_i}{\sum_{j=1, j\neq i}^{N} g_{jk} p_j + \eta} \leq R_{i,\max} \\
\sum_{k=1}^{M} c_{ik} &= 1 \\
c_{ik} &= b_{ik} \\
h &= \sum_{i=1}^{N}\sum_{k=1}^{M} \left| c_{ik} - b_{ik}^{-} \right| \\
h &\leq h_{\max}
\end{aligned}$$

Figure 5.13: MINLP problem

well known LSA. It is expected, and will be validated in our simulations, that LSA initial points outperform random initializations, thus the term improved MINLP (I-MINLP) algorithm.

The standard DICOPT software starts its algorithm by solving the problem as an RMINLP problem. Then, the relaxed values of the binary variables are input to the MILP master problem. It is, therefore, insensitive to any initial assignment for the binary variables and lacks the flexibility of examining initial assignments such as LSA. One way to remove this limitation is to modify the DICOPT code and change the sequence of solving the NLP and MILP subproblems. This modification

Solve Figure 5.13 problem as MINLP using DICOPT solver
If integer solution exists, for every user i
    Assign user i to base station k if $b_{ik} = 1$,
    Allocate floor(r_{ik}) and floor(p_i)
Else
    Assign user i to base station k
        when $c_{ik} \geq c_{ij}\ \forall\ j \neq k$
    Allocate floor(r_{ik}) and floor(p_i)
End

Figure 5.14: MINLP algorithm

requires further research beyond the scope of this work. However, the particular structure of our problem can be exploited to compensate for this inflexibility in DICOPT. That is, the initialization can be effectively performed by applying the desired initial assignment to the auxiliary variable c_{ik}.

5.2.3 Numerical Results

First Simulation

We conduct a Monte Carlo simulation to study the comparative performance of the resource management algorithms. This experiment is intended to simulate a uniform traffic in the network in a statistical average sense. Table 5.2 summarizes the simulation parameters and assumptions. The target SIR per bit for the first

Table 5.2: Simulation parameters and assumptions.

No of users	1 to 70	Max power	1 watt
No of BS's	9	Max rate	256 ksps
Network coverage	6×6 km^2	Min rate	0
User distribution	Uniform	Handoff cost	0
Propagation law	$10^{\zeta/10}\ r^{-4}$	Bandwidth	5 MHz
	$\mu_\zeta = 0,\ \sigma_\zeta = 8$ dB	Chip rate	4.096 Mcps
Path gains	Known	E_b/I_0	8.2 dB
Background noise	.001 watt		

simulation is 8.2 dB which can achieve a BER performance of 10^{-6} using a (23,12) Golay code and soft decision decoding [23].

For any i users in the network, $i \in \{1, \cdots, 70\}$, 100 random configurations are generated. The location of each user is a two-dimensional uniform random variable in a rectangular area of 6×6 km^2. The 9 base stations are located on a 2km grid centered in this area. The path gain for a user is calculated based on a path-loss exponent of 4 and a standard deviation of 8 dB for shadowing. Thus, each configuration is characterized by a path gain matrix whose elements are g_{ik} with the size 70×9. We solve the resource management problem using the RMINLP, LSA, MINLP and I-MINLP algorithms, as developed in Section 5.2.2. Figure 5.15 illustrates the resulting network throughputs versus the number of users for each simulation run and in comparison with the network capacity (see Definition 1 on page 46). The cell capacity C_c is calculated according to Corollary 3 on page 65 as (5 MHz/8.2 dB)=757 ksps, thus resulting in a network capacity of $C = 9C_c = 6813$ ksps. The mean and standard deviation of the network throughput over 100 results

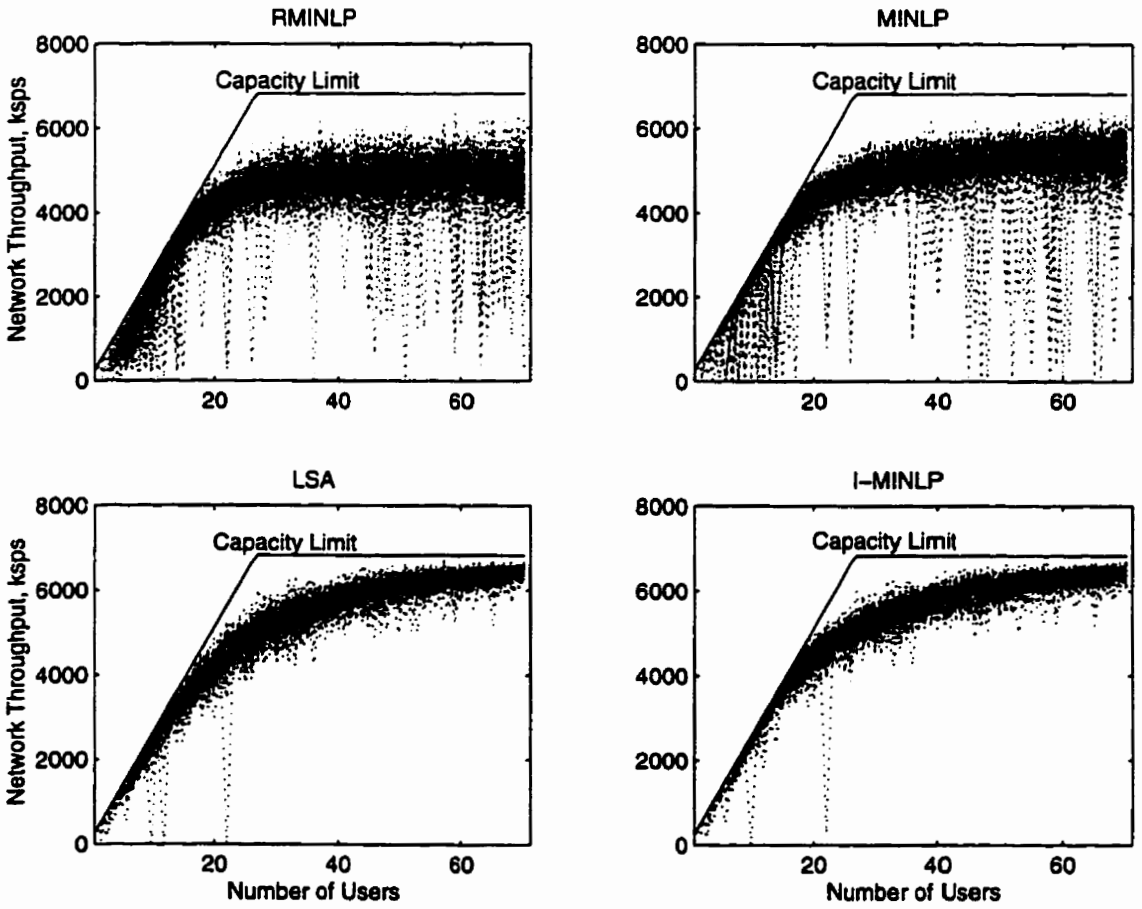


Figure 5.15: Network throughput for different algorithms.

are sketched in Figures 5.16 and 5.17 on page 93. The observations and results are as follows.

1. The largest average network throughput with the lowest standard deviation is achieved by the I-MINLP algorithm. When 70 users are in the network, the average network throughput for each algorithm is given in Table 5.3.

2. The achievable throughput is bounded by two factors: the number of users, and the network capacity. That is, up to a certain number of users n, the

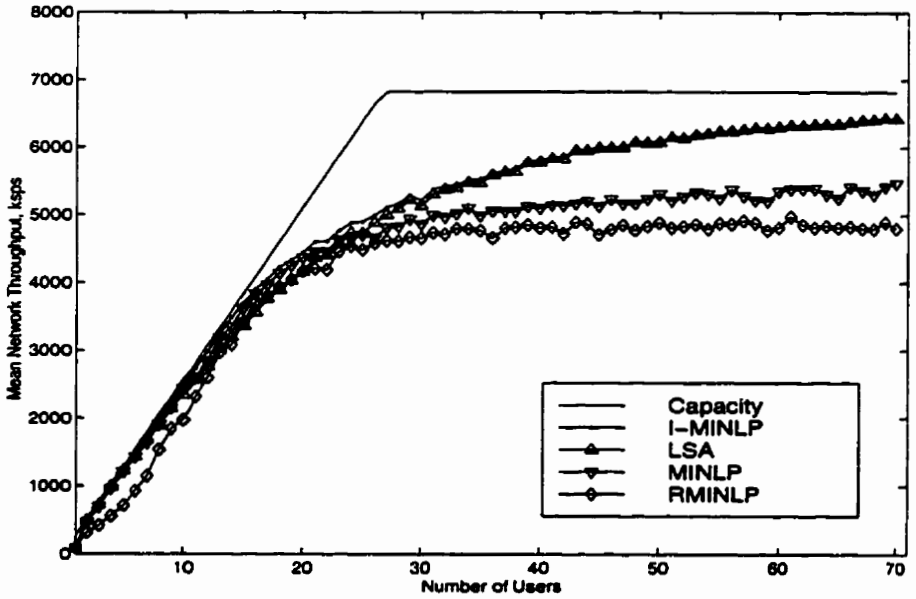


Figure 5.16: Mean network throughput of different resource management algorithms.

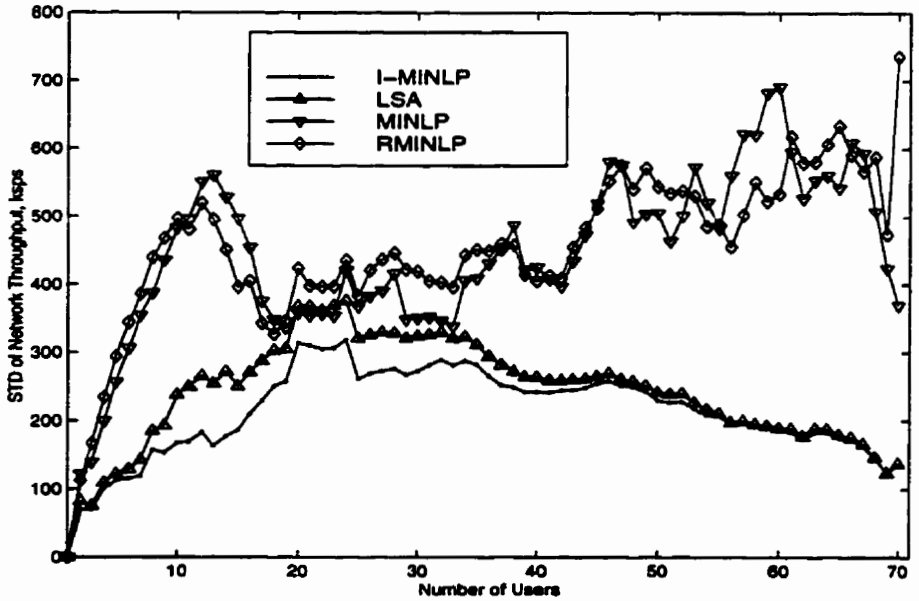


Figure 5.17: Standard deviation of network throughputs.

Table 5.3: Capacity utilization (N=70, uniform user distribution).

Algorithm	Network Throughput (ksps)	Utilized Capacity (%)
Network Capacity	6813	
I-MINLP	6424	94.3
LSA	6422	94.3
MINLP	5475	80.4
RMINLP	4802	70.5

capacity increases linearly and beyond that limit, it stays constant. This number can be found as

$$n = \left\lceil \frac{\text{Network Capacity}}{\text{Maximum Rate}} \right\rceil \tag{5.29}$$

or $n = \lceil 6822/256 \rceil = 27$. The result of this observation is that spectrum efficiency must be evaluated with a sufficiently high number of users in the network unless there is no maximum rate limit.

3. The role of the initialization of the algorithms is significant. The MINLP algorithm with a random initial assignment utilizes 14% less capacity than the case when initialized by the LSA (see Table 5.3).

4. The mean network throughputs for the I-MINLP and LSA algorithms are very close. This is expected since users are located uniformly in the network and throughputs are averaged over the 100 runs. In other words, the aggregate result tends to remove nonuniformities in the interference seen by each base station and the LSA solution gets closer to the optimal solution. A comparison between the performance of the I-MINLP and LSA algorithms should be

made based on individual results of each run. Let Δ be the difference in the network throughputs in a specific configuration. An important observation is that $\Delta \geq 0$ is always true, meaning that the I-MINLP outperforms the LSA algorithm. The average distribution of Δ is shown in Figure 5.18 for the 100 random configurations. It shows that Δ varies from 0 to 967 ksps but is less than 100 ksps 70% of time.

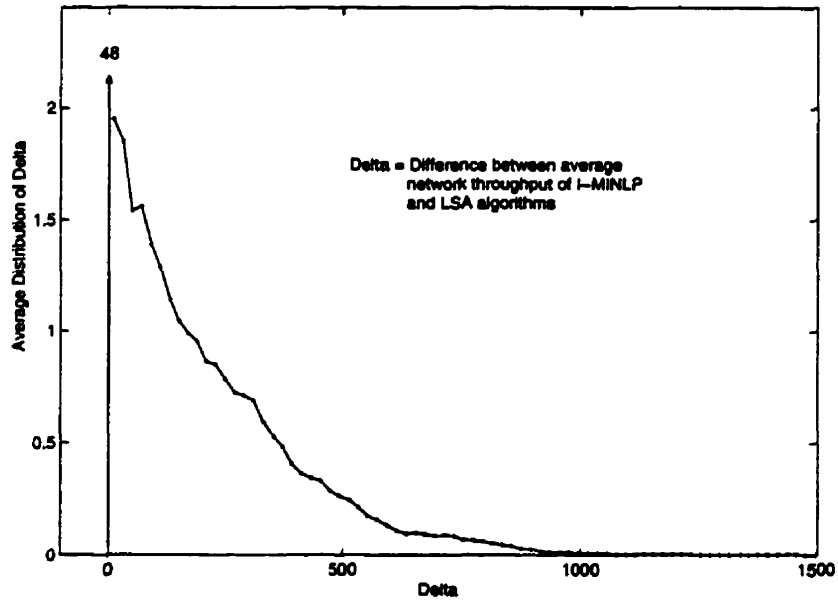


Figure 5.18: Average distribution of the difference in the network throughputs.

Second Simulation

The objective of the second simulation is to examine the effect of uneven traffic on the performance of LSA and I-MINLP algorithms. For this purpose, 100 Class III users are located based on a 2-dimensional Gaussian distribution in a network with

9 base stations. The concentration of users is around the center of the network. Starting with such a configuration, mobile users move with random speeds in an outward direction. This simulation is performed for 200 frames under the condition as summarized in Table 5.4. The loglinear propagation law is selected for illustration

Table 5.4: Second simulation parameters and assumptions.

No of users	70	Max power	1 watt
No of BS's	9	Max rate	256 ksps
Network coverage	8×8 km^2	Min rate	0
User distribution	Gaussian	Max no. handoffs	10
Propagation law	r^{-4}	Bandwidth	5 MHz
Path gains	Known	Chip rate	4.096 Mcps
Background noise	.001 watt	E_b/I_0	3.3 dB

purposes. The target SIR per bit is similar to the LCD service requirement in [10] which can maintain a BER of 10^{-6} by using a turbo code with constraint length of 3, QPSK modulation, 2-antenna diversity, rake receiver and soft decision decoding. The symbol rate in the physical layer for a 64-kbps LCD service is 256 ksps. Figures 5.19-5.21 illustrate 3 snapshots of the network for I-MINLP and LSA algorithms under the same condition. An explanation of these figures is given as follows:

- Each mobile user is represented by a rectangle, colored based on its allocated rate. The colorbar is shown vertically on the right hand side of the network and scales allocated rates from 0 (dark blue) to 256 (dark red) ksps.

- Assignment of each user is shown by a dotted line originating from the user to

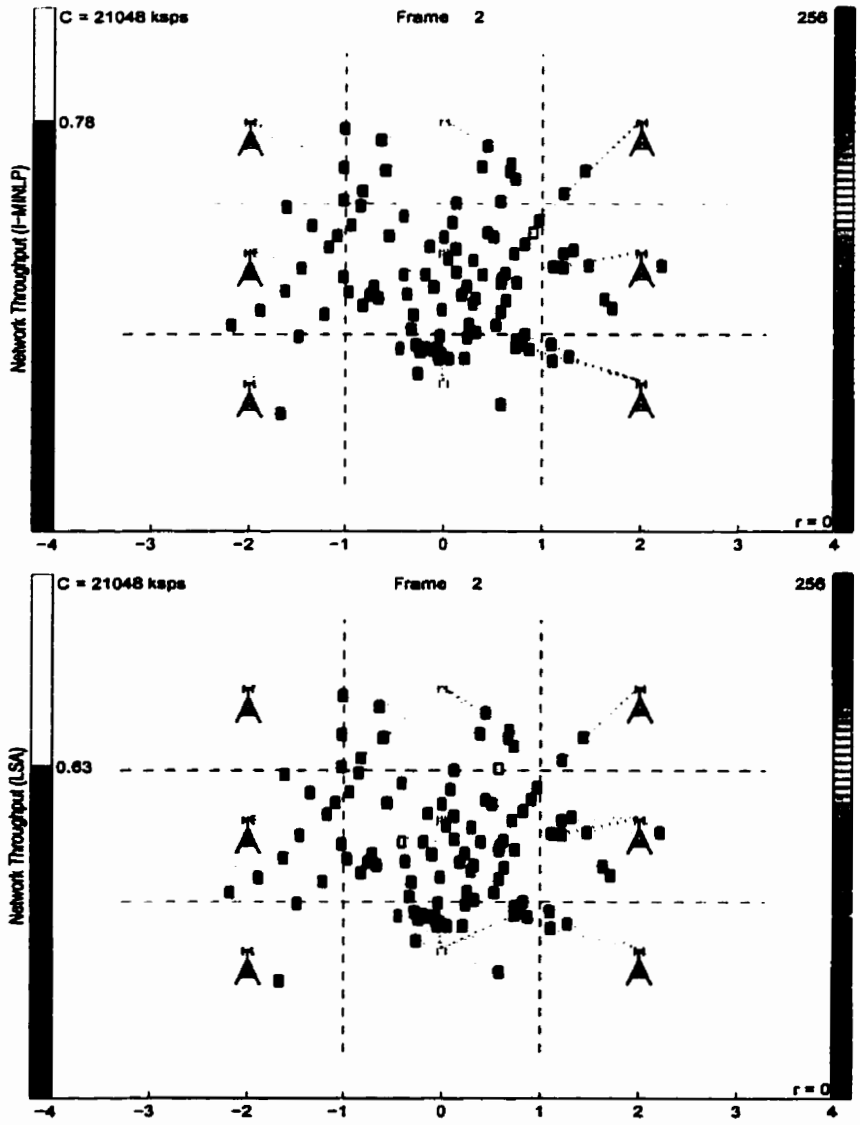


Figure 5.19: Comparison of I-MINLP and LSA for unevenly distributed traffic (frame 2).

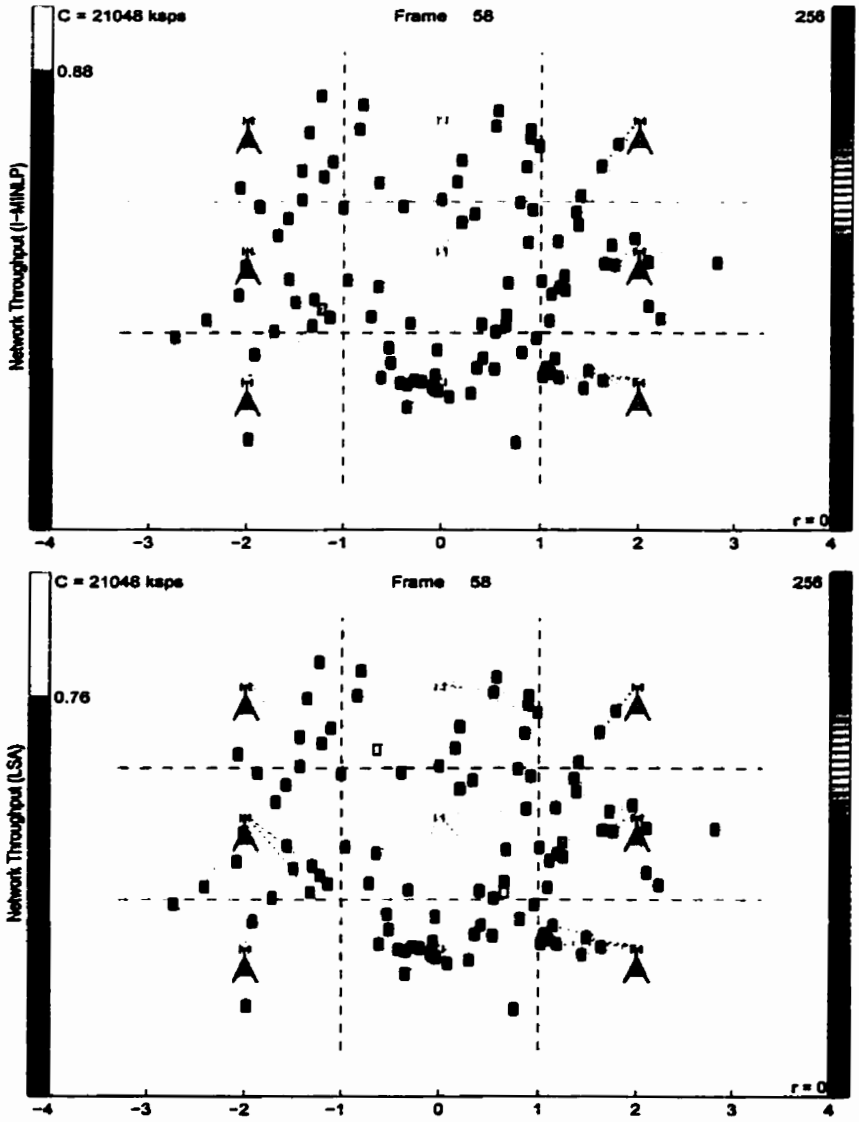


Figure 5.20: Comparison of I-MINLP and LSA for unevenly distributed traffic (frame 58).

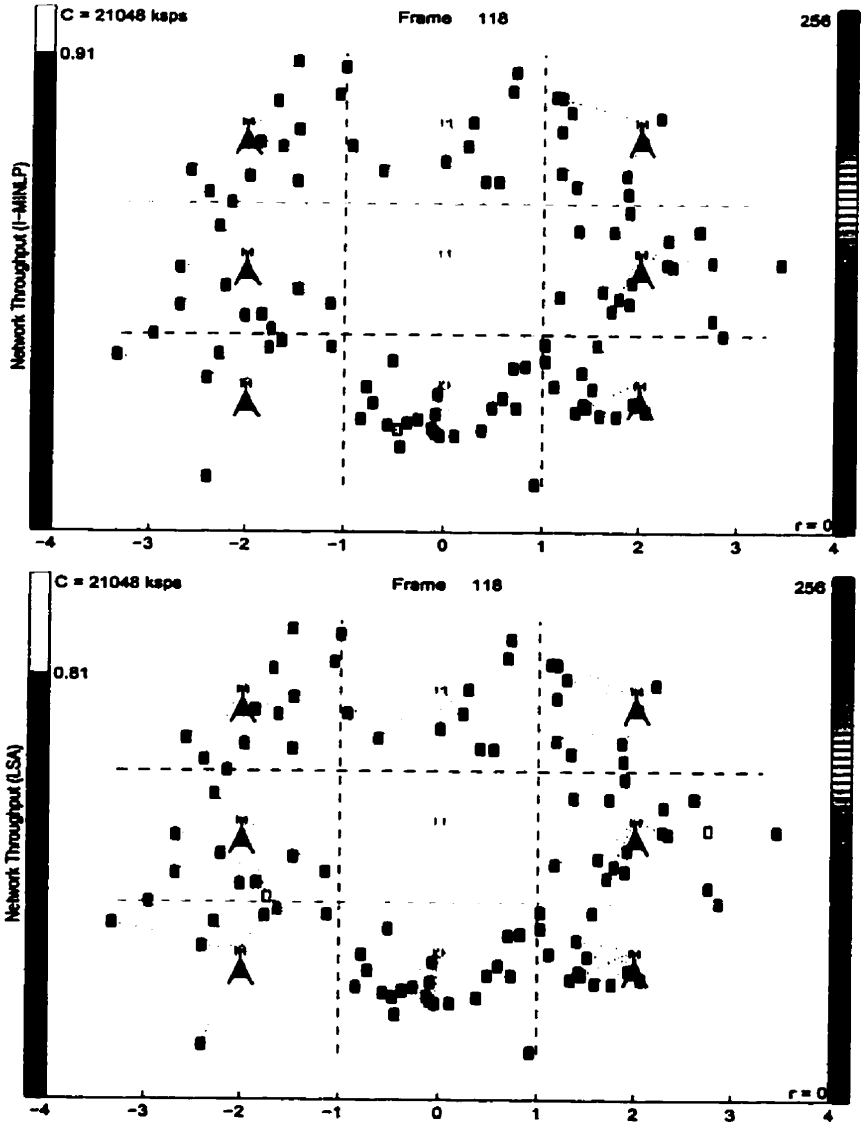


Figure 5.21: Comparison of I-MINLP and LSA for unevenly distributed traffic (frame 118).

the assigned base station. No connection line implies that at the particular frame the user has not been allocated any resources.

- The network throughput is shown with respect to the capacity by another bar graph on the left hand side of the figures. The capacity and fraction of the utilized capacity are printed beside the bar. The following observations and results are noteworthy.

1. LSA assignment is not the best assignment when traffic is non-uniform. The I-MINLP algorithm achieves a higher throughput by sharing the traffic load among all base stations.

2. Network throughput of both algorithms vary with time as the traffic pattern varies; the higher concentration of users in the network, the less total network throughput.

3. On average, 10 percent improvement is gained by optimal base station assignments. Figure 5.22 illustrates the network throughputs in 200 successive time frames.

4. In Figure 5.22, infeasible solutions appear in the form of small gaps in the results, as throughput is zero when there is no feasible solution. It can be seen that I-MINLP experiences much less infeasibility than LSA.

Third Simulation

Another simulation is performed to study the reuse efficiency and the effect of the number of base stations in a fixed area using the I-MINLP algorithm with a fixed number of users (100) and a target SIR per bit of 3.3 dB as in the LCD service.

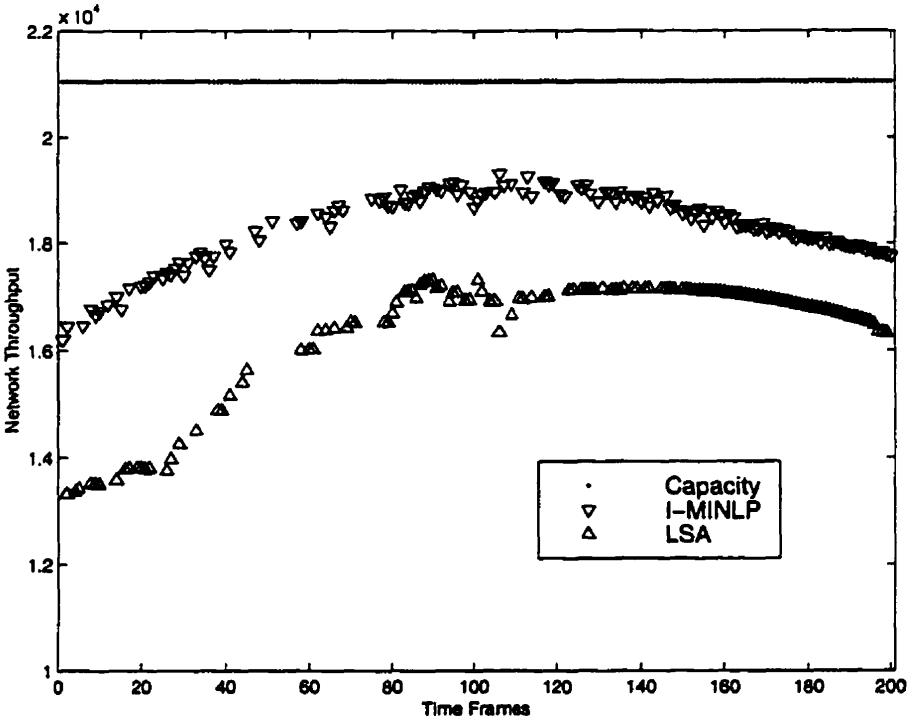


Figure 5.22: Network throughputs (unevenly distributed traffic).

The cell capacity in this case is 2338 ksps. Figure 5.23 illustrates the network and cell throughputs for 1 to 25 base stations in the 6 × 6 km^2 area. Two limits on the achievable capacity are shown in the upper figure: the capacity limit ($M \times C_c$) which linearly increases with the number of cells, and the population limit which limits the achievable throughput to $N \times R_{\max}$. These limits are almost equal for $M = 11$. As real indications of the performance, the results for $M \leq 9$ can be used as an indication of the capacity utilization and related to the **reuse efficiency** or **reuse factor** of the network. Table 5.5 presents the reuse factor using I-MINLP resource allocation algorithm. A reuse efficiency of greater than 0.94 compared to the theoretical reuse factor of 1 in CDMA systems is quite satisfactory and sufficiently high.

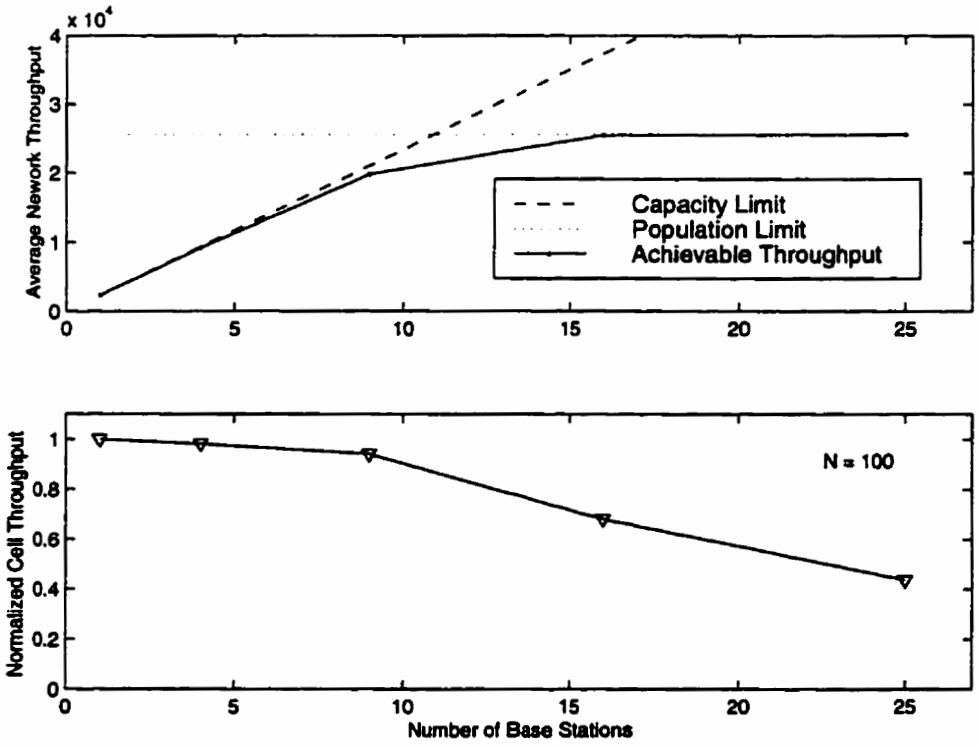


Figure 5.23: Average cell and network throughput for 100 users versus the number of base stations.

5.3 Summary

In this chapter, we have developed techniques and algorithms to solve the resource management problem for single- and multi-cell environments. The single-cell algorithm finds an exact global optimum and determines the maximum achievable throughput per cell for a given maximum power and a specific SIR per bit requirement. This value represents a new capacity bound and can be used as a benchmark for evaluation purposes. The multi-cell solution for fixed assignments such as LSA assignment is globally optimal. When assignment decisions are included in the

Table 5.5: Network reuse factor with I-MINLP algorithm.

M	Capacity (ksps)	Network Throughput (ksps)	Reuse Factor
1	2338	2327 ksps	0.995
4	9352	9175 ksps	0.981
9	21042	19798 ksps	0.941

mathematical programming problem, an I-MINLP algorithm has been developed which utilizes the network capacity almost completely, i.e., more than 94% in a network with less than 9 base stations.

Chapter 6

Implementation Issues

Resource allocation algorithms face a number of problems in their implementation in practice. This chapter is devoted to several related issues. One is in conjunction with the input path gain estimations and their inaccuracies. We are interested in performing a sensitivity analysis to investigate the effect of estimation error on the performance of the algorithms. The other issue is related to the control strategy in implementing the resource management algorithms. We study a centralized implementation as a primary choice when the algorithm performs based on network-wide information. Some simplification has been made in the centralized method to reduce the computational complexity and information flow in the network at a cost of lower throughput. This is followed by a discussion of decentralization methods where decentralizations at three different levels are studied. This chapter concludes with a compatibility analysis of our algorithms with respect to the recent major proposals for IMT-2000 as the third generation of wireless systems for the near future.

6.1 Sensitivity Analysis

The knowledge of path gains is required in all algorithms that have been developed and described so far. These path gains in the links between users and base stations should be estimated and input to the resource management center. These estimations are presumably prone to errors and could result in invalid resource values. The purpose of this section is to study the sensitivity of our resource management algorithms to the errors in the path gain estimations. In other words, we intend to investigate how path gain estimates instead of the accurate values affect the performance of the algorithms.

A well-known method to measure the path gain from a mobile to a base station is to use the knowledge of transmitted and received power of a reference pilot signal [53] in the forward link. Using the path gain of the forward link in the reverse link, however, is erroneous due to the carrier frequency difference, resulting in different path loss [72] and different fading rates[1] in the two links. Another method to estimate the path gains in our model is a direct use of the mobile transmitted power, already known by the base station, and the received signal power at its receiver. This method does not suffer from the problem in previous method, however, constant channel variation will affect the accuracy in both methods. An important factor in fast channel variation is high mobile speed which increases path gain estimation error.

Let $\hat{g}_{ik}$ be the estimate of the actual path gain g_{ik} that is experienced by the transmitted signal from user i to base station k. While the signal undergoes the path

[1]The normalized fading rate is defined as the product of the doppler frequency and data symbol duration. Since forward and reverse links use different carrier frequencies, the doppler frequency in each link will be different for a given mobile speed, resulting in different fading rates.

with gain g_{ik}, the resource management algorithm uses $\hat{g}_{ik}$ in the computations and allocates the resources accordingly. The question is whether the allocated resources under the actual channel state are feasible. Obviously, the answer to this question is related to the error level. We are interested to find a mapping between the standard deviation of the error and the probability of an infeasible solution.

Assuming independent fading in each path, g_{ik} can be modeled as an independent random variable. It is related to the path gain estimate by the relation

$$g_{ik} = \hat{g}_{ik} + e_{ik} \tag{6.1}$$

where e_{ik} is the error estimation with zero mean and variance σ_g^2. Experimental data, e.g. in [73, 74], and theoretical studies on the short term average of a radio signal in fading channels [54] suggest that the received signal power at the base station has a lognormal distribution. Given p_i as the allocated power to user i, the path gain distribution is lognormal with mean $\mu_{g_{ik}}$ and variance σ_g^2. The distribution of g_{ik} is given as

$$f(g_{ik}) = \frac{1}{\sqrt{2\pi}\sigma g_{ik}\beta} \exp\left\{-\frac{10\log(g_{ik}-\mu)^2}{2\sigma^2}\right\}, \tag{6.2}$$

where $\beta = \ln 10/10$, and σ and μ are in dB and related to the mean and variance of the lognormal distribution as

$$\mu_{g_{ik}} = \hat{g}_{ik} = e^{\beta\mu+\beta^2\sigma^2/2} \tag{6.3}$$

$$\sigma_g^2 = e^{2\beta\mu+2\beta^2\sigma^2} - e^{2\beta\mu+\beta^2\sigma^2}. \tag{6.4}$$

For user i assigned to base station k and allocated power p_i and data rate r_{ik}, the probability $P_{inf,i}$ that the allocated resources are infeasible is

$$P_{inf,i} = \Pr\left(r_{ik} > \frac{w_i g_{ik} p_i}{\sum_{j\neq i} g_{jk} p_j + \eta}\right) \tag{6.5}$$

$$= \Pr\left(I_{ik} > \frac{w_i g_{ik} p_i}{r_{ik}}\right) \tag{6.6}$$

where

$$I_{ik} = \sum_{j \neq i} g_{jk} p_j + \eta. \tag{6.7}$$

Sum of lognormal random variables is often approximated by another lognormal random variable [21]. A Gaussian approximation is also valid when lognormal random variables are independent and their number is large enough. The latter conditions hold in our problem, therefore, applying the central limit theorem, the distribution of I_{ik} is approximated as Gaussian with the following statistics.

$$\mu_{I_{ik}} = E[I_{ik}] \tag{6.8}$$

$$= E[\sum_{j \neq i} g_{jk} p_j + \eta] \tag{6.9}$$

$$= \sum_{j \neq i} \hat{g}_{ik} p_j \tag{6.10}$$

$$\sigma^2_{I_{ik}} = \mathrm{Var}[\sum_{j \neq i} g_{jk} p_j + \eta] \tag{6.11}$$

$$= \sum_{j \neq i} \sigma^2_g p_j^2. \tag{6.12}$$

Therefore, $P_{inf,i}$ will be equal to

$$P_{inf,i} = \int_{g_{ik}} Q\left(\frac{w_i g_{ik} p_i / r_{ik} - \mu_{I_{ik}}}{\sigma_{I_{ik}}}\right) f(g_{ik}) dg_{ik}. \tag{6.13}$$

In the derivation of $\sigma^2_{I_{ik}}$, the variance of η is assumed to be negligible.

We are interested in the probability that no user in the network encounters infeasibility. Based on the above definition, $1 - P_{inf,i}$ is the probability that user i is in a feasible state. Thus, the probability that all users are in feasible states is $\Pi_i(1 - P_{inf,i})$ due to the independence of g_{ik}'s and $P_{inf,i}$'s. The desired probability, therefore, becomes $1 - \Pi_i(1 - P_{inf,i})$. Figure 6.1 presents the average probability of infeasible state versus the standard deviation of error σ_g for various numbers of discretization levels. Each subplot presents the result for a particular number of

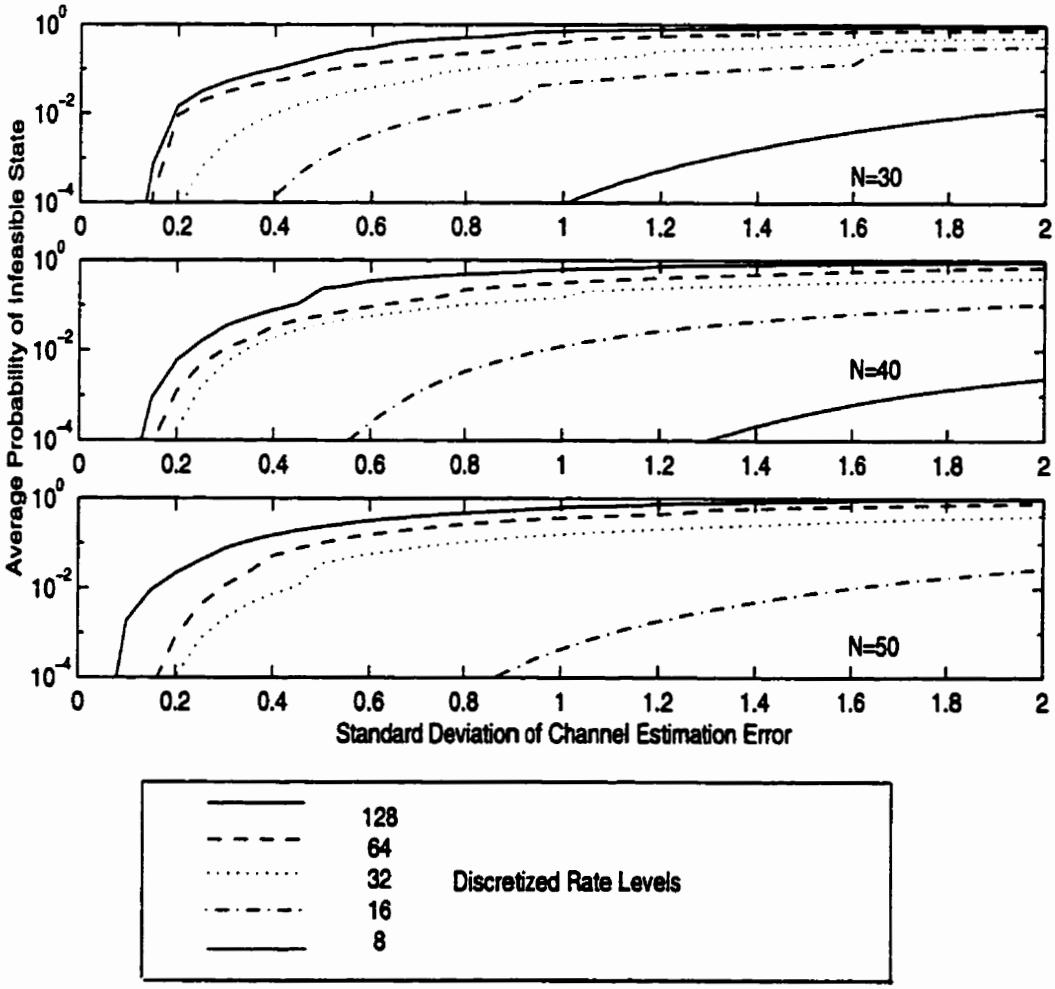


Figure 6.1: Sensitivity of the resource allocation algorithms to the path gain estimation error.

users. It can be observed that the probability of infeasible solution increases as the error increases. When the number of rate levels is large, the estimation error has a more adverse effect. This is justifiable due to the fact that a larger number of rate levels yields a higher capacity utilization and is closer to the optimal solution, thus, there is a small margin for error tolerance. Furthermore, the error tolerance for a lower number of rate levels increases with N while it remains unaffected or gets slightly worse for higher numbers of rate levels.

6.2 Centralized Implementation

Centralized implementation is applicable when an algorithm manages and performs resource allocation in a network-wide scale. That is, all necessary information is made available to an RMC for processing and decisions are broadcast to base stations and users. Figure 6.2 illustrates a centralized implementation. Base station k measures the path gain g_{ik} for all i and reports them to the RMC. User information including service types and qualities are stored at a user data base upon the admission of each user. Fixed user parameters including λ, λ_h, $P_{\max}$, $R_{\max}$, and E_b/I_0's are made available to the RMC by the data base. Using channel and user information, the RMC runs the resource management algorithm and outputs the new resource allocation values including p, r, and b and passes these to the base stations and users. In such a centralized implementation, the following information must be available at each base station:

1. Transmitted powers of all users: this information together with the received signal powers is required for measuring the path gains. This method is the preference against using the pilot signal in the forward link because measurements are more accurate and no relevant control information is needed over

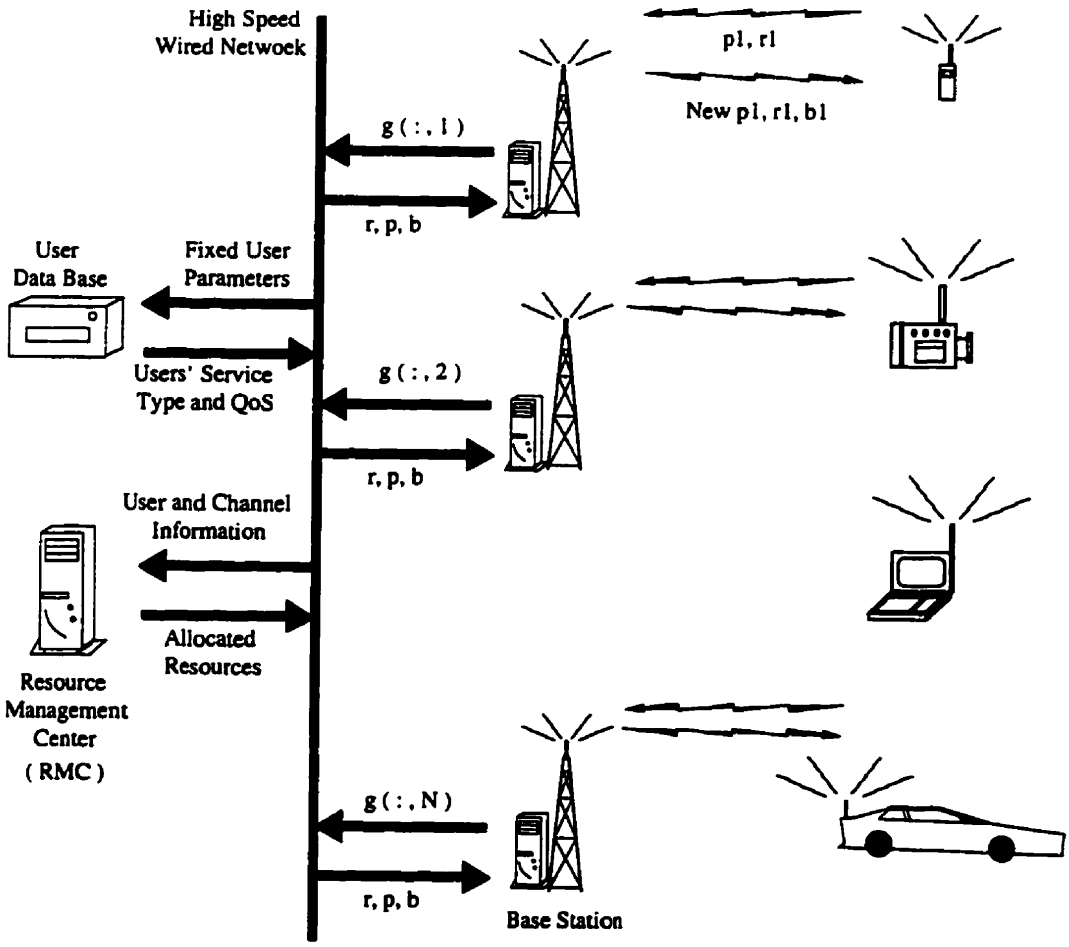


Figure 6.2: A centralized implementation.

the wireless link, as the knowledge of the transmitted powers already exists at the base station. Hence, in a network with N users and M base stations, NM measurements must be performed and reported to the RMC.

2. Transmission rates of the users being served by the base station: The cell-site receivers need the data symbol duration (T_s) of the users in their integrators and samplers, as depicted in Figure 6.3.

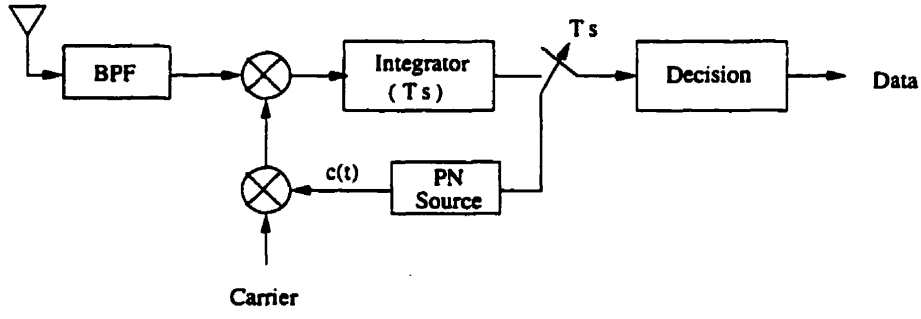


Figure 6.3: A simplified CDMA receiver.

Centralized algorithms, in general, are difficult to implement due to their need for global information and control and, consequently, high amount of computations and communications in the network. For example, in dynamic channel allocation algorithms in narrowband wireless systems, a centralized scheme requires global rearrangement of all radio channels assigned to the users whenever a new user is admitted or a handoff takes place. This shows the algorithm to be impractical [21]. In CDMA, however, centralized resource management is less complex because the channel (PN code) assignment is fixed during a call and all users share the same radio channels. A centralized implementation of our resource management algorithms has the following advantages:

1. Resource allocation is based on global information and optimization, thus a higher throughput is achievable.

2. Local congestion due to uneven traffic conditions is avoided by optimal base station assignments.

3. Despite increasing the information flow among base stations and the RMC in the high speed wired network, there is no additional corresponding information flow on the wireless section.

Any step to reduce the amount of information flow in the network and lessen the complexity of the mathematical programming problem would eventually facilitate the implementation of our algorithms. Under certain reasonable assumptions, a more realizable centralized algorithm can be developed as follows.

Let I_k be the total received signal at base station k which can be expressed as

$$I_k = \sum_{j=1}^{N} g_{jk} p_j + \eta \approx \sum_{j=1, j\neq i}^{N} g_{jk} p_j + \eta \tag{6.14}$$

where $k \in \{1, ..., M\}$. At any time, I_k is known to the base station by measuring the radio signal at the receiver input. Assuming that the variation of I_k between two consecutive time frames is sufficiently small, the following approximation holds:

$$I_k(n) \approx I_k(n-1) \tag{6.15}$$

Thus, using the measured received signal at base station k, the mathematical programming problem in Figure 5.13 on page 89 can be written as in Figure 6.4. The first constraint in this figure results from combining the following constraints:

$$0 \leq p_i \leq P_{i,\max} \tag{6.16}$$

$$R_{i,\min} \leq \frac{w_i g_{ik} p_i}{I_k} \leq R_{i,\max}. \tag{6.17}$$

This algorithm has the following advantages over the algorithm in Figure 5.13:

$$\underset{p}{\text{Maximize}} \quad \sum_{i=1}^{N}\sum_{k=1}^{M} \frac{\lambda_i w_i g_{ik}}{I_k} c_{ik} p_i - \lambda_h h$$

subject to

$$\begin{aligned}
\frac{I_k}{w_i g_{ik}} R_{i,\min} &\leq p_i \leq \min\left(P_{i,\max}, \frac{I_k}{w_i g_{ik}} R_{i,\max}\right) \\
\sum_{k=1}^{M} c_{ik} &= 1 \\
c_{ik} &= b_{ik} \\
h &= \sum_{i=1}^{N}\sum_{k=1}^{M} \left|b_{ik} - \bar{b_{ik}}\right| \\
h &\leq h_{\max}
\end{aligned}$$

Figure 6.4: Simplified MINLP algorithm

1. The required number of measurements (path gains and interference levels) by the base stations reduces from NM to $N + M$. Each base station measures the path gain of the users that it serves plus total interference at its receiver.

2. The information required at the RMC will be reduced by the same portion.

3. The number of active constraints is reduced by $2N$.

4. The objective function is converted from a sum of linear fractions to a simple quadratic form.

The performance of the simplified MINLP (S-MINLP) problem in Figure 6.4 in comparison with the I-MINLP is highly affected by the variation of I_k over time. One effective factor in this variation is the standard deviation σ of the shadowing.

When σ is low, I_k is closer to its previous value I_k^- with a higher probability and vise versa. When $\sigma = 0$ (no shadowing), the distance from a base station is the only effective factor in determining the path gain. Variation in the path gain due to the distance is very limited and negligible during a time frame and so the same is true for I_k. Therefore, a high performance is expected. Conversely, with a high σ, the difference between I_k and I_k^- is larger with a higher probability. Figure 6.5 presents comparative throughputs and the relative throughput loss for S-MINLP with respect to I-MINLP. Regarding the numerical results and the fact that in an outdoor mobile environment shadow standard deviation is near 8 dB, S-MINLP can be employed only if it outperforms other algorithms, in particular, LSA. It is noteworthy that we have assumed independent propagation conditions in successive time frames. In practice, however, the existing correlation between the propagation medium for a single user in 2 successive time frames will affect the result favorably. To illustrate this fact in a more practical condition, we repeat the second simulation in Section 5.2.3 on page 95 to compare S-MINLP and I-MINLP algorithms. Figures 6.6-6.8 similarly illustrate 3 snapshots of the network. It can be observed that:

1. With S-MINLP, the number of users that are serviced in a time frame is larger.

2. S-MINLP has a lower performance with respect to I-MINLP (6% on average) but it outperforms LSA. Figure 6.9 compares network throughputs for the three algorithms in a plot similar to Figure 5.22.

3. There is no infeasible solution for S-MINLP throughout the simulation (200 frames).

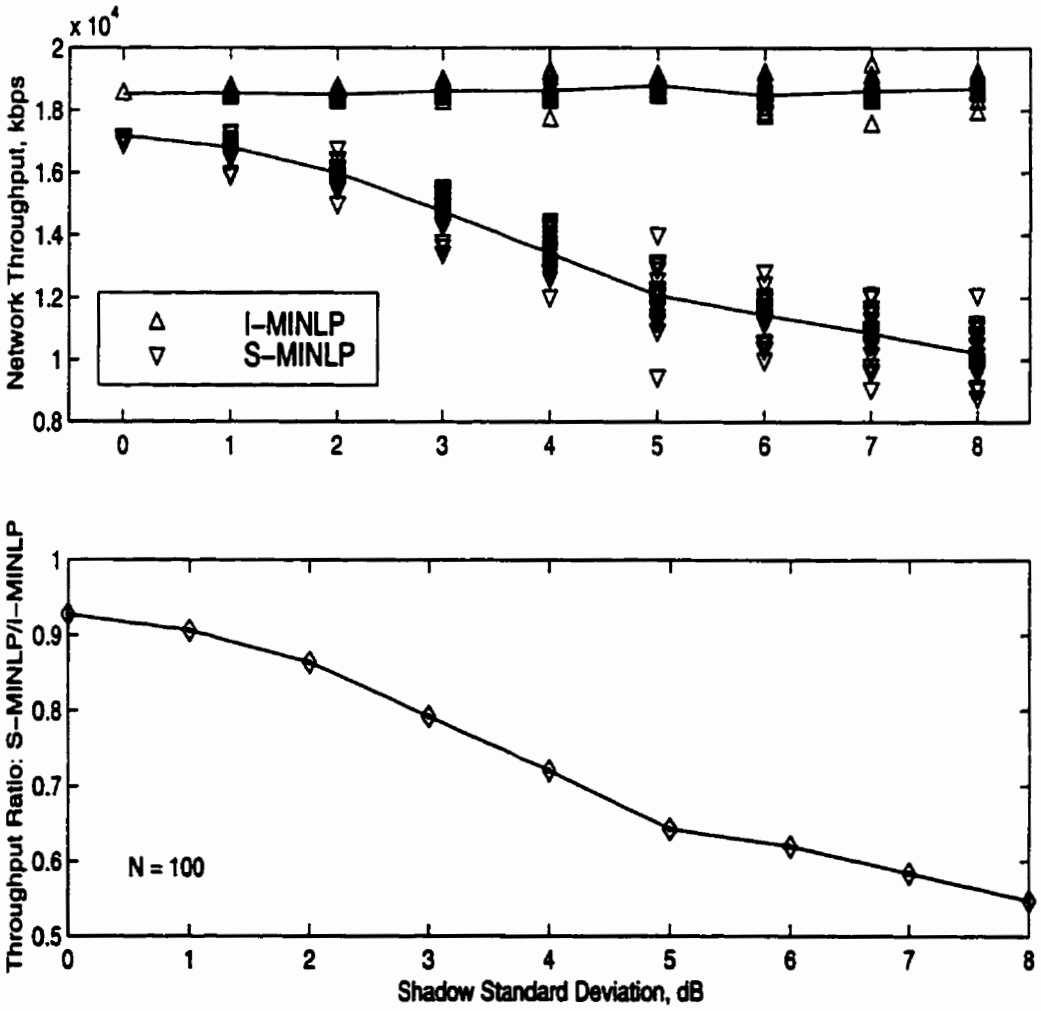


Figure 6.5: Relative Performance of I-MINLP and S-MINLP.

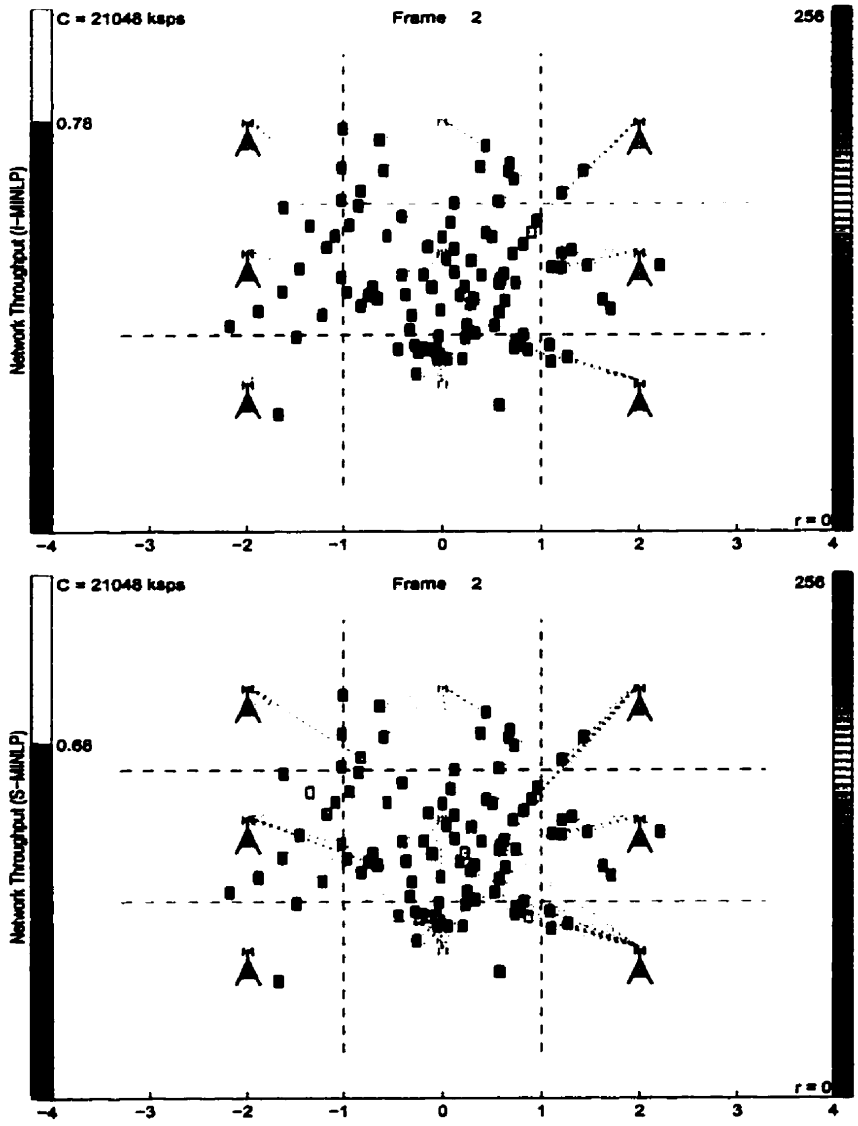


Figure 6.6: Comparison of I-MINLP and S-MINLP for unevenly distributed traffic (frame 2).

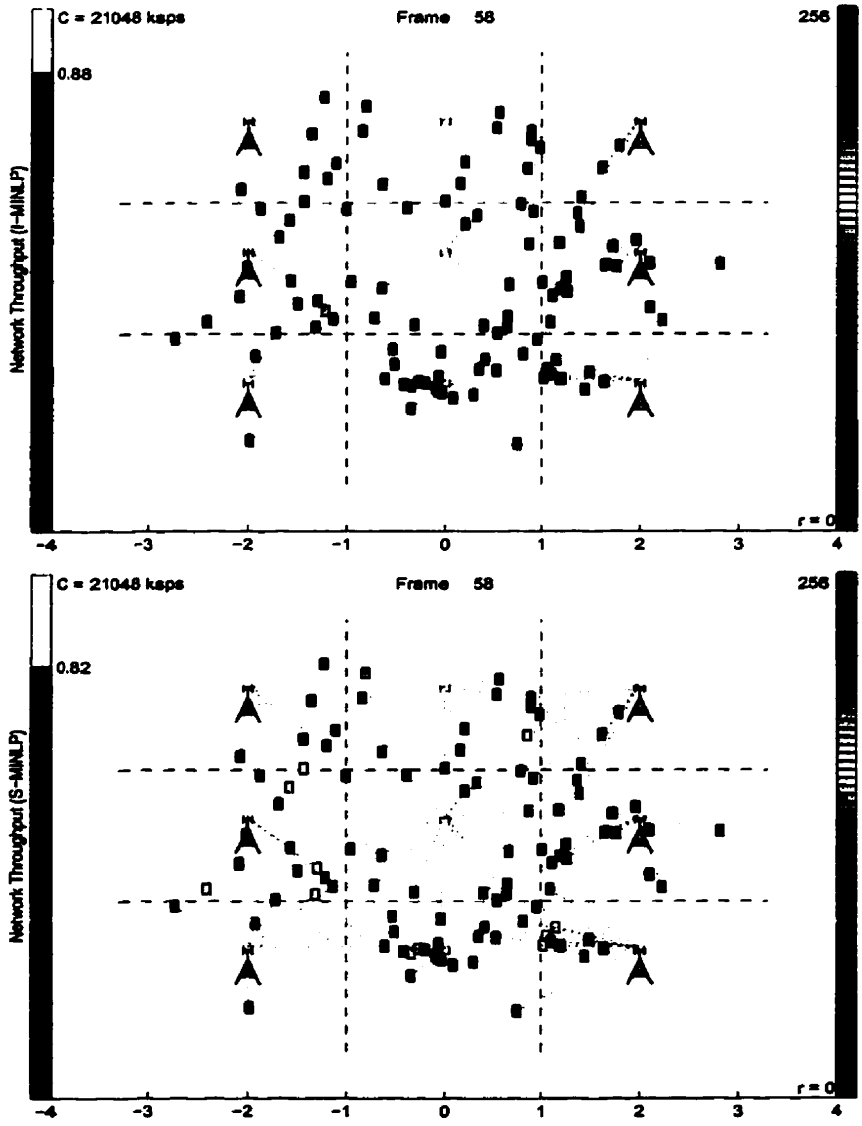


Figure 6.7: Comparison of I-MINLP and S-MINLP for unevenly distributed traffic (frame 58).

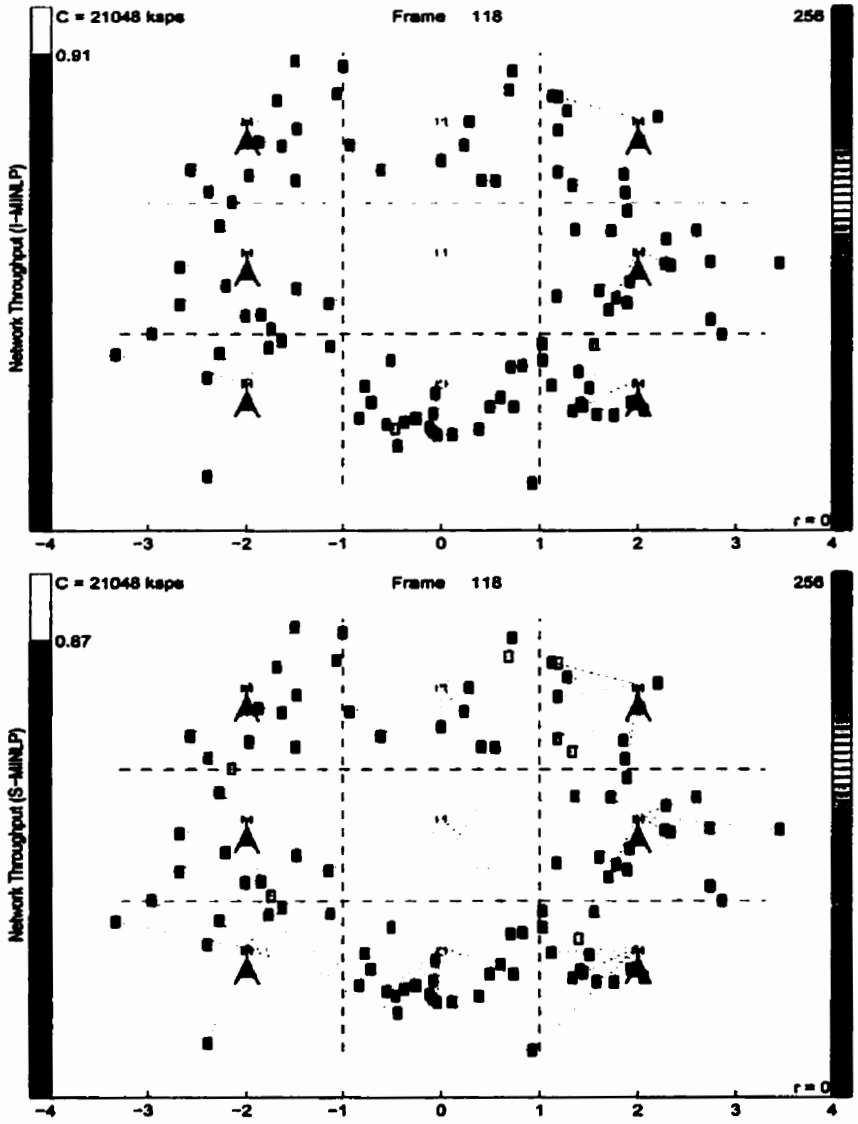


Figure 6.8: Comparison of I-MINLP and S-MINLP for unevenly distributed traffic (frame 118).

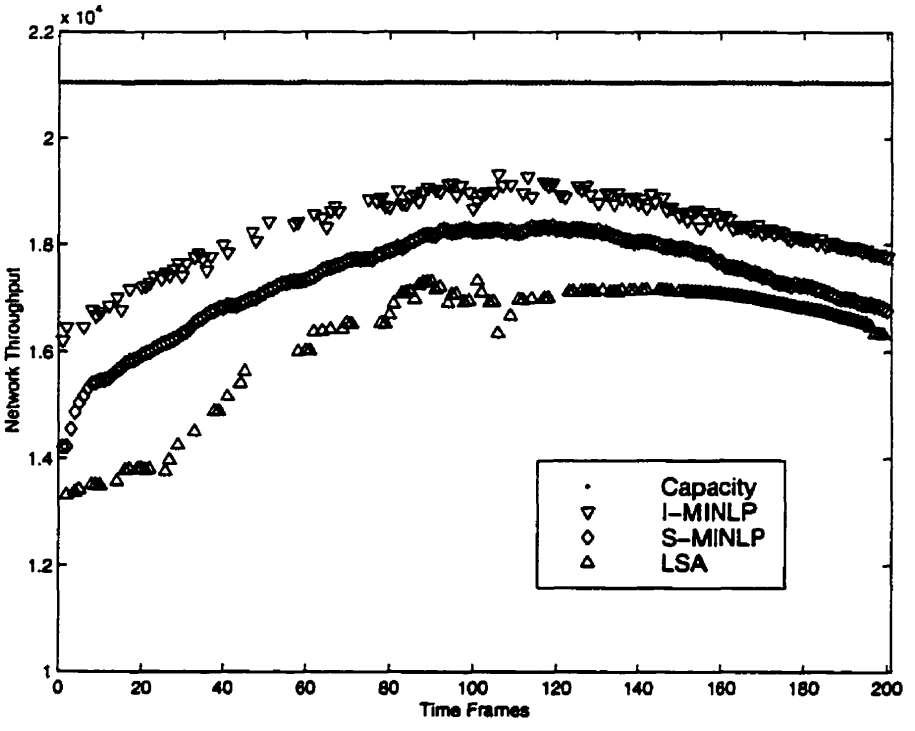


Figure 6.9: Network throughputs (unevenly distributed traffic).

Computational Complexity

Computation time is essential for implementation of the resource management algorithms. For both I-MINLP and S-MINLP, a large portion of the computations is performed by DICOPT which is called from within a GAMS program. Our main simulation program is written in MATLAB. To solve the mathematical programming problem, the GAMS program is called from within the MATLAB environment. For this purpose, we have used an interface program developed by Ferris in [71]. This structure adds the power of MATLAB to GAMS, thus providing a powerful tool for our simulation purposes. This combination, however, is subject to computation and time overhead due to exchange of data between the programs. In order to get an estimate of the computation time, one way is to add up the cpu-times in

different stages. This summation is not accurate enough as it excludes the overhead. As an over-estimate, we measure the elapsed computation time for different numbers of users in a network with 9 base stations. The simulation condition is similar to Table 5.4 except for the propagation condition and hansoff constraint which are the same as the condition in Table 5.2. The program runs on a SUN Ultra-10 workstation. Figure 6.10 illustrates the elapsed computation time to run

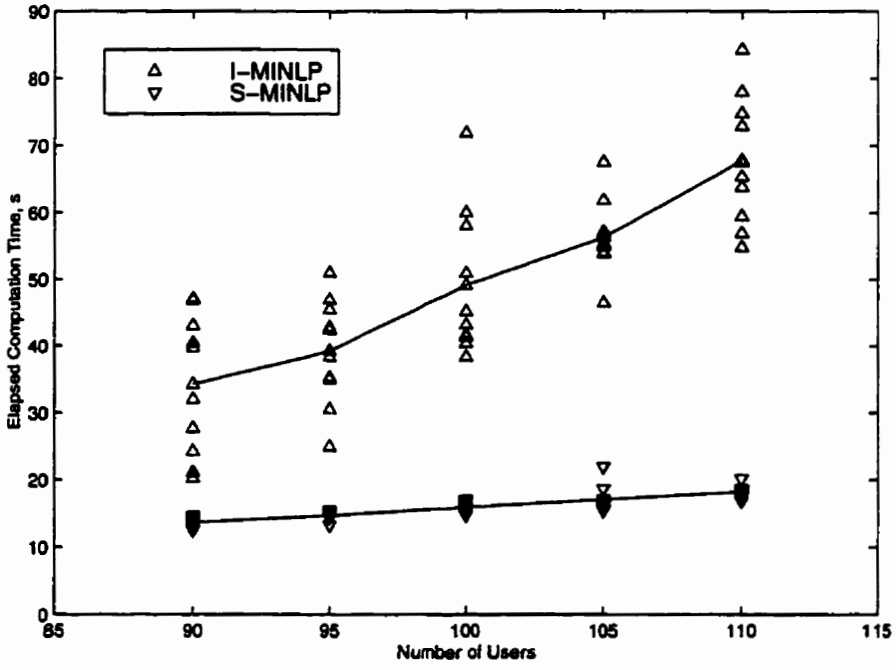


Figure 6.10: Elapsed computation time for the I-MINLP and S-MINLP algorithms.

I-MINLP and S-MINLP algorithms for populations of 90 to 110 users in the network. It can be seen that the time increases almost linearly with respect to N. The elapsed computation time is measured while running the program on a multi-user workstation with shared resources in a local area network. Obviously, the computation time on a dedicated powerful machine at the resource management center will

save time considerably. Moreover, if the program code is optimized, the overhead time to transfer data between subprograms will be reduced significantly.

6.3 Decentralization

Resource management for the uplink of a wireless system can be performed in the network level (centralized), sub-network level (partially decentralized), base-station level (decentralized), and user level (fully decentralized). Figure 6.11 visualizes the decentralization at different levels.

6.3.1 Centralized Resource Management

The outcome of our research provides detailed models and algorithmic solutions for network-level resource management. Our centralized algorithms do not put a burden of overhead information on the wireless part except allocation of the rate, power, and base station assignment to each user. Besides, with today's high speed broadband networks, the overhead information flow between base stations and the RMC can easily be taken care of. The solutions not only provide ideal performance limits, benchmarks and capacity bounds for evaluation purposes but also can be considered as viable candidates for resource management in future systems. Ways of implementing these algorithms in the framework of the IMT-2000 proposals are discussed at the end of this chapter.

6.3.2 Partially Decentralized Resource Management

The partially decentralized implementation is very similar to the centralized algorithm except for receiving additional interference from other sub-networks in the

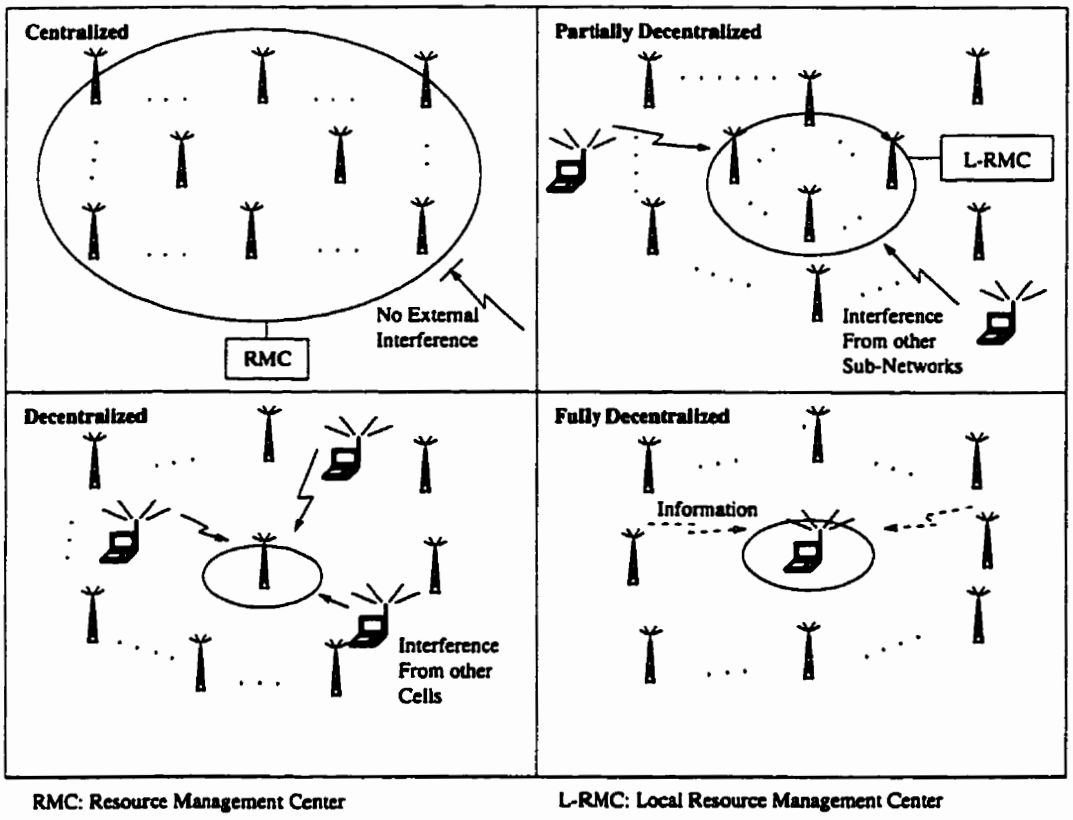


Figure 6.11: Decentralization Levels.

area. We use the terms intra-network and inter-network interference to refer to the interference from within the sub-network and from other sub-networks, respectively. With a slight modification, previously developed mathematical programming problems and corresponding algorithms can be extended for a partially decentralized resource management. Let f be the *interference ratio* defined as the ratio of the inter-network to the intra-network interference. The mathematical programming that can model resource management in the system will be similar to those in Figures 5.13 and 6.4 with the following modifications:

1. M, N and other parameters refer to the corresponding values in the sub-network.

2. The interference term $\sum_{j=1,j\neq i}^{N} g_{jk}p_j + \eta$ or I_k will be multiplied by a factor $(1+f)$, assuming that the background noise is negligible with respect to the interference term.

A thorough investigation is performed on the interference ratio f in [6, 76, 74]. The propagation condition in the study is similar to that of our simulations as described in Table 5.2. It has been shown both analytically and by simulation that the interference factor at a base station varies between 1/2 and 1/3 by users who are power controlled by other base stations. Under a similar condition, the interference ratio for a sub-network is less than that of a single base station because part of the interference has been already included in the intra-network interference. In a more sophisticated design, the size and user population of sub-networks can be dynamically changed according to the instantaneous state of the network. Further investigation on evaluation of the interference ratio and ways of dividing the network into sub-networks is part of the proposed future work. The resource allocation in a partially decentralized network will be suboptimal and less efficient than in a centralized network. The advantages would be less information flow and computational complexity as well as higher practicability when the network size is very large.

6.3.3 Decentralized Resource Management

When resource management is performed at the base station level, the assignment of mobiles automatically is excluded from the mathematical programming problem and should be decided independently, thus resulting in a suboptimal solution.

For the user assignments, the most popular scheme is the LSA assignment. If we apply LSA assignments to every user, decide and allocate power and rate values locally at each base station, the problem to solve will be similar to a single-cell system which has already been efficiently solved. The only difference is an additional $(1 + f)$ factor to account for the external interference at the base station. Extensive research on decentralized resource management schemes has been carried out. In comparison with other recent works which maximize the throughput using decentralized schemes, e.g. [80], our algorithm solves the problem using linear programming methods which are very efficient and provide unique global solutions. A different objective function from throughput maximization is suggested in the literature to minimize the total allocated power to the users. This objective is intended to minimize the interference to other cells, however, it eventually reduces the system throughput, as well. In fact, in the SIR constraint given in (4.4), when we equate the constraint, it is implied that a certain rate is achieved with the minimum required power, or that with a certain allocated power the rate is maximum. Since we have used the SIR constraint with its equality sign, the throughput is maximized with the minimum required power. If there is a need to control the total allocated power, one can set a constraint as $\sum_i p_i \leq P_{sum}$ where P_{sum} can be selected such that the resource utilization by the users in a particular cell is balanced with its neighboring cell. Thus, over-utilization of resources in a particular cell can be prevented in favor of other cells.

6.3.4 Fully Decentralized Resource Management

In a fully decentralized resource utilization scheme, a user collects the required data from neighboring base stations, performs necessary computations, and decides on its power and rate. An example of such an algorithm is given in [39]. According to

this proposed fully decentralized algorithm, total interference at each base station is collected from the broadcast data by each base station and path gains to different base stations are measured by using the pilot signals in the forward links. Then, the required transmit power to all base stations is computed and the one with the minimum required power is selected for its communications. A single service with a fixed rate is assumed in this algorithm. In a multi-service environment, selection of both rate and power is not an easy task unless everybody is forced to transmit at their minimum rate, resulting in a low throughput efficiency. Another disadvantage is a higher complexity in the mobile terminal.

6.4 Compatibility with IMT-2000

The recent proposals for third generation mobile systems under the IMT-2000 project have many common features with our system model. This coincidence provides a ground for compatibility of our resource management algorithms with the next generation wireless systems. In what follows, we discuss a way of implementing our algorithms in the proposed systems.

Consensus in the definition of a standard and its system design aspects has been very rare in the communications field in the past. The recent proposals for IMT-2000, e.g. [10, 11, 12], are good steps in the direction of globalization of the third-generation wireless standard. These proposals are similar in many details of the system-level design, however, differences still exist mostly in the adaptability with second-generation systems. As far as reverse-link resource management is concerned, there is no basic difference between the second generation standard IS-95 and the new proposals. However, as there are a number of undecided and unfinalized system design issues, it is very likely that the resource management section will

be modified soon. This fact is explicitly mentioned in the European proposal [11] that "optimal resource management" is to be considered in the next version of their proposal. So far, the proposals have considered decentralized resource management. Users are assigned to the base stations based on the LSA assignment method. Power control is maintained in both open-loop and closed-loop forms. Open-loop power control (OLPC) is used on channels that cannot employ closed-loop power control (CLPC), e.g., on the random access channel (RACH), or at the beginning of a CLPC cycle. OLPC is performed as follows:

1. The receiver measures the path gain over a certain time period using the pilot signal in the PERCH (pilot) channel and takes the average.
2. The transmitted power is calculated based on the measured path gain.

CLPC's basic functions are:

1. The received SIR at the base station is measured periodically.
2. If the measured SIR is less than the target value, a "1" bit is sent to the user. Otherwise, a "0" is transmitted.
3. The user terminal will decrease/increase its output power by 1 dB upon receiving a "0"/"1" power control bit. Both the target SIR and the 1 dB step size can vary dynamically during the operation. An outer-loop power control is considered to adjust the SIR threshold level.
4. When a power control bit is not received, the power level remains constant.

Variable-rate transmission is performed using a few specified rates and multi-code transmission. Decisions on the transmission rate are made based on the negotiations with users.

The frame structure is composed of slots (0.625 μs), frames (10 ms), and super-frames (720 ms). Every 16 slots are multiplexed to form a frame and a super-frame is made of 72 frames. Rate variations and open-loop power control are updated in every frame. Closed-loop power control is performed on a slot-based period. The proposed frame structures of the forward and reverse links for third generation wireless systems [10, 11] are illustrated in Figures B.1 and B.2 in Appendix B.

Potential application of this research in IMT-2000 - A combination of our resource management algorithms with the CLPC can be applied to modify and improve the FDD-mode reverse-link resource management in IMT-2000 proposals. In this combination, OLPC is excluded and any of our centralized, partially decentralized, and decentralized algorithms can be employed, however, the latter version being the most compatible. The basic operation of the proposed resource management for IMT-2000 using a decentralized algorithm is as follows. It is assumed that at the beginning of frame n, user i is transmitting at power p_i and rate r_i, and assigned to base station k.

1. The user rate and assignment remains unchanged during the frame.

2. Upon receiving the power control bit in slot 1 of frame n, the user's output power at slot 2 is changed by 1 dB. This power adjustment will continue until the last slot (slot 16) of frame n.

3. The path gain measurement is performed in each slot and averaged over the frame. This can be done in two ways:

   - by the user using the information in the forward link, similar to [10, 11, 12],
   - by the base station using the information in the reverse link, as proposed in this thesis.

In the former case, the path gain information should be made available to the base station through control channels. In the latter case, the output power is known based on the initial power at the first slot and all previous power control bits.

4. The new assignment is found at frame $n+1$ using the measured path gains to the nearby base stations.

5. The allocated power and rate at frame $n+1$, slot one is provided by the base station and sent to the user via forward-link control channels.

Figure 6.12 illustrates the frame structure of the forward link dedicated physical channel (DPCH) for the proposed resource management scheme.

The centralized algorithm can also be employed in the above combination. In this case, assignments will be the outcome of the algorithm rather than being selected independently.

6.5 Summary

In this chapter, we studied some practical aspects of resource management. Regarding the path gain measurement error, it became evident that for a limited number of rate levels, there will be a good error tolerance. Quantitative data to support this feature has been provided. A less complex version of the I-MINLP algorithm (S-MINLP) has been developed to facilitate a centralized implementation of optimal resource management. It has been shown that S-MINLP outperforms LSA and, with respect to I-MINLP, runs faster with lower infeasibility at a cost of 6% less throughput. The other important issue is the way the algorithms are

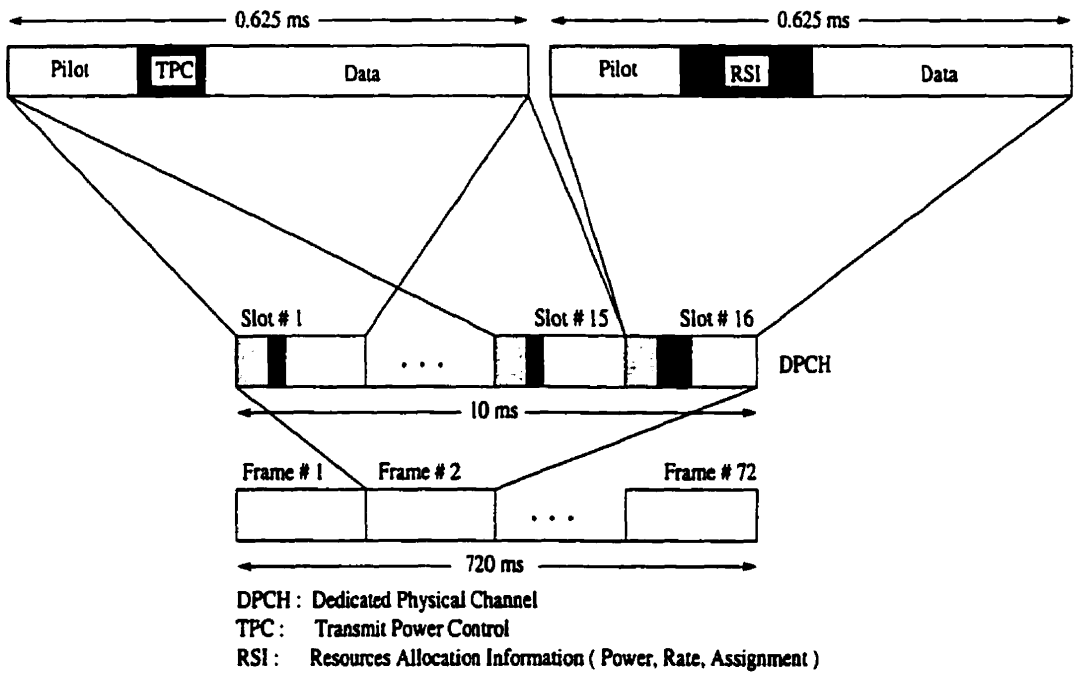


Figure 6.12: Forward link DPCH frame structure for the proposed resource management scheme.

implemented. Among 4 different schemes, centralized and partially decentralized algorithms are the favorable ones with a high throughput efficiency. A combination of our algorithms and the CLPC is suggested to improve resource management of the IMT-2000 proposals.

Chapter 7

Conclusions and Future Work

This research is aimed at developing algorithms for optimal resource management in a wireless multimedia WCDMA system to maximize the network resource utilization and provide satisfactory QoS for each user. Resource management is modeled as a mathematical programming problem based on the objective and the proposed system model. This problem is analyzed, restructured, and solved for single- and multi-cell systems. Resource management algorithms are developed to adopt solution procedures.

This chapter summarizes the results and concludes this research. An outline for the proposed future work follows at the end.

- **Wireless Multimedia CDMA Model**

  A comprehensive model for wireless multimedia CDMA is the first requirement for our research. In the literature, several models have been proposed which lack the generality feature in the sense that they neither include all service classes, as defined in this thesis, nor can they be applied to a general

multi-cell environment. We have developed an inclusive model as presented in Chapter 3. The key features of this model are: supporting a wide range of delay-tolerant and real-time services, employing power and rate control to accommodate a wide range of QoS's, and combining resource management and base station assignment. The mathematical description of this model involves a set of control variables including mobile transmitted powers, data rates, base station assignments, and the number of handoffs. To solve the problem, two sets of input data are required: user-supplied data and network-supplied data. The user-supplied data includes: BER requirement, delay bound, maximum rate, maximum power, source rate for VBR class I services, and the size of the data for class II file-transmission applications. The network-supplied parameters are: maximum number of handoffs, and path gains. The minimum-rate limits are also required for solving the problem. We have derived mathematical expressions for the minimum-rate limits for different applications, as reported in Section 3.5.2.

- **Optimal Resource Management Problem**

  Using the proposed system model, we have developed a mathematical programming problem, i.e., an objective function and a set of constraints, based on our research objective as presented in Section 4.1. We have adopted a wireless service provider's perspective to materialize our objective. A pricing scheme has been developed to map the network throughput onto the network revenue, and handoff switching overhead onto a certain cost. This pricing scheme is a novel approach where every user pays proportional to the amount of consumed resources, which in turn is related to its QoS requirement and air-time. The structure of the mathematical programming problem is primarily nonlinear with a non-convex feasible space. Since such problems are difficult

to solve both for global and local optima [58, 77], we have attempted to find equivalent problems with certain preferences in their structures, as discussed in Chapter 4. We have developed two alternative models for our problem. The first one has linear constraints with a linear multi-fractional objective function, as presented in Figure 4.2. The second alternative model is presented in Appendix A.4 and provides a less complex nonlinear objective function with linear constraints. Further investigation has proved that the first alternative is a better choice and leads to interesting solutions. Despite the fact that power and rate variables in the mathematical programming problem have a discrete nature, we have shown that we can treat them conveniently as continuous variables and, thus, save a significant amount of computational cost. Then, in practice we can allocate close discrete values to the users. It has also been shown that if the number of discrete levels exceeds 64 and 128 for power and rate variables, respectively, the throughput loss will be negligible.

- **Single-Cell Solution**

  In Chapter 5, we have proved mathematically that there exists an equivalent linear programming problem for the resource management model in Figure 4.2 for a populated single cell. The solution of this problem is a global optimum and can be found efficiently and fast. The single-cell algorithm is developed for the resource management in a single cell. For a set of target SIR per bits, there is a capacity limit beyond which is not achievable. This limit defines a new measure for the network capacity in terms of the throughput weighted by the service qualities. We believe that this measure of the system capacity is more suitable for multimedia applications than counting the number of users in each service. The latter one, adopted in the literature e.g. in [80], is not a real indication of the capacity if the network is supposed to support

services as defined in our system model. This capacity is a benchmark and an upper bound for performance evaluation of any network in terms of its throughput per cell. The capacity of a multi-cell network must be multiplied by soft-handoff gain [81] if soft handoff is employed.

- **Multi-Cell Solution**

  A general solution for the multi-cell environment requires solving a nonlinear subproblem for a large set of feasible assignments. This process has been facilitated by developing an equivalent linear problem for the nonlinear subproblem and reducing the size of the feasible assignments set by using SIR and handoff constraints. We have shown that this approach is not viable when the network size gets large. For decentralized resource management, where base station assignment is decided independently and only the nonlinear subproblem remains to be solved, Theorem 1 on page 78 is very helpful to find a unique globally optimum solution.

  In order to overcome the assignment problem and its large scale, the problem has been reformulated from a max-max structure to a single MINLP problem. Despite being very difficult in nature, we have been able to use an available tool to solve the MINLP problem. Minor changes have been necessary in the structure of the problem in using the tool. Simulation results suggest that the best throughput performance can be achieved when we select the LSA assignment as the starting point for the assignment variables (I-MINLP algorithm). In comparison with random initial assignment (MINLP algorithm), I-MINLP has gained a 14% throughput improvement. Overall, a throughput efficiency better than 94% is attained (reuse factor of 0.94) for a network with up to 9 base stations. The reuse factor will decrease as the number of base stations

increases. Furthermore, it has been shown that the computation time varies almost linearly with the number of users in the network.

- **Sensitivity Analysis**

  It is important to know how robust the developed algorithms are to the estimation error in path gain measurements. The result of our investigation shows that the more we utilize the network capacity, the more vulnerable the algorithm is to the estimation error. Therefore, with a reasonable margin, an acceptable robustness is achievable. If the number of discrete rate levels is not too high, e.g. less than 32, the margin will automatically be there (see Figure 4.3 on page 59).

- **Centralized versus Decentralized Algorithms**

  A centralized implementation attains maximum performance due to its global outlook in the solution. No matter how practical it is, the centralized algorithm provides an upper bound performance for the system and can be used as a benchmark for evaluation purposes. It has been argued, however, that contrary to the centralized resource allocation algorithms in conventional wireless systems, e.g. dynamic channel allocation, which are very difficult in practice to implement, WCDMA system does not face similar problems. For example, in centralized dynamic channel allocation algorithms, frequency management is a very complex task and users may have to use different radio channels as a requirement of the network-wide solution. In contrast, WCDMA system users use the same radio channel and PN codes throughout their calls. Nevertheless, a similar complexity may occur in assigning the users to the base stations in successive time frames, i.e., the so-called ping-pong effect, as described in Chapter 2 on page 25. This problem has already been taken care of by in-

cluding the handoff switching cost in the objective function and controlling its effect by the cost per handoff λ_h and maximum number of handoffs $h_{\max}$. The assignment of the rate and power only adds corresponding information flow from the RMC to base stations through broadband wired links and to users through dedicated control channels. For a more effective realization, a simplified version of the centralized algorithm (S-MINLP) has been developed with less computational complexity and information flow among base stations and the RMC at the cost of a lower achievable throughput. The comparative performance of S-MINLP and I-MINLP has shown that for the simulation condition in Table 5.4 on page 96, on average, the throughput loss is 10%.

Decentralization at the sub-network, base station, and user levels has also been discussed. With small modifications, the centralized algorithms can be extended to a lower level resource management. A good compromise among different schemes is the partially decentralized algorithm which favors the advantages of both centralized and decentarlized algorithms. It is flexible enough to range from a centralized algorithm when the sub-network consists of all existing base stations to a decentralized algorithm by having only one base station in its sub-network. It has been argued that a fully centralized scheme is not a good choice in wireless multimedia CDMA systems.

- **Compatibility with IMT-2000**

  Since there is no significant improvement in the reverse-link resource management of the current proposals for IMT-2000, we believe that the results of this research can be applied to the proposals for further upgrade and performance improvement. In this thesis, we have proposed a resource management scheme which is a combination of our developed algorithms and the

well known closed-loop power control. According to this scheme, our resource management algorithm allocates data rates, powers and base station assignments to all users based on a partially decentralized scheme at every 10-ms frame. During that frame, data rate and base station assignments remain unchanged. In order to capture fast channel variation, closed-loop power control is employed to update the power level by ± 1 dB in accordance with a "1" or "0" power control bit.

7.1 Concluding Remarks

This thesis presents the result of a research on modeling and optimal resource management in wireless multimedia WCDMA systems. In addition to the research objective, background material, and literature survey on the subject, a system model including a wide range of multimedia services is proposed. The problem of resource management in a multimedia environment is formulated as a mathematical programming problem which maximizes the profit gained by a wireless multimedia service provider subject to satisfying the service and QoS requirement for each user. The problem is solved with a great deal of effort for single- and multi-cell systems. The sensitivity of the solution to erroneous input data is examined. Different algorithms for the resource management are developed. The simulation results validate the viability of the algorithms in high resource utilization. Resource management operation at different operation levels are discussed and a partially decentralized scheme is introduced. A resource management scheme is suggested for the reverse-link FDD-mode resource management of IMT-2000 systems.

The main contributions of this work can be enumerated as:

1. Introduction of an inclusive model for wireless multimedia communications including the new concept of minimum-rate limit for delay control.
2. Introduction of a pricing scheme and a throughput and capacity measure for multimedia wireless services.
3. Solution of a complex large-scale nonlinear programming problem resulting in:
   - Developing an equivalent LP problem for a sum of linear fractions (extension of LFP theorem).
   - development of resource management algorithms for single- and multi-cell systems with high performance and low complexity,
   - derivation of conditions for robustness against path gain estimation errors.
4. Introduction of a flexible partially decentralized scheme for resource management.
5. Suggestion of a combined optimal resource management and CLPC for the reverse-link FDD-mode resource management in IMT-2000 systems.

This research, in part, has been published in [1, 2].

7.2 Future Work

The outline of the suggested future work is as follows:

- In the formulation of the partially decentralized resource management, the interference ratio f is an important factor. Despite availability of mathematical expressions and simulation results for the interference ratio in the IS-95 system, we need similar expressions and results for WCDMA multimedia applications. Derivation of the desired expressions is a requirement for partially decentralized implementation of the resource management algorithms.

- In this research, we have dealt with the base station assignment in the form of hard handoff. As a basic feature of CDMA systems, the algorithms should include soft handoff in the formulation of the resource management.

- It is desirable to extend resource management techniques to the forward link. The extension and combination of the resource management in both links can be another trend in the future work.

- The TDD mode in IMT-2000 proposals [10, 11] is proposed as a hybrid technique (CDMA/TDMA) where each WCDMA channel is split into 16 TDMA time slots. In this mode, the forward and reverse link use the same frequency band. A thorough study of this mode and associated problems involved in resource management is another interesting topic for the future work.

- Further work is required to optimize the programs developed for implementation of the resource management algorithms to minimize the computation time.

Appendix A

Optimization Theory and Techniques

Many problems of both practical and theoretical importance concern themselves with the choice of a "best" configuration or set of variables to achieve some goal. Mathematical programming employs a mathematical model to describe an optimization problem and includes three basic sets of elements: decision or *control variables*, *constraints*, and the *objective function*. The control variables are the unknowns which are to be determined from the solution of the model. The set of constraints limits the control variables to their *feasible* or permissible values. A common constraint is the so-called nonnegativity constraint which requires all control variables to be either zero or positive. The objective function specifies the goal in terms of the control variables. An *optimal solution* of the model is obtained if the control variables yield the best value of the objective function, while satisfying all the constraints.

Depending on the characteristics of the control variables, the objective function

and the constraints, most mathematical programming problems fall in one of the wide categories of linear/non-linear, integer/continuous-variable and single/multi-objective problems.

In what follows, several definitions and theorems of optimization theory and linear algebra are introduced for further use in the thesis. A brief description of the simplex method is provided, as well.

A.1 Definitions and Theorems

Definition 2 The *gradient* of a continuous function $f : \Re^n \rightarrow \Re$ at $x \in \Re^n$ is defined as

$$\nabla f(x) = \left[\frac{\partial f}{\partial x_1}(x_1), \cdots, \frac{\partial f}{\partial x_n}(x_n)\right]'. \tag{A.1}$$

The second-order partial derivative of f, the so-called *Hessian matrix*, is defined as

$$\begin{bmatrix} \frac{\partial^2 f}{\partial^2 x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial^2 x_n} \end{bmatrix}. \tag{A.2}$$

Definition 3 f is *convex* (*concave*) if, whenever $x, y \in \Re^n$, and $0 \leq \alpha \leq 1$,

$$f(\alpha x + (1-\alpha)y) \leq (\geq)\alpha f(x) + (1-\alpha)f(y). \tag{A.3}$$

This means, geometrically, that the line segment drawn between any two points on the graph of the function never lies below (above) the graph. A set $E \subset \Re^n$ is a *convex set* if $\alpha x + (1-\alpha)y \in E$ for any $x, y \in E$ and $0 \leq \alpha \leq 1$ [27]. A *convex*

programming problem is one which minimizes a convex function (or maximizes a concave function) over a convex constraint set [28].

A continuous function $f(x)$ is quasi-convex over a set S if either of the following equivalent statements holds [28]:

(a) $\{x|f(x) \leq c, x \in S\}$ is convex for all c.

(b) $x_1, x_2 \in S,\ f(x_2) < f(x_1) \Rightarrow f(\alpha x_2 + (1-\alpha)x_1) \leq f(x_1)$.

Definition 4 Let $A \in \Re^{n\times n}$ be an $n \times n$ symmetric matrix. A is said to be *positive (negative) definite* if $x'Ax > 0 (< 0)$ for every nonzero $x \in \Re^n$. A is *positive (negative) semidefinite* if $x'Ax \geq 0 (\leq)$ for all nonzero $x \in \Re^n$. A is *indefinite* if it is neither positive semidefinite nor negative semidefinite. All eigenvalues of a positive (negative) semidefinite matrix are nonnegative (nonpositive) real numbers [26].

Definition 5 The two mathematical programming problems

(I) *Maximize* $F(x)$ *subject to* $x \in \Delta$

(II) *Maximize* $G(x)$ *subject to* $x \in \Gamma$

where Δ and Γ are some feasible sets, are called *equivalent* if there is a one-to-one map $y(.)$ of the feasible set Δ of (I) onto the feasible set Γ of (II) such that $F(x) = G[y(x)]$ for each $x \in \Delta$. Thus, (I) is equivalent to (II) if and only if (II) is equivalent to (I) [27].

Definition 6 An *affine* function of $x \in \Re^N$ is a linear function of x plus a constant. If the objective function is the ratio of two affine functions of x and the constraints are linear, the problem is called *linear fractional programming.*

Definition 7 A mathematical programming problem has a *standard form* if [29]

1. All the constraints are equations except for the nonnegativity constraints which remain "$\geq$" inequalities.
2. The right hand side element of each constraint equation is nonnegative.
3. All the variables are nonnegative.
4. The objective function is of the maximization or the minimization type.

Theorem 2 Any local minimum of a convex programming problem is a global minimum [28].

Theorem 3 A *positive semidefinite quadratic form* is convex [28].

Theorem 4 The linear fractional programming

$$\underset{p}{\text{Maximize}} \left\{\frac{m'p}{g'p+\eta}\right\} \text{ subject to } p \geq 0,\ Ap = b,\ g'p+\eta > 0 \qquad \text{(A.4)}$$

has an equivalent linear program with one additional variable and constraint given as

$$\underset{y,u}{\text{Maximize}} \{m'y\} \text{ subject to } y \geq 0,\ u > 0,\ g'y+\eta u = 1,\ Ay - bu = 0 \qquad \text{(A.5)}$$

where p, y, and b belong to $\Re^N$, $A \in \Re^{N\times N}$, and $u \in \Re$, and it is assumed that no point $(y, 0)$ with $y \geq 0$ is feasible for (A.5) [27].

Proof: Let

$$u = \frac{1}{g'p + \eta} \tag{A.6}$$

$$y = up \tag{A.7}$$

then, $y \geq 0$, $u > 0$, and

$$Ay - ub = u(Ap - b) = 0 \tag{A.8}$$

$$g'y + u\eta = u(g'p + \eta) = 1 \tag{A.9}$$

Thus, the point (y, u) is feasible for (A.5). Conversely, if (y, u) is feasible for (A.5), and no point $(y, 0)$ is feasible for (A.5), then, $u > 0$, and $p = y/u$ satisfies $p \geq 0$, and $Ap - b = u^{-1}(Ay - ub) = 0$. Therefore, (A.6) and (A.7) map the feasible set of (A.4) one-to-one onto the feasible set (A.5). Moreover, the objective functions are related by

$$\frac{m'p}{g'p + \eta} = \frac{m'y}{g'y + \eta u} = m'y. \tag{A.10}$$

□

A.2 Simplex Method

Given a linear programming problem where constraints have "=", "$\leq$", and "$\geq$" signs, the optimal solution (if one exists) is obtained by the simplex method in the following steps [29]:

1. *Express the problem in the standard form*

   To convert inequality constraints to equalities, slack variables and artificial variables are used as described in the following two-variable examples. All

variables are nonnegative.

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 \leq b_1 \quad &\Longrightarrow \quad a_{11}x_1 + a_{12}x_2 + u_1 = b_1 \\ a_{21}x_1 + a_{22}x_2 = b_2 \quad &\Longrightarrow \quad a_{21}x_1 + a_{22}x_2 + v_2 = b_2 \\ a_{31}x_1 + a_{32}x_2 \geq b_3 \quad &\Longrightarrow \quad a_{31}x_1 + a_{32}x_2 + v_3 - u_3 = b_3 \end{aligned}$$

where b_i, $i \in \{1, 2, 3\}$ is nonnegative, u_i and v_i are the slack and artificial variables, respectively.

2. *Select a starting basic feasible solution*

   A system of n linear equations and m variables (unknowns), where $m > n$, has an infinite number of solutions. A feasible solution can be obtained by setting any $m - n$ unknowns to zero and solving for the remaining n unknowns. In the simplex method, there is a selective iterative procedure which will yield the optimum solution in a finite number of iterations. The variables set to zero are called *non-basic variables* and the rest are called *basic variables.*

   Determination of a starting feasible solution involves two cases: for inequalities with $\leq$, the slack variables are basic variables and all other variables are set to zero; in the constraints with $=$ or $\geq$, artificial variables are considered as the basic variables.

3. *Generate new basic feasible solutions using the optimality and the feasibility conditions*

   In this step, slack and artificial variables are replaced such that the new basic variable both improves the objective function and satisfies the constraints.

The feasibility condition of the simplex method, called the first phase of the simplex method, is used in nonlinear problems with linear constraints for the feasibility test.

A.3 Proof of Corollary 5

For the feasibility condition of a system of 2 users and 2 base stations, we use the first phase of the simplex method. Using the rate constraint in Figure 4.2, the constraints of this system are

$$R_{1,\min} \leq \frac{w_1 g^l_{1a_1} p_1}{g^l_{2a_1} p_2 + \eta} \leq R_{1,\max} \tag{A.11}$$

$$R_{2,\min} \leq \frac{w_2 g^l_{2a_2} p_2}{g^l_{1a_2} p_1 + \eta} \leq R_{2,\max} \tag{A.12}$$

$$p_1 \leq P_{1,\max}, \quad p_2 \leq P_{2,\max} \tag{A.13}$$

which set six inequalities. To perform the simplex feasibility test, we need to express the constraints in the standard form, as defined in Appendix A.1 . That is, to alter the above inequalities into equalities. For this purpose, we add the slack variable u_i and the artificial variable v_i to the inequalities [29]. Thus, the constraints become

$$\frac{w_1}{R_{1,\max}} g^l_{1a_1} p_1 - g^l_{2a_1} p_2 + u_1 = \eta \tag{A.14}$$

$$\frac{w_2}{R_{2,\max}} g^l_{2a_2} p_2 - g^l_{1a_2} p_1 + u_2 = \eta \tag{A.15}$$

$$\frac{w_1}{R_{1,\min}} g^l_{1a_1} p_1 - g^l_{2a_1} p_2 + v_3 - u_3 = \eta \tag{A.16}$$

$$\frac{w_2}{R_{2,\min}} g^l_{2a_2} p_2 - g^l_{1a_2} p_1 + v_4 - u_4 = \eta \tag{A.17}$$

$$p_1 - P_{1,\max} + u_5 = 0 \tag{A.18}$$

$$p_2 - P_{2,\max} + u_6 = 0 \tag{A.19}$$

We solve this system of linear equations symbolically for p_1, p_2, and different sets of four Slack and artificial variables. The desired feasibility condition is derived by applying the nonnegativity property of p_1 and p_2 to the solution of the linear system with the entering variables p_1, p_2, v_3, v_4, u_5 and u_6, as given in (5.23) on page 82. □

A.4 Equivalent Mathematical Programming

Another attempt to linearize the constraints with a simpler objective function is proposed as follows. Let

$$p_i = e_i r_i, \tag{A.20}$$

where $e_i > 0$, then, the constraint in (4.5) can be written as (time indices are omitted)

$$w_i g^l_{ia_i} e_i - \sum_{j=1, j\neq i}^{N} g^l_{ja_i} p_j = \eta. \tag{A.21}$$

Thus, the objective function becomes

$$\underset{a^l}{\text{Maximize}} \quad \left\{ \max_{p,e,h} \sum_{i=1}^{N} \lambda_i \frac{p_i}{e_i} - \lambda_h h \right\}, \quad e_i \neq 0 \tag{A.22}$$

subject to

$$0 \leq p_i \leq P_{i,\max} \tag{A.23}$$

$$R_{i,\min} e_i \leq p_i \leq R_{i,\max} e_i \tag{A.24}$$

$$h \leq h_{\max} \tag{A.25}$$

$$w_i g^l_{ia_i} e_i - \sum_{j=1, j\neq i}^{N} g^l_{ja_i} p_j = \eta. \tag{A.26}$$

Our new mathematical programming has a nonlinear objective function and linear constraints with both equality and inequality signs. The Hessian matrix of the nonlinear term in the objective function, i.e., $\sum_{i=1}^{N} \lambda_i \frac{p_i}{e_i} = \sum_{i=1}^{N} \lambda_i \frac{x_i}{x_{n+i}}$ where $x = [p|e]$, can be shown to be

$$H = \begin{bmatrix} \emptyset & H_1 \\ H_1 & H_2 \end{bmatrix} \tag{A.27}$$

where partitions are $n \times n$, $\emptyset$ is an all zero partition, and H_1 and H_2 are diagonal with diagonal elements $-\lambda_i/x_{n+i}^2$ and $2\lambda_i x_i/x_{n+i}^3$, respectively. Matrix H has a sparse structure with diagonal partitions which can be exploited to reduce the computational complexity.

A.5 Mixed Integer Nonlinear Programming

MINLP problems, in general, consist of both continuous and integer variables which appear in nonlinear terms in the objective function and constraints. A particular class of MINLP problems has been appeared in many branches of engineering and extensive effort has been exerted for their solution. In this class of MINLP, integer variables are binary and in linear terms, and nonlinearities are confined in continuous variables in separate terms. There are many problem-specific methods to solve this class of MINLP problems, however, general-purpose algorithms also exist. Some of the important algorithms are branch and bound [70] and the outer approximation/equality relaxation method [66, 67]. The latter method has been developed as a solver called DICOPT for GAMS modeling language. In the following more information on GAMS/DICOPT MINLP solver is provided.

GAMS/DICOPT Algorithm

DICOPT is a program for solving MINLP problems with linear binary and continuous nonlinear variables. It has been developed by Dr. Grossmann's group at the Engineering Design Research Center (EDRC) at Carnegie Mellon University. DICOPT is designed based on the outer approximation for the equality relaxation strategy. It solves a number of NLP (nonlinear programming) and MILP (mixed-integer linear programming) subproblems. Any NLP and MILP solver can be used

for solving these subproblems. The DICOPT algorithm has some provisions to find global optimum, however, it is not guaranteed.

The GAMS/DICOPT algorithm has been designed based on the following two objectives:

1. To use the existing modeling language GAMS.
2. To employ existing NLP and MILP solvers so that any improvement in any of the solvers is available to MINLP problems.

DICOPT typically solves the following type of problems:

$$\text{Maximize/Minimize} \quad f(x) + c'y \tag{A.28}$$

subject to

$$g(x) + Hy \,(\leq, =, \geq)\, b \tag{A.29}$$

$$l \leq x \leq u \tag{A.30}$$

$$y \in \{0,1\} \tag{A.31}$$

The vector of binary variables y appears only in linear terms while the vector of continuous variables x can be in nonlinear terms in the objective function and constraints.

The DICOPT algorithm has been developed based on the following techniques [69]:

1. **Outer Approximation** refers to the fact that the surface described by a convex function lies above the tangent hyper-plane at any interior point of the surface. In the DICOPT algorithm outer approximations are obtained

by generating linear approximations of nonlinear convex functions at each iteration. These estimations should underestimate the objective function and overestimate the feasible region.

2. **Equality Relaxation** refers to relaxing an equality constraint to be an inequality constraint. This property is used in the MILP master problem to accumulate linear approximations.

3. **Augmented Penalty** refers to the introduction of non-negative slack variables on the right hand sides of the inequalities obtained by the equality relaxation technique. It also refers to the modification of the objective function in non-convex cases.

The DICOPT algorithm starts by solving the relaxed MINLP (RMINLP) problem, i.e., the binary variables are treated as continuous variables having values between 0 and 1. If the solution yields binary values for the corresponding variables, it is considered as the optimum solution and the algorithm stops. Otherwise, it continues with an alternating sequence of NLP subproblems and MILP master problems. The NLP subproblems are solved for fixed binary variables that are predicted by the MILP master problem at each iteration. The algorithm by default stops when NLP subproblem starts worsening. The algorithm for the given problem

$$\text{Minimize } c'y + f(x) \tag{A.32}$$

subject to

$$Ay + h(x) = 0 \tag{A.33}$$

$$By + g(x) \leq 0 \tag{A.34}$$

$$l \leq x \leq u, \quad y \in \{0,1\} \tag{A.35}$$

works as follows:

1. Solve the problem as an RMINLP. If $y^0 = y$ is binary, stop. Binary solution is found. Otherwise go to step 2. (The superscript denotes the sequence of intermediate solutions.)

2. Find a binary point y^1 with an MILP master problem that features an augmented penalty function to find the minimum over the convex hull determined by the half-spaces at the solution (x^0, y^0).

3. Fix the binary variables $y = y^1$ and solve the resulting NLP. The new solution would be (x^1, y^1).

4. Find a binary solution y^2 with an MILP master problem that corresponds to the minimization over the intersection of the convex hulls described by the half-spaces of KKT points[1] at y^0 and y^1.

5. Repeat step 3 and 4 until the NLP subproblem starts worsening (increasing). Repeating step 4 means augmenting the set over which the minimization is performed with additional linearizations at the new KKT point.

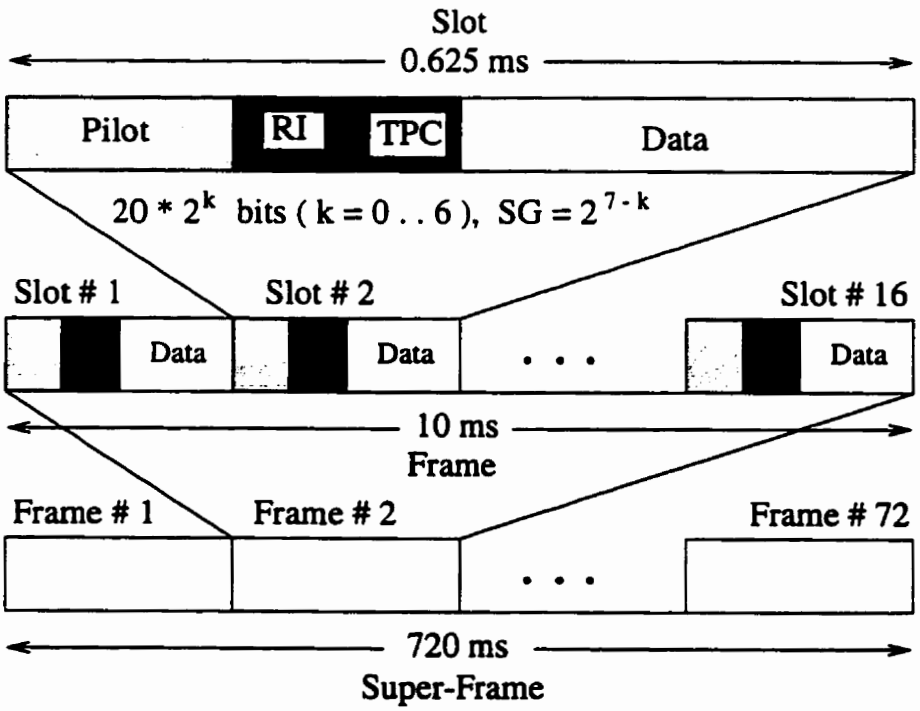
[1]A feasible point that satisfies the Karush-Kuhn-Tucker conditions. This condition is the Lagrange Multiplier Rule, which Kuhn and Tucker published as an extension to allow inequality constraints.

Appendix B

IMT-2000 Frame Structures

The dedicated physical channels (DPCH) in the forward and reverse link are used to carry user and control information between the network and users. These channels correspond to two category of channels as dedicated physical data channels (DPDCH) and dedicated physical control channels (DPCCH) [10]. In the following, the frame structure of these channels is plotted in order to provide the necessary background for the proposed resource management algorithm for the IMT-2000 standard.

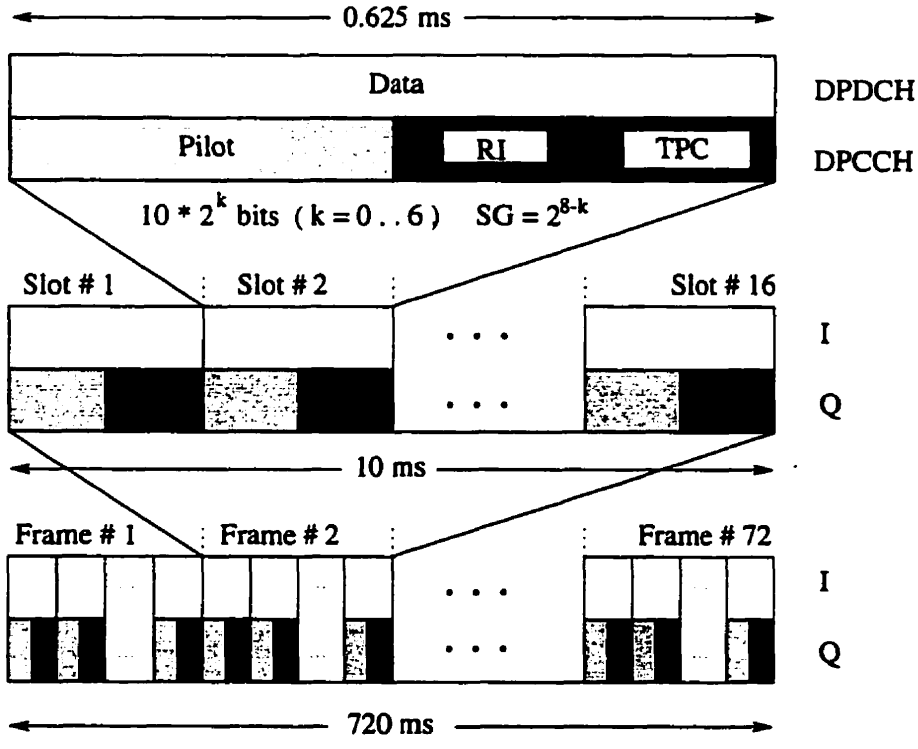
B.1 Forward Link



TPC : Transmit Power Control
RI : Rate Information
SG : Spreading Gain

Figure B.1: Forward link DPCH frame structure in IMT-2000 proposals.

B.2 Reverse Link



DPDCH : Dedicated Physical Data Channel (I - channel)
DPCCH : Dedicated Physical Control Channel (Q - channel)

TPC : Transmit Power Control
RI : Rate Information
SG : Spreading Gain

Figure B.2: Reverse link DPDCH and DPCCH frame structure in IMT-2000 proposals.

Bibliography

[1] M. Soleimanipour, W. Zhuang, and G. H. Freeman, "Modeling and Resource Allocation in Wireless Multimedia CDMA Systems," *Proceedings IEEE VTS 48th Vehicular Technology Conference,* pp. 1279-1283, Ottawa, May 1998.

[2] M. Soleimanipour, G. H. Freeman, and W. Zhuang, "An Algorithm for Maximal Resource Utilization in Wireless Multimedia CDMA Communications," *Proceedings IEEE VTS 48th Vehicular Technology Conference,* pp. 2594-2598, Ottawa, May 1998.

[3] D. C. Cox, "Wireless Personal Communications: A Perspective," *The Mobile Communications Handbook,* Edited by J. D. Gibson, CRC Press Inc., 1996.

[4] R. O. Onvural, *Asynchronous Transfer Mode Networks: Performance Issues,* Artech House Inc., 1994.

[5] W. Honcharenko, J. P. Kruys, D. Y. Lee, and N. J. Shah, "Broadband Wireless Access," *IEEE Communications Magazine,* vol. 35, no. 1, pp. 20-27, January 1997.

[6] K. S. Gilhousen, I. M. Jacobs, R. Padovani, A. J. Viterbi, L. A. Weaver and C. E. Wheatly, "On the Capacity of a Cellular CDMA System," *IEEE Transactions on Vehicular Technology,* vol. 40, no. 2, pp. 303-312, May 1991.

[7] W. C. Y. Lee, *Mobile Communications Design Principals*, John Wiley & Sons, Second Edition, 1993.

[8] M. Soleimanipour, *A Study of Mobile Cellular Systems Capacity with Emphasis on Cellular CDMA Systems*, M.A.Sc. thesis, Dept. of Electrical and Computer Engineering, University of Waterloo, Canada, 1995.

[9] W. C. Y. Lee, "Estimate of Channel Capacity in Rayleigh Fading Environment," *IEEE Transactions on Vehicular Technology*, vol. 39, no. 3, pp. 187-189, August 1990.

[10] ARIB IMT-2000 Study Committee, *Japan's Proposal for Candidate Radio Transmission Technology on IMT-2000: W-CDMA*, Association of Radio Industries and Businesses (ARIB), Japan, June 1998.

[11] ETSI/UTRA, *The ETSI UMTS Terrestrial Radio Access (UTRA) ITU-R RTT Candidate Submission*, European Telecommunication Standard Institution (ETSI), June 1998.

[12] TIA/TR45.5.4, *The cdma2000 ITU-R RTT Candidate Submission (0.18)*, Telecommunications Industry Association (TIA), USA, 1998.

[13] J. Zander, "Performance of Optimum Transmitter Power Control in Cellular Radio Systems," *IEEE Transactions on Vehicular Technology*, vol. 41, no. 1, pp. 57-62, February 1992.

[14] R. D. Yates, "A Framework for Uplink Power Control in Cellular Radio Systems," *IEEE Journal of Selected Areas in Communications*, vol. 13, no. 7, pp. 1341-1347, September 1995.

[15] A. J. Viterbi, "When Not to Spread Spectrum - a Sequel," *IEEE Communications Magazine*, vol. 23, no. 4, pp. 12-17, April 1985.

[16] L. C. Yun and D. G. Messerschmitt, "Power Control for Variable QOS on a CDMA Channel," *Proceedings IEEE MILCOM '94*, Long Branch, NJ, pp. 178-182, October 1994.

[17] G. R. Cooper and R. W. Nettleton, "A Spread Spectrum Technique for High-Capacity Mobile Communications," *IEEE Transactions on Vehicular Technology*, vol. VT-27, November 1978.

[18] R. L. Pickholtz, L. B. Milstein and D. L. Schilling, "Spread Spectrum for Mobile Communications," *IEEE Transactions on Vehicular Technology*, vol. 40, no. 2, pp. 86-93, May 1991.

[19] A. J. Viterbi, *CDMA - Principles of Spread Spectrum Communications*, Addison-Wesley Publishing Company, 1995.

[20] E. Kudoh and T. Matsumoto, "Effects of Power Control Error on the System User Capacity of DS/CDMA Cellular Mobile Radios," *IEICE Transactions on Communications*, vol. E75-B, no. 6, pp. 524-529, June 1992.

[21] G. L. Stuber, *Principles of Mobile Communication*, Kluwer Academic Publishers, Second Printing, 1997.

[22] M. Soleimanipour and G. H. Freeman, "A Realistic Approach to the Capacity of Cellular CDMA Systems," *Proceedings IEEE VTS 46th Vehicular Technology Conference*, vol. 2, pp. 1125-1129, Atlanta, GA, May 1996.

[23] J. G. Proakis, *Digital Communications*, McGraw-Hill International Editions, 1989.

[24] M. B. Pursley, "Performance Evaluation for Phase-Coded Spread Spectrum Multiple Access Communication-Part I: System Analysis," *IEEE Transactions on Communications,* pp. 795-799, Aug. 1977.

[25] M. O. Sunay and D. J. McLane, "Calculating Error Probabilities for DS CDMA Systems: When Not to Use the Gaussian Approximation," *Proceedings IEEE International Symposium on Personal, Indoor and Mobile Radio Communications,* Taipei, Taiwan, vol. 3, pp. 1744-1749, 1996.

[26] J. E. Dennis, Jr. and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations,* Prentice Hall Series in Computational Mathematics, 1983.

[27] B. D. Craven, *Fractional Programming,* Heldermann Verlag Berlin, Sigma Series in Applied Mathematics, 1988.

[28] L. S. Lasdon, *Optimization Theory for Large Systems,* The Macmillan Company, 1970.

[29] H. A. Taha, *Operations Research, An Introduction,* The Macmillan Company, 1971.

[30] J. M. Aein, "Power Balancing in System Employing Frequency Reuse," *COMSAT Technical Review,* vol. 3, pp. 277-300, Fall 1973.

[31] R. W. Nettleton and H. Alavi, "Power Control for a Spread Spectrum System," *IEEE Vehicular Technology Conference VTC-83,* pp. 242-246, 1983.

[32] W. Tschirks, "Effect of Transmission Power Control on the Cochannel Interference in Cellular Radio Networks," *Elektrotechnik Inform.,* vol. 106, no. 5, 1989.

[33] J. Zander, "Transmitter Power Control for Co-channel Interference Management in Cellular Radio Systems," *WINLAB Workshop on Third Generation Wireless Information Networks*, pp. 241-247, 1993.

[34] S. C. Chen, N. Bambos, and G. J. Pottie, "On Distributed Power Control for Radio Networks," *International Conference on Communications*, New Orleans, LA, pp. 1281-1285, May 1994.

[35] G. J. Foschini and Z. Miljanic, "A Simple Distributed Autonomous Power Control Algorithm and its Convergence," *IEEE Transactions on Vehicular Technology*, vol. 42, no. 4, pp. 641-646, November 1993.

[36] D. Mitra, "An Asynchronous Distributed Algorithm for Power Control in Cellular Radio Systems," *WINLAB Workshop on Third Generation Wireless Information Networks*, pp. 249-259, 1993.

[37] S. Ulukus and R. D. Yates, "Stochastic Power Control for Cellular Radio Systems," *IEEE Transactions on Communications*, vol. 46, no. 6, pp. 784-798, June 1998.

[38] R. D. Yates and C. Huang, "Integrated Power Control and Base Station Assignment," *IEEE Transactions on Vehicular Technology*, vol. 44, no. 3, pp. 1-7, August 1995.

[39] S. V. Hanly, "An Algorithm for Combined Cell-Site Selection and Power Control to Maximize Cellular Spread Spectrum Capacity," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 7, pp. 1332-1340, September 1995.

[40] J. X. Qiu and J. W. Mark, "A Dynamic Load Sharing Algorithm Through Power Control in Cellular CDMA," *Proceedings IEEE International Sym-*

posium on Personal, Indoor and Mobile Radio Communications, PIMRC, Boston, MA, vol. 3, pp. 1280-1284, September 1998.

[41] S. Papavassiliou and L. Tassiulas, "Joint Optimal Channel, Base station and Power Assignment for Wireless Access," *IEEE/ACM Transactions on Networking,* vol. 4, no. 6, pp. 857-872, December 1996.

[42] S. Kim and D. Kim, "Optimum Transmitter Power Control in Cellular Radio Systems," *Information Systems and Operational Research, INFOR,* vol. 35, no. 1, February 1997.

[43] E. Seneta, *Non-negative Matrices and Markov Chains, Second Edition,* Springer-Verlag, 1981.

[44] R. Rezaiifar, A. M. Makowski, and S. P. Kumar, "Stochastic Control of Handoffs in Cellular Networks," *IEEE Journal on Selected Areas in Communications,* vol. 13, no. 7, pp. 1348-1362, September 1995.

[45] M. Asawa and W. E. Stark, "Optimal Scheduling of Handoffs in Cellular Networks," *IEEE/ACM Transactions on Networking,* vol. 4, no. 3, pp. 428-441, June 1996.

[46] C. I. Cook, "Development of Air Interference Standards for PCS," *IEEE Personal Communications,* vol. 1, no. 4, pp. 30-34, Fourth Quarter 1994.

[47] C. I and K. K. Sabnani, "Variable Spreading Gain CDMA with Adaptive Control for True Packet Switching Wireless Network," *Proceedings IEEE International Conference on Communications,* Seattle, WA, pp. 725-730, June 1995.

[48] A. Sampath, P. S. Kumar, and J. M. Holtzman, "Power Control and Resource Management for a Multimedia CDMA Wireless System," *Proceedings IEEE International Symposium on Personal, Indoor and Mobile Radio Communications,* vol. 1, pp. 21-25, Toronto, Canada, 1995.

[49] M. Aoki, *Introduction to Optimization Techniques,* The Mcmillan Book Co., New York, 1971.

[50] A. S. Acampora and M. Naghshineh, "Control and Quality-of-Service Provisioning in High-Speed Microcellular Networks," *IEEE Personal Communications,* vol. 1, no. 2, pp. 36-43, Second Quarter 1994.

[51] S. Manji, *Reverse Link Power Control for Multimedia Packetized DS-CDMA in a Slowly Rayleigh Fading Environment,* M.A.Sc. thesis, Dept. of Electrical and Computer Engineering, University of Waterloo, Canada, 1997.

[52] N. D. Wilson, R. Ganesh, K. Joseph, and D. Raychaudhuri, "Packet CDMA Versus Dynamic TDMA for Multiple Access in an Integrated Voice/Data PCN," *IEEE Journal of Selected Areas in Communications,* vol. 11, no. 6, pp. 870-885, August 1993.

[53] TIA/EIA/IS-95 Interim Standard, *Mobile Station-Base Station Compatibility Standard for Dual-Mode Wideband Spread Spectrum Cellular System,* Telecommunication Industry Association, July 1993.

[54] W. C. Y. Lee, *Mobile Communications Engineering,* McGraw-Hill Book Company, 1982.

[55] Chin-Lin I and R. D. Gitlin, "Multi-Code CDMA Wireless Personal Communications Networks," *Proceedings IEEE International Conference on Communications,* Seattle, WA, vol. 2, pp. 1060-1064, 1995.

[56] A. H. Aghvami, "Future CDMA Cellular mobile systems Supporting Multi-Service Operation," *Proceedings IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '94)*, September 18-22, 1994.

[57] J. C-I Chuang, K. B. Lataief, K. F. Cheung, C.C. Ling, R. D. Murch, and C. T. Nguyen, "Wireless Personal Communications in Hong Kong: A University Perspective," *IEEE Personal Communications,* vol. 4, no. 2, pp. 30-42, April 1997.

[58] J. D. Pintér, *Global Optimization in Action,* Kluwer Academic Publishers, 1996.

[59] H. P. Williams, *Model Building in Mathematical Programming,* John Wiley & Sons Ltd., England, 1990.

[60] F. Forgo, *Nonconvex Programming,* Akademiai Kiado, Budapest, 1988.

[61] P. E. Gill, W. Murray, B. A. Murtagh, M. Saunders, and M. H. Wright, "GAMS/MINOS," Appendix D in Brooke A., Kendrick D. and Meeraus A., *GAMS: A User's Guide,* The Scientific Press, Redwood City, California, 1988.

[62] J. Abadie and J. Carpentier, "Generalization of the Wolfe Reduced Gradient Method to the case of Nonlinear Constraints," in *Optimization,* R. Fletcher (ed.), Academic Press, New York, pp. 37-47, 1969.

[63] A. S. Drud, "A GRG Code for Large Sparse Dynamic Nonlinear Optimization Problems," *Mathematical Programming,* no. 13, pp. 153-191, 1985.

[64] A. S. Drud, "CONOPT - A Large-Scale GRG Code," *ORSA Journal on Computing,* vol 6, no 2, pp. 207-216, 1994.

[65] D. Kendrick A. Meeraus, *GAMS, An Introduction,* Technical Report, Development and Research Department at the World Bank, Washington, DC, 1985.

[66] M. A. Duran and I. E. Grossmann, "An Outer-Approximation for a Class of Mixed-Integer Nonlinear Programs," *Mathematical Programming,* vol. 36, pp. 307-339, 1986.

[67] G. R. Kocis and I. E. Grossmann, "Relaxation Strategy for the Structural Optimization for Process Flowsheets," *Industrial and Engineering Chemistry Research,* vol. 26, pp. 1869-1880, 1987.

[68] G. R. Kocis and I. E. Grossmann, "Computational Experience with DICOPT Solving MINLP Problems in Process Systems Engineering," *Computer and Chemical Engineering,* vol. 13, no. 3, pp. 307-315, 1989.

[69] J. Viswanathan and I. E. Grossmann, "A Combined Penalty Function and Outer Approximation Method for MINLP Optimization," *Computer and Chemical Engineering,* vol. 14, no. 7, pp. 769-782, 1990.

[70] E. M. L. Beale, "Integer Programming," *The State of the Art in Numerical Analysis,* (D. Jacobs, Eds), pp. 409-448. Academic Press, London, 1977.

[71] M. C. Ferris, *MATLAB and GAMS,"* University of Wisconsin, Wisconsin, USA, July 1998.

[72] J. D. Parsons, *The Mobile Radio Propagation Channel,* Wiley, New York, 1992.

[73] R. Padovani, "Reverse Link Performance of IS-95 Based Cellular Systems," *IEEE Personal Communications,* vol. 1, no. 3, pp. 28-34, Third Quarter 1994.

[74] A. M. Viterbi and A. J. Viterbi, "Erlang Capacity of a Power Controlled CDMA System," *IEEE Journal on Selected Areas in Communications*, vol. 11, no. 6, pp. 892-900, August 1993.

[75] M. Ho and G. L. Stuber, "Capacity and Power Control for CDMA Microcells," *Wireless Networks*, vol. 1, no. 1, pp. 355-363, 1995.

[76] A. J. Viterbi, A. M. Viterbi, and E. Zehavi, "Other-Cell Interference in Cellular Power-Controlled CDMA," *IEEE Transactions on on Communications*, vol. 42, no. 4, pp. 1501-1504, April 1994.

[77] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, Massachusetts, 1995.

[78] C. A. Floudas, *Nonlinear and Mixed-integer optimization*, Oxford University Press, 1995.

[79] C. H. Papadimitriou and C. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*, Prentice-Hall Inc., 1982.

[80] S. Ramakrishna and J. M. Holtzman, "A Scheme for Throughput Maximization in a Dual-Class CDMA System," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 6, pp. 830-844, August 1998.

[81] A. J. Viterbi, A. M. Viterbi, K. G. Gilhousen, and E. Zehavi, "Soft Handoff Extends CDMA Cell Coverage and Increase Reverse Link Capacity," *IEEE Journal on Selected Areas in Communications*, vol. 12, no. 8, pp. 1281-1287, October 1994.