INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

Medium Access Control with Congestion Feedback in CDMA Based Networks

Xumin Sun

Department of Electrical and Computer Engineering
McGill University
Montreal, Canada

October 2000

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Master of Engineering.

Abstract

The research reported here deals with the design of the uplink flow control in a CDMA–based wireless access network. Demand from each source is assumed infinitely divisible, the control is rate–based and the service model is ATM/ABR (Available Bit Rate). The flow control problem in general is to manage — dynamically, and subject to prescribed constraints on transmitter power and signal–to–interference ratio — the instantaneous allocation of rate to individual sources. Our particular interest is in settings where the controller has access to information on downstream congestion (real or virtual) for each connection, and where quality-of–service is specified on time scales that are slow relative to the rate of channel variation. Our objective is to exploit the congestion feedback, as well as the temporal flexibility in the quality–of–service specification, to refine the match between resource allocation and need. We propose a framework in which the problem can be posed precisely, and provide a solution in the case that there is but a single base station. The solution has two components. One describes the set of rate allocations that are consistent with the power and SIR constraints. The other uses the congestion feedback, modeled by the states of certain reference buffers downstream of the base station, to select a specific rate allocation within the admissible rate region. The benefit in terms of call–carrying capacity is indicated through simulations.

Sommaire

Ce mémoire découle d'un projet de recherche sur le contrôle des flots "source — station d'antenne" dans un réseau d'accès sans fil CDMA. On suppose que le trafic est continu (infiniement divisible), que le contrôle s'éffectue par une modulation du débit instantané de chaque source and que le service offert aux abonnés se conforme au modèle ATM/ABR (Méthode de transmission asynchrone / Available Bit Rate). Le problème est de gérer, de façon dynamique et en respectant des contraintes préscrites sur la puissance des émetteurs et sur les rapports "signal-interférence", l'allocation instantanée des taux de transmission des sources. Nous nous intéressons en particulier aux instances du problème où le contrôleur peut s'informer sur les états des tampons mémoire d'aval (réels ou virtuels) associés aux connections en vigeur, et où l'échelle du temps à laquelle on évalue la qualité de service est lente par rapport à celle qui caractérise le processes des évanouissements. L'objectif est de constater comment tirer profit de la flexibilité ainsi mise à la disposition du contrôleur pour mieux raccorder la gestion de la largeur de bande et les exigeances des sources. Nous proposons une formulation précise du problème et une solution qui s'applique au cas qu'il n'y a qu'une seule antenne. La solution a deux composants, dont l'un décrit l'ensemble des allocations qui respectent les contraintes, et l'autre, sous forme d'un algorithme, selectionne une allocation admissible particulière selon les informations retournées. L'approche est mise en valeur à l'aide de simulations.

Acknowledgments

I would like to express my deepest gratitude to my supervisor, Professor Michael Kaplan, for his continuous support and guidance throughout my process of this thesis. His intelligence, extreme patience, and valuable advice did more than bring this work to a successful conclusion.

I would also like to thank Professor Peter Kabal, for his resourceful advice on editing my thesis. My special thanks go to my fellow graduate students in the Telecommunications and Signal Processing Laboratory for their fruitful suggestions, encouragement and companionship. They are: Joachim Thiemann, Dorothy Okello, Khaled-El-Maleh, Tamanna Islam, Liqing Zhang, etc.

I am thankful to Sylviane Duval, for her effort of correcting my humble English. I would also thank all my friends for their all kinds of helps, encouragement and best wishes.

Contents

1 Introduction **1**
1.1 Overview of the Uplink Multi–Access Problem 2
1.2 Position of the Problem Relative to the Literature 4
1.3 Thesis Plan 6

2 Background Studies **7**
2.1 Power Control in CDMA System 7
2.2 Throughput Maximization in Voice/Data Systems 9
2.3 Available Bit Rate Service in ATM networks 11
2.4 Flow Control in Wireless ATM Networks 13
2.5 Conclusion 14

3 Problem Formulation and Rationale **15**
3.1 Introduction 15
3.2 The Network Model 20
3.3 the Traffic Model 21
3.4 The Channel Model and SIR 21
3.5 The Admissible Rate Region 23
3.6 The Shape of $\mathcal{R}$ 25
3.7 Dynamic Rate Control 29

4 Performance Analysis **36**
4.1 Simulation Model 36
4.1.1 Log-normal Shadowing 36
4.1.2 The ABR Service Model 37

4.1.3 Rate Matching 38
4.1.4 Choice of Parameters 39
4.2 Rate Tracking 40
4.2.1 Choosing the Control Parameters $b^1, b^2, w, \overline{\mu}$ 41
4.2.2 Choosing $\overline{\mu}$ and w 50
4.3 System Capacity 50
4.3.1 System Comparison 55
4.3.2 Capacity Limit 56

5 Summary and Future Work **63**

References **66**

List of Figures

2.1 Control loop for ABR traffic 12

3.1 The Network Model 16
3.2 The Simplified Network Model 17
3.3 Diagram of coordinate transformation from $\mathcal{P} \rightarrow \mathcal{G}$, M=2. 27
3.4 System Model 30
3.5 Backlog feedback on $\mathbf{R}'$ assignment. 35

4.1 Flow diagram of ABR rate assignment 37
4.2 Sensitivity to $\overline{\mu}$ and (b^1, w) for 2 user case. 42
4.3 Sensitivity to $\overline{\mu}$ and (b^1, w) for 3 user case. 43
4.4 Sensitivity to $\overline{\mu}$ and (b^1, w) for 4 user case. 44
4.5 Sensitivity to $\overline{\mu}$ and (b^1, w) for 5 user case. 45
4.6 Sensitivity to b^2 and $\overline{\mu}$, w for $\overline{\mu} = \lambda_0$ 48
4.7 Sensitivity to b^2 and $\overline{\mu}$, w for $\overline{\mu} = 1.5\lambda_0$. 49
4.8 Throughput versus buffer length B for $\overline{\mu} = \lambda_0$. 51
4.9 Packet loss-rate versus buffer length B for $\overline{\mu} = \lambda_0$. 52
4.10 Throughput versus buffer length B for $\overline{\mu} = 1.5\lambda_0$. 53
4.11 Packet loss-rate versus buffer length B for $\overline{\mu} = 1.5\lambda_0$. 54
4.12 Rate comparison between MAC, PPA and IS-95. 60
4.13 Rate comparison between $\boldsymbol{R}$, $\overline{\boldsymbol{R}}_{server}$, and $\overline{\boldsymbol{R}}$. 61
4.14 System capacity versus different $\overline{\mu}$. 62

List of Tables

4.1 System Parameters 40
4.2 Performance gains for 2-user case. 46
4.3 Performance gains for 3-user case. 46
4.4 Performance gains for 4-user case. 46
4.5 Performance gains for 5-user case. 46
4.6 Average packet loss-rate. 55
4.7 Average transmit power. 56
4.8 Effect of a on transmit power. 58
4.9 Effect of a on transmit rate. 58

Chapter 1

Introduction

We describe an approach to the management of *uplink* radio bandwidth in a CDMA–based, terrestrial, wireless access network. The main features of the network model (see Figure 3.4 in Chapter 3.) include the terminal population, whose individuals generate the traffic to be carried; the base stations, which supply the interface between wireless access and wireline transport; and the data buffers in the base stations, the backlogs in which are viewed as summarizing the history of the mismatch between the rate at which data is written onto the radio channel and the rate at which it is read off downstream of the radio channel. We use those backlogs to guide the allocation of radio bandwidth to individual terminals. Assuming that individual QoS (quality–of–service) objectives are *rate–based* and *elastic*, in the specific sense of being defined through time averages over non–trivial time windows, we seek a *rate–based* medium access control strategy with the following properties: (1) it responds to performance feedback; (2) it takes into account that users are not all equal — both in terms of the services they need and in terms of the propagation environment in which those services are to be delivered; (3) it exploits the elasticity in the definition of individual QoS targets to accommodate time variations in the propagation environment. Our work differs from much of the literature on wireless multiple access in terms of the way in which congestion feedback is used to define fairness and efficiency of the medium access control.

1.1 Overview of the Uplink Multi–Access Problem

The signal received at a base station in a wireless network is an aggregate — the sum of transmissions that are intended for the base station in question, of transmissions that are intended for other base stations, of echoes (due to multipath) of all of these, and of noise. The business of the receiver is to separate the signals that it wants from interference and noise, and to isolate from the aggregate of desired signals the data flows of individual users. The key activities are signal separation — referring to the separation of the desired signals from the undesired ones — and signal decoding, by which a message written onto the channel by a particular user is extracted from the distorted, noise–corrupted facsimile actually presented to the base station.

Medium access control is deployed for the purpose of ensuring that the receiver can do its job within prescribed bounds on error rate. It locates the various user transmissions in *frequency*, in *time* and in *codespace* so as to facilitate demultiplexing. The various approaches described in the literature on medium–access control differ in the extent to which flexibility in frequency, time and signal format are actually exploited. As the name suggests, frequency–based medium–access control effects user separation through *frequency planning*, also known as Frequency Division Multiple Access (FDMA), which in the context of modern terrestrial wireless means *frequency reuse*: different users may be co–located in frequency, provided (i) that they are assigned to different base stations and (ii) that the effect on signal quality due to in–band interference is acceptably small. The cellular idea in its simplest form is a representation of static (precomputed) frequency planning, in which the frequency allocated to a particular user is a function of his geographical position relative to the base stations; to the extent that the allocation is insensitive to temporal variations in demand, the strategy is a form of resource reservation and suffers from the potential for inefficient characteristic of the class. Frequency planning can be made dynamic, with corresponding improvements in efficiency, but at the cost of a large increase in the complexity of implementation.

Time–based approaches (Time Division Multiple Accesses, TDMAs) to medium access control [1] typically combine frequency planning and *scheduling* for user sepa-

ration. Users co–located in space, in the sense of belonging to the same cell or to nearby cells, can use a common set of frequencies provided that their transmissions are disjoint in time. To the extent that both temporal and spectral allocations are precomputed and thus insensitive to demand (the case, for example, when time is allocated on a circuit–switched basis), there remains the potential for inefficiency. The main advantage to time over frequency as the access control variable is that the guard bands required in the implementation of standard FDM[1] waste bandwidth.

Code Division Multiple Access (CDMA) [2] substitutes codespace separation for frequency planning, thereby moving complexity from frequency management to code and receiver design. In its simplest form — the form designated *Spread Spectrum* CDMA (SS–CDMA, [3, 4]) — each signal is formatted to look like white noise to all the others, and processed at the base station by a receiver designed for white Gaussian noise. The result is a system which at least in the case that fading is slow and flat — assumptions adopted in our own work — is relatively easy to analyze; the corresponding formula for signal–to–interference ratio is supplied in Chapter 3. But the spread spectrum approach to user separation is not as effective as it might be: partly because the randomization by which interference is converted to noise actually makes the decoding problem harder in terms of achievable error rate, and partly because the receiver, in executing the decoding function, is in fact using less than the total amount of information available to it. More modern approaches to CDMA involve careful design of the formatting (coding) mechanism by which users are (at least partially) separated, and systematic exploitation, in the form of receiver architectures capable of *Multi–User Detection* [5], of *all* the information available in the aggregate received signal. What all CDMA systems have in common — a point that is key to the design of the medium access control — is that they allow multiple users, co–located in time, frequency and space, to transmit concurrently.

Our interest is in the SS–CDMA approach to medium access control. We assume that the source data is tolerant of delays on the order of a fraction of a second. The controls we study act by modulating the rate at which the data generated by any

[1] As opposed to so–called Orthogonal FDM, which at the cost of some processing obviates the guard bands.

particular source is permitted to enter the radio channel. Because the instantaneous rates allocated to the various sources depend as much on network dynamics as on instantaneous demand, there is a need for buffering. The controls we propose can be viewed as buffer management systems.

The OSI protocol stack locates medium–access control between the physical and link layers. But in fact the problem of designing an effective MAC strategy bestrides the whole bottom half of the stack, from physical layer through link to network layer. In the CDMA context in particular, medium access control is effected at the physical layer through multi–user coding and antenna design; at the MAC layer, through packet–level flow control and scheduling; at the link layer, through error management; and at the network layer, through end-to–end QoS management. Our work deals with the MAC and network layers, and with the possibility of a useful collaboration between them for the purpose of managing QoS while retaining flexibility in the allocation of radio bandwidth. Our physical layer model is standard and simple — single–user detection, classical AWGN receiver. The link layer is ignored; a constraint on received signal-to–interference ratio is assumed sufficient to ensure that error control remains a low–profile activity.

1.2 Position of the Problem Relative to the Literature

Literature related to the thesis research is reviewed in Chapter 2. This section provides a qualitative description of the problem to be solved. The objective is to locate the problem relative to what has been done before.

The idea is that the dynamic operation of the wireless access network is specified by an *operating point* — by a set of bit-rates, one such for each terminal and for each instant (or slot) of time. The operating point is a vector, each coordinate of which determines the bit–rate of an individual terminal. The goal of the medium access control enterprise is to select an appropriate operating point. Because transmissions from different terminals, whether in the same or in different cells, mutually interfere, the individual bit–rates are optimally decided jointly, the selection of any one of them impacting the quality of reception for all the others.

The set of all possible operating points can be visualized as a region in an appropriate multi–dimensional space. The space will be Euclidean in the case that the rate assignment is *static*, and a function space otherwise. That set, denoted $\mathcal{R}$ in Chapter 3, is delimited by a number of constraints. One set of constraints formalizes the quality–of–service objectives for the various calls in progress; quality of service in our model is assumed represented by signal–to–interference ratio (SIR). A second set of constraints flows from a limitation on maximum power. All the constraints together determine the precise shape of $\mathcal{R}$. There are reasonably efficient algorithms for computing the boundaries of $\mathcal{R}$. The main point is that $\mathcal{R}$ is a multi–dimensional continuum. Some additional criterion, above and beyond constraints on power and SIR, is needed in order to select a *particular* operating point from the continuum of possibilities.

It is here that our work diverges from the literature. Individual data rates are typically assumed in the literature to be fixed by the sources, meaning that the operating point, once determined to be feasible in the sense of belonging to $\mathcal{R}$, is *unique*; the single degree of freedom remaining — the assignment of terminals to base stations — is optimized so as to minimize the power required to achieve those rates at the target SIR levels.

Our point of view is different. The data rates themselves (in addition to the assignment of terminals to base stations) are assumed amenable to control — subject to constraints that are formulated, not in terms of *instantaneous* values, but rather in terms of certain *averages*. The averaging is on a time scale that is determined by the service requirements. The idea is to exploit the flexibility implied by the averaging in the service specification to compensate for fluctuations in channel capacity due to fading; basically, bandwidth is assigned on a priority basis to connections that on the one hand most urgently need it (possibly because of earlier starvation due to poor propagation conditions) *and* that on the other hand can actually use it (meaning that flow downstream of the base station is suitably unimpeded). The problem is thus

(1) To select a suitable definition for the averaging in question;

(2) To devise a bandwidth management algorithm adapted to the particular form of

the quality–of–service specification, and which at the same time is capable of responding to congestion feedback originating downstream of the access network.

The notion of *effectiveness* in this connection is quantified by the number of *calls* — at a specified level of quality–of–service — that can be accommodated simultaneously per base station; or by the total throughput when the number of calls is fixed.

1.3 Thesis Plan

Chapter 2 reviews the literature that precedes the thesis research. Chapter 3 provides a precise formulation of problem and of the notion of averaging employed in the definition of quality–of–service, and proposes a specific solution in the form of a parametric family of bandwidth management algorithms applicable to the case of a single base station. Chapter 4 describes an end-to–end connection model suggested by the ABR (Available Bit Rate) service paradigm, and the results of simulations undertaken to test the effectiveness of our algorithm. Chapter 5 summarizes and suggests possible extensions.

Chapter 2

Background Studies

Various control schemes for CDMA system have been reported in recent literature. One branch of this research focuses on the power control needed to reduce interference and improve system capacity. Another branch concentrates on voice/data integrated service networks. Enhancement of the data throughput while maintaining acceptable QoS requirement is the central issue. In this chapter, we review these studies, and describe the basic ABR service model. Since our goal is to build up an integrated MAC layer connecting the CDMA-based wireless with an ABR-based wireline network, a certain feedback mechanism is needed. This is emphasized in the last part of the chapter, where we review a flow control algorithm that uses buffering in the base station to control the mobile transmission bandwidth.

2.1 Power Control in CDMA System

In a CDMA system, the mobile users transmit concurrently in both time and frequency. The result is multiple access interference (MAI) — the various user signals interfere with each other. The problem of signal detection/decoding in the CDMA environment is complicated by a combination of MAI, Doppler effects and loss of signal energy due to fading. Power control is one of the techniques used to combat fading. In a different form and on a different scale, it can be used as well to control MAI. Therefore, the design of the power control algorithms is not a trivial matter.

Early work on power control proposed two approaches. One of them [6, 7] centers on

schemes which keep the received power at a constant level. These schemes have the advantage of controlling the effects of fading. They indicate an increase of system capacity by a factor of roughly 2 relative to schemes with constant transmission power. The power control in the second generation PCS system, IS-95, uses this proposal. Another approach ([8, 9, 10]) is derived from the concept of *SIR balancing*, first used in [11] for satellite systems. This yields a "fair" distribution of the interference in the sense that all mobile users experience the same SIR level. Zander [12] and his group [13] further investigated the *SIR balancing* technique and proved that it is optimum in the sense that it minimizes the threshold SIR for a given QoS requirement. Numerical results presented by this algorithm show capacity gains in the order of a factor of 4, compared with a system using fixed transmission power. However, these studies neglect the effect of crosstalk among cells and assume the assignment of users to base stations to be fixed.

Recent research on power control [14, 15, 16, 17] has focused more on dynamically assigning users to a base station, while regulating transmitter power. The overlapping topography among cells in modern wireless communication provides an opportunity to implement load sharing. For example, so-called *cell breathing* is achieved by dynamically adjusting the virtual size of a cell. When the load in a cell increases, users in the boundary region may be re-assigned to less heavily-loaded neighbor cells. When the load decreases, the base station accepts some boundary users from more heavily-loaded neighbors.

Hanly [16] describes an algorithm to achieve cell breathing. It operates as follows. Each base station measures the total interference received from all mobiles in the network and broadcasts this value via control channels. A mobile user collects the information from its surrounding base stations. It is assumed that a mechanism exists by which the user can determine the fading coefficients to the adjacent base stations; for example, the user can measure the strength of pilot signals broadcast by these base stations and compare them with unfaded original values. It then removes its own contribution from the received interference measurements, and computes the transmit power needed to counteract the interference at each base station. Based on this, it selects the base station which requires minimum power. The algorithm is decentral-

ized in the sense that each user computes for himself the minimum transmit power at time (n+1), under the assumption that all the other users remain fixed at their time (n) power values. It is found to be optimal in the sense that interference is minimized. However, the algorithm needs extensive computational, storage and signaling facilities at both base station and mobile, and is therefore difficult to implement.

A practical pilot power control algorithm is presented in [17], using the same concept of cell breathing. The basic concept is that when the SIR falls below a certain threshold, the base station implicitly directs some users to more lightly loaded neighbor cells by lowering the pilot signal power. This is implemented as follows. Since the mobile users select the base station with the strongest pilot signal power among all those received, some users in the overlapping region handoff to another cell when the original base station's pilot power is sufficiently reduced. However, the pilot power cannot be decreased indefinitely. A minimum level is required to guarantee sufficient overlap to conduct the soft handoff. Alternatively, the base station automatically increases the pilot power to some upper bound when the SIR is larger than the threshold. In this way, it provides a friendly interface for the users in the heavily-loaded surrounding cells to handoff. In comparison with Hanly's algorithm, both the signaling and computation are much reduced. Furthermore, the upgrade is needed only within the base station itself, while no change is needed on the mobile side.

2.2 Throughput Maximization in Voice/Data Systems

In data communications, both the transmitter power of the mobiles and their transmit rates may be controllable resources. In the voice/data mixed traffic environment, real-time voice has a stringent delay time constraint, whereas data often does not. Consequently, voice connections are handled by the call-level admission control. The central station is required to ensure that the system has sufficient bandwidth for new and existing calls before another new incoming call is accepted. Since voice sources may not always transmit information at their peak rates, there will occasionally be some unused bandwidth. This residual bandwidth can be used to serve data traffic.

To date, numerous studies have been published on the enhancement of data through-

put in an integrated voice/data network. In [18], a feedback-driven congestion control is proposed. A data user experiences three states in the full transmission process. The first is determined at call-level when there is empty space in the central buffer. Currently, only a virtual connection is used and no spreading codes are assigned. The second state is *standby* state. That is, the first M_d users in the queue are assigned C_d CDMA codes. These data users start transmitting their messages until transfer is done with probability p_{star}, which corresponds to the third — *active* — state. p_{star} is a monotonic decreasing function of the current load when the load is below the threshold T. For the optimal solutions of (M_d, C_d, T), the author argued that M_d can be fixed at a reasonable number. A single spreading code brings better throughput. Multiple codes are applicable when traffic is light and interference is correspondingly low, and delay time is much reduced in this way. The bandwidth threshold T is more sensitive to the voice traffic level than the other two parameters, which can be adaptively adjusted according to the traffic load. This algorithm has certain limitations. First, the transmit power constraint is not considered. Second, the QoS requirements — *packet error rate* (PER) for both voice and data — are discussed as an average, which is inappropriate because it cannot guarantee individual PER requirement even should the average PER be met.

Same problem of maximizing data throughput while limiting the probability of outage (P_{out}) was also investigated in [19]. Based on the feasible transmit power inequalities deduced in [20, 21], the author presented a 1-bit broadcast feedback access control scheme for the integrated CDMA systems. The basic idea is to limit P_{out} by reducing the permission probability for data when the load is 'high', and increasing the permission probability when the load is 'low'. The permission probability is triggered by the *persistence* state, $T(t)$, which is defined as

$$T(t+1) = \begin{cases} T(t) + K, & \text{if load} \geq \Omega \\ T(t) - 1, & \text{if load} < \Omega \end{cases} \tag{2.1}$$

When the mobiles receive the state message broadcast by the base station, each data-user will transmit data with probability $\pi^{T(t)}$, or remain in the buffer with probability $1-\pi^{T(t)}$. This model is quite similar to [18]. Both adjust the data user's transmission

probability as a decreasing function of the current load. The peak power constraint is not considered in either case. However, [19] potentially has more flexibility than [18] when power constraint is implemented, at which point the load is also a function of the fading. This is because [19] controls the transmission probability in every time-slot interval; whereas in [18], once a user initiates data transfer, it will not cease even if the load violates the threshold. The only solution, at that point, has to rely on refraining p_{star} for the rest of the data users in the queuing buffer.

Based on the mathematical model in [20], Ramakrishna [22] compares two transmission schemes for delay-tolerant data. The first uses conventional CDMA. The second is a TDMA-type scheme that schedules data users so that, at any given moment, only a limited number are allowed to transmit data, while the remaining maintain synchronization. The conclusion is that throughput is maximized by allowing only one user to transmit at a time. Performance is described for a variety of channel models. The time-varying fading channels are implicitly modeled by the Log-normal distributed variant requirement for SIR level. Unfortunately, this exploration only applies to the non-power-constraint case. A more complicated system with both power constraint, and time-varying fading is presently non-existent, and therefore a problem that requires further consideration.

2.3 Available Bit Rate Service in ATM networks

The available Bit Rate (ABR) service [23, 24] is a rate-based, packet level flow control mechanism. It is implemented via a stream of resource management (RM) cells, generated by the source (*forward* RM) and looped back by the destination (*backward* RM). The RM cells are generated proportional to the data rate (Fig. 2.1). The rate of the data cells at which an ABR source is allowed to transmit is called the allowed cell rate (ACR). Initially, it is set at the initial cell rate (ICR), which is always between the minimum cell rate (MCR) and the peak cell rate (PCR).

There are two congestion markings for ABR traffic. For EFCI marking, when a data cell travels forward through the network, switches can set the EFCI bit in the data cell that informs the destination of congestion. For the ER marking, switches do not

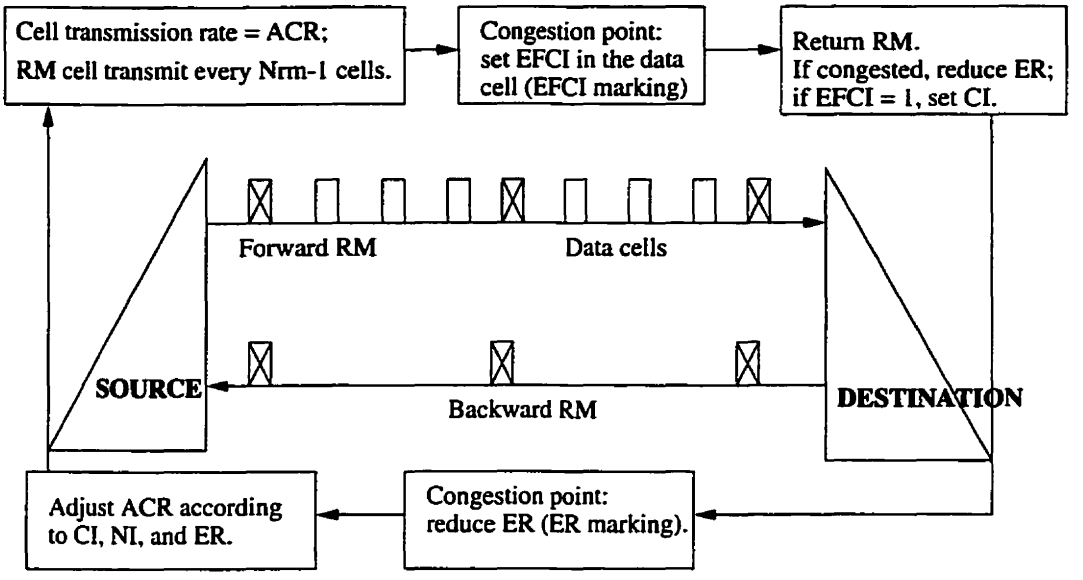


Fig. 2.1 Control loop for ABR traffic

need to set the EFCI bit; instead, they can directly indicate the bandwidth that can be assigned to the connection by setting the ER field in the RM cell. The source continues to send RM cells after every $(Nrm - 1)$ data cells. The source rate is controlled by the return of these RM cells to the destination.

At the beginning of every control cycle, the source first fills up the *forward* RM with the current cell rate, and the rate at which it wishes to transmit in the ER field. When the RM cell arrives at the destination, it becomes the *backward* RM. If the destination is congested to the extent that the rate in the ER field cannot be supported, the destination reduces the ER to a rate that it can support, and also sets the CI bit. If EFCI marking is applied instead and the destination finds that the EFCI bit is set in the last data cell, it will set the RM's CI bit as well to indicate congestion.

On the way back through the network, each switch can reduce the ER field whenever it finds that it cannot support the rate, but not the other way round since this will cause a bottleneck for the previous switch.

When the RM finally arrives the source, the source modifies its ACR based on the information in the RM. If the CI is not set, the source can increase its ACR by a fixed increment, up to the ER value returned, but never exceeding the PCR. This fixed increment is RIF*PCR where RIF stands for the rate increase factor. When CI is set (CI=1), the ACR is decreased by an amount equal to a proportion of its current ACR, which is the rate decrease factor (RDF). If it is still greater than the returned ER, the ACR is then set to the value in ER field, though beyond the MCR. When the *no-increase* (NI) bit is set, it tells the source to note the CI and ER fields in the RM cell, but to keep the current ACR value.

In practice, some networks exist which are incapable of both EFCI and ER markings. The rate is then totally dependent on the CI and NI bits in the RM cell, which is called *relative rate marking.*

2.4 Flow Control in Wireless ATM Networks

The models [18, 19, 22] discussed in section 2.2 focused on maximizing throughput. This approach can create a bottleneck at the wireless–wireline interface. It is therefore necessary to consider issues related to flow control. Francis [25, 26] first realized this problem and proposed a rate allocation scheme for ABR service in wireless ATM.

Francis's rate allocation scheme considered N ABR mobile traffic sources connected with a single bottleneck node (or link). The ATM cells generated by the sources share the same buffer in the base station before passing through the node. A binary feedback scheme is used at the base station to tell the mobiles about congestion conditions in the wireline link. When the backlogs in the buffer overflow a predetermined threshold, the base station will set a feedback bit (similar to a CI bit in wireline ATM) and broadcast to all mobile users. The essential difference between the wireless ATM and the wired one is that the rate assigned to each mobile is a function of channel conditions as well as of the congestion state in the wireline node. For example, if CI=0 and fading is severe for a particular user, the base station reduces its transmission rate during that period and allocates the extra bandwidth to users with better channel conditions. In this way, the system achieves fairness among mobile users and

bandwidth balance at the bottleneck node.

2.5 Conclusion

We will see in Chapter 3 that our MAC algorithm is stimulated by the same motivation as Francis's flow control scheme, except that:

- Francis suggested TDMA based uplink transmissions, whereas our baseline is CDMA wireless networks.
- Francis dealt with one single node sharing one traffic buffer in the wireline entrance, while in our system model, each mobile user connects to the wireline network using a dedicated virtual connection (VC) and buffer.
- The power constraint while absent in [25, 26] is an essential factor in our study. Our mathematical model develops from the same SIR equation as that in [19, 22].

Chapter 3

Problem Formulation and Rationale

This chapter has two objectives. First, it formulates the problem of interest and justifies the choices made in arriving at that formulation. Second, it offers an algorithm that solves the problem in the particular case that there is only one base station. The problem is thus defined at a level of generality which exceeds that of the problem actually solved. We feel that the problem, though only partially solved, is nonetheless worth stating in full generality; and that the solution, while incomplete, is of interest both in and of itself and as a guide to the search for a solution to the general case.

3.1 Introduction

The network we have in mind, as previously described, amounts to two wireless access networks interconnected by a wireline transport network. The routing mechanism is assumed VC (virtual circuit) based, meaning that a particular source seeking connection to a particular destination is assigned a path from end to end that does not change during the lifetime of the call. An example of such a path is shown in Fig. 3.1; it has wireless segments at each end and a wireline segment in between. The essential features of the path are the two terminals, the two base stations (one serving the originating terminal and the other, the destination) and the wireline circuit connecting the base stations. In the discussion to follow, we focus on the particular

virtual circuit represented by the path in Fig. 3.1. Inasmuch as each virtual circuit is uniquely identified with a particular *flow*, a particular *terminal* and a particular *source–destination pair*, we use all four terms interchangeably.

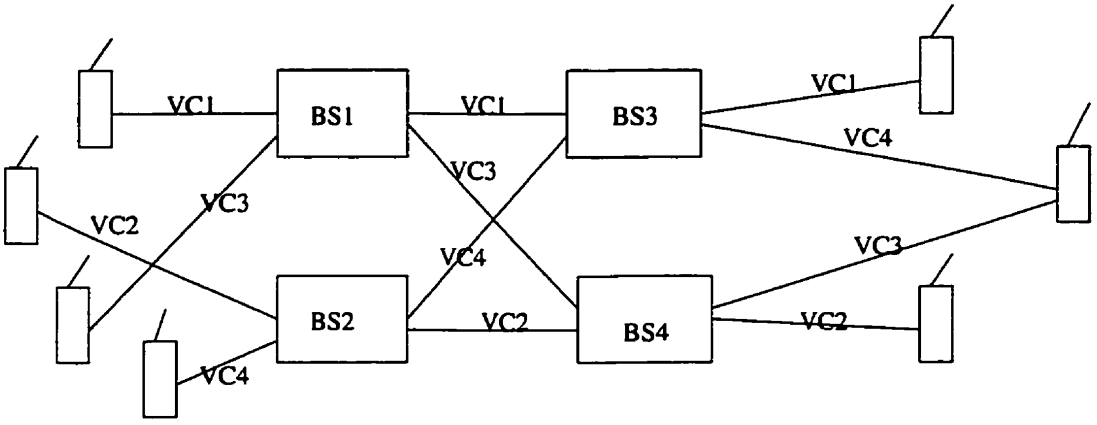


Fig. 3.1 The Network Model

The data *rate* in the virtual circuit to which we refer is affected by several factors: by variations in the capacities of the radio channels at each end, by the rates at which the transmitting end can generate data and at which the receiving end can absorb it, and by flow regulation mechanisms installed at the various interfaces and switching nodes along the way. Because of the stochastic nature of the demand at the various multiplexing points along the route, and because of the stochastic variability in the channel capacity at the two ends, the instantaneous flow rates, as measured at different sites in the circuit, are typically unequal. It is this inequality — its duration, extent and the potential for data overflow and underflow associated with it — that needs to be managed. Buffering handles the transients. Intelligence, in the form of packet–level flow control, manages the buffers. We are specifically interested in the *uplink* radio channel (terminal to base station) and in the design of the strategy, implemented in the *base stations*, by which packet–level access to the uplink is controlled.

Congestion, manifested by long and rising buffer backlogs, can occur anywhere along the path, up to and including the receiving terminal. It is most likely to occur in

the air interface at the two ends — simply because instantaneous capacity is less predictable there than in the intervening wireline segments. Nonetheless, we assume — mostly for simplicity — that the critical buffer is located at the interface between the originating radio channel and the wireline network immediately downstream of it. There is one such buffer for *each* active flow. The assumption is justified in two ways. First, because congestion tends to propagate backward, it can be thought of as presenting itself to the source in the form of back - pressure at the terminating end of the radio link. Second, and more important to us, is that we view the interface buffer as embodying a kind of *virtual queue* whose purpose is to *define* the *average* capacity of the originating radio link — the average capacity as seen by the virtual circuit to which that buffer is assigned. We return to this point in a moment in connection with the formulation of the control problem. Whatever the interpretation of the role of the buffer, the effect of the assumption is to narrow the focus of our study, from the multi–hop trajectory depicted in Fig.3.1 to the single–hop segment, consisting of the originating radio access network, its base station and the buffer at the wireless/wireline interface, shown in Fig. 3.2.

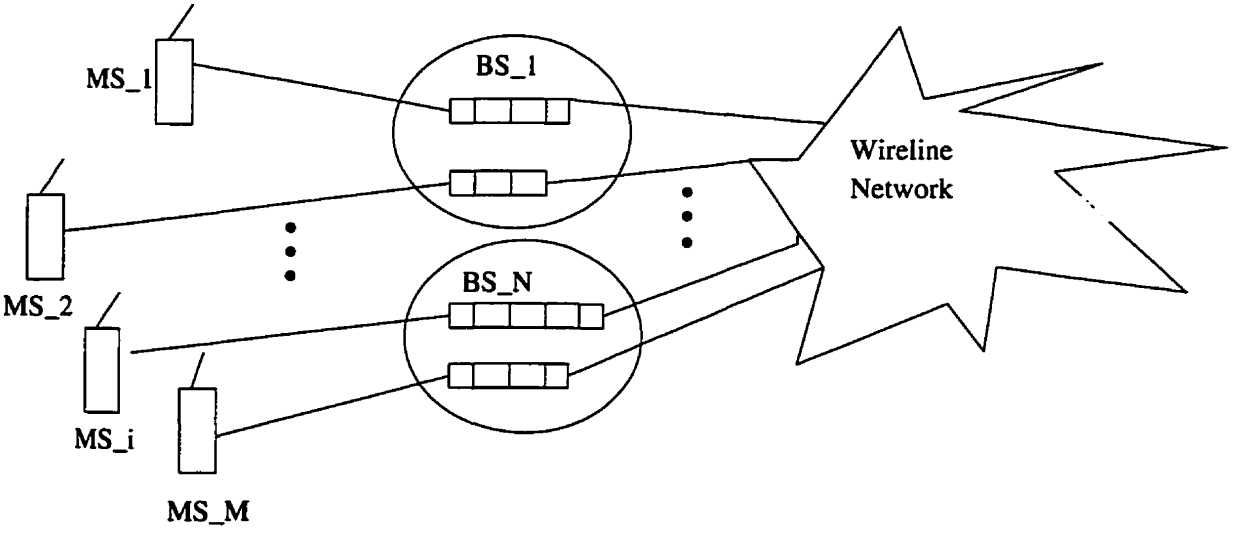


Fig. 3.2 The Simplified Network Model

Our problem is to manage that buffer. Inflow to the buffer is provided by the radio

channel; outflow is mediated by the wireline network and by the radio channel at the far end. Inflow and outflow are thus stochastic, timevarying and generally unequal. The buffer is controlled by modulating the input rate; that is, by controlling access to the radio channel. The selection of a particular rate control policy is to be guided by the quality–of–service (QoS) objectives associated with the various VC's. We assume, referring to a particular VC, that QoS has three coordinates:

(1) The incidence of data loss due to buffer overflow.

(2) Average data rate, the notion of *average* in this connection remaining to be defined.

(3) Received signal–to–interference ratio (SIR), measured at the base station.

They are made precise with the help of some simplifying assumptions. The first such assumption, already described and motivated in the preceding paragraph, is that data loss, to the extent that it occurs at all, occurs at the wireless/wireline interface at the base station. So the first objective in designing the medium access control in the radio access network is to bound the backlog in the buffer at the base station.

Regarding Item (2) in the list, the objective is to provide some sort of underbound to QoS that is *flexible*; specifically, that acts on a time scale which is long enough to allow for recovery from short–term fading effects. We propose a mechanism that is similar in spirit to the Leaky Bucket, except that its role is to *sustain* data rate, rather than to limit it. The mechanism in question takes the form of a bank of *virtual* buffers, one for each flow on the originating radio channel. The parameters of each buffer — its length and depletion rate — are determined as part of the connection setup process. The objective of medium access control is to avoid *underflow* in the virtual buffer, and *overflow* in all buffers, real and virtual. The effect is to impose both upper and lower bounds to data rate, suitably averaged. We assume for simplicity that both overflow and underflow in each virtual circuit are defined relative to the *same* buffer — which brings us to the system described in Fig. 3.2. More generally, each of the two bounds can be formulated in terms of overflow and underflow of *different* buffers, of different lengths and with possibly different depletion rates.

Item (3) is related to the *fidelity* of the signal received at the base station. The SIR associated with a particular VC is the ratio of *bit energy* (in joules, say) to *interference power per unit bandwidth* (watts per Hertz). Our basic requirement in this connection is that probability of error for a given signaling format (BPSK, for example) be a function of SIR. Such is the case when the receiver technology is based on single–user (as opposed to multi–user) detection, and when the inter–user interference is approximately white and Gaussian. Following much of existing literature on the performance modeling of spread–spectrum CDMA, we assume that both conditions are met. The fidelity constraint for VC (i) can thus be expressed in the form

$$\mathrm{SIR_i} \geq \gamma_i, \tag{3.1}$$

where $\mathrm{SIR_i}$ denotes the SIR for VC (i) and γ_i is a threshold depending on the modulation and reflecting the level of fidelity required.

The three coordinates together define the radio segment of our VC. Inasmuch as the first two coordinates are defined in terms of overflow and underflow of the base station buffer, we can regard the bandwidth requirements of the VC as defined jointly by the buffer and by the SIR threshold γ_i. Indeed, the buffer, like the Leaky Bucket, can be thought of simply as an artifice inserted into the data path specifically in support of bandwidth management; so interpreted, it would be designed jointly with the medium access control to realize particular rate, loss and SIR objectives. The important thing, whatever the interpretation, is that the buffers be implemented at the base station, rather than at the source, it being at the base stations that the uplink controllers are located and that the information provided by the buffer backlogs is actually used.

The system model has three components: network, channel and traffic. We proceed to describe each one separately, and then our approach to the design of a medium access control that realizes the three objectives cited above. That approach has two parts. The first characterizes the class of controls that are consistent with the given SIR constraints. The second provides a particular control within that class which in

the case of a single base station seems to perform favorably in terms of buffer overflow and underflow.

3.2 The Network Model

The network model includes the terminals, the base stations providing near–end network access and the interface buffers in the base stations. There are M terminals (and thus M flows) and N base stations. Each flow is assigned a VC and a base station, and each such VC is provided a buffer at the base station. We write A_k for the set of flows (terminals) assigned to base station (k). The i–th VC is characterized by three processes, all of them functions of time: the rate $R_i(\cdot)$, the *transmitter* power $P_i(\cdot)$ and the backlog $x_i(\cdot)$ (referring to the buffer at the base station). Power is assumed subject to a limitation of the form

$$P_i(t) \leq p_i, \tag{3.2}$$

where p_i is a given constant. The buffer associated with the i–th VC has length L_i (packets) and depletion rate process $\mu_i(\cdot)$; the statistics of μ_i are determined by the nature of the service commitment at connection setup time (ABR is an example) and by the availability of bandwidth downstream of the base station (corresponding to the wireline and wireless far–end segments of the overall path). The role of the medium access control is to set $R_i(t)$ and $P_i(t)$ for all t and i, taking into account the buffer lengths L_i, the SIR constraints γ_i and the backlogs $x_i(\cdot)$.

Remark: More generally, the A_k — the allocation of terminals to particular base stations — are subject to control as well, and optimized *jointly* with the R_i and P_i. In the case that the R_i are known and fixed, this is done, for example, in [22]. Our problem is different in the respect that here the R_i and P_i are *both* controllable. We simplify the optimization by assuming that the A_k are fixed.

Restriction: We assume from now on that the power limits p_j are all equal; that is,

that $p_j = p$ for all j.

3.3 the Traffic Model

The traffic model, though simple, is not uncommon in the flow control literature. It is defined by two assumptions: first, that each terminal is capable of filling whatever bandwidth is allocated to it; second, that units of traffic are infinitely divisible (the hallmark of the so–called *fluid* model). Each VC can thus be regarded as permanently backlogged and transmitting at precisely the maximum rate allotted to it. It follows that the parameter $R_i(t)$ defined above refers both to the rate set by the controller and to the rate at which transmission actually occurs. The effect is to remove uncertainty at the controller as to the state of the (remote) source buffers, and thus to reduce — significantly — the complexity of the control algorithm. One way of dealing with more bursty traffic processes is to suppose that the sources provide some kind of local state information to the controller, thereby essentially reducing the more general problem to the one we have studied. That said, the precise role of source state information in the formulation and performance of the medium access control remains outside the scope of the thesis.

3.4 The Channel Model and SIR

Our channel model (referring to the uplink) is simple but standard: flat passband of bandwidth W (Hertz), additive white Gaussian noise, flat slow fading. We write σ_k^2 for the intensity (in watts per Hertz) of the white noise at the receiver in the base station (k), and α_{ik} for the factor by which energy emitted at terminal (i) is attenuated en route to base station (k). The signal from terminal (i) is accordingly received at base station (k) at power $P_i\alpha_{ik}$. The channel model is characterized by the N–vector (σ_i^2) of noise intensities and by the $M \times N$ matrix (α_{ik}) of fading coefficients.

The medium access control proposed below is *slotted.* In particular, it assumes that the fading coefficients, while timevarying, in fact vary sufficiently slowly as to be essentially constant over one time slot. Because we shall assume that the channel parameters are known to the controller at the start of every slot, the law of the time

variation is immaterial.

Regarding the computation of SIR, recall that the quantity in question is the ratio of *per–bit energy* in the desired signal to total interference *intensity* (power per Hertz). In the case that i is chosen from A_k (meaning that the VC from terminal (i) passes through base station (k)), the denominator for VC (i) is

$$\sigma_i^2 + \frac{1}{W}\sum_{j\neq i} P_j \alpha_{jk}.$$

The numerator — the energy at base station (k) per received bit on VC (i) — can be written in the form

$$\frac{P_i \alpha_{ik}}{R_i}.$$

Dividing numerator by denominator gives

$$\mathrm{SIR_i} = \frac{\mathrm{P_i}\alpha_{\mathrm{ik}}/\mathrm{R_i}}{\sigma_{\mathrm{k}}^2 + \frac{1}{\mathrm{W}}\sum_{\mathrm{j}\neq\mathrm{i}} \mathrm{P_j}\alpha_{\mathrm{jk}}} \quad (\mathrm{i} \in \mathrm{A_k}). \tag{3.3}$$

which in turn (following renormalization of α_{ik} by the noise power $\sigma_k^2 W$) yields

$$\frac{R_i \mathrm{SIR_i}}{W} = \frac{P_i \alpha_{ik}}{1 + \sum_{j\neq i} P_j \alpha_{jk}} \quad (i \in A_k). \tag{3.4}$$

This is the expression with which we work. It is not new; see, for example, [22]. It *is* occasionally seen without the factor R_i in the numerator and without the normalization by W in the denominator; that is, as a ratio of signal power to interference power, rather than as a ratio of signal energy to interference intensity. In the case that the R_i's are all the same, which is the one most often encountered, the difference

is immaterial.

3.5 The Admissible Rate Region

The thresholds γ_i on SIR constrain the rate and power allocation. This section studies the action of those constraints, ignoring both the buffer backlog processes and the time variability in the channel; both are brought in later, in selecting a *specific* rate and power allocation at a *specific* time from among the *set* of allocations consistent with the SIR constraints. The problem in this section is completely time–invariant. A particular rate and power allocation is described by time–independent M–vectors $\mathrm{R} = (R_1, \ldots, R_M)$, $\mathrm{P} = (P_1, \ldots, P_M)$.

Definitions:

(1) The pair (P, R) is *admissible* if and only if it is compatible with the *SIR constraints* $\mathrm{SIR_i} \geq \gamma_i$ and the *power constraints* $P_i \leq p$, $i = 1, \ldots, M$.

(2) P is admissible if and only if (P, R) is admissible for some R, and R is admissible if and only if (P, R) is admissible for some P.

(3) P is *efficient for* R if and only if (i) (P, R) is admissible *AND* (ii) there does *not* exist admissible $(\mathrm{P}', \mathrm{R})$ for which $\mathrm{P}' < \mathrm{P}$.[1]

The set of admissible *power* vectors, and the set of admissible *rate* vectors, form subsets of Euclidean M–space. They are denoted $\mathcal{P}$ and $\mathcal{R}$ respectively.

Our goal is to use knowledge of the $x_i(\cdot)$'s to guide the choice of a particular $\mathrm{P} \in \mathcal{P}$ and $\mathrm{R} \in \mathcal{R}$. We note first certain simple properties of $\mathcal{P}$ and $\mathcal{R}$ that are evident by inspection of the expression recorded above for SIR.

Fact:

[1]The *weak* vector inequality ($\leq$) is to be understood as applying separately and simultaneously in all coordinates; the *strict* inequality ($<$) means weak inequality in all coordinates and strict inequality in at least one of them.

(1) For *any* P satisfying the power constraints, there exists R satisfying the SIR constraints. It follows that $\mathcal{P}$ is the entire M–dimensional cube $[0,p]^M$.

(2) If (P,R) is admissible, then so is (P,R') for any $\mathrm{R}' \leq \mathrm{R}$.

(3) If (P,R) is admissible and if $\mathrm{SIR_i} > \gamma_\mathrm{i}$ for some i, then there exists $\mathrm{P}' \leq \mathrm{P}$, with $P_i' < P_i$, such that (P',R) is also admissible. It follows in particular that P' can be chosen so that in fact $\mathrm{SIR_i} = \gamma_\mathrm{i}$ for all i.

We are naturally interested in power allocations P that are efficient in the sense described in the definition. In light of Fact (3), any such P is characterized by the property that the SIR constraints are in fact *equality* constraints.

Summary: R is admissible ($\mathrm{R} \in \mathcal{R}$) if and only if

$$\frac{R_i\gamma_i}{W} = g_i(\mathrm{P},\alpha) \quad (i = 1,\ldots,M) \tag{3.5}$$

for some P in the rectangle $\mathcal{P}$, where α denotes the matrix (α_{ik}) (normalized, as described above, by the noise powers) and where

$$g_i(\mathrm{P},\alpha) \triangleq \frac{P_i\alpha_{ik}}{1+\sum_{j\neq i} P_j\alpha_{jk}} \quad (i \in A_k).$$

The admissible rate region $\mathcal{R}$ is thus (to within a simple scaling of the coordinate axes) identical to the *range* of the function

$$g(\mathrm{P},\alpha) \triangleq (g_1(\mathrm{P},\alpha),\ldots,g_M(\mathrm{P},\alpha)) \quad (\mathrm{P} \in \mathcal{P},\ \alpha \text{ fixed}).$$

Recall from Section (3.1) the three objectives which are to guide construction of the medium access control. The third — accommodation of the various SIR constraints

— is exactly equivalent, given α, to the requirement $\mathrm{R} \in \mathcal{R}$. Formulated in terms of $\mathcal{R}$, our problem is to design an algorithm which selects $\mathrm{R} \in \mathcal{R}$ as a function of α and the base station buffer backlogs, and in such a way as to avoid the overflow and underflow events referred to in Section (3.1). To that end it would be useful to have a simple test for verifying membership in $\mathcal{R}$. The next section summarizes what we know on this point. The situation is simplest in the case of a single base station, corresponding to $N = 1$.

Remark: The mapping g, regarded as a function of P for given α, is often, but not always, invertible on $\mathcal{P}$. Invertibility depends on α. We think of the controller as specifying R first, from which an efficient P is computed *via* Equation 3.5.

3.6 The Shape of $\mathcal{R}$

Let $\Gamma = (\Gamma_1, \ldots, \Gamma_M)$ be the M–vector defined by

$$\frac{1}{\Gamma_i} \triangleq 1 + \frac{W}{R_i \mathrm{SIR}_i}. \tag{3.6}$$

It is easy to check, starting from Equation 3.5, that

$$\Gamma_j = \frac{P_j \alpha_{jk}}{1 + \sum_{i=1}^{M} P_i \alpha_{ik}} \quad (j \in A_k). \tag{3.7}$$

It is also easy to see that Γ_i increases monotonically in the product $R_i \mathrm{SIR}_i$. One approach to $\mathcal{R}$ is *via* the set — call it $\mathcal{G}$ — of feasible Γ's, from which $\mathcal{R}$ is obtained by simple coordinate transformations.

Equation 3.6 for given α defines a nonlinear deformation $\Gamma = G(\mathrm{P}, \alpha)$ of $\mathcal{P}$ (G being a simple transformation of the g defined above). Since $G(\cdot, \alpha)$ is continuous, $\mathcal{G}$ is connected, closed and bounded. The shape of $\mathcal{G}$ can be inferred from the action of $G(\cdot, \alpha)$ on the $2M$ $(M-1)$–dimensional hyperplanes that together form the boundary

of $\mathcal{P}$. It is immediate that the coordinate hyperplane $P_j = 0$ is mapped into the coordinate hyperplane $\Gamma_j = 0$, $j = 0, \ldots, M$. So $\mathcal{G}$ is an M–dimensional solid in Γ–space bounded by the M coordinate hyperplanes and by the images under $G(\cdot, \alpha)$ of the M hyperplanes $P_j = p$. The latter remain to be determined. It can be seen from Equation 3.7 that $\mathcal{G}$, by definition contained within the unit M–cube, is strictly bounded by the hyperplanes

$$\sum_{j \in A_k} \Gamma_j = 1. \tag{3.8}$$

There are two simple special cases, corresponding respectively to $N = 1$, $N = M$. In the first case, all terminals speak to the same base station; in the second case, there is a unique correspondence between terminals and base stations, terminal (i) being associated with base station (i). Equation 3.7 in the first case becomes

$$\Gamma_j = \frac{P_j \alpha_j}{1 + \sum_{i=1}^{M} P_i \alpha_i} \quad j = 1, \ldots, M, \tag{3.9}$$

the second subscript in the α_{jk} being unnecessary. In the second case,

$$\Gamma_j = \frac{P_j \alpha_{jj}}{1 + \sum_{i=1}^{M} P_i \alpha_{ij}} \quad j = 1, \ldots, M. \tag{3.10}$$

In each of the two cases, we examine the transformation, under the mapping $G(\cdot, \alpha)$, of the hyperplane $P_j = p$. The discussion is specialized, for convenience and without loss of generality, to the sub–case $j = 1$.

The case $N = 1$:

The situation here is rather trivial and in a slightly different form has been considered elsewhere ([22]). There is a quickly accessible, closed–form solution to everything.

Replace P_1 by p in Equation 3.9, then solve the equations corresponding to $j \geq 2$ to get $P_2, \ldots, P_M$ *explicitly* in terms of $\Gamma_2, \ldots, \Gamma_M$. Apply that solution to the equation for Γ_1, eliminating $P_2, \ldots, P_M$ there in favour of $\Gamma_2, \ldots, \Gamma_M$; you get

$$\left(\frac{1+p\alpha_1}{p\alpha_1}\right)\Gamma_1 + \Gamma_2 + \cdots + \Gamma_M = 1, \tag{3.11}$$

which describes an $(M-1)$–dimensional hyperplane in Γ–space — the image under $G(\cdot, \alpha)$ of the set $P_1 = p$. The image of the set $P_j = p$ is described similarly. It emerges that $\mathcal{G}$ is an M–dimensional *convex polygon* (Fig. 3.3) illustrates for$M = 2$). The planarity of the bounding surfaces is unique to the case $N = 1$.

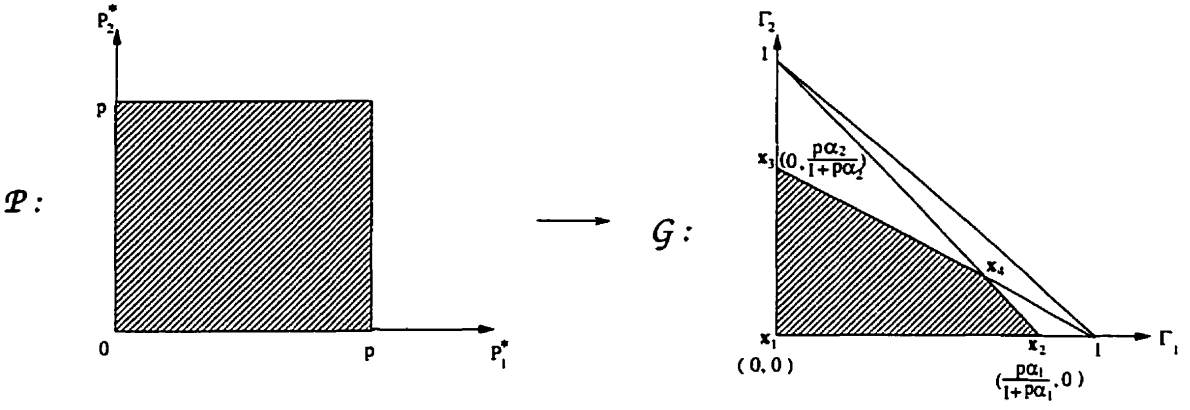


Fig. 3.3 Diagram of coordinate transformation from $\mathcal{P} \rightarrow \mathcal{G}$, M=2.

Remark: The (set–theoretic) convexity of $\mathcal{G}$ has immediate implications for system optimization. If the objective function $f(\cdot)$ is in fact a *convex* function of Γ, then it takes its maximum on the *extreme points* of $\mathcal{G}$. These in turn correspond to fixed points of $\mathcal{P}$, which are none other than the M–vectors whose coordinates are either 0 or p. So policies that are optimal relative to such an objective function are of the "bang–bang" variety — each of the M terminals transmits either at full power, or else not at all. Since for each i the rate R_i is convex in Γ_i, the sum-rate criterion

$$f(\Gamma) = R_1 + \cdots + R_M$$

(a sum of convex functions) is itself convex in the vector argument Γ, and thus a example of the sort of f to which the bang–bang characterization applies. Which of the $2^M - 1$ extreme points are indeed the optimal ones depends on the fading and noise coefficients; there are cases in which the sum of the rates is maximized by giving access to just one terminal, and denying access to all the others, a strategy which can be viewed as a static precursor to TDMA. In any event, the sum–rate criterion in a system where the terminals are capable of unbounded greediness (the traffic model assumed here) is hardly appropriate; the systems we have in mind are to be organized so as to provide adequate service for as many as possible, rather than excellent service for a few.

Remark: When the maximal power p is infinite, meaning that power is unconstrained, the hyperplane described by Equation 3.8 in fact forms the actual boundary of the closure of $\mathcal{G}$; that is, the condition $\sum_j \Gamma_j < 1$ is sufficient as well as necessary in order that Γ be feasible. In the more general case $N > 1$, the N conditions $\sum_{A_k} \Gamma_j < 1$ are necessary but *NOT* sufficient for all p — even for $p = \infty$.

The case $N = M$:

The program here is essentially the same as before, with Equation 3.10 replacing Equation 3.9 as the point of departure, except that the intermediate step — obtaining $P_2, \ldots, P_M$ in terms of $\Gamma_2, \ldots, \Gamma_M$ when $P_1 = p$ — cannot be carried out explicitly. Instead, you get

$$(P_2, \ldots, P_M)\,\Delta_1 = -\left((1 + p\alpha_{12}), \ldots, (1 + p\alpha_{1,M})\right),$$

where Δ_1 is the $(M-1) \times (M-1)$ matrix constructed from (α_{ij}) by deleting the first

row and column and by multiplying the diagonal elements α_{jj} $(j \geq 2)$ by

$$\left(1 - \frac{1}{\Gamma_j}\right).$$

The equation in (3.10) corresponding to $j = 1$, $P_1 = p$ can be written in the form

$$\left(\frac{1 + p\alpha_{11}}{p\alpha_{11}}\right)\Gamma_1 + \sum_{i=2}^{M} \Gamma_1 \cdot \frac{P_i \alpha_{i1}}{p\alpha_{11}} = 1.$$

In the previous case ($N = 1$), the i-th summand on the RHS is just Γ_i, which is exactly what makes the surface described by this equation planar; in the present case, the summands are complicated nonlinear, non-explicit functions of $\Gamma_2, \ldots, \Gamma_M$. In the simplest case $M = N = 2$, the bounding surfaces work out to be hyperbolic and $\mathcal{G}$ is accordingly convex. We have not attempted to determine whether or not the convexity property is found in more general settings as well.

Remark: The situation when $M > N$ ($M < N$ does not happen) is essentially the same as when $M = N$, except that the computational burden rises rapidly with M.

3.7 Dynamic Rate Control

The $2M$ constraints on power and SIR together determine the admissible rate region $\mathcal{R}$. It remains to decide how a particular rate vector $R \in \mathcal{R}$ is to be selected. Our approach, as described above, is to allow the dynamics of the congestion processes along the various paths — processes that in our simplified model are represented by the backlogs in the interface buffers — to guide the choice of R. The variation of backlog in the buffer associated with the i-th flow is a function of the buffer input process, the depletion rate process and the buffer length. The first of these is just the packet stream coming off the radio channel, as determined by the medium access control. The depletion rate process, defining the instantaneous rate at which data leaves the buffer, is determined by the dynamics of the wireline segment of

the connection, which in turn are constrained by QoS commitments agreed upon at the time that the connection was established. The symbols B_i, $\mu_i(t)$, introduced in Section (3.2), denote, respectively, the length of the i–th buffer and its instantaneous output rate at time t. In our simulations, reported in Chapter 4, $\mu_i(\cdot)$ is random, its statistics constructed in emulation of the ABR (Available Bit Rate) service class. We assume in any case that the long–run *mean* depletion rate $\bar{\mu}_i$ is a fixed, known attribute of the connection. As noted above, the parameters B_i, $\bar{\mu}_i$, in the presence of medium access control designed to avoid buffer overflow and underflow, can be viewed as defining the quality–of–service commitment relative to Source (i).

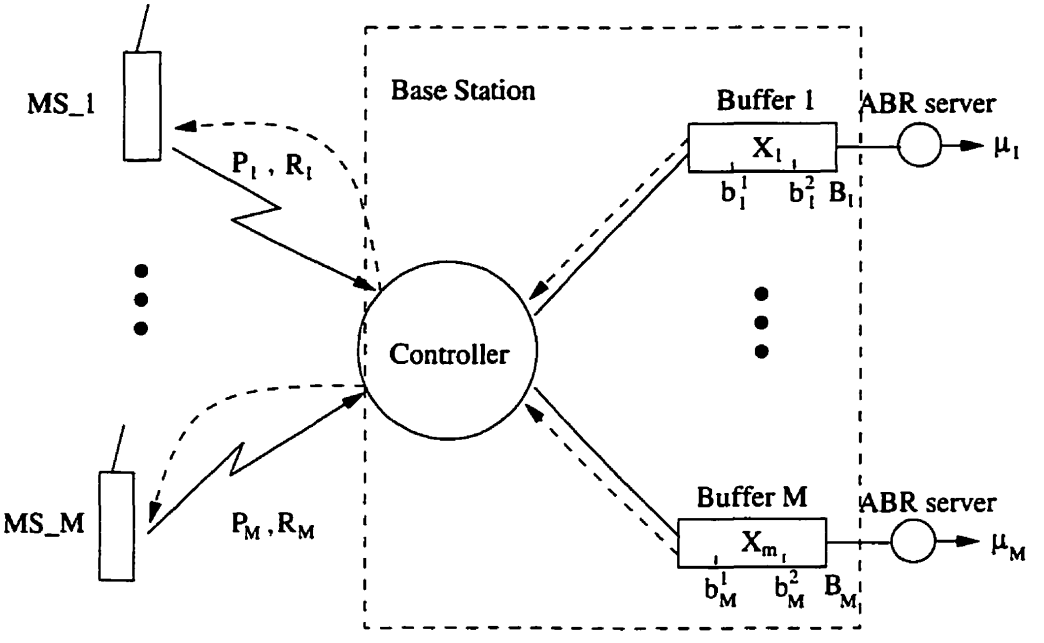


Fig. 3.4 System Model

The system we wish to control is shown in Fig. 3.4. System operation is assumed *slotted*, meaning that rate allocations are updated on a strictly periodic basis, remaining constant between updates. A slot, or frame, is the interval between updates. The fading coefficients α_{ik} are assumed constant as well over one slot and known to the controller; the slots should thus be small enough that the assumption is approximately valid. Recall that the packets in our model are infinitely small and unrelated

to slot size.

Remark: Medium access control in the system we have in mind is implemented in the form of rate modulation, which in turn is effected through power control (see Equation 3.9). We see two kinds of power control in force, undertaken at different time scales and for different purposes. *Slow* power control, enacted in every slot (at the beginning) is for medium access control; *fast* power control, acting almost continuously, counters the effect of fluctuations in fading that occur within a slot. The (almost) continuous exchange of data between terminals and base station for purposes of *fast* power control provides the basis for the estimation of the fading coefficients at the slot boundaries.

The control problem is described by its *state variables*, its *control variables* and its *performance variables*. The state variables are the buffer backlog processes $x_i(\cdot)$, $i = 1, \ldots, M$. The control variables are the rates $R_1, \ldots, R_M$, computed and updated at slot boundaries; in light of the relationship between R and P implied by the SIR constraints, the power allocation $P_1, \ldots, P_M$ forms an essentially *equivalent* set of control variables[2] The performance variables are the long–run throughputs achieved, and the corresponding data loss rates due to buffer overflow.

The control we propose is presented in two stages. In the first stage, we compute a rate allocation R designed exclusively to avoid underflow and overflow. In the second stage, we show how to correct it so as to ensure compliance with the power and SIR constraints.

Stage (1):

The algorithm is described parametrically. The parameters corresponding to flow (i) are denoted b_i^1, b_i^2, $\bar{\mu}_i$ and w. The first two are backlog thresholds:

[2]Assuming that the assignment of terminals to base stations is given; in the case of multiple base stations, the computation of powers from rates involves an additional step — the optimization of the terminal assignments.

$$0 \leq b_i^1 \leq b_i^2 \leq B_i,$$

where B_i is the length of the base station buffer associated with flow (i). The parameter $\bar{\mu}_i$ refers to the mean rate at which the base station buffer is depleted; its value is fixed at the time that the connection is set up, and remains unchanged for the duration of the call. The parameter w, as will be seen, determines the sensitivity of the rate allocation to fading.

Consider a particular slot. Abusing notation, we write x_i for the flow (i) backlog in the base station buffer, as measured at the beginning of the slot. The effect of the backlog parameters b_i^1, b_i^2 is to partition the buffer into three regions: Region (I) corresponds to $x_i < b_i^1$; Region (II), to $b_i^1 \leq x_i \leq b_i^2$; and Region (III), to $x_i > b_i^2$. Define

$$\bar{\alpha} \triangleq \frac{1}{M} \sum_1^M \alpha_i, \qquad Q_i \triangleq \left(\frac{\alpha_i}{\bar{\alpha}}\right)^w \bar{\mu}_i.$$

The rate assigned to flow (i) in the slot in question depends on which of the three backlog regimes just described in fact prevails at the beginning of the slot:

- **Region (III):** In this case, flow (i) is assigned *zero* rate in the slot.
- **Region (II):** The allocation in this case is $R_i' \triangleq Q_i$, where Q_i is defined above.
- **Region (I):** The assigned rate in this case, selected to avoid underflow, varies approximately inversely with backlog. The particular form of the variation selected in our simulations is given by

$$R_i' \triangleq \begin{cases} Q_i \cdot \dfrac{b^2}{x_i} & \frac{1}{10}b^1 \le x_i \le b^1 \\ \\ Q_i \cdot \dfrac{10b^2}{b^1} & 0 \le x_i \le \frac{1}{10}b^1. \end{cases}$$

The quantity R_i' so specified is the *tentative* rate assigned to flow (i) for the duration of the slot. Fig. 3.5 shows the variation of R_i' with x_i in the particular case that the three regions are of equal length; that is, corresponding to $b_i^1 = b_i^2/2 = B_i/3$.

Observe that the SIR and power constraints have so far played no part in the rate computation. In particular, the allocation $R' = (R_1', \ldots, R_M')$ need not be admissible in the sense of belonging to $\mathcal{R}$. The purpose of the second stage of the algorithm, described next, is to revise the computation where necessary to ensure compliance with the constraints.

Stage (2):

Start with the tentative allocation R' produced by Stage (1). Check admissibility by calculating the corresponding Γ (*via* Equation 3.6), and then (recalling Equation 3.11) verify

$$\left(\frac{1+p\alpha_i}{p\alpha_i}\right)\Gamma_i + \sum_{j\neq i}\Gamma_j \leq 1$$

for $i = 1, \ldots, M$. Provided that each of the M inequalities is valid, R' is admissible and

$$R \triangleq R'$$

defines the rate allocation in the present slot as determined by the base station. If one or more of the inequalities *fails*, meaning that R' falls *outside* of $\mathcal{R}$, then we successively decrement the various R_i''s, testing at each step for admissibility and stopping at the first iteration which is admissible.

There are many different ways in which to decrement the components of R' in pursuit of admissibility. We chose to proceed indirectly, in terms of Γ rather than R', simply because our test for admissibility has been formulated in terms of Γ. Recall from Equation 3.6 that Γ_i is a simple, increasing function of R_i. Our algorithm is expressed in terms of a single "step–size" parameter $a \in (0,1)$. A single iteration of the algorithm is described by the recursion

$$\Gamma_i \longleftarrow (1-a^{r_i})\,\Gamma_i, \quad i = 1, \ldots, M, \tag{3.12}$$

where r_i denotes the *rank* of terminal (i) when the Γ's are sorted in *ascending* order.

The proportional change in Γ_i thus varies *inversely* with tentative rate, being *largest* for that terminal whose tentative rate is *smallest* and *vice versa*. The reasoning is as

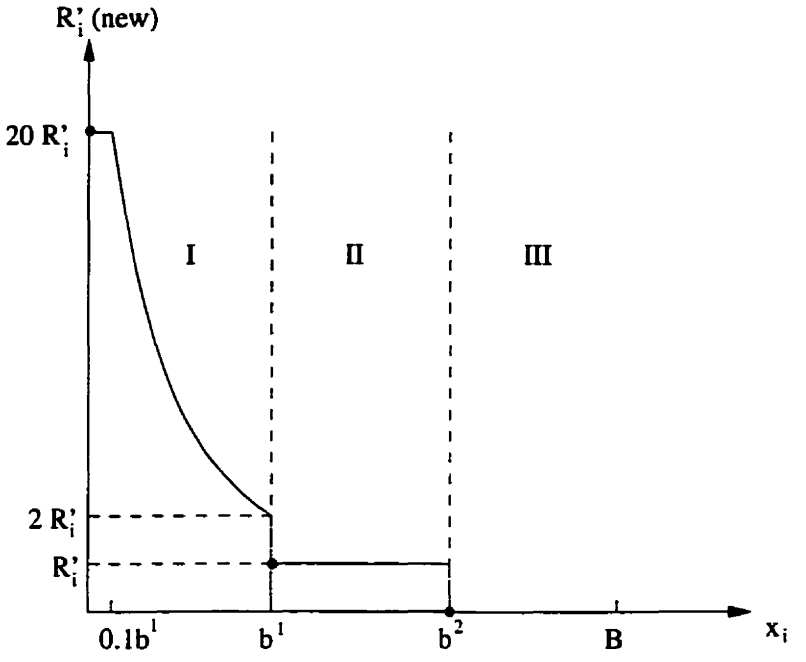


Fig. 3.5 Backlog feedback on $\mathbf{R}'$ assignment.

follows: a terminal whose tentative allocation is small is one for whom buffer overflow is a more imminent possibility than buffer underflow; for such a terminal, network performance, as indicated by average rates over the last little while, is likely to be less sensitive to bandwidth penalties that flow from the SIR and power constraints, than for a terminal whose average–rate performance has been less favorable. We view our approach as taking from the rich to give to the poor.

The actual value of the parameter a is chosen so as to achieve an acceptable compromise between level of resolution and the number of iterations required.

Chapter 4

Performance Analysis

In this chapter, we investigate the numerical results by selecting different control parameters. First the simulation model is presented, and interaction between call-level admission and burst-level performance is investigated.

4.1 Simulation Model

For the purpose of the simulation, the fading coefficient vector $\boldsymbol{\alpha}$ and the outgoing depletion rates, $\boldsymbol{\mu}$, are generated randomly on a time-slot basis.

4.1.1 Log-normal Shadowing

In general, the propagation models for radio channel can be categorized into: small-scale fading and large scale shadowing. Small-scale fading, normally modeled by Rayleigh or Ricean distribution, are typically measured in units of μs. The size of one time-slot, however, is chosen at 20 ms in our simulation. Because in our simulations we assume that α is in fact known at the start of each slot, the actual distribution of α has no role in the detailed execution of the algorithm. For the sake of simplicity, we ignore the variation of $\boldsymbol{\alpha}$ within one time-slot and simulate it following a Log-normal shadowing distribution at every 20 ms. Then, α_j has a Gaussian distribution related to the mean path-loss$\overline{PL}(d)$, that is:

$$-10log_{10}\alpha = \overline{PL}(d) + X_\delta, \tag{4.1}$$

where X_δ is a zero-mean Gaussian random variable with standard deviation δ in units of dB.

4.1.2 The ABR Service Model

The ABR service model mentioned in Chapter 2 is implemented for the depletion process in our research. Specifically, the *relative rate marking* scheme, in which the serving rate is mediated by the CI and NI bits in the RM cell, has been used. Fig. 4.1 shows the flow diagram of the ABR rate assignment.

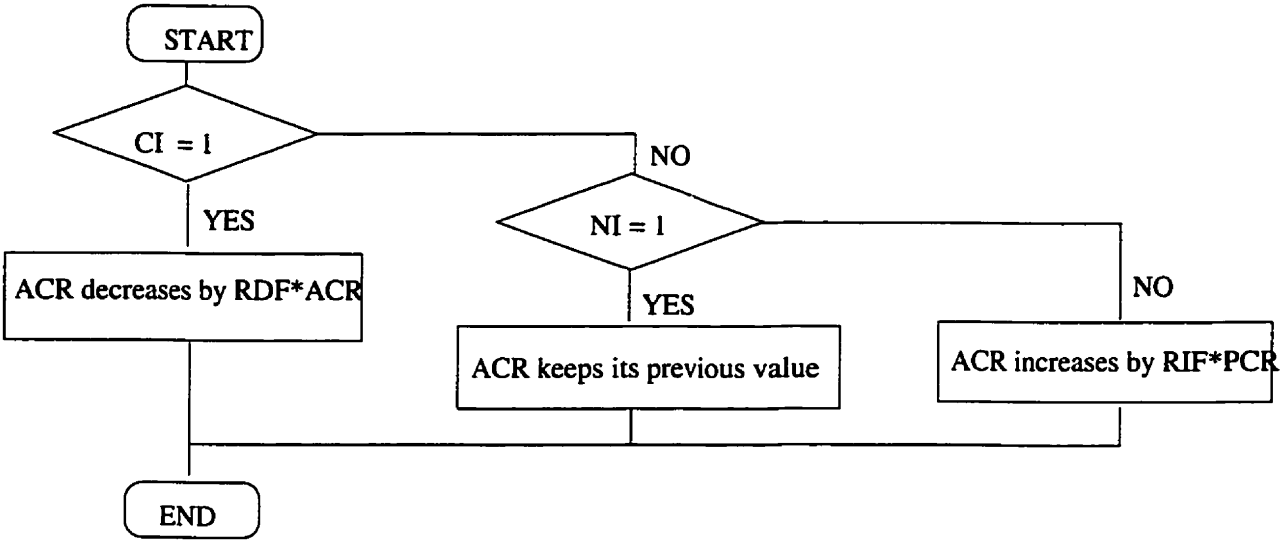


Fig. 4.1 Flow diagram of ABR rate assignment

In the simulation, we set the MCR equal to 0. CI and NI are generated by uniformly distributed random variables — Pr{CI=0} = Pr{CI=1} = 0.5, and similarly for NI. We can then determine the PCR according to $\overline{\mu}$, the mean of ACR:

$$\mu_{n+1} = \begin{cases} \mu_n - RDF \times \mu_n, & CI = 1; \\ \mu_n, & CI = 0, NI = 1; \\ \mu_n + RIF \times PCR, & CI = 0, NI = 0. \end{cases}$$
$$\Rightarrow \quad PCR = 2 \cdot \overline{\mu} \qquad \text{if } RDF = RIF \tag{4.2}$$

4.1.3 Rate Matching

The remaining simulation problem is how to determine a sensible depletion mean rate $\overline{\mu}$. There are two aspects to this in terms of rate matching. One is flow control at the MAC layer. The other is the system capacity at the network layer. For the first aspect, we are interested in the rate tracking ability of the MAC algorithm. The initial value of the mean depletion rate is determined by negotiation between the user and the network. For the second aspect, the depletion rate is predetermined by the network from its congestion status. Here the MAC algorithm works to serve as many mobile users as possible with fixed outgoing bandwidth to the wireline network.

Call Admission Control with Rate Tracking

We now focus on the initial bandwidth allocation in terms of rate tracking. M mobile terminals send out connection request messages through the uplink *Random Access Channel* (RACH) using the slotted ALOHA scheme. The message includes the SIR threshold and the transmit power each terminal is using. The latter is used to determine the fading coefficient. The base station then establishes uplink *Dedicated Data Channels* (DDCH) for the data transmission and downlink *Dedicated Control Channels* (DCCH) for the rate and power assignment for each mobile terminal.

The base station regards the initial fading condition as the average level of the radio channel: $\overline{PL}(d)$ is the difference between the received power (in dB) and the transmit power (again, in dB) in the connection request message. Together with the SIR criterion provided in the message, the base station determines the average service rate $\overline{\mu_i}$ for each mobile terminal. We assume no rate restraint from the mobile source and that the terminal always has data to transmit. The following describes the

procedure for assigning $\overline{\mu_i}$ at the call-level admission control:

STEP 1 Given the fading coefficients α_i and SIR_i for i=1,...,M, the base station computes $\mathbf{\Gamma}$. Power is assumed to be at peak and hence $\mathbf{\Gamma}$ is an extreme point in the admissible region.

STEP 2 Applying the inverse function of Eq. (3.6), we obtain

$$R_i = \frac{W}{SIR_i} \cdot \frac{\Gamma_i}{1-\Gamma_i} \qquad i = 1, ..., M \tag{4.3}$$

This defines a rate vector, denoted $\boldsymbol{\lambda}_0$, which is admissible and extremal in the sense of corresponding to our extremal power assignment.

STEP 3 We then define

$$\overline{\mu_i} \triangleq c \cdot \lambda_{0i}, \tag{4.4}$$

Where c is chosen large enough that throughput is reasonably high and low enough that the wireless segment can track the depletion rate.

4.1.4 Choice of Parameters

The system parameters used in the simulation are summarized in Table. 4.1. The control variables are: $\overline{\mu}$ — the mean service rate; B — the buffer length; b^1 — the lower backlog threshold; b^2 — the upper backlog threshold; w — the power index; and a — the convergence step factor.

We assume that traffic storage in the buffers is *stream-type* as opposed to *packet-type*. For the sake of simplicity, the SIR criteria and the mean path-loss $\overline{PL}$ are the same for all mobile users. Therefore, we expect the same control parameters for each connection as well.

The simulation starts with empty backlogs in the buffer and runs for a duration of 2000 time-slots. Our goal for *rate tracking* is to balance efficiency and fairness. The

Table 4.1 System Parameters

Item	Symbol	Value
Bandwidth (MHz)	W	1.25
Time slot (msec)		20
Noise spectral density	No	10^{-8}
Packet size (bits)		1024
Peak transmit power (mW)	p	200
Signal-to-Interference ratio (dB)	SIR	8.5
Simulation time (time-slot)		2000
Data message arrival rate	λ	variable
Log-normal fading channel		
Log-normal channel path-loss mean (dB)	$\overline{PL}$	5
Log-normal channel standard deviation (dB)	X_δ	5
ABR-type outgoing traffic		
Number of cells between RM cells	Nrm	16
Rate increase factor	RIF	1/16
Rate decrease factor	RDF	1/16
Minimum cell rate	MCR	0
Allowed cell rate	ACR	μ, variable
Peak cell rate	PCR	$2\overline{\mu}$

results and analysis will be presented in Section 4.2. In Section 4.3, we compare the numerical performance with the standard IS-95 system and with another system model aimed at maximizing total throughput. The gains with respect to throughput and transmit power will be discussed. We then investigate the *system capacity*. Interactions between the control parameters and the wireline traffic rate will be addressed.

4.2 Rate Tracking

The control variables — $\overline{\mu}, B, b^1, b^2, w, a$ — form a 6-dimensional parameter space. We explore the effect of each variable separately. The convergence step factor a mainly

concerns computational complexity; we set it to be 0.2 for this experiment. Second, we set the buffer length B to a large value; its impact will be discussed in section 4.2.2. Now, we can concentrate on the interaction between (b^1, b^2, w) and $\overline{\mu}$.

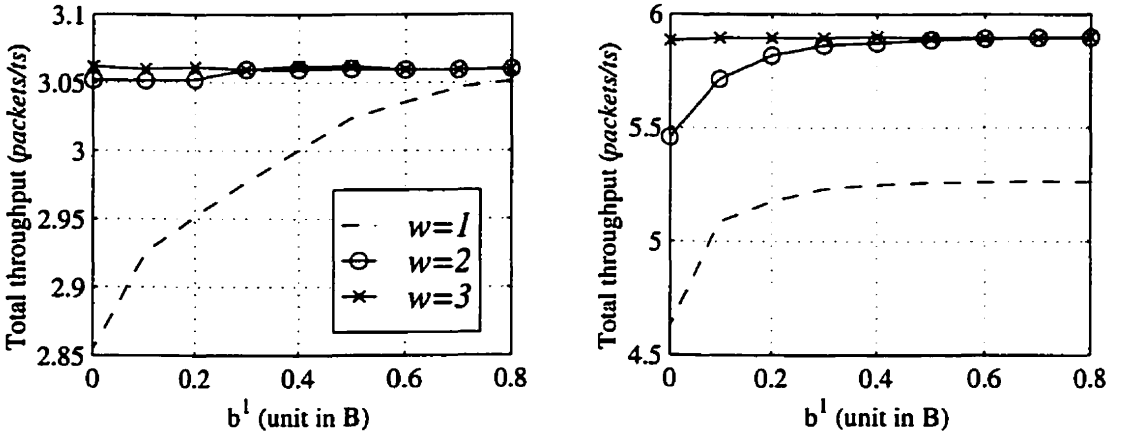
4.2.1 Choosing the Control Parameters $b^1, b^2, w, \overline{\mu}$

The goal, roughly stated, is to give a kind of rate advantage to terminals with light fading, while maintaining service of an appropriate lower level for terminals experiencing strong fading. Packet loss-rate is not an issue at this point because the buffer length B is set sufficiently large. The approximate bandwidth consumed by the ABR traffic is determined at the call-level setup stage by $\overline{\mu}$. We first fixed $b^2 = 0.8B$ and explored the relation between b^1 and $\overline{\mu}, w$, then fixed b^1 and observed the impact of changing b^2.

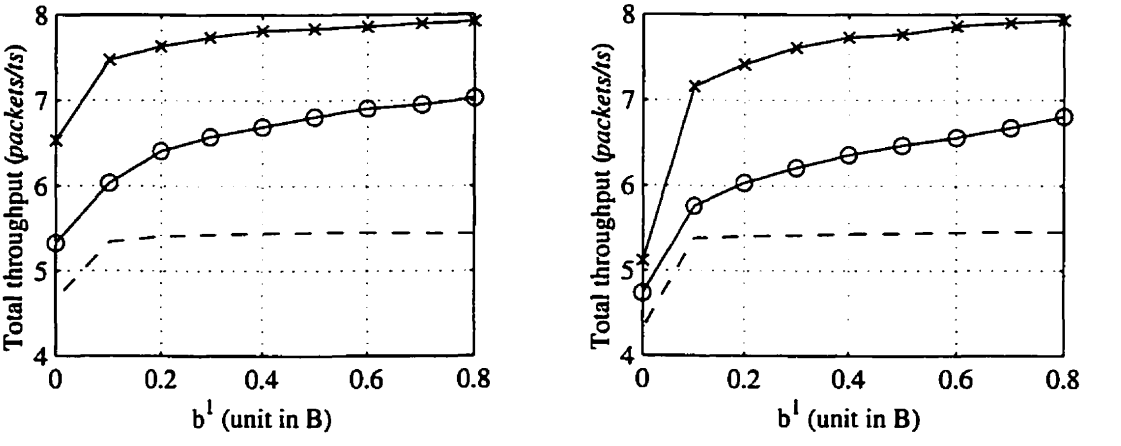
Relation between b^1 and $\overline{\mu}, w$

The first experiment runs for 2 to 5-users. The numerical results are given in Fig. 4.2 – Fig. 4.5 and Table. 4.2 – Table. 4.5. From the figures, it can be seen that for $w = 1$, the total throughput depends greatly on the lower backlog boundary b^1. This is especially so when low values of initial $\overline{\mu}$ are chosen. In contrast, when $w = 2$ or $w = 3$, the performance is almost the same and changing b^1 does not influence the total throughput much. It can be explained thus. When $w = 1$, the rate adjustment cannot be handled merely by the fading conditions; instead, b^1 is a necessary control parameter, When $w > 1$, the algorithm can better distinguish good channels from bad channels and allows mobiles to transmit data with full awareness of their current good fading conditions. In the latter case, the system can choose a small b^1 to allow a large feedback-free domain, where it assigns bandwidth simply according to the fading coefficients. For a proper b^1, it can be chosen at any values on the flat part of the curves in Fig. 4.2 – Fig. 4.5.

The next question is how to choose $\overline{\mu}$ when w is fixed at 2 or 3. This can be answered by Table. 4.2 – Table. 4.5. The columns illustrate the actual ABR mean rates and the performance gains when the estimated bandwidth $\overline{\mu}$ is determined to be 1/2, 1, 1.5 and 2 times the peak arrival rate λ_0 at the initial stage. The $\overline{\mu}$ being larger, the

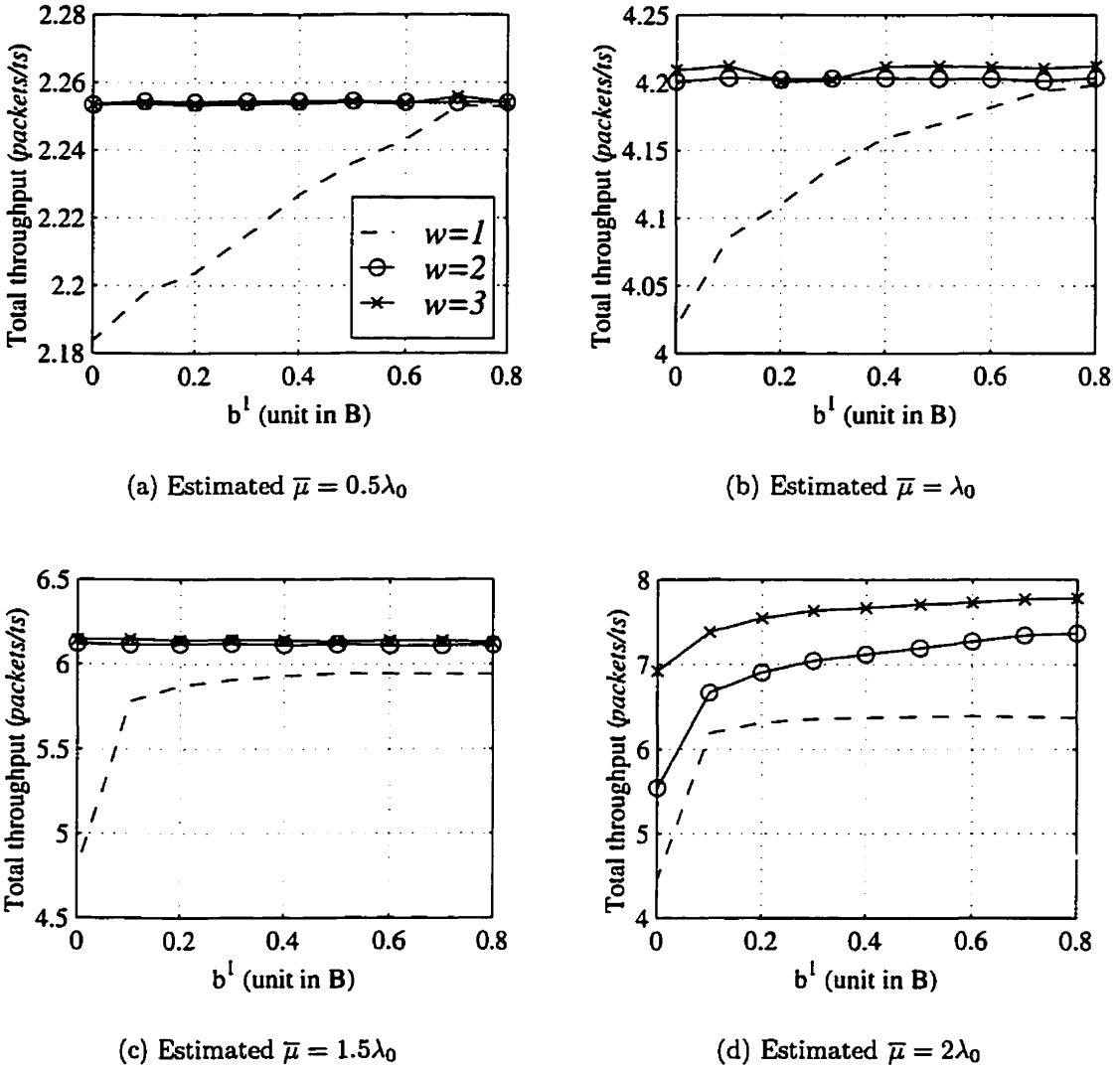


(a) Estimated $\overline{\mu} = 0.5\lambda_0$ (b) Estimated $\overline{\mu} = \lambda_0$



(c) Estimated $\overline{\mu} = 1.5\lambda_0$ (d) Estimated $\overline{\mu} = 2\lambda_0$

Fig. 4.2 Sensitivity to $\overline{\mu}$ and (b^1, w) for 2 user case.
B=250, a=0.2, b^2=0.8B



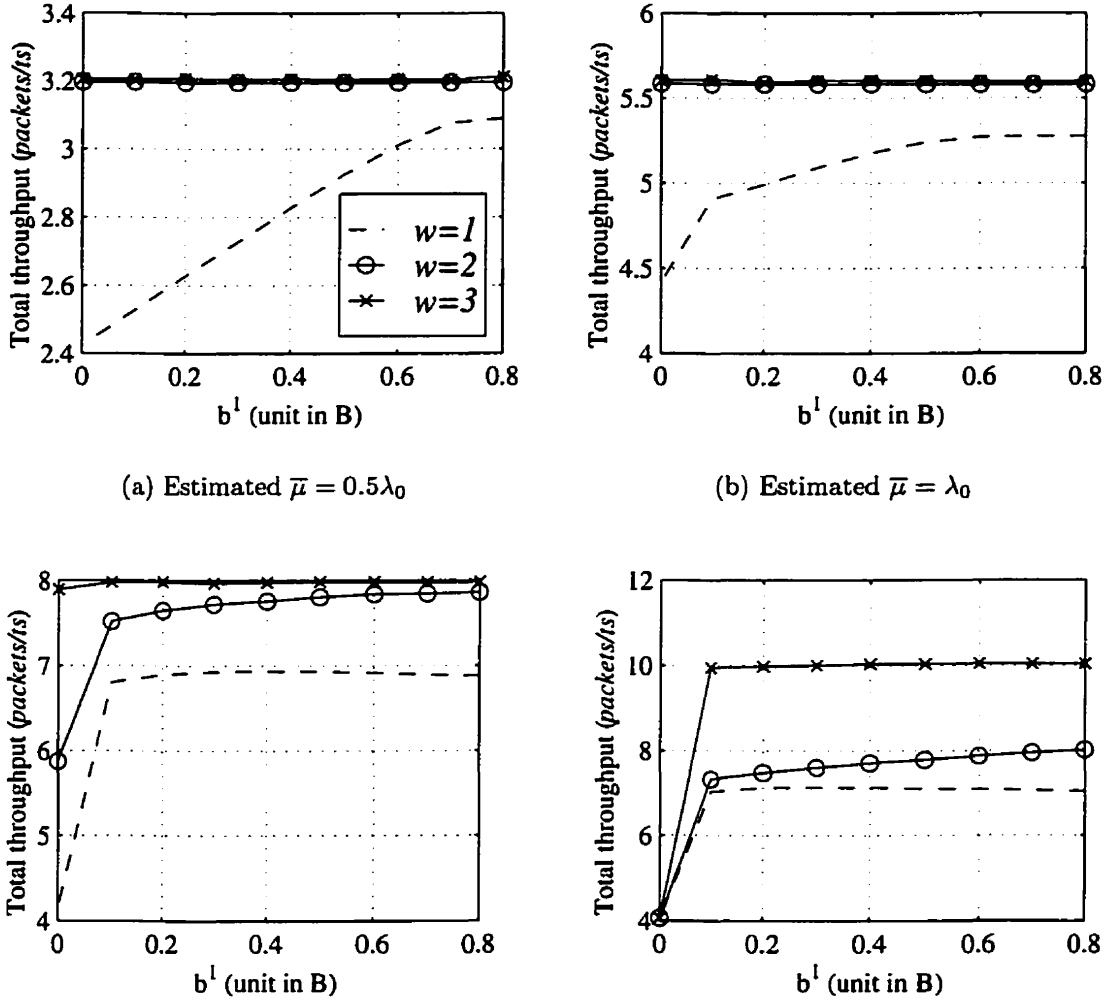
(a) Estimated $\overline{\mu} = 0.5\lambda_0$

(b) Estimated $\overline{\mu} = \lambda_0$

(c) Estimated $\overline{\mu} = 1.5\lambda_0$

(d) Estimated $\overline{\mu} = 2\lambda_0$

Fig. 4.3 Sensitivity to $\overline{\mu}$ and (b^1, w) for 3 user case.
B=250, a=0.2, b^2=0.8B



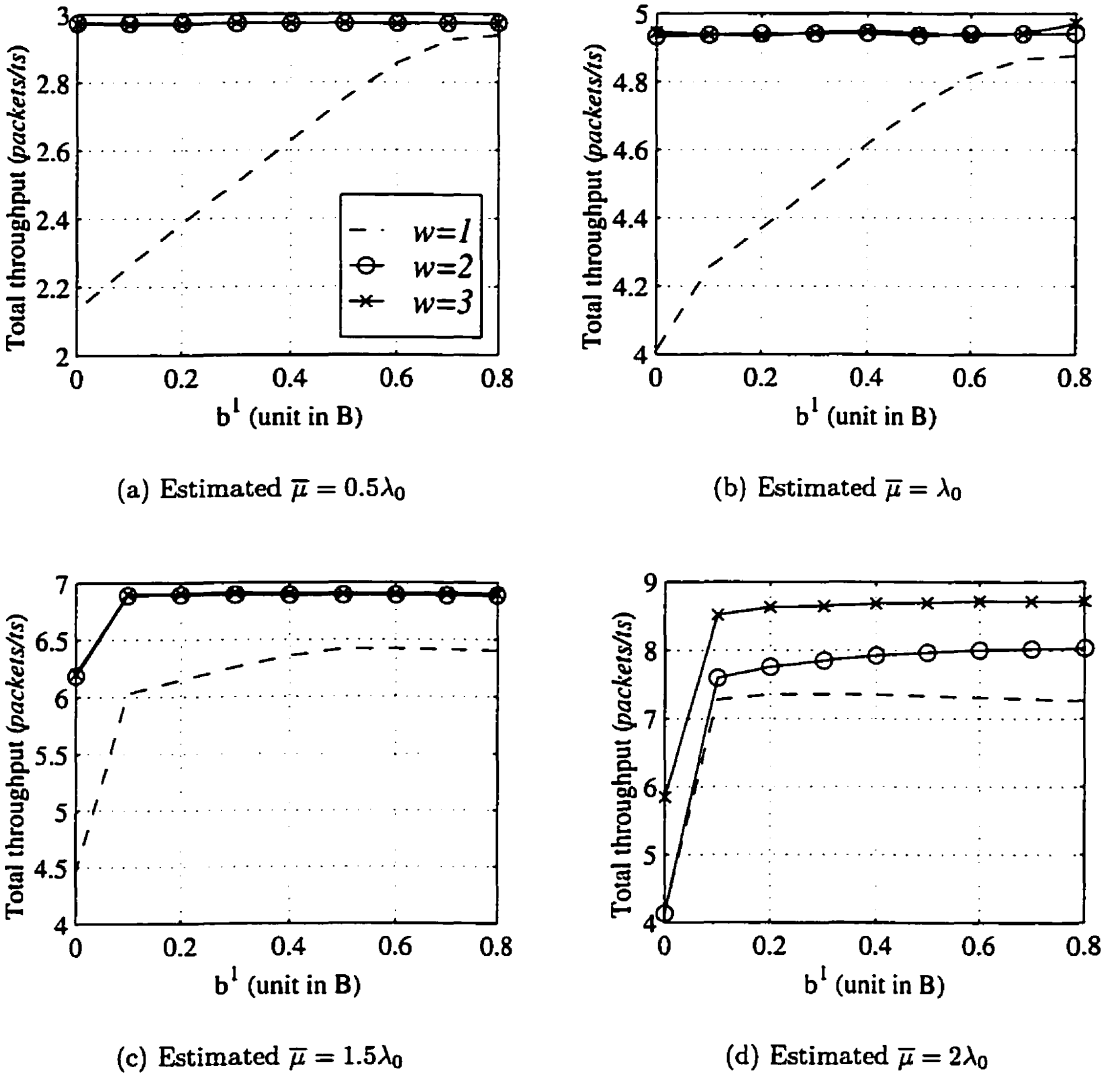
(a) Estimated $\overline{\mu} = 0.5\lambda_0$

(b) Estimated $\overline{\mu} = \lambda_0$

(c) Estimated $\overline{\mu} = 1.5\lambda_0$

(d) Estimated $\overline{\mu} = 2\lambda_0$

Fig. 4.4 Sensitivity to $\overline{\mu}$ and (b^1, w) for 4 user case.
B=500, a=0.2, b^2=0.8B



(a) Estimated $\overline{\mu} = 0.5\lambda_0$

(b) Estimated $\overline{\mu} = \lambda_0$

(c) Estimated $\overline{\mu} = 1.5\lambda_0$

(d) Estimated $\overline{\mu} = 2\lambda_0$

Fig. 4.5 Sensitivity to $\overline{\mu}$ and (b^1, w) for 5 user case.
B=500, a=0.2, b^2=0.8B

Table 4.2 Performance gains for 2-user case.
Performance gain $\stackrel{\text{def}}{=}$ Total throughput/Actual $\overline{\mu}$
$b^1 = b^2 = 0.8B, B = 250 packets, a = 0.2$

Estimated $\overline{\mu}$ at call setup stage	$0.5\lambda_0$	λ_0	$1.5\lambda_0$	$2\lambda_0$
Actual $\overline{\mu}$ (packets/ts)	2.8615	5.7230	8.5845	11.4460
Performance gain ($w = 2$)	1.07	1.03	0.82	0.59
Performance gain ($w = 3$)	1.07	1.03	0.92	0.69

Table 4.3 Performance gains for 3-user case.
$b^1 = b^2 = 0.8B, B = 250, a = 0.2$.

Estimated $\overline{\mu}$	$0.5\lambda_0$	λ_0	$1.5\lambda_0$	$2\lambda_0$
Actual $\overline{\mu}$ (packets/ts)	1.9545	3.9090	5.8635	7.8180
Performance gain ($w = 2$)	1.15	1.07	1.04	0.94
Performance gain ($w = 3$)	1.15	1.08	1.04	1.00

Table 4.4 Performance gains for 4-user case.
$b^1 = b^2 = 0.8B, B = 500, a = 0.2$.

Estimated $\overline{\mu}$	$0.5\lambda_0$	λ_0	$1.5\lambda_0$	$2\lambda_0$
Actual $\overline{\mu}$ (packets/ts)	2.399	4.798	7.197	9.596
Performance gain ($w = 2$)	1.33	1.17	1.09	0.84
Performance gain ($w = 3$)	1.33	1.17	1.11	1.05

Table 4.5 Performance gains for 5-user case.
$b^1 = b^2 = 0.8B, B = 500, a = 0.2$.

Estimated $\overline{\mu}$	$0.5\lambda_0$	λ_0	$1.5\lambda_0$	$2\lambda_0$
Actual $\overline{\mu}$ (packets/ts)	1.9745	3.9490	5.9235	7.8980
Performance gain ($w = 2$)	1.51	1.25	1.16	1.02
Performance gain ($w = 3$)	1.51	1.25	1.16	1.10

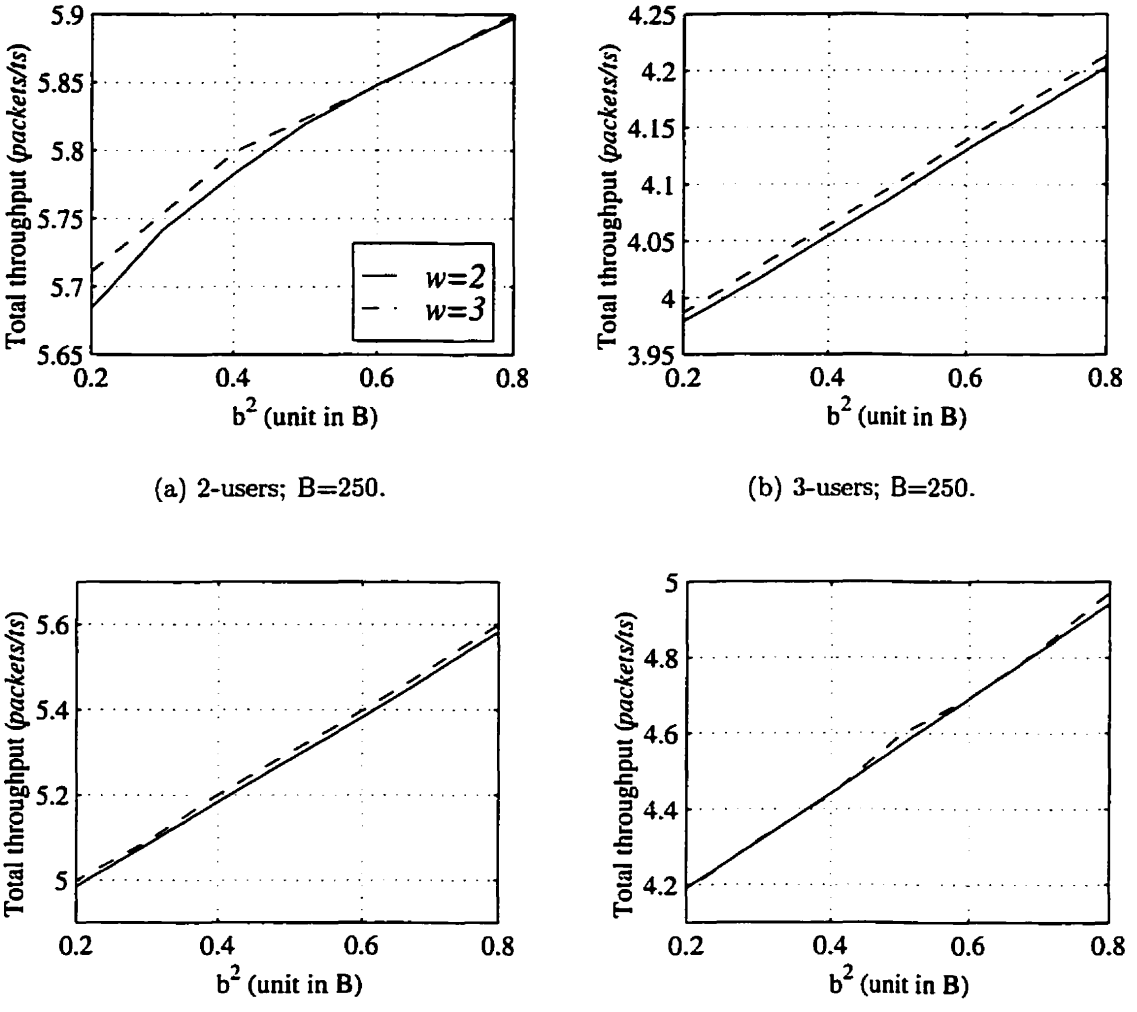
corresponding Γ' transformed from $\boldsymbol{R}'$ is further away from the admissible region $\mathcal{G}$, and therefore after adjustment from Stage (2) (see section 3.7), the final transmit rates granted by the base station are much lower than $\boldsymbol{R}'$. This is especially true when $2\lambda_0$ is applied for the 2-user case (Table. 4.2). This time, the adjustment to the admissible region leads to a bottleneck effect that brings a low performance gain. In other words, the power constraint and the existence of fading make it impossible for the mobile users to match the outgoing rates, a difficulty which cannot be solved by any MAC algorithm.

Another interesting phenomenon is observed when the population of mobile terminals increases, causing the performance gains to have a stable rise. Since interference is a big issue in a CDMA system, the fewer the connections in a cell, the more significant the contribution of each user to the interference. This gives us a simple rule: if the factor c is set to a value between $\overline{\mu}$ and λ_0 so that the algorithm works for the 2-user case, the same c works for multiple users. As can be seen in Fig. 4.2 and Table. 4.2, the largest $\overline{\mu}$ can be assigned in the call setup stage and is therefore equal to $1.5\lambda_0$.

How to choose b^2

To investigate sensitivity to the control parameter b^2, Figs. 4.6- 4.7 illustrate the total throughput versus b^2 under $\overline{\mu} = \lambda_0$ and $\overline{\mu} = 1.5\lambda_0$. It shows that by increasing the upper boundary b^2, the throughput increases almost linearly (except for Fig. 4.7(a)). Furthermore, the change of w from 2 to 3 does not influence the performance.

Decreasing b^2 makes for smaller backlogs. This prevents overflow, but can lead to inefficiency due to underflow. Since b^2 is designed to prevent overflow by cutting off bandwidth when backlogs exceed it — $x_i \geq b^2$ — call it the "dangerous" region, it can ideally be set close to the buffer limit — B. However, the operation in the "safe" region — $b^1 \leq x_i < b^2$ — may exceed the buffer length because the operator is not aware of the upper boundary. In conclusion, the increase of b^2 will bring better throughput and buffering efficiency, however the tradeoff is that it enhances the possibility of packet loss. Therefore, the highest we can set it at is 0.8B, and it should be dynamically adjusted according to the buffer size.



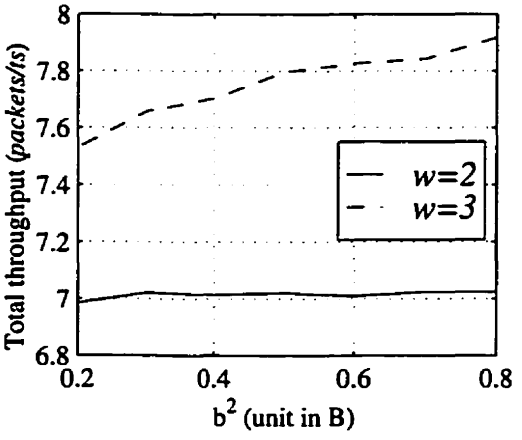
(a) 2-users; B=250.

(b) 3-users; B=250.

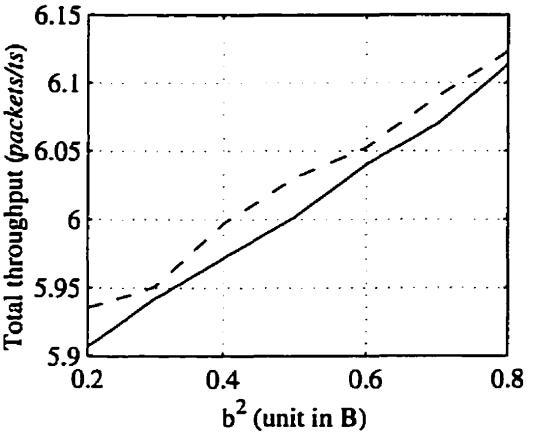
(c) 4-users; B=500.

(d) 5-users; B=500.

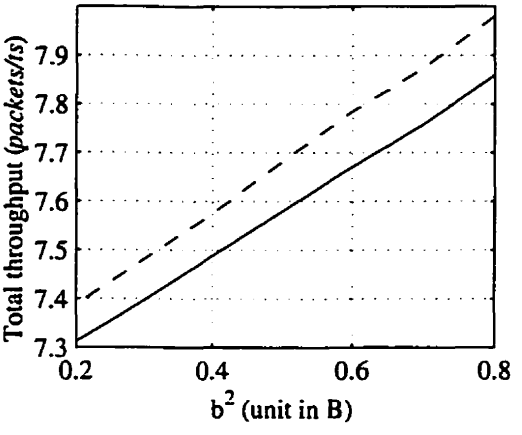
Fig. 4.6 Sensitivity to b^2 and $\overline{\mu}$, w for $\overline{\mu} = \lambda_0$ $a = 0.2, b^1 = b^2$.



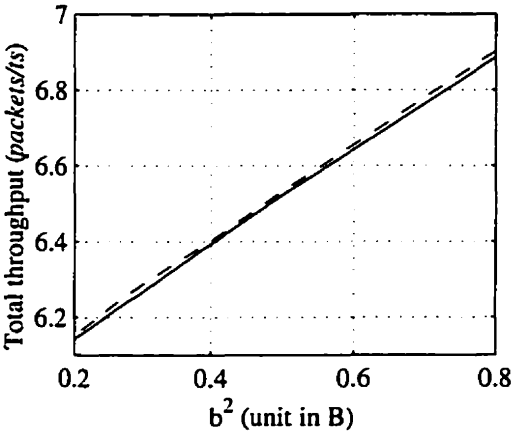
(a) 2-users; B=250.



(b) 3-users; B=250.



(c) 4-users; B=500.



(d) 5-users; B=500.

Fig. 4.7 Sensitivity to b^2 and $\overline{\mu}$, w for $\overline{\mu} = 1.5\lambda_0$. $a = 0.2, b^1 = b^2$.

4.2.2 Choosing $\overline{\mu}$ and w

The previous section discussed the interaction between b^1, b^2, w and $\overline{\mu}$ determined how to choose backlog thresholds b^1 and b^2. The remaining work explores the impact of buffer length B with $\overline{\mu}$ and w, and seeks to rationalize the choice of these parameters.

The Parameter w

Figs. 4.8–4.11 compare performance for two settings of w and for various B. obviously, a larger buffer increases data throughput and prevents packet-loss. The choice of $w = 3$ is immediately out of the question. As shown in Fig. 4.9 and 4.11, the minimum buffer size required for $w = 3$ is much larger than that of $w = 2$ in order to prevent packet loss, while there is no throughput improvement by setting $w = 3$ instead of 2.

The parameter $\overline{\mu}$

The next question is how to choose B and $\overline{\mu}$. This can be answered by a closer look at the figures. The tradeoff between B and $\overline{\mu}$ is that when $\overline{\mu}$ rises, more buffering is required to prevent packet-loss. For example, when w is set at 2, and $\overline{\mu}$ goes from λ_0 to $1.5\lambda_0$, the minimum size of buffer required to prevent overflow rises from 45 packets to 85, 60 to 80, 75 to 100, and 90 to 125, for two, three, four and five-users respectively. This is equal to an increase in buffer size of 1.89, 1.23, 1.23, and 1.39 times. Consequently, the actual throughput is enhanced from 5.64 packets/time-slot to 7.03, 3.975 to 5.925, 4.905 to 7.28, and 4.13 to 6.15, which corresponds to an actual throughput increase of 1.25, 1.49, 1.48, and 1.49 times. As the increase in the throughput is larger than that of the buffer required, we still benefit from raising the admission control variable $\overline{\mu}$ and therefore would choose $\overline{\mu} = 1.5\lambda_0$ except in the 2-user case where the performance gain is less than one when $1.5\lambda_0$ is chosen for $\overline{\mu}$.

4.3 System Capacity

We now move on to investigate the system capacity. Performance is compared with that of the *peak-power* algorithm (PPA) and of the IS-95 standard.

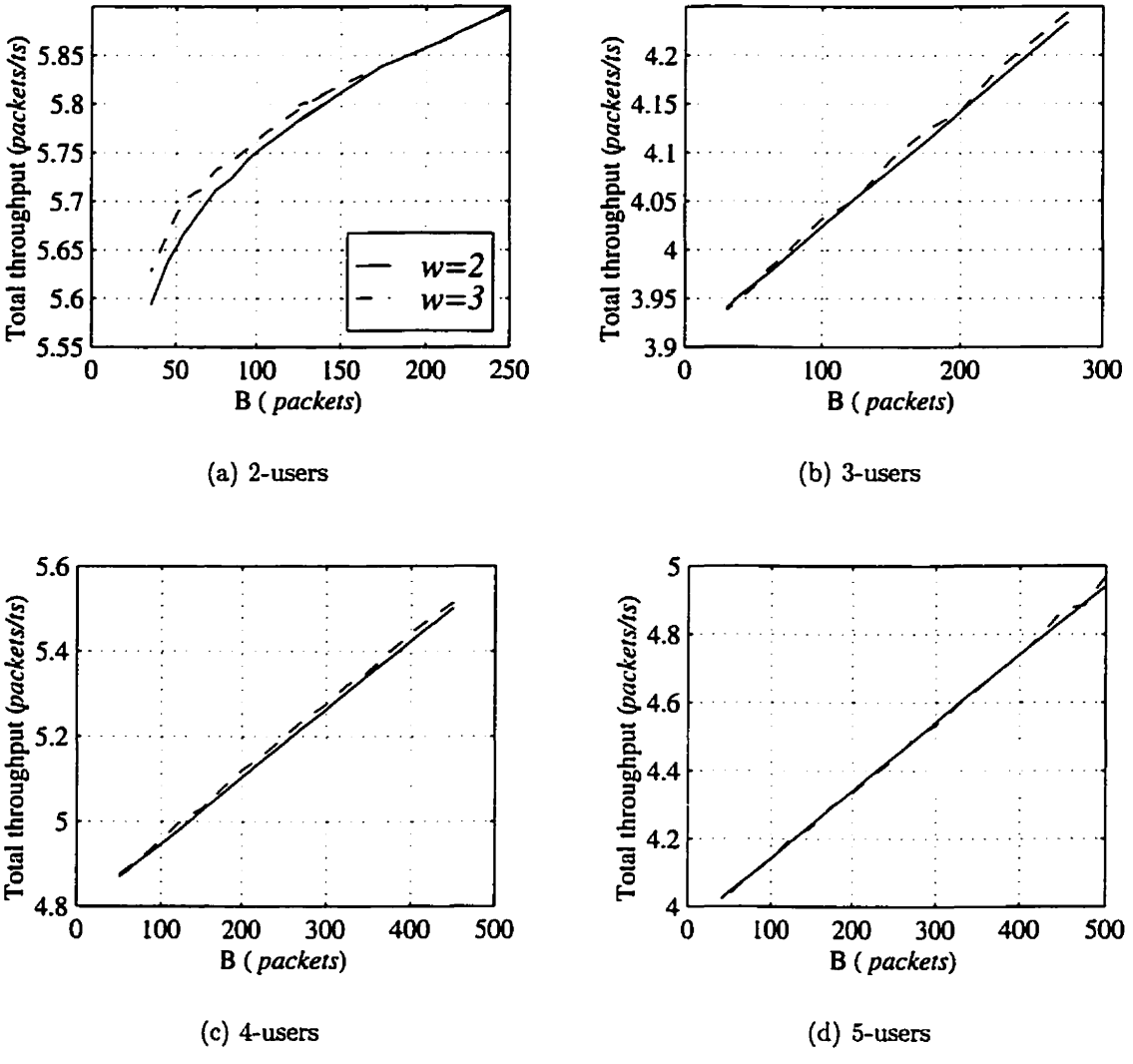
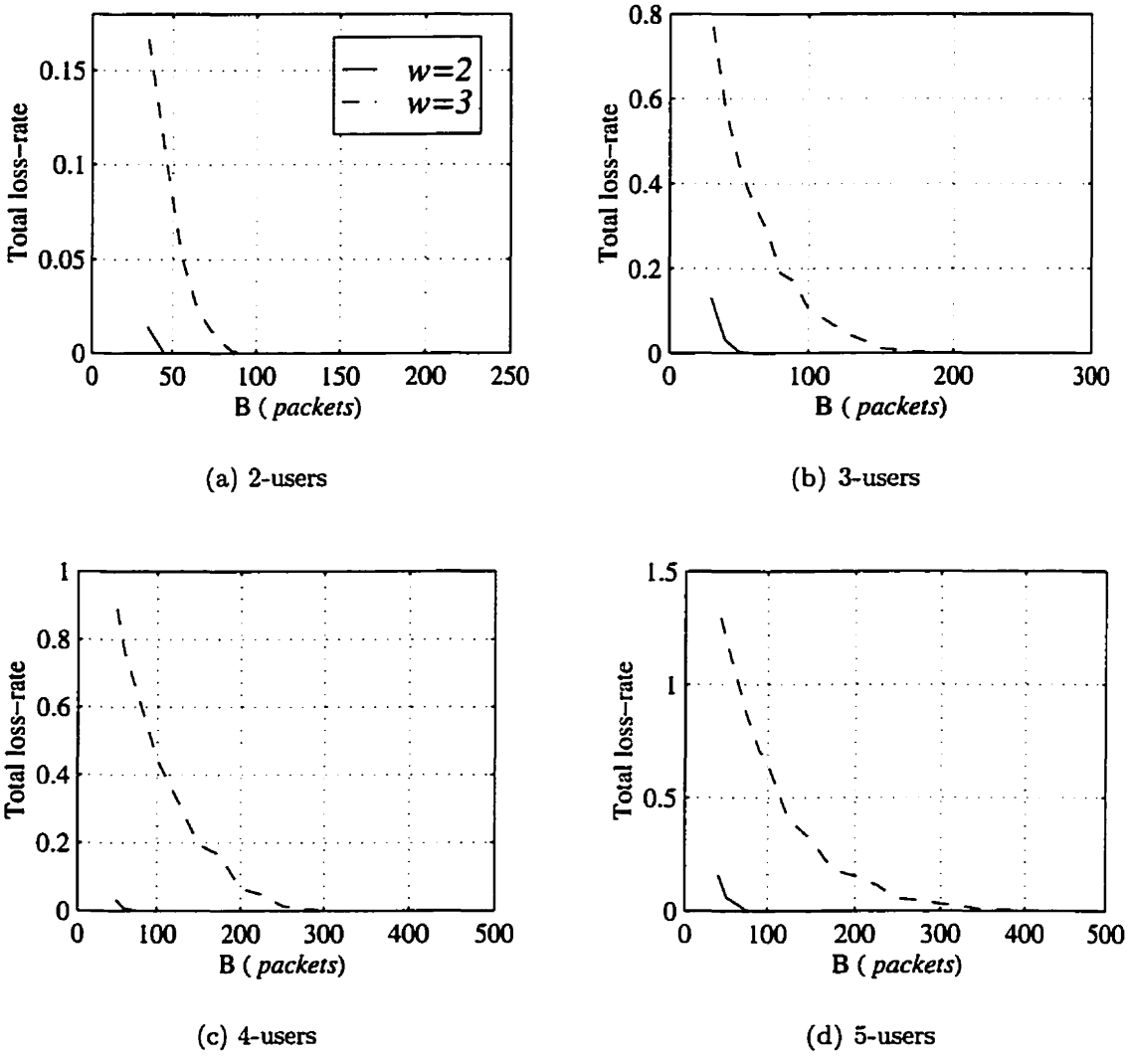


Fig. 4.8 Throughput versus buffer length B for $\overline{\mu} = \lambda_0$. $a = 0.2, b^1 = b^2 = 0.8B$.



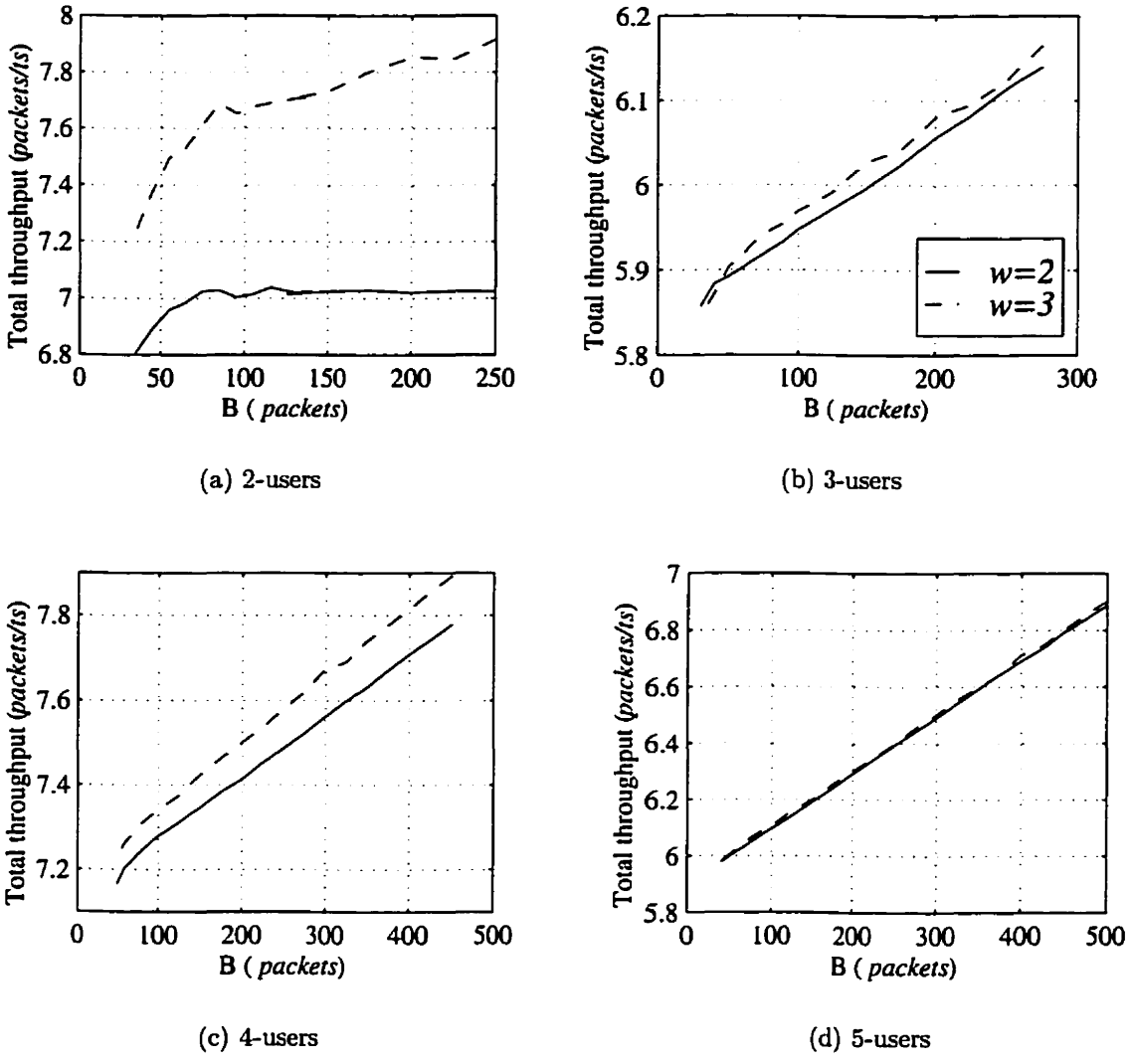
(a) 2-users

(b) 3-users

(c) 4-users

(d) 5-users

Fig. 4.9 Packet loss-rate versus buffer length B for $\overline{\mu} = \lambda_0$.
$a = 0.2, b^1 = b^2 = 0.8B$



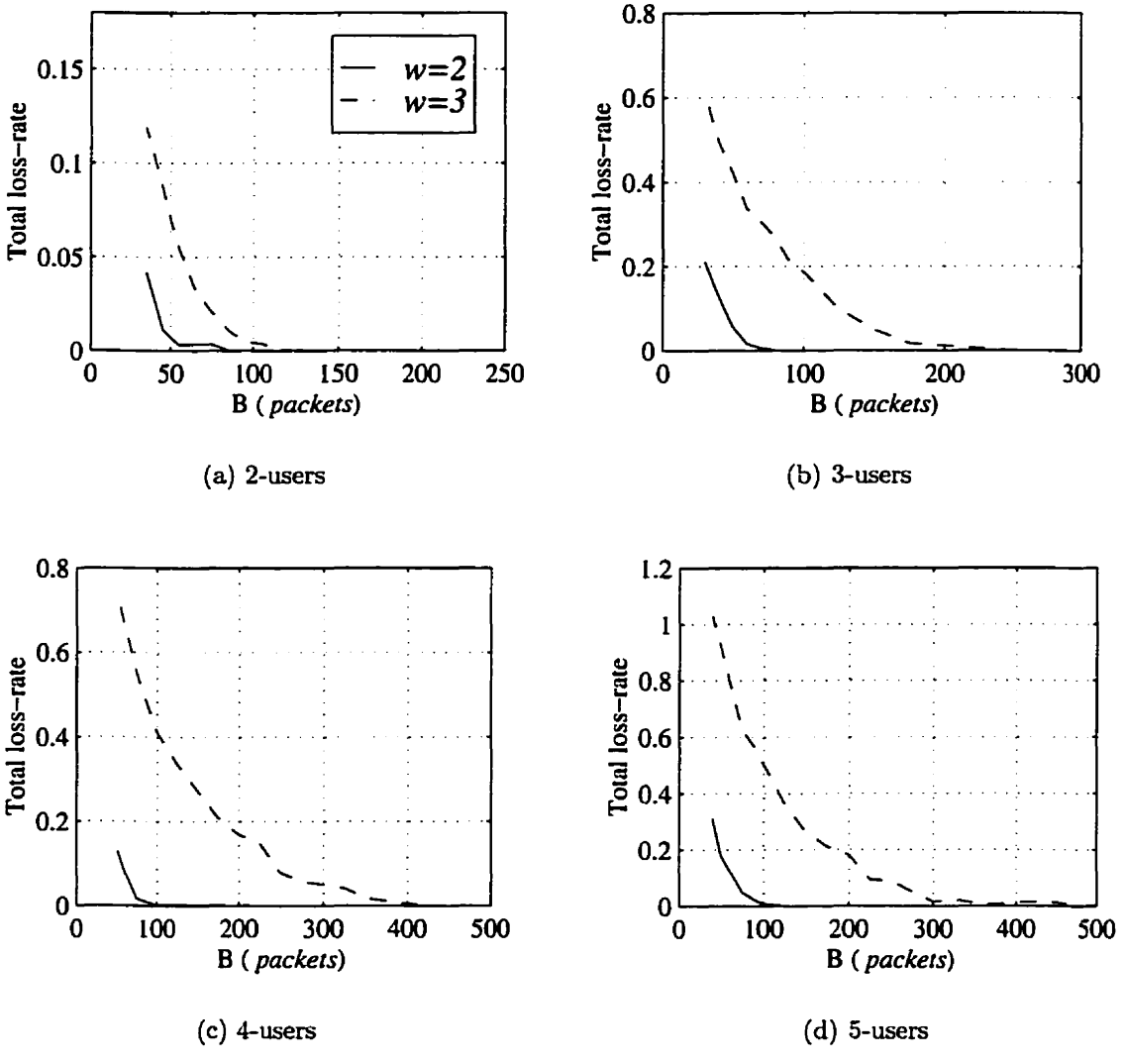
(a) 2-users

(b) 3-users

(c) 4-users

(d) 5-users

Fig. 4.10 Throughput versus buffer length B for $\overline{\mu} = 1.5\lambda_0$. $a = 0.2, b^1 = b^2 = 0.8B$.



(a) 2-users

(b) 3-users

(c) 4-users

(d) 5-users

Fig. 4.11 Packet loss-rate versus buffer length B for $\overline{\mu} = 1.5\lambda_0$. $a = 0.2, b^1 = b^2 = 0.8B$.

- **PPA:**
  The mobile terminals transmit data at their peak power. Such a strategy is not optimal in terms of transmit power, but it may be competitive in the sense of maximizing throughput.

- **IS-95:**
  Power control acts to ensure that the received power from each terminal in the cell is the same.

The above two schemes have in common that the congestion state from the wireline network is not considered. Our goal is to compare our MAC algorithm with these conventional ones in order to discuss the role of buffering in the whole process.

4.3.1 System Comparison

Fig. 4.12 plots the mean rates of varying numbers of users, ranging from 10 to 25. At this point, $\overline{\mu}$ is set to be 9.6 kbps exclusively. Except for the 10-user case, the MAC algorithm surpasses the other two in terms of average rate. Table. 4.6 shows that the loss-rate of PPA is far beyond the QoS requirement in the 10-user case (normally QoS requires a packet loss-rate of less than 1%). In terms of power consumption, Table. 4.7 illustrates the average transmit power in the MAC and IS-95 systems. As the mobile population increases, the average power for each user is reduced so that the interference remains stable to satisfy the SIR criteria. If both aspects, throughput and power consumed are combined, it is evident that the MAC algorithm performs favorably.

Table 4.6 Average packet loss-rate.

	Number of users			
	10	15	20	25
MAC	0	0	0	0
PPA	0.3284	0.0050	0	0
IS-95	0.0062	0	0	0

Table 4.7 Average transmit power.

	Number of users			
	10	15	20	25
MAC (mV)	49.4856	41.0385	33.0552	24.3603
IS-95 (mV)	62.6446	51.9697	45.2822	41.5590

Examining the figures, we notice that under IS-95 the arrival rates are equal for all the mobile users. This is because, as the received powers are all the same, Γs are identical and therefore induce the equal transmit rates. Since the good channels have to waste their current light fading conditions, and instead transmit at a low power budget as well as a low rate, IS-95 is a rather conservative method of preventing packet loss by passively admitting low throughput.

Conversely, PPA provides an aggressive way to improve throughput. By exclusively transmitting at peak power, it forms an *extreme point* in the admissible region $\mathcal{G}$, which is a candidate to achieve the maximum throughput. However, when the number of users is small, the limitation of outgoing rate causes a bottleneck effect. At this point, PPA is apparently inappropriate since large rates lead to high possibility of packet loss. Although the loss rate is no longer a problem as a result of the transmission rates being reduced in accordance with the interference (see Figs. 4.12(b) – (d)), the average throughput by PPA is lower and less robust than that of the MAC algorithm, due to its insensitivity to its insensitivity to backlog and the absence of a mechanism giving priority to those in good channels.

4.3.2 Capacity Limit

Moving to the MAC algorithm itself, we find that there is a capacity limit. Figs. 4.12 demonstrate that when the system has accepted between 10 to 20 connections, it can track the outgoing rate precisely. Besides, it obtains performance gain from the effect of buffering. As the population grows, some of the connections can no longer follow the service rate. In order to determine the extent of the adjustment by the MAC, we calculate the average fading coefficients $\overline{\boldsymbol{\alpha}}$, and estimate the average throughput $\overline{\boldsymbol{R}}$

by explicitly plugging $\overline{\alpha}$ into the CDMA basic equation(Eq. 3.3), with p assigned as the transmit power for each connection. Fig. 4.13 compares the average rate $\boldsymbol{R}$ using MAC with $\overline{\boldsymbol{R}}$ (dashed lines) and the actual mean ABR service rates $\overline{\boldsymbol{R}}_{server}$.

In Fig. 4.13(a), the $\overline{\boldsymbol{R}}$ curve is above $\overline{\boldsymbol{R}}_{server}$. This indicates that the bottleneck exists at the wireline entrance. Consequently, we expect the MAC to work mostly against buffer overflow. The same can be said for Fig. 4.13(b). When the fading is statistically matched with the service rate (Fig. 4.13(c)), MAC is working in a relatively light-loaded way and only needs to adjust the rate to the burst. As the population expands to 25 users, the bottleneck moves to the wireless end as illustrated by the $\overline{\boldsymbol{R}}$ and $\overline{\boldsymbol{R}}_{server}$ curves in Fig. 4.13(d). At this moment, MAC is to compensate the underflow effect induced by the wireless segment, and it can only adjust to a certain degree.

Two questions remain. One is whether this incapability of rate tracking is due to improper choice of control parameters. This doubt exists because the computation complexity, the *convergence step factor,* a, is relaxed as the number of users increases. The second question is how much the system limit is related to the average outgoing rate, $\overline{\mu}$.

Influence of a

Transmit power and rate are measured under different a for the 10-user case as shown in Table. 4.8 and 4.9. The rate barely changes. The transmit power, however, increases with a. For different connections, this increase in power is different. One reason for the heterogeneousness is that in the numerical simulation, the computation is chosen in a FIFO-like style, favoring connections with small user indexes. The same trend is observed in Fig. 4.13(d): mobile users with indexes beyond 20 suffer from weak tracking abilities, while those others can still well respond to the outgoing rate. Calculating the power and rate allocation in random fashion instead of according to the connection index number may solve this. Since it cannot obtain the total throughput gain and improve tracking ability as a whole, we will not discuss this issue further.

Table 4.8 Effect of a on transmit power.

Transmit power at a=0.4 (mV)	relative power comparison to a=0.4				
	a=0.5	a=0.6	a=0.7	a=0.8	a=0.9
62.8300	0.9835	1.0322	1.0620	1.1335	1.3241
70.7846	1.0243	1.0347	1.0914	1.1467	1.3096
64.7051	1.0358	1.0623	1.1035	1.1574	1.3108
39.6035	0.9597	1.0057	1.0428	1.0706	1.2977
53.7944	0.9772	1.0411	1.1018	1.1902	1.3640
35.2011	1.1097	0.9613	1.1783	1.2555	1.7012
60.4866	0.9972	0.9704	1.0469	1.0388	1.2637
41.8222	1.0051	1.0498	1.1321	1.2999	1.7322
38.4337	1.0328	1.0283	1.1145	1.2379	1.6787
39.8594	1.0297	1.1275	1.2078	1.2451	1.6446

Table 4.9 Effect of a on transmit rate.

Transmit rate at a=0.4 (kbps)	relative rate comparison to a=0.4				
	a=0.5	a=0.6	a=0.7	a=0.8	a=0.9
19.3718	1.0006	1.0008	1.0010	1.0012	0.9983
21.9339	1.0003	1.0004	1.0005	1.0006	0.9721
21.7637	0.9999	0.9999	1.0000	1.0001	0.9891
14.5378	1.0008	1.0009	1.0011	1.0015	1.0002
16.6085	0.9996	0.9995	0.9994	0.9994	0.9995
15.6800	0.9994	1.0000	0.9994	1.0000	1.0000
18.1898	1.0005	1.0010	1.0017	1.0025	1.0016
17.0802	0.9999	1.0000	0.9999	0.9999	1.0000
16.0635	1.0004	1.0008	1.0016	1.0021	1.0018
14.1047	0.9988	1.0000	1.0000	0.9986	1.0000

Impact of $\overline{\mu}$ on system capacity

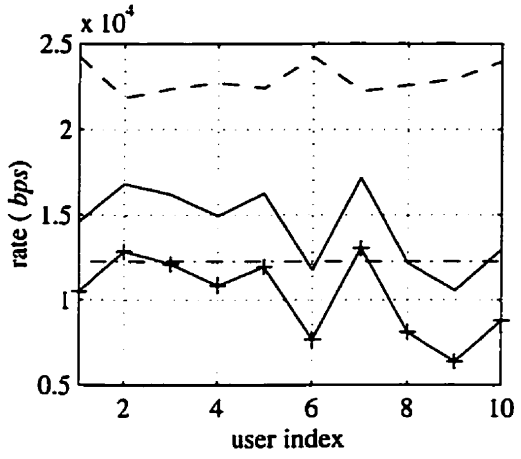
We have shown that a is not a critical control parameter to throughput. Fig. 4.14 illustrates the influence of $\overline{\mu}$ on the capacity limit.

As shown above, the ratio between the outgoing rate, $\overline{\mu}$, and the transmit rate from the terminal, R is around unity, which indicates that the wireless and wireline components of the transmission path are well-matched. Later on, we use $\overline{\mu}$ to represent the transmission rate using our algorithm. From Fig. 4.14, we see that when $\overline{\mu} = 9.6$kbps, the base station can accept about 21 mobile connections, at which — without any control — the system can only obtain average rates $\overline{R}$ around 8.63 kbps. We define:

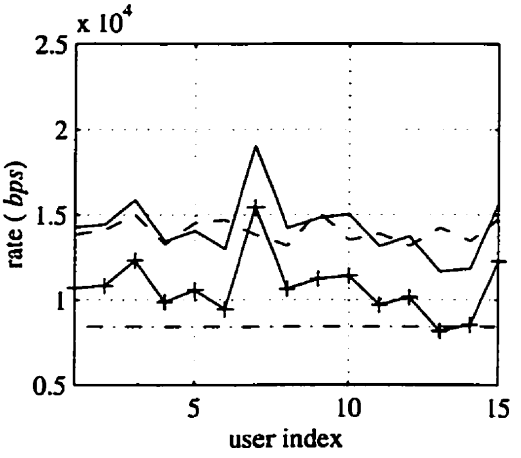
$$\text{system gain} = \frac{\overline{\mu}}{\overline{R}} \tag{4.5}$$

We found that the system gains are 1.11, 1.26, 1.35, and 1.56 for $\overline{\mu} = 9.6, 14.4, 19.2,$ and 24 kbps respectively, and that the system gain is enhanced as the average throughput allowed to the wireline network rises. In other words, the algorithm works more efficiently for a high-rate integrated network.

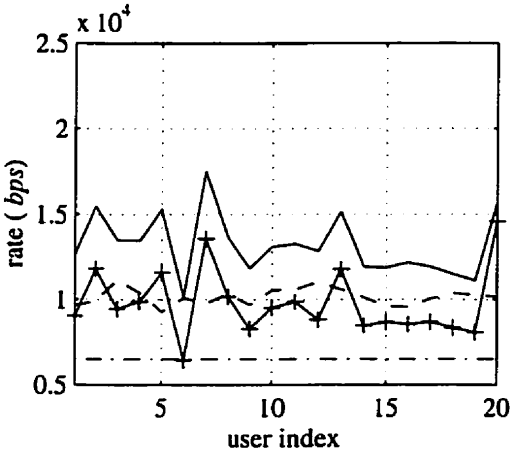
In real implementation, the fading coefficient and the corresponding average rate $\overline{R}$ should be estimated first, and then the system gain computed. If it is beyond the limit of the system gain for this particular $\overline{\mu}$, the connection is granted. Otherwise, it brings degradation even if the MAC algorithm is applied, and therefore the request of connection should be denied.



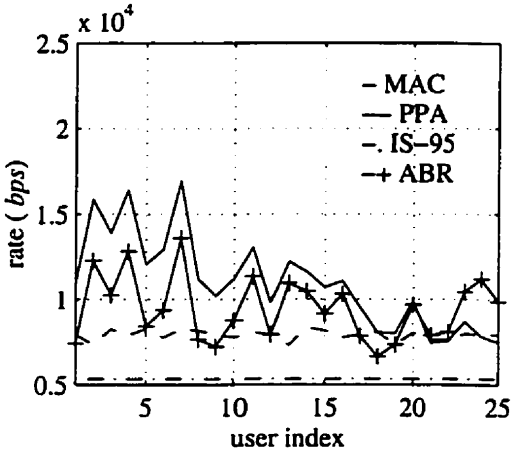
(a) 10-user case. $b^1 = b^2 = 0.8B, B = 200, w = 2, a = 0.4$



(b) 15-user case. $b^1 = b^2 = 0.7B, B = 200, w = 2, a = 0.5$

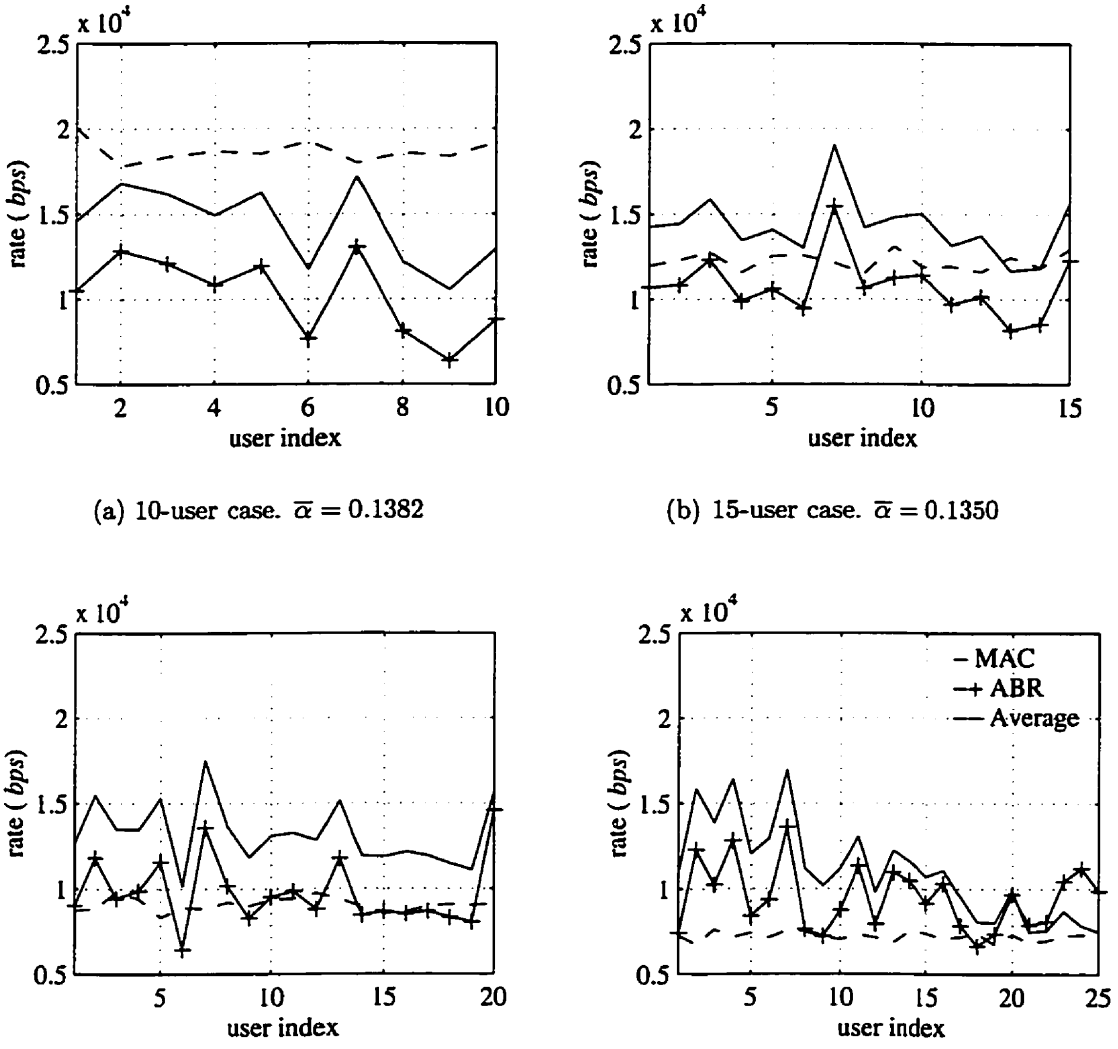


(c) 20-user case. $b^1 = b^2 = 0.7B, B = 200, w = 2, a = 0.6$



(d) 25-user case. $b^1 = 0.5B, b^2 = 0.7B, B = 200, w = 2, a = 0.65$

Fig. 4.12 Rate comparison between MAC, PPA and IS-95.



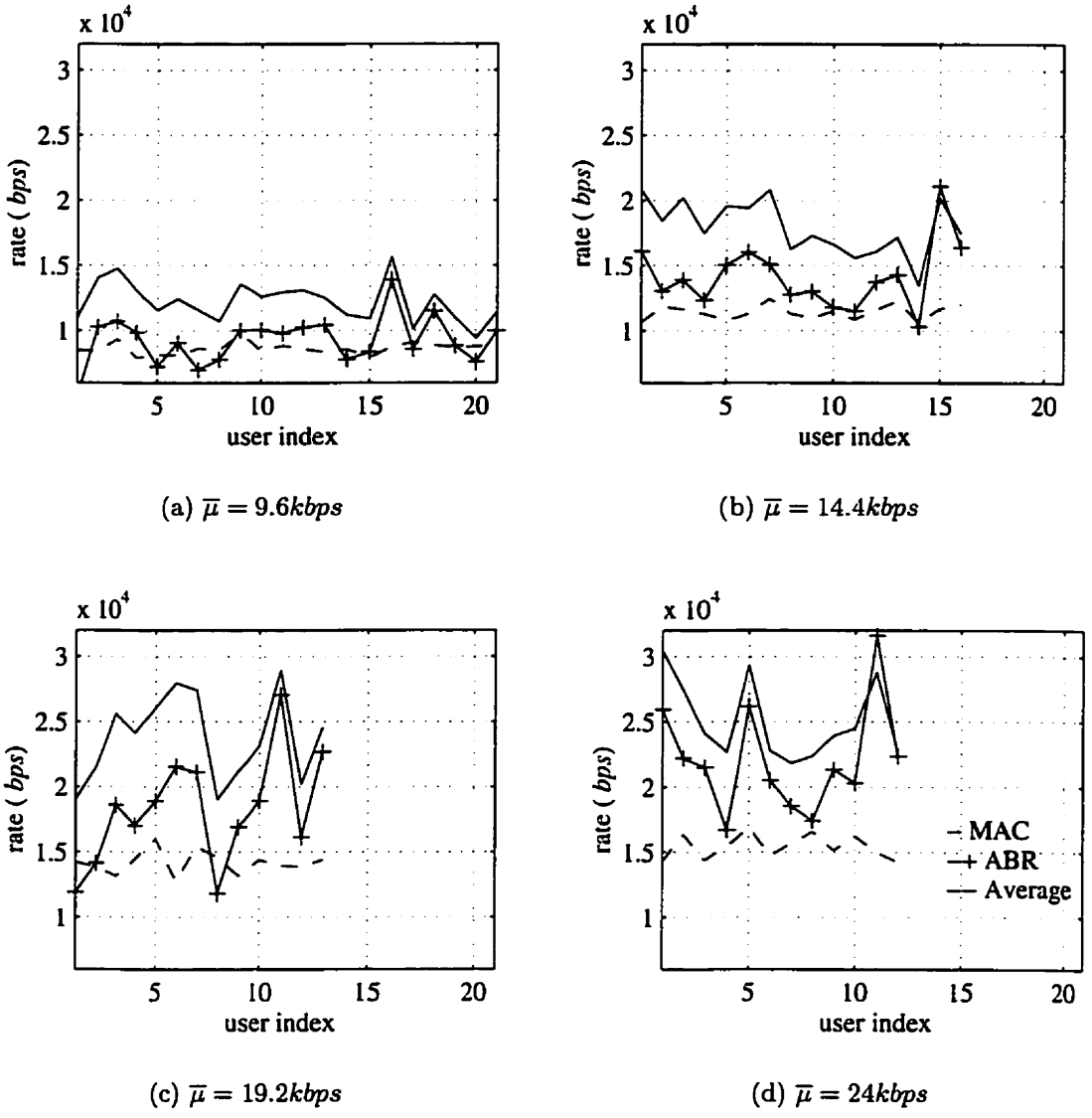
(a) 10-user case. $\overline{\alpha} = 0.1382$

(b) 15-user case. $\overline{\alpha} = 0.1350$

(c) 20-user case. $\overline{\alpha} = 0.1375$

(d) 25-user case. $\overline{\alpha} = 0.1367$

Fig. 4.13 Rate comparison between $\boldsymbol{R}$, $\overline{\boldsymbol{R}}_{server}$, and $\overline{\boldsymbol{R}}$.



(a) $\overline{\mu} = 9.6kbps$

(b) $\overline{\mu} = 14.4kbps$

(c) $\overline{\mu} = 19.2kbps$

(d) $\overline{\mu} = 24kbps$

Fig. 4.14 System capacity versus different $\overline{\mu}$.

Chapter 5

Summary and Future Work

We proposed an algorithm which uses network–layer congestion feedback to guide the dynamic assignment of rates to active calls in a CDMA–based wireless uplink. The idea was to tie MAC–layer bandwidth management explicitly to end–to–end QoS measurements and objectives. The objectives in question were formulated in terms of overflow and underflow events in certain reference buffers defined by their lengths and depletion rates; the buffers can be viewed either as elements of the physical end–to–end path, or else as artifices inserted into the end–to–end path for the purpose of defining the associated resource requirements. The effect in any case is an implicit, soft constraint on *average* rate, the averaging and its time scale being determined by the buffer parameters. Motivation was two–fold: on the one hand, to exploit the fact that QoS for at least some data services is typically evaluated on time scales that are slow relative to those associated with fluctuations in radio channel capacity; and on the other hand, to identify a performance measure that can usefully guide the design of the access control while bypassing the fairness issue, it being tricky to find a definition that feels right in settings where different users see different channel capacities.

Our algorithm was formulated in two stages. The first stage identifies the set of *admissible* rate vectors, referring to those allocations that are consistent with given power and SIR constraints and from which the particular rate assignment is to be selected. The second phase, at a rate determined by the fading rate in the channel,

periodically selects a point in the admissible rate region according to measured values of the fading coefficients and the states of the reference buffers. The algorithm was designed and tested for the single–cell environment.

Testing was by simulation. Our simulation results suggest that the algorithm does what it was designed to do; namely, to ensure for each connection that the mismatch between wireless rate and wireline rate remains suitably small. The results also suggest that such mismatch control has the desired effect in terms of resource management: implicitly, it prioritizes access to bandwidth, allocating it most generously to connections in danger of starvation and least generously to connections close to saturation — in both cases, as indicated by the states of the reference buffers. Efficiency ultimately is measured by the number of connections that can be sustained concurrently and within the associated constraints on power, SIR and QoS; the number of sustainable connections in our simulation model, using a value of 9.6 kbits/sec for the buffer depletion rates, was approximately doubled by applying the algorithm described.

While the results seem promising, much more needs to be done in order to determine conclusively just how congestion feedback can be exploited usefully for radio bandwidth management. The following are examples of directions in which our research can be continued:

(1) *Multi–cell networks:* Our problem formulation includes the multi–cell case. What is needed is a computationally effective description of the admissible rate region. It turns out (Kaplan, private communication) that there is a fast computational technique for deciding whether or not a *given* rate vector is in fact admissible; this provides the mechanism needed to extend the algorithm described in Chapter (3) to the multi–cell case, assuming that the assignment of terminals to base stations is *given.* It remains to evaluate the performance of the algorithm in the more general setting, and to determine whether useful improvements are obtainable by *optimizing* the assignment of terminals to base stations, as in [16], for example. Such optimization, implemented dynamically, would entail, as part of the handoff procedure, setting up and maintaining the reference buffers upon which our algorithm depends — no small matter in practical networks.

(2) *Optimization:* The algorithm that we designed is really a parametric family of algorithms of a particular structure. While that structure was observed to do the job required of it, it almost certainly is not unique. More generally, one could hope to discover a structure that is *optimal* in the sense, say, of minimizing the total mean power required to meet all SIR and QoS constraints. Dynamic Programming, while typically computationally demanding, might provide a conceptual framework (at least) in which to search systematically for an optimal strategy.

(3) *Bursty sources:* An obvious drawback in our formulation of the bandwidth management problem is that the sources are assumed willing to fill any bandwidth allotted to them. In more realistic settings, the sources are *bursty*, meaning that the controller, viewed as resident in the base stations, is uncertain as to who needs what at any particular time. The problem is bring the states of the *sources*, along with the states of the downstream reference buffers, into the control calculation. The ultimate objective is to figure out how best to match resources and need, resource volumes being defined by their instantaneous values and "need", by a combination of traffic and QoS specification so averaged as to provide for an appropriate level of temporal elasticity in the service model and maximum flexibility in the bandwidth allocation.

References

[1] David D. Falconer, F. Adachi, and B. Gudmundson, "Time Division Multiple Access Methods for Wireless Personal Communications", *IEEE Communications Magazine*, pp. 50-57, Jan. 1995.

[2] William C. Y. Lee, "Overview of Cellular CDMA", *IEEE Transactions on Vehicular Technology*, Vol. 40, No.2, pp. 291-302, May 1991.

[3] Ryuji Kohno et al., "Spread Spectrum Access Methods for Wireless Communications", *IEEE Communications Magazine*, pp. 58-67, Jan. 1995.

[4] M. K. Simon, J. K. Omura, R. A. Scholtz, and B. K. Levitt, *Spread Spectrum Communications*. Rockville, MD: Computer Science Press, 1985.

[5] Ehsan Rezaaifar, and Ahmed K. Elhakeem, "Signature code selection for multiuser detection scheme in CDMA systems", *Electrical and Computer Engineering, 1995. Canadian Conference*, pp. 661-663 vol.2.

[6] W. Tschirks, "Effects of transmission power control on the cochannel interference in cellular radio networks", *Elektrotechnik inform.*, Vol. 106, No. 5, 1989.

[7] T. Fujii and M. Sakamoto, "Reduction of cochannel interference in cellular systems by intra-zone channel reassignment and adaptive transmitter power control", *Proc. IEEE Veh. Technol. Conf., VTC-88*, 1988, pp.668-672.

[8] R.W. Nettleton, "Traffic theory and interference management for a spread spectrum cellular radio system", *Proc. ICC-80*, Seattle, WA, 1980.

[9] R.W. Nettleton and H. Alavi, "Downstream power control for spread-spectrum cellular mobile radio system", *Proc. Globecom'82*, Miami, FL, 1982.

[10] R.W. Nettleton and H. Alavi, "Power control for spread-spectrum cellular mobile radio system", *Proc. IEEE Veh. Technol. Conf., VTC-83*, 1983, pp.242-246.

[11] J.M. Aein, "Power balancing in systems employing frequency reuse", *COMSAT Tech. Rev.*, vol. 3, no. 2, Fall 1973.

[12] Jens Zander, "Performance of optimum Transmitter power control in cellular radio systems", *IEEE Trans. Veh. Technol.*, vol. 41, no. 1, pp.57-62, Feb. 1992.

[13] Sudheer A. Grandhi, R. Vijayan, D. J. Goodman, and J. Zander, "Centralized power control in celular radio systems", *IEEE Trans. Veh. Technol.*, vol.42, no.4, pp.466-468, Nov. 1993.

[14] Roy D. Yates and Ching-Yao Huang, "Integrated power control and base station assignment", *IEEE Trans. Veh. Technol.*, vol. 44, no. 3, pp.638-644, Aug. 1995.

[15] Symeon Papavassiliou and L. Tassiulas, "Joint optimal channel base station and power assignment for wireless access", *IEEE/ACM Trans. on Networking*, vol. 4, no. 6, pp.857-872, Dec. 1996.

[16] Stephen V. Hanly, "An algorithm for combined cell-site selection and power control to maximize cellular spread spectrum capacity", *IEEE J. Select. Areas Commun.*, vol. 13, no. 7, pp.1332-1340, Sept. 1995.

[17] James X. Qiu and J. W. Mark, "A dynamic load sharing algorithm through power control in cellular CDMA", *Personal, Indoor and Mobile Radio Commun., 1998. 9^{th} IEEE International Symposiu*, vol. 3, pp.1280-1284, 1998.

[18] Te-Kai Liu and J. A. Silvester, "Joint admission/congestion control for wireless CDMA systems supporting integrated services", *IEEE J. Select. Areas Commun.*, vol. 16, no. 6, pp.845-857, Aug. 1998.

[19] Ashwin Sampath and J. M. Holtzman, "Access Control of Data in Integrated Voice/Data CDMA Systems: Benefits and Tradeoffs", *IEEE J. Select. Areas Commun.*, Vol. 15, No. 8, pp. 1511-1526, Oct. 1997.

[20] A. Sampath, P. S. Kumar, and J. M. Holtzman, "Power control and resource management for a multimedia wireless CDMA system", *Proc. PIMRC'95*, Toronto, Canada, pp.21-25, Sept. 1995.

[21] A. Sampath, N. B. Mandayam, and J. M. Holtzman, "Erlang capacity of a power controlled integrated voice/data CDMA system", *Proc. IEEE Veh. Technol. Conf.*, Phoenix, AZ, May 1997.

[22] Sudhir Ramakrishna and J.M. Holtzman, "A Scheme for Throughput Maximization in a Dual-Class CDMA System", *IEEE J. Select. Areas Commun.*, Vol. 16, No. 6, pp. 830-844, Aug. 1998.

[23] Kerry W. Rendick, "Evolution of controls for the available bit rate service", *IEEE Commun. Mag.*, pp.35-39, Nov. 1996.

[24] Hiroshi Saito et al., "Performance issues in public ABR service", *IEEE Commun. Mag.*, pp.40-48, Nov. 1996.

[25] Malathi Francis, Weihua Zhuang, "Rate control for ABR service in wireless ATM networks", *Veh. Technol. Conference, 1998. VTC 98. 48th IEEE*, vol. 3, pp.1905-1909, 1998.

[26] Malathi Francis, Weihua Zhuang, "A flow control framework for ABR services in wireless/wired ATM networks", *Veh. Technol. Conference, 1999 IEEE 49th*, vol. 2, pp.1156-1160, July 1999.