Optimal Resource Management in Wireless Multimedia Wideband CDMA Systems

Majid Soleimanipour, Weihua Zhuang, *Senior Member*, *IEEE*, and George H. Freeman

Abstract—This paper proposes a scheme of optimal resource management for reverse-link transmissions in multimedia wideband code-division multiple-access (WCDMA) communications. It is to guarantee quality-of-service (QoS) by resource (transmit power and rate) allocation and to achieve high spectral efficiency by base-station assignment. This approach takes the form of a nonlinear-programming large-scale optimization problem: maximizing an abstraction for the profit of a service provider subject to QoS satisfaction. Solutions for both single-cell and multicell systems are investigated. The single-cell solution has the advantage of low complexity and global convergence in comparison with the previous work. Maximum achievable throughput (capacity) of a single cell is mathematically evaluated and used as the benchmark for performance measure of multicell systems. For multicell systems, due to its max-max structure, solving the optimization problem directly entails a high-computational complexity. Instead, the problem is reformulated to a mixed integer nonlinear-programming (MINLP) problem. Then, binary variables indicating base-station assignments are relaxed to their continuous analogs to make a computer solution feasible. Furthermore, approximations can be made to make the resource-management scheme less computationally complex and allow its partial decentralization. The sensitivity of the proposed scheme to path-gain estimation error is studied. Simulation results are presented to demonstrate the performance of the proposed scheme and the throughput improvement achieved by combining resource allocation with base station assignment.

Index Terms—Wideband code-division multiple access (WCDMA), resource management, quality of service (QoS), base station assignment, multimedia services.

◆

---

1 INTRODUCTION

DIRECT-SEQUENCE wideband code-division multiple access (WCDMA) has been proposed as the major multiple-access technique for the third generation of wireless systems [1], [2], [3] due to its potential for high-capacity reliable mobile communication over fading channels and its ability to accommodate multimedia services. In WCDMA systems, transmit power and rate can be controlled to accommodate the various bit-error rate (BER) and delay requirements of multimedia users. The allocated resources may vary significantly with the base-station assignment. If traffic is evenly distributed over the whole network and, consequently, each signal received at a base station sees the same total multiple-access interference, the conventional least-signal attenuation (LSA) assignment provides the best performance. Each mobile is connected to the base station with the strongest pilot signal. With unevenly-distributed traffic, a base station with higher local traffic, despite being the choice of LSA, may receive a mobile signal with a lower signal-to-interference ratio than would a nearby base station with lighter local traffic. Therefore, an assignment decision based on the global traffic (reasonably, a cluster of nearby base stations) is expected to perform better.

Combining power control with base-station assignment has been formulated as an optimization problem for single-service (voice) systems [4], [5], [6], [7]. The tradeoff between handoff-switching cost and connection quality has been investigated separately from power-control optimization [8], [9]. For multimedia communications, variable spreading gain and power control have been proposed to handle mixed traffic with different rates and QoS requirements [10], [11]. In [12], minimizing the total power and maximizing the total rate have been treated as separate optimizations on the reverse link of a single-cell system. As future wireless systems will employ packet-switching techniques to provide multimedia services in an Internet Protocol based network infrastructure [13], intensive research on resource allocation for packet-switching WCDMA systems has been carried out in recent years [14], [15]. In this paper, we investigate optimal resource management for the reverse link of a wireless multimedia WCDMA system. The proposed scheme combines power and rate control with base-station assignment in a nonlinear-programming (NLP) large-scale optimization problem. It maximizes an abstraction for the profit of a service provider subject to QoS satisfaction.

In Section 2, we describe the multimedia WCDMA system model and optimal resource-management problem. In Section 3, we present the solution to the optimization problem for a single-cell system. The single cell capacity is derived and used as a benchmark for performance measure of multicell systems. In Section 4, the solution for a multicell system is investigated. The max-max structure of the optimization implies a high-computational complexity in solving it directly. As a result, the original problem is reformulated as a mixed integer nonlinear-programming (MINLP) problem, then improved by relaxing the integer

---

- *M. Soleimanipour is with the Department of Electrical Engineering, Faculty of Engineering, Imam Hossein University, P.O. Box 16535-187, Tehran 16698, Iran. E-mail: msolmani@ihu.ac.ir.*
- *W. Zhuang and G.H. Freeman are with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1.*
  E-mail: wzhuang@bbcr.uwaterloo.ca and freeman@pce.uwaterloo.ca.

TABLE 1
Summary of Important Mathematical Symbols

symbol	definition
a_i^l	the base station for mobile i in assignment l
$\mathbf{a}$	base station assignment vector
$\mathbf{b}$	binary assignment matrix
$\mathbf{c}$	continuous version of $\mathbf{b}$
$C\ (C_c)$	network (cell) capacity
E_b/I_0	received bit energy to interference density ratio
g_{ik}	path gain from mobile user i to base station k
$\mathbf{g}$	path gain matrix
$h\ (h_{\max})$	number of handoffs per time frame (the maximum number)
i	index for a mobile user
I_k	the total received power at base station k
k	index for a base station
l	index for a base station assignment
M	number of base stations
n	index for a time frame
N	number of mobile users
$p\ (\mathbf{p})$	transmission power (vector)
$P_{i,\max}\ (\mathbf{P}_{\max})$	maximum transmission power of mobile i (vector)
$r\ (\mathbf{r})$	transmission rate (vector)
$R\ (R_c)$	network (cell) throughput
$R_{i,\max}\ (R_{i,\min})$	maximum (minimum) rate for mobile i
$\mathbf{R}_{\max}(\mathbf{R}_{\min})$	maximum (minimum) rate vector
$\mathbf{S}$	set of all feasible assignments
$|S|$	the cardinality of $\mathbf{S}$
T_f	time frame duration
w	W/γ
W	total available frequency bandwidth
λ	transmission cost per data unit
λ_h	cost per handoff switching
$\gamma\ (\bar{\gamma})$	target E_b/I_0 value (the average value)
η	background noise power
τ	delay bound
τ_r	residual delay bound

variables and allowing LSA assignment initialization. Furthermore, a less computationally complex version of the improved MINLP (I-MINLP) approach, called simplified MINLP (S-MINLP), is proposed which can facilitate centralized or partially decentralized implementation. In Section 5, preliminary simulation results demonstrate that both the I-MINLP algorithm and S-MINLP algorithm can significantly outperform LSA assignment when the distribution of traffic is nonuniform. Using a resource-management algorithm requires knowledge of the propagation-path gains. The effect of path-gain estimation error on resource management is investigated in Section 6. Section 7 gives the conclusions of the research. The proofs of the results, given as lemmas, corollaries, and theorems, can be found in the appendix. As there are many variables used in this paper, Table 1 gives a summary of the important symbols.

2 SYSTEM MODEL AND RESOURCE MANAGEMENT

There are N mobile users and M base stations in the system and a total bandwidth of W (for the reverse link) with a fixed chip rate. Packetized information from mobiles is transmitted in synchronized time frames, each having a constant period T_f. Repetition codes [16] are used for different transmission rates to preserve perfect multiplexing while using a fixed symbol duration. We assume channel variations over time T_f (e.g., 10 ms) are small so channel characteristics are treated as constant during each frame.

The system accommodates three classes of service similar to those in [17]. Class I are highly delay-sensitive real-time connections with zero delay tolerance such as voice or low-rate video. Class II are non-real-time delay-sensitive services with a small delay bound such as remote log-in, file transfer protocol (FTP), and similar applications associated with transport control protocol (TCP). Class III are delay-tolerant services such as paging, electronic mail, voice mail, facsimile, and data-file transfer. Both constant-bit-rate and variable-bit-rate services are supported in each class. The QoS parameters under consideration are BER and delay bound. BER is related to the ratio E_b/I_0 of the average signal energy per information bit to the interference-plus-noise spectral density seen at the receiver. The relation between BER and E_b/I_0 is one-to-one and dependent on channel coding, modulation, diversity, etc. The target BERs for voice and data are typically 10^{-3} and 10^{-6}, respectively

[1], [2], [3]. The transmission-delay QoS requirement is specified by a maximal tolerable delay τ.

The control variables for resource management are the transmit powers and rates of the mobiles and the base-station assignment. These variables are to be updated at the discrete times $n \in \{0, 1, \cdots\}$ of the next frame. Let $p_i(n)$ and $r_i(n)$ denote the transmit power and rate of mobile i at time n, where $i \in \{1, 2, \cdots, N\}$. Let $a_i(n) = k$ denote the assignment of mobile i to base-station k, where $k \in \{1, 2, \cdots, M\}$. There are M^N distinct assignments which we distinguish by superscripts so the ℓth assignment is represented by the vector $\mathbf{a}^\ell(n) = [a_1^\ell(n), \cdots, a_N^\ell(n)]$, where $\ell \in \{1, 2, \cdots, M^N\}$. The path gain from user i to base-station k is denoted by $g_{ik}(n)$. The number of handoffs is $h^\ell(n)$ and depends both on the next assignment $\mathbf{a}^\ell(n)$ and the current (say jth) assignment $\mathbf{a}^j(n-1)$ since a handoff is initiated if $a_i^\ell(n) \neq a_i^j(n-1)$.

The network throughput is not properly represented by $\sum_i r_i(n)$ as it does not capture the fact that different users may have different QoS requirements. When the total interference at a base station is much larger than the signal from user i, the required signal power at the base station from that mobile is approximately proportional to its E_b/I_0 requirement. Let γ_i be that E_b/I_0 requirement. Since WCDMA is interference limited, the network resources used for each bit of user i are approximately proportional to γ_i. Thus, the product $\gamma_i r_i(n)$ is a better indication of the resources given to user i than the rate $r_i(n)$ alone. In comparison with the conventional $\sum_i r_i(n)$ for single-service applications, $\sum_i \gamma_i r_i(n)$ is quantitatively different by a scaling factor and is not directly compatible. To avoid this problem, we divide by the user-average E_b/I_0 requirements (γ_is) and define network throughput by

$$R(n) = \frac{1}{\bar{\gamma}} \sum_{i=1}^N \gamma_i r_i(n), \text{ where } \bar{\gamma} = \frac{1}{N} \sum_{i=1}^N \gamma_i. \quad (1)$$

It reduces to the usual $\sum_i r_i(n)$ if all users have the same required E_b/I_0. The network capacity $C(n)$ is defined as the maximal achievable throughput.

Having knowledge of the user types, QoS requirements, and path gains, a resource-management algorithm is to update, at the beginning of each frame, the powers $\mathbf{p}(n) = [p_1(n), \cdots, p_N(n)]$ and rates $\mathbf{r}(n) = [r_1(n), \cdots, r_N(n)]$ for the best base-station assignment $\mathbf{a}^\ell(n)$, such that 1) the BER and delay requirements of each user can be guaranteed, 2) the network throughput $R(n)$ is maximized, and 3) the number of handoffs $h^\ell(n)$ is controlled to not exceed a predefined threshold. As described above, BER requirements are expressed as target E_b/I_0s and are indirectly given by the vector $\gamma = [\gamma_1, \cdots, \gamma_N]$. A target E_b/I_0 can be achieved by controlling the power and rate. Controlling the rate can also satisfy a delay requirement.

Without going into details of commercial issues, we abstract that the network revenue is proportional to the total rate weighted by the QoSs provided in the network, so we (arbitrarily) define it by $\sum_{i=1}^N \lambda_i(n) r_i(n)$, where

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

Here, A, B, and D are positive real constants and $\tau_{r,i}(n)$ is the residual delay, i.e., the time remaining if we are to guarantee the delay requirement for user i at time n. The constant A is used to prevent the function from being 0, while the constants B and D are used to associate the cost with the delay requirement. Thus, a real-time service is more expensive than a delay-insensitive service by a factor of $(A + B)/A$ for the same BER. The rate of change in price is controlled by D. For smaller values of D, the price of delay-insensitive services drops faster. The residual delay is mapped to a requirement of minimal transmission rate $R_{i,\min}(n)$ [18].

Each handoff is associated with additional control signaling and possible buffered-data transfer for establishing a new connection [19] and for transmission of the user's information (such as the new base-station assignment) to the network database. We define λ_h as a fixed cost per handoff, which is assumed the same for all services. The overall profit at time n is then

$$\sum_{i=1}^N \lambda_i(n) r_i(n) - \lambda_h h^\ell(n)$$

for the lth assignment. Our objective is to determine the control variables $\mathbf{p}(n)$, $\mathbf{r}(n)$, and $\mathbf{a}^\ell(n)$, which maximize profit subject to satisfying QoS requirements for all users.

The optimizing of resource allocation for maximal total profit can be formulated as shown in Fig. 1, where $P_{i,\max}$ is the mobile's maximal power, $R_{i,\max}$ is its maximal rate that preserves an acceptable processing gain, $h_{\max}$ is the maximal number of handoffs allowed, η is the average power of the background noise, and $w_i = W/\gamma_i$. The assignment $\mathbf{a}^\ell(n)$ determines the number of handoffs. Rate only appears in the second constraint. Otherwise, from the resource manager's point of view, it is related through the path gains and BER requirements to power. In attempting to solve the mathematical-programming problem shown in Fig. 1, a fundamental question to be addressed is whether the model variables should be integer or real. Technologically and practically, the allocated power and data rate are discrete. However, it is well-known that integer-valued problems are inherently much harder to solve than the corresponding real-valued problems and a great deal of effort is exerted to avoid integer programming in model building [20]. One way out of this problem is to solve the relaxed (real-valued) problem and in some way round or truncate noninteger results. Simulations in [21] indicate that power and rate variables may be handled this way and, if the number of discretization levels is sufficiently large, the resulting throughput loss should be negligible. To simplify the notation, we henceforth omit time indices from time-dependent variables (unless it is necessary to keep the indices) and the assignment superscript ℓ.

The preceding resource allocation formulation aims at a maximal utilization of the radio spectrum by using the newly defined network throughput. As the CDMA system is interference limited, the new throughput definition, $(1/\bar{\gamma}) \sum_i \gamma_i r_i(n)$, is a better indication of how the resources are utilized than the conventional throughput $\sum_i r_i(n)$. On the other hand, the objective of maximizing the profit does not necessarily translate to a maximum value of the

$$\text{Maximize}_{\mathbf{a}^\ell(n)} \left\{ \text{Maximize}_{\mathbf{p}(n)} \left\{ \sum_{i=1}^{N} \frac{\lambda_i(n)\, w_i g_{ia_i^\ell}(n)\, p_i(n)}{\displaystyle\sum_{\substack{j=1 \\ j \neq i}}^{N} g_{ja_i^\ell}(n)\, p_j(n) + \eta} - \lambda_h h^\ell(n) \right\} \right\}$$

Subject to

$$0 \le p_i(n) \le P_{i,\max}$$

$$R_{i,\min}(n) \le \frac{w_i g_{ia_i^\ell}(n)\, p_i(n)}{\displaystyle\sum_{\substack{j=1 \\ j \neq i}}^{N} g_{ja_i^\ell}(n)\, p_j(n) + \eta} \le R_{i,\max}$$

$$h^\ell(n) \le h_{\max}$$

Fig. 1. Resource-management optimization problem.

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

subject to

$$0 \quad \le \quad y_i \quad \le \quad P_{i,\max}\, u$$

$$\frac{R_{i,\min}}{w_i g_i} \le y_i \le \frac{R_{i,\max}}{w_i g_i}$$

$$\mathbf{g} y' + \eta u \quad = \quad 1$$

Fig. 2. Linear programming model for a single-cell system.

conventional throughput in the case of multiclass services, especially with Classes II and III services. If the objective were to maximize the conventional throughput, a user with a lower transmission accuracy requirement would be allocated more resources than a user requiring higher transmission accuracy. Because of the nonlinear relation between E_b/I_0 and BER, a smaller amount of resources is required to achieve the same conventional throughput for a lower transmission accuracy requirement. A better compromise between the maximal conventional throughput and the maximal radio resource utilization may be achieved via techniques such as automatic retransmission request (ARQ) protocols for error correction [22] for non-real-time services. The formulation of resource allocation with ARQ is extremely complex as the extra transmission delay incurred in the retransmission needs to be taken into account in ensuring the QoS satisfaction.

3 SINGLE-CELL SOLUTION

3.1 Linear Programming Model

In a single-cell system ($M = 1$), there is only one base station and, therefore, only one base station assignment ($M^N = 1$). The path gain from mobile i to the base station at time n is denoted by g_i. The value of h^l is zero because no handoff takes place. In the following lemma, we prove the existence of an efficient and convex solution for the mathematical programming problem in Fig. 1 for a single

cell if the number of users is large enough ($N \gg 1$). In other words, our optimization problem can be translated into a linear programming problem for a single-cell environment.

Lemma 1. *For $N \gg 1$ and $M = 1$, the optimization problem defined in Fig. 1 has an equivalent linear programming (LP) problem as shown in Fig. 2, where $\mathbf{y}$, $\mathbf{g}$, and u are defined as*

$$\mathbf{y} = u\mathbf{p}, \tag{2}$$

$$\mathbf{g} = [g_1, g_2, \ldots, g_n], \tag{3}$$

$$\mathbf{m} = [\lambda_1 w_1 g_1, \cdots, \lambda_N w_N g_N], \tag{4}$$

$$u = \frac{1}{\mathbf{g} p' + \eta}. \tag{5}$$

Corollary 1. *The capacity of a single cell is $C_c \approx W/\bar{\gamma}$.*

Thus, in a populated and interference-limited cell, the capacity is independent of the system parameters such as

$$\mathbf{P}_{\max} = [P_{1,\max}, P_{2,\max}, \ldots, P_{N,\max}],$$

$\mathbf{R}_{\max} = [R_{1,\max}, R_{2,\max}, \ldots, R_{N,\max}]$, N, and $\mathbf{g}$. It is affected by the available bandwidth and the error performance requirements. In a single service case, where γ_i is the same for all users, the above capacity is equivalent to what has been given in [24] in terms of the number of users in the cell.

3.2 Simulation of the Multiclass Services

Consider that there are 50 mobile users in service at time $n = 0$ and there is no new call arrival. The users are randomly located in a 4-km-wide square cell with a uniform distribution and the base station is located at the cell center. Under the assumption that each and every user is at a standstill during its call duration, the frame duration is chosen to be 1 s. Other system parameters are set at: $W = 5$ MHz, $\gamma_i = 10$ dB, $R_{i,\max} = 128$ kbps, and $P_{i,\max} = 1$ watt. The transmission channel exhibits a fourth-order log-linear propagation law with log-normal shadowing. Each user has a stored data or image file for transmission from time $n = 0$. The file size, after encoding and using a modulation with the spectral efficiency of 1 bit/s/Hz, is uniformly distributed with mean 500 kilobits and standard deviation 96 kilobits. Simulation results are presented for three users: users #2 (U2), #13 (U13), and #29 (U29), with a path gain (normalized to the average path attenuation among all the users) of $g_2 = 0.11$, $g_{13} = 0.03$, and $g_{29} = 90.3$, respectively. Among all the users, users #13 and #29 have the lowest and highest path gains, respectively. For comparison, we consider three scenarios:

A. all the connections are Class III,
B. user #2 requires Class II service with a delay bound of 20 frames and the rest connections are Class III, and
C. user #13 requires Class I service with a rate of 64 kbps and the rest connections are Class III.

Figs. 3a and 3b illustrate the allocated rates and the residual amounts of data for scenario A. It takes 57 frames to complete the transmissions for all the users. The service time of user #29 is the shortest (three frames) due to the high path gain and, therefore, high allocated rate; while that of user #13 is the longest (57 frames) due to the low path gain. From the simulation, it is observed that:

1. if a mobile is able to transmit at the maximum rate, it is allocated the minimum power that satisfies the target BER,
2. if a mobile is not able to transmit at the maximum rate, it is allocated maximum power to achieve the highest rate that satisfies the target BER, and
3. all the users have a share of the network resources and communicate reliably although some transmission rates may be very low. (For example, Fig. 3a shows that the initial allocated rates to users #2 and #13 are in the range of a few kbps.)

Figs. 3c and 3d depict the results for scenario B. It can be seen that by the end of the 20th time frame, the data from user #2 has been transferred successfully. The service class change of user #2 slightly increases the transmission duration of user #13 (by one frame), but does not affect the transmission of user #29. Figs. 3e and 3f show the results for scenario C. While satisfactory service is offered to user #13 (i.e., a constant 64 kbps rate during the service time), the service to user #29 remains the same as that in scenario A, but the service time for user #2 is increased slightly as compared with that in scenario A. In both scenarios B and C, when more resources are allocated to a specific user than those in scenario A, other users in the cell may experience a longer service duration due to the decrease in the available remaining resources. The resource allocation algorithm indeed gives a higher priority to the more expansive services of Class I and II to satisfy the QoS requirements, at the cost of service quality degradation of other less expansive connections. It is observed that the QoS requirements of all the 50 users are satisfied in the simulation.

3.3 Comparison with Previous Work

Optimal resource management for a single-cell system, supporting only Class I services, is addressed in [12]. Their formulation of the problem is a special case of our mathematical model described in Fig. 1. This special case is solved in [12] employing the gradient projection method for nonlinear problems. It is reported that the algorithm converges to local minima in certain cases. These local minima imply that the problem is nonconvex. No specific solution is proposed in the literature to overcome this problem except trying different initial values. For further performance evaluation, it is helpful to compare our results from the LP algorithm with what has been reported in [12]. We adapt our system parameters to the simulation condition of [12] to maximize $\sum_i r_i$: bandwidth $W = 1.25$ MHz, $R_{i,\min} = 8$ kbps, $\gamma_i = 5$, and the maximum received power at the base station $q_{i,\max} = 1$ watt for voice users, $R_{i,\min} = 4$ kbps, $\gamma_i = 8$, and the maximum received power $q_{i,\max} = 0.5$ watt for data users. Table 2 presents the results given in [12] together with ours, where subscript v and d are used to refer to voice and data users, respectively. In general, the global solution of our LP algorithm provides a higher performance in terms of the sum of allocated rates $(\sum_i r_i)$ in kbps, except for the first result. The exception is mainly due to the approximation made in linearizing the problem in Lemma 1, as given in (19). The approximation may degrade the performance of the single-cell solution when the number of users in the cell is small, depending on the path gains and allocated powers of all the users. In terms of computational complexity, in general, an NLP problem is more complex than an LP problem of the same size. In particular, gradient projection method is a feasible direction method to project the gradient into the feasible space. In [12], it is said that 40 to 100 iterations are needed for their NLP algorithm to converge to a local maximum. To project the gradient onto the feasible space, a number of matrix multiplications and inversions are required. If the linear feasible space is defined by $Hx = b$, $(HH')^{-1}$ is one of the necessary computations in each iteration [23]. Using MATLAB, this operation for $N_v = 25$ and $N_d = 1$ needs at least 83 kflop (floating point operations). Regarding the number of iterations, 3.3 to 8.3 Mflop computation is needed in total. The same problem is solved by our LP algorithm with 766.1 kflop in MATLAB.

4 MULTICELL SOLUTIONS

In a multicell environment (general case), the optimization problem of Fig. 1 has the max-max structure with its inner NLP problem and outer assignment maximization over a huge set of possible assignments. The complexity of this problem is extremely high and derivation of an efficient and accurate solution is very challenging. In this section, we first try to solve the problem as is and then reformulate the max-max form into a single problem.
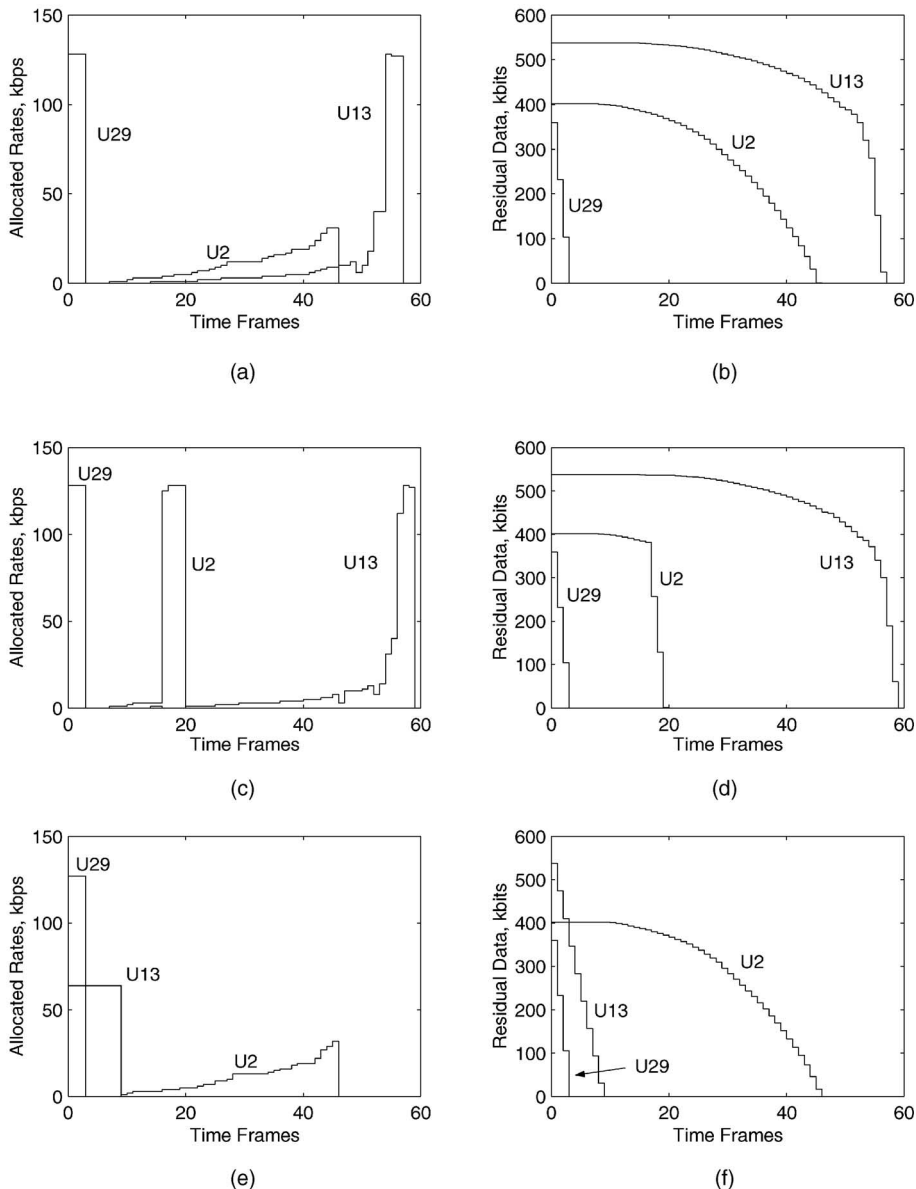
Fig. 3. Allocated rates and residual data amounts in the single-cell multiclass services.

4.1 Solution of the Max-Max Problem

A straightforward approach to solve the max-max problem in Fig. 1 requires an efficient and accurate solution for the NLP subproblem and some criteria to significantly reduce the size of the feasible assignment set. The NLP subproblem is nonconvex. It is well-known that, in nonconvex programming, different approaches may prove to be best fitted to different problems [25]. One way to solve the NLP subproblem is to use available solvers directly, such as MINOS (modular in-core nonlinear optimization system) [26], CONOPT [27], [28], [29], and the optimization toolbox in MATLAB which is based on the sequential quadratic programming (SQP) method [23]. Many optimization algorithms, including those in the above packages, have been developed to find at least one local optimum for nonlinear problems. None of the existing algorithms, however, guarantees a global optimal solution unless the

problem is convex or quasiconvex. Since we are interested in a global solution and our problem is a nonlinear nonconvex problem, an alternative to using NLP solvers is to linearize or convexify the problem, at least in an approximate sense [30], as in the case for a single-cell system. This can be done by exploiting the fractional structure of the objective functions in the alternative models, using an approach similar to that in Theorem A1 of the appendix. Given an assignment vector $\mathbf{a}^l$, the optimization problem converts to the form shown in Fig. 4. Here, without loss of generality, we have assumed the cost per handoff $\lambda_h = 0$ for simplicity of the analysis. The NLP problem is linearized in the following generalized version of Theorem A1 and can be solved by using LP methods.

Theorem 1. *The linear multifractional programming problem in Fig. 4 has an equivalent LP problem as shown in Fig. 5, where*

TABLE 2
Comparison between Our Results and those Reported in [12]

Number of users		Results from [12]					Our results				
N_v	N_d	q_v	q_d	r_v	r_d	$\sum_i r_i$	q_v	q_d	r_v	r_d	$\sum_i r_i$
10	1	1	0.3	22.1	4	225	1	0.3	21.6	4	220
25	1	0.7	0.5	9	4	229	0.7	0.5	9.1	4	231.5
1	5	1	0.1	102	4	122	1	0.5	52.6	16.4	134.6
1	20	1	0.5	20.8	6.25	145.8	1	0.5	20.4	6.3	146.4

$$u_{a_i} = \frac{1}{\sum_{j=1}^{N} g_{ja_i} p_j + \eta}, \qquad (6)$$

$$\mathbf{y}_{a_i} = u_{a_i} \mathbf{p}. \qquad (7)$$

Using the equivalent LP problem, a simulation is carried out for a small scale network with two base stations and three users [21]. To find the optimal throughput, the LP problem is solved for all eight possible assignments and the maximum throughput over all assignments is selected in every time frame. The results are obtained for more than 50 frames and compared with the case of nearest base station assignment. On average, optimal assignments result in 11 percent higher throughput.

The NLP subproblem or its equivalent LP problem in Figs. 4 and 5 should be solved for each base station assignment $\mathbf{a}^l \in \mathbf{S}$, where $\mathbf{S}$, the set of feasible assignments, is a subset of M^N possible assignments. In this case, the cardinality of $\mathbf{S}$, $|S|$, has a significant impact on the complexity of the solution. Thus, it is very important to eliminate infeasible and invalid assignments and avoid unnecessary computations. The E_b/I_0 and handoff constraints can have a significant role in reducing $|S|$. The E_b/I_0 constraint limits $|S|$ due to the fact that reliable communications can usually take place only within a certain range and through a number of nearby base stations.

Corollary 2. *There exists a lower bound on the path gain g_{ik} beyond which reliable communications from user i to base station k is not possible.*

It is also desirable to develop an analytical expression for the feasibility condition when the nonlinear problem has linear constraints. Having such an expression derived, it is possible to find out whether an assignment is feasible by performing the first phase of the simplex method. As an example, the following corollary provides an analytical feasibility condition for a system of two base stations and two users.

Corollary 3. *Let $M = 2$ and $N = 2$. The assignment $\mathbf{a}^l$, where $l = 1, \cdots, 4$, is feasible if*

$$\frac{g_{1a_1} g_{2a_2}}{g_{1a_2} g_{2a_1}} > \frac{\gamma_1 \gamma_2 R_{1,\max} R_{2,\max}}{W^2}. \qquad (8)$$

This condition relates the locations and propagation media of the users to their service qualities. Having the lower bound in Corollary 2, all assignments to base stations with a path gain below the lower bound or being invalid in the feasibility condition can be removed. Similarly, extending condition (8) to other values of N and M, we can perform a feasibility test for each assignment before going through the optimization process.

The other factor in reducing $|S|$ is the limited number of handoffs, $h_{\max}$. With this constraint, the cardinality of $\mathbf{S}$ drops significantly. If we let at most $h_{\max}$ users switch to new base stations, we have

$$|S| \leq \sum_{j=0}^{h_{\max}} \binom{N}{j} (M-1)^j. \qquad (9)$$

This value is derived based on the fact that there are $(M-1)^j$ different assignment vectors with j handoffs. Obviously, for $h_{\max} = N$, $|S|$ is equal to M^N. As an example, let $N = 20$ and $M = 5$. If $h_{\max} = 4$, the number of assignments reduces from $5^{20} = 9.54 \times 10^{13}$ to 425×10^3. This number will

$$
\begin{array}{cc}
\underset{\mathbf{P}}{\textbf{Maximize}} & \displaystyle\sum_{i=1}^{N} \frac{\lambda_i w_i g_{ia_i} p_i}{\sum_{j=1, j \neq i}^{N} g_{ja_i} p_j + \eta}
\end{array}
$$

subject to

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

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

Fig. 4. The NLP subproblem for a typical assignment.

$$\text{Maximize} \atop \mathbf{y}_{a_i}, u_{a_i}} \quad \sum_{i=1}^{N} \lambda_i w_i g_{ia_i} y_{ia_i}$$

subject to

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

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

Fig. 5. The equivalent LP problem for the NLP subproblem.

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

Subject to

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

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

$$\sum_{k=1}^{M} b_{ik} = 1 \;\text{ and }\; h = \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{M}\left| b_{ik} - b_{ik}^{-} \right| \leq h_{\max}$$

Fig. 6. Reformulated problem for a multicell system.

further be reduced to less than 6.2×10^3 if each mobile finds its best assignment from the two nearby base stations.

The solution of the max-max problem suffers from the limit on the number of users and base stations as the computational complexity increases exponentially with N and M, no matter how efficiently the NLP subproblem is solved. A completely different approach to solving the problem is to reformulate it to a less complex problem, preferably changing the structure from the max-max form to a single problem.

4.2 Problem Reformulation: I-MINLP Algorithm

The optimization problem of Fig. 1 can be reformulated if we introduce binary assignment variables determined from the assignment $\mathbf{a}^\ell(n)$ by

$$b_{ik}^\ell(n) = \begin{cases} 1, & \text{if user } i \text{ assigned to base-station } k \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

Let b_{ik} denote $b_{ik}(n)$ and b_{ik}^{-} denote $b_{ik}(n-1)$. Using these, we can reformulate the problem to that of Fig. 6, where $\mathbf{b} = [b_{ik}]$ is an $N \times M$ matrix. This is the same optimization problem expressed in a MINLP form but it is still a difficult problem to solve. Several approaches were tried in [21] and simulation results there indicate that a direct solution starting from an initial LSA assignment (what we call I-MINLP) yields the best allocations so only details related to that method are discussed.

MINLP problems include the complexities of both NLP and integer programming problems and have proven to be very difficult to solve. Some helpful developments since the mid 1980's include the outer-approximation algorithm [31] and its extension with the equality-relaxation strategy [32]. These are available in a program called DICOPT (DIscrete Continuous OPTimizer) [33], [34], available as a solver within the GAMS (General Algebraic Modeling System) package [35]. To solve a MINLP problem using DICOPT, its integer variables must be binary and they must appear linearly. In our problem, the binary variables b_{ik} are involved nonlinearly in the objective function. Thus, we modify the problem by introducing $c_{ik} \in [0, 1]$ as a continuous version of the assignment variable and replacing b_{ik} by c_{ik} in the objective function. Correspondingly, the handoff variable h has a new representation (approximate unless the c_{ik}s solve to binary integers) as

$$h = \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{M}\left| c_{ik} - b_{ik}^{-} \right|. \quad (11)$$

Fig. 7 shows this version of the problem, where $\mathbf{c} = [c_{ik}]$ is the $N \times M$ assignment matrix. The resulting assignments are binary integers if the problem has an integer solution. Otherwise, since a user is to be connected to only one base station (assuming hard handoff), it is reasonable to assign mobile i to base-station k ($b_{ik} = 1$) if $c_{ik} \geq c_{ij}$ for all $j \neq k$. Fig. 8 summarizes the algorithm for the solution of the

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

$$\text{Subject to}$$

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

$$R_{i,\min} \leq \frac{w_i g_{ik} \, p_i}{\sum\limits_{\substack{j=1 \\ j \neq i}}^{N} g_{jk} \, p_j + \eta} \leq R_{i,\max}$$

$$\sum_{k=1}^{M} c_{ik} = 1 \quad \text{and} \quad h = \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{M} \left| c_{ik} - b_{ik}^{-} \right| \leq h_{\max}$$

Fig. 7. The MINLP problem.

SOLVE the MINLP problem of Fig. 7 using the DICOPT solver

IF an integer solution exists for every user i

THEN
 Wherever $c_{ik} = 1$, assign user i to base-station k
 with allocated rate $\lfloor r_{ik} \rfloor$ and power $\lfloor p_i \rfloor$

ELSE
 Wherever $c_{ik} \geq c_{ij}$ for all $j \neq k$,
 assign user i to base-station k with allocated rate $\lfloor r_{ik} \rfloor$ and power $\lfloor p_i \rfloor$

Fig. 8. The MINLP algorithm.

MINLP problem. The floors are taken over the sets of available discrete rates and powers.

Our optimization problem is nonconvex so finding a good solution depends on having a good starting point, which LSA assignment seems to provide. The DICOPT software starts its algorithm by solving the problem as a relaxed MINLP optimization. Then, the relaxed values of the binary variables are input to the mixed integer linear-programming (MILP) master problem. It is, therefore, insensitive to any initial assignment for the binary variables. By changing the assignment variables from discrete to continuous, we have also made the assignment initialization possible.

4.3 Implementation: S-MINLP Algorithm

Centralized resource management needs information on a network-wide scale and is difficult due to the resulting communication overhead (both for collecting the information and for disseminating the resource allocations) and due to the computational complexity. WCDMA systems are a bit less complex than narrowband wireless systems in that the channel assignment (the spreading code) is fixed during a call and all users share the same radio channels. Fig. 9 illustrates a centralized implementation, where all necessary information is made available to the resource-management center (RMC) for processing and the resulting decisions are transmitted to base stations and users. Each base-station k ($\in \{1, \cdots, M\}$) measures the path gains g_{ik} for all mobiles $i = 1, \cdots, N$ and reports them to the RMC. User

information, including service types and QoS requirements, is stored in a user database updated on the admission of each new user. Fixed user parameters, including λ_i, λ_h, $P_{i,\max}$, $R_{i,\max}$, and the E_b/I_0-requirement γ_i, are made available to the RMC by the database. Having the path gains and user information, the RMC runs a resource-management algorithm to determine the new resource-allocation values $\mathbf{p}$, $\mathbf{r}$, and $\mathbf{b}$ which are sent to the base stations and users. Base stations need the following data: 1) allocated power for each user in the network—these, together with the received signal powers, are required for measuring path gains. This method is preferred over using the pilot signal in the forward link because the measurements are more accurate and, given knowledge of the users' power levels at the base station, that control information is not sent over the wireless link; and 2) allocated rate for each user of the base station—the cell-site receivers need the data-transmission-symbol durations for their matched filters.

A centralized implementation of the I-MINLP resource-management algorithm can modify base-station assignments to reduce local congestion due to uneven traffic distribution (while maintaining QoS) and it does not impose additional control signaling on the wireless links beyond telling each mobile its allocated power, rate, and base station. Still, we would like to reduce the amount of control information flowing in the wired network and reduce the computational complexity of the approach. Toward that
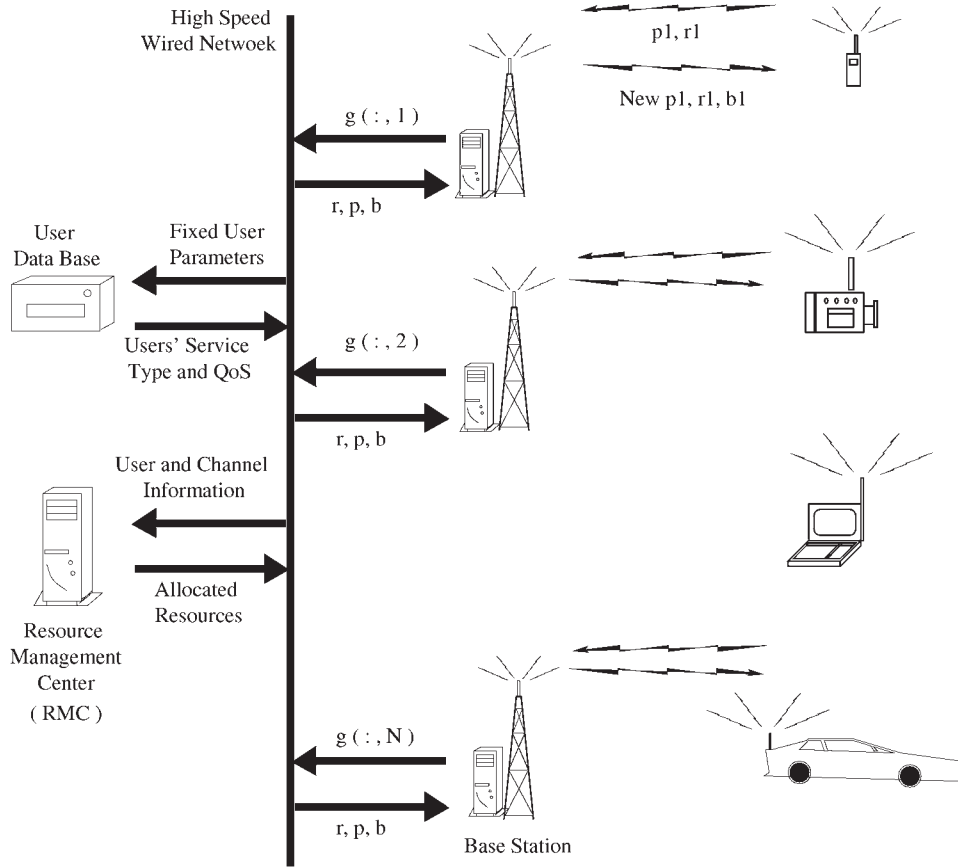
Fig. 9. A centralized implementation of the I-MINLP algorithm.

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

Subject to

$$\frac{I_k}{w_i g_{ik}} R_{i,\min} \leq p_i \leq \min \left\{ P_{i,\max}, \frac{I_k}{w_i g_{ik}} R_{i,\max} \right\}$$

$$\sum_{k=1}^{M} c_{ik} = 1 \ \text{ and } \ h = \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{M} \left| c_{ik} - b_{ik}^- \right| \leq h_{\max}$$

Fig. 10. The S-MINLP algorithm.

end, we introduce approximations as follows: Let I_k be the total received power at base-station k (at time n), we have

$$I_k = \sum_{j=1}^{N} g_{jk} p_j + \eta \approx \sum_{\substack{j=1 \\ j \neq i}}^{N} g_{jk} p_j + \eta \qquad (12)$$

for WCDMA with a reasonably large number of users in the system and we use this to represent the interference plus noise for mobile user i. This simplifies the interference sums in the objective function and second constraint of Fig. 7. Also, I_k is easily measured at the base station. If its variation between two consecutive time frames is sufficiently small, we can obtain an approximation of I_k at frame n by $\hat{I}_k \approx I_k^m(n-1)$, where $I_k^m(n-1)$ is the measured total

received power at base-station k in the previous frame and is updated from frame to frame. This eliminates the interaction of the current power allocation with the interference-plus-noise parts of the objective function and second constraint, whereby the latter becomes

$$R_{i,\min} \leq \frac{w_i g_{ik} p_i}{I_k} \leq R_{i,\max} \qquad (13)$$

and it can be combined with the first constraint. Putting this all together, the MINLP problem of Fig. 7 becomes the S-MINLP problem of Fig. 10. The algorithm of Fig. 8 is used with S-MINLP instead of MINLP in the first step. Resource allocation based on S-MINLP has the following advantages with respect to one based on MINLP:

TABLE 3
Simulation Parameters and Assumptions

Number of users	100 class III ($\tau = \infty$)	Number of base stations	9 (3×3)
Propagation law	r^{-4}	Path-gain estimation error	None
Background noise power	0.001 W	Maximal power	1 W
Maximal rate	256 ksps	Minimal rate	0
Rate revenue form	$[1 + \exp(-\tau_r)]E_b/I_0$	Maximal number of handoffs	10
Time frame duration	10 ms	Bandwidth	5 MHz
Chip rate	4.096 Mcps	Target E_b/I_0	3.3 dB

1. A base station measures the path gains of its own users and the total received signal. Overall, $N + M$ instead of NM such items are sent to the RMC.
2. The number of active constraints is reduced by $2N$.
3. The objective function changes from a sum of linear fractions to a quadratic form (the $c_{ik}p_i$ part).

Thus, there is less signaling overhead on the wired network and the optimization problem is computationally easier.

So far, we have focused on the centralized resource allocation. However, partially decentralized resource management will reduce the computational complexity for a system with large numbers of base stations and mobile users. In this case, we can partition the network coverage into cell clusters, each cluster consisting of a small number of radio cells. At the cluster level, resource management is distributed, while within each cluster resource management is centralized. The intercluster interference from all other clusters should be taken into account in determining the E_b/N_0 value. With slight modification, the S-MINLP problem/algorithm can be extended to be partially decentralized. Define the interference ratio f of intercluster interference to intracluster interference. In Figs. 7 and 10, assume the following modifications: 1) Parameters such as M and N refer to the cluster. 2) Interference $\sum_{j \neq i} g_{jk}p_j + \eta$ or I_k is multiplied by $1 + f$, assuming that the background noise η is small compared with the total interference.

Other authors [24], [36], [37] have studied the suitability of using the interference ratio f in conditions similar to those of our simulations as described in Section 5. It has been shown both analytically and by simulation that $f \approx 1/3$ to $1/2$ due to users who are power controlled by other base stations. Under similar conditions, the interference ratio for a base station in a cluster will be less than this because part of the interference from users assigned to other base stations is already included in the intracluster interference (i.e., some nearby base stations are part of the same cluster). In a more sophisticated design, the size and user population of clusters could be dynamically changed according to the instantaneous state of the network. Further investigation on evaluating the interference ratio and dividing the network into subnetworks could be considered for future work. Overall, resource allocation in a partially decentralized network will be worse in terms of spectral efficiency but require less signaling in the wired network and less computational effort.

5 PERFORMANCE ANALYSIS

Improvement by optimal base station assignment. To compare, consider the LSA algorithm where 1) the base station assignment is based on the conventional LSA criterion and is independent of the resource allocation, 2) the optimal resource allocation is defined by the NLP subproblem as given in Fig. 4, and 3) the unique solution of the NLP subproblem is solved exactly by using the equivalent LP problem given in Fig. 5. For a uniform distribution of mobile traffic, simulation results in [21] demonstrate that LSA algorithm performs very close to the I-MINLP algorithm. Here, we compare them when the mobiles are not uniformly spread around the network. In the simulation, 100 Class III users are randomly located according to a two-dimensional radially symmetric Gaussian distribution centered (with a standard deviation of 1 km) on a network with nine (3×3 on a 2-km grid) base stations in the middle of an 8×8-km square. The mobile users move in a radially outward direction and each has a random data-transfer requirement (uniform between 0 and 1,024 kb). A logarithmically linear propagation law (without shadowing) is used and the path gains are known accurately. To demonstrate the effect of user mobility on resource allocation and base station assignment, we assume large random values for mobile speeds. At the start, mobiles are mostly clustered near the middle cell. By the end, they are concentrated in the outer eight cells. Table 3 summarizes these and other simulation parameters and assumptions. The target E_b/I_0 is similar to that of the long constrained delay (LCD) data-bearer service described in [1] which can maintain a BER of 10^{-6} by using a turbo code with constraint length 3, quadrature phase-shift keyed (QPSK) modulation, 2-antenna diversity, RAKE receiver, and soft-decision decoding. The symbol rate in the physical layer for a 64-kbps LCD service is 256 ksps. The simulation is carried out for 200 frames. To address handoff cost, the number of handoffs is limited to 10. With a small number of handoffs, it is observed from simulation that the value of the handoff cost λ_h has a negligible impact on the resource allocation when the overall handoff cost is small as compared with the revenue. As a result, in the following, we present the simulation results for the case of $\lambda_h = 0$. Figs. 11 and 12 illustrate two snapshots of the network for I-MINLP and LSA algorithms under the same conditions respectively. An explanation of these figures is given as follows:
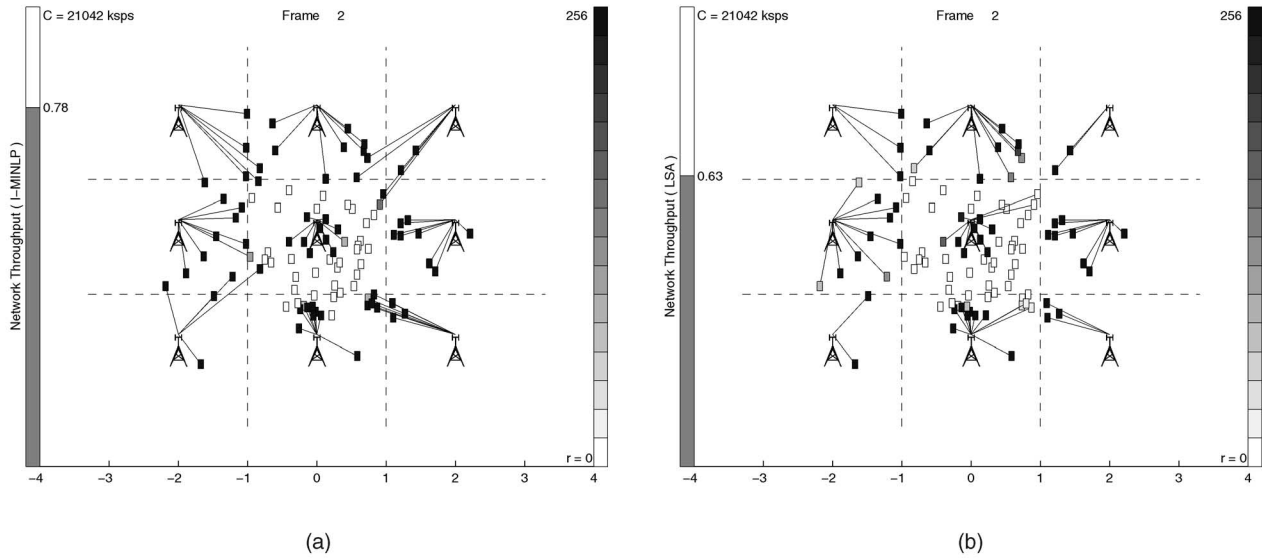
Fig. 11. Comparison of the I-MINLP and LSA algorithms for unevenly distributed traffic (frame 2). (a) I-MINLP and (b) LSA.
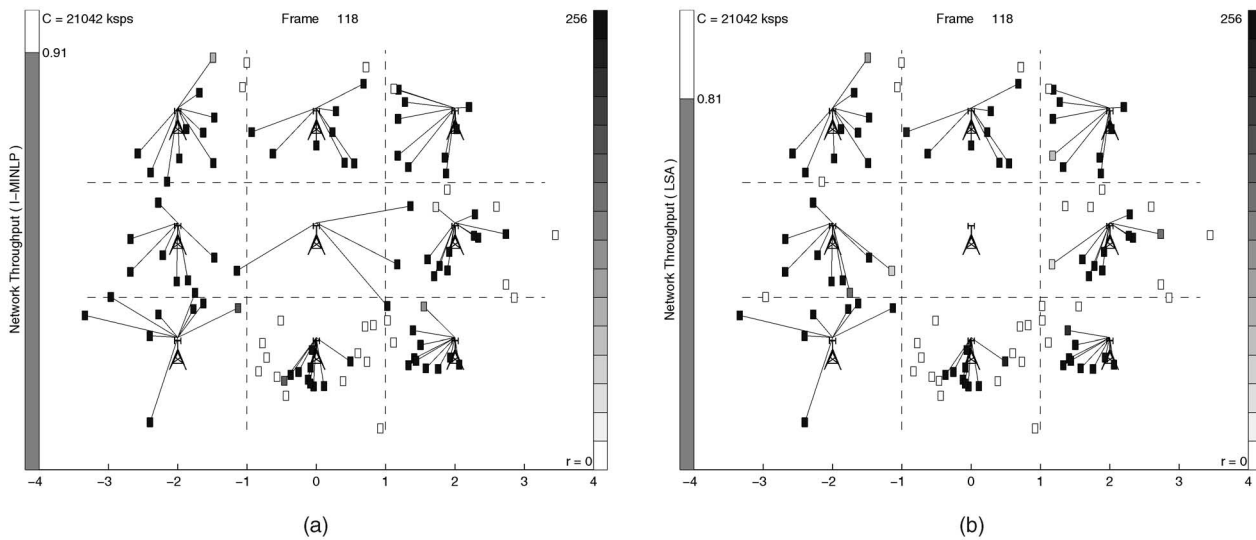

Fig. 12. Comparison of the I-MINLP and LSA algorithms for unevenly distributed traffic (frame 118). (a) I-MINLP and (b) LSA.

1. Each mobile user is represented by a rectangle, shaded based on its allocated rate. The vertical bar shown on the right-hand side scales allocated rates from 0 (light) to 256 (dark) ksps.
2. Assignment of each user is shown by a solid line originating from the user to the assigned base station. No connection line implies that at the particular frame the user has not been allocated any resources (including the case that the data transfer is completed).
3. The network throughput is shown with respect to the capacity by the vertical bar on the left-hand side.

The capacity and fraction of the utilized capacity are printed beside the bar. Note that, as all the Class III users have the same E_b/I_o requirement, the network throughput given in the following simulation results is the actual throughput defined in the conventional way (i.e., the sum of the transmission rates from all the mobiles). Fig. 14 shows the network throughput in 200 successive time frames for the I-MINLP, S-MINLP, and LSA algorithms. The network capacity is calculated based on the single-cell solution which has global convergence. In the figure, infeasible solutions appear in the form of small gaps in the results, as throughput is zero when there is no feasible solution. We make the following observations:

1. LSA assignment is not the best assignment when traffic is nonuniform. The I-MINLP and S-MINLP algorithms achieve a higher throughput by sharing the traffic load among all base stations.
2. Network throughput with any of the algorithms varies with time as the traffic pattern varies: the higher the concentration of users in the network, the lower the total network throughput.
3. On average, the I-MINLP algorithm has about 10 percent improvement over the LSA algorithm by optimal base-station assignment and has 6 percent

TABLE 4
Network-Reuse Factor Using I-MINLP

M	C (ksps)	R (ksps)	ρ
1	2338	2327	0.995
4 (2×2)	9352	9175	0.981
9 (3×3)	21042	19798	0.941

improvement over the S-MINLP algorithm. The S-MINLP algorithm outperforms the LSA algorithm.

4. The S-MINLP algorithm produces no infeasible solutions in the simulation over 200 frames and the I-MINLP algorithm experiences much less infeasibility than the LSA algorithm.

Frequency reuse factor with uniform traffic: Another simulation is carried out to study the frequency reuse efficiency and the effect of the number of base stations in a fixed area using the I-MINLP algorithm in otherwise the same conditions described in Table 3. The cell capacity in this case is 2,338 ksps. Table 4 presents the network capacity C and network throughput R, both in ksps, for $M = 1$, 4, and 9, respectively. The frequency-reuse factor ρ, defined as R/C, is also given. As M increases, the reuse factor decreases due to the increased intercell interference. However, reuse factors larger than 0.94 for all the M values are sufficiently high, as compared with the theoretical reuse factor of one in WCDMA systems.

Performance comparison between the I-MINLP and S-MINLP algorithms. The simulation environment is the same as that given in Table 3 except here we consider propagation shadowing with a standard deviation of 0 to 8 dB. It is assumed that the shadowing is independent from frame to frame (the worst-case) and there is no path-gain estimation error. S-MINLP will be affected by variations in I_k across

two frame periods ($2T_f$). If the standard deviation σ of the log-normal shadowing process increases, the path-gain variations due to shadowing increase, and there is a higher probability of large differences between I_k and $\hat{I}_k$. This is verified by the simulation results shown in Fig. 13. Fortunately, in practice, the strong autocorrelation of the shadowing process for each mobile over a period of $2T_f$ would improve the performance of the S-MINLP algorithm, as that demonstrated in Fig. 14 in the case of no shadowing.

Computational complexity. Speed of computation is crucial if we are to implement such a resource-management algorithm. For both the I-MINLP and S-MINLP approaches, a large amount of the computation is done in DICOPT which is called from GAMS which is called from the main simulation program running in MATLAB using an interface described in [38]. Although this is a nicely flexible structure, it adds significant overhead in the exchange of data between program environments. Still, to get some idea of the computational complexity, we measure the elapsed computation time for different numbers of users when the simulations are run on a Sun Ultra-10 workstation (networked multiuser environment). Fig. 15 illustrates the elapsed computation time to run the I-MINLP and S-MINLP algorithms for nine base stations and 90 to 110 users in the network. It appears that the time increases almost linearly with N. The actual times would be much less on a dedicated powerful computer at the RMC or if the program code were rewritten as a single dedicated application. We have not determined the theoretical complexity.

6 EFFECT OF PATH-GAIN ESTIMATION ERROR

Knowledge of the reverse-link path gains (signal attenuations) is required in the resource-management algorithms but these are subject to estimation error. Let $\hat{g}_{ik}$ be the estimate of the actual path gain g_{ik} so, while the signal actually experiences g_{ik}, the resource-management algorithm uses
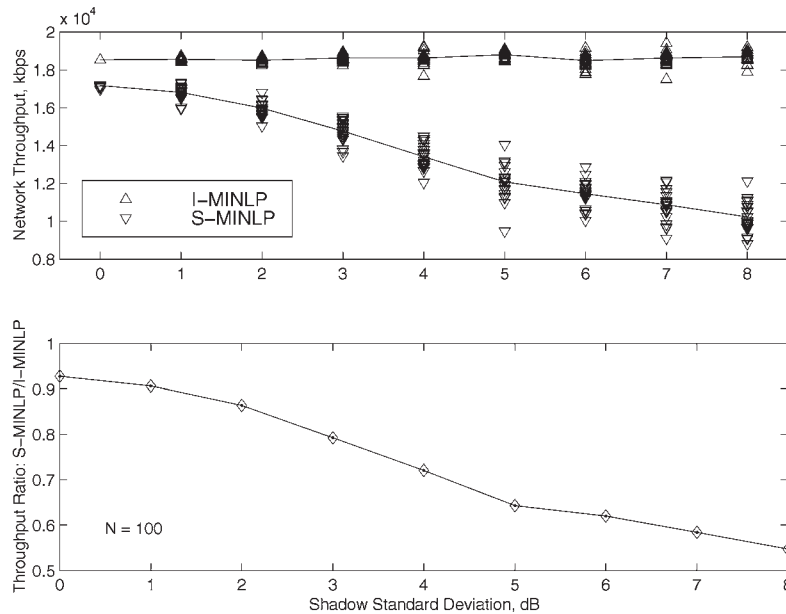

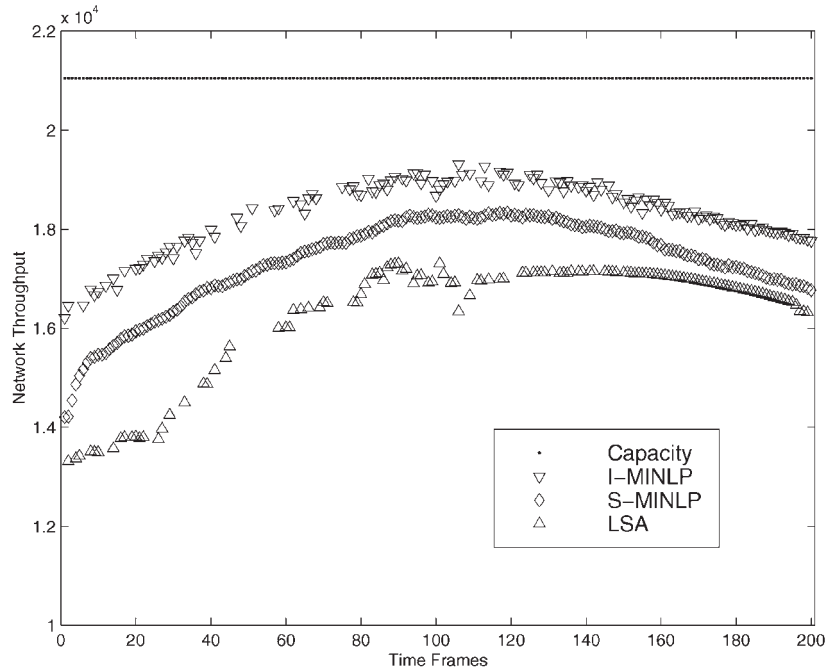Fig. 13. Throughput comparison of I-MINLP and S-MINLP.

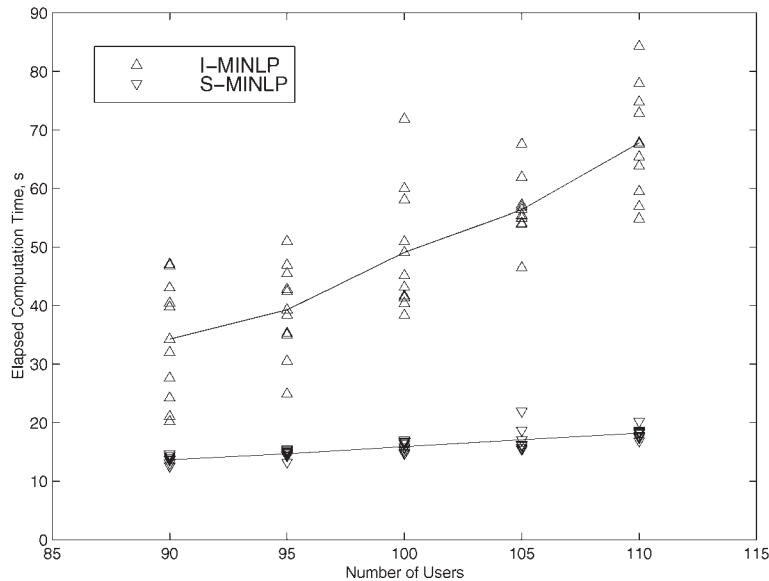Fig. 14. Throughput versus time with unevenly distributed traffic.

Fig. 15. Elapsed computation time versus number of users.

$\hat{g}_{ik}$ in the computations and allocates its resources accordingly. The question is whether the allocated resources remain feasible. The relationship between the standard deviation of the estimation error and the probability of an infeasible solution is investigated in the following.

Assuming independent fading in each path, the g_{ik}s can be modeled as independent random variables and related to the estimated path gains by $g_{ik} = \hat{g}_{ik} + e_{ik}$, where e_{ik} is an estimation error with zero mean and variance σ_g^2. Experimental data [37], [39] and theoretical studies [40] on the short term averages of radio signals on fading channels suggest that the received signal power at the base station has

a log-normal distribution. Thus, given p_i, the path-gain distribution is log-normal with mean $\mu_{g_{ik}}$ and variance σ_g^2. The distribution of g_{ik} becomes

$$f_{g_{ik}}(z) = \frac{1}{\sqrt{2\pi}\sigma\beta\,z}\exp\left\{-\frac{10\log(z-\mu)^2}{2\sigma^2}\right\}, \qquad (14)$$

where $\beta = (\ln 10)/10$ and the parameters σ and μ (in dB) are related to the mean and variance of the distribution through

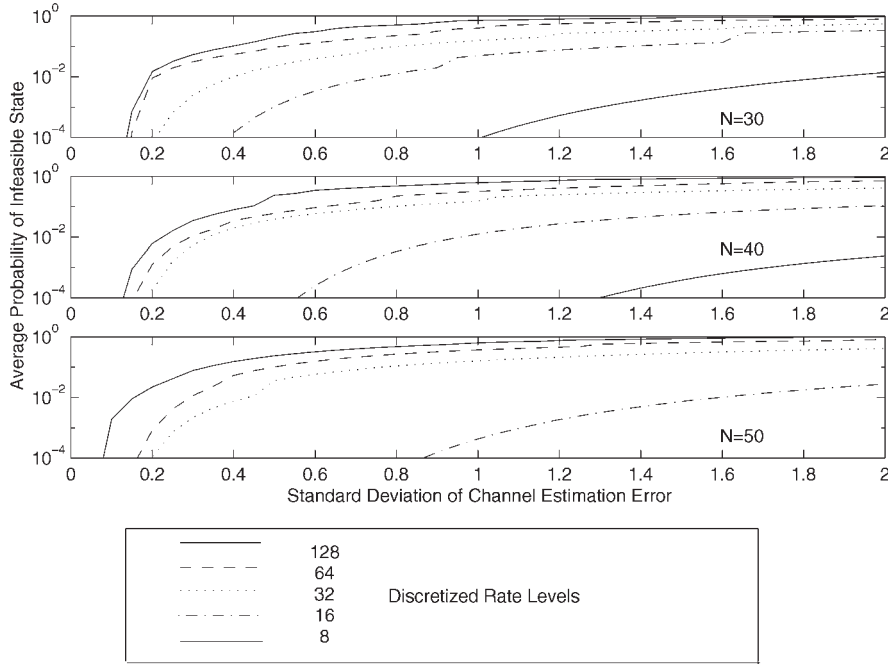$$\mu_{g_{ik}} = \hat{g}_{ik} = e^{\beta\mu + \beta^2\sigma^2/2} \qquad (15)$$

and

Fig. 16. Probability of infeasibility versus path-gain estimation error.

$$\sigma_g^2 = e^{2\beta\mu + 2\beta^2\sigma^2} - e^{2\beta\mu + \beta^2\sigma^2}. \tag{16}$$

For user i assigned to base-station k with power p_i and rate r_i, the probability that the allocated resources are infeasible is

$$P_{\text{inf},i} = \Pr\left(r_i > \frac{w_i g_{ik} p_i}{\sum\limits_{j \neq i} g_{jk} p_j + \eta} \right) = \Pr\left(\xi_{ik} > \frac{w_i g_{ik} p_i}{r_i} \right), \tag{17}$$

where

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

Sums of log-normal random variables are often approximated by another log-normal random variable [41]. A Gaussian approximation is also valid when log-normal random variables are independent and their number is large enough. The latter conditions hold in our problem, therefore, the distribution of ξ_{ik} is approximated as Gaussian with parameters

$$\mu_{\xi_{ik}} = E[\xi_{ik}] = E\left[\sum_{j \neq i} g_{jk} p_j + \eta \right] = \sum_{j \neq i} \hat{g}_{ik} p_j$$

$$\sigma_{\xi_{ik}}^2 = \text{Var}\left[\sum_{j \neq i} g_{jk} p_j + \eta \right] = \sum_{j \neq i} \sigma_g^2 p_j^2$$

assuming the variance of η is negligible. As a result,

$$P_{\text{inf},i} = \int Q\left(\frac{w_i z p_i / r_i - \mu_{\xi_{ik}}}{\sigma_{\xi_{ik}}} \right) f_{g_{ik}}(z) \, dz. \tag{18}$$

Infeasibility occurs with probability $1 - \prod_i (1 - P_{\text{inf},i})$.

Fig. 16 shows the probability of infeasibility versus estimation-error standard deviation for different numbers of discrete rate levels and $N = 30, 40, 50$. We observe the following:

1. A larger number of rate levels (which should yield a higher capacity utilization closer to optimal and have a smaller margin for error) makes infeasibility more likely.
2. Error tolerance increases with the number of users if the number of rate levels is small.
3. Error tolerance is unaffected or slightly worse as the number of users increases if the number of rate levels is large.

In the system, the path gain can be estimated directly from the mobile transmit power (known to the base station) and the received power measured at the base station receiver. The estimation error mainly results from the fact that the channel changes with time. With a small time frame duration (e.g., $T_f = 10$ ms), channel variations over the period should not be significant and, therefore, high estimation accuracy can be expected.

7 CONCLUSIONS

In this paper, we have developed techniques and algorithms to solve the optimal resource management problem for single-cell and multicell systems. The single-cell algorithm finds the exact global optimum and determines the maximum achievable throughput per cell for specific E_b/I_0 requirements. This value represents a new cell capacity bound and is used as a benchmark for evaluation purposes. In a multicell system, for fixed assignments such as LSA assignment, the optimization problem has an equivalent LP

$$\text{Maximize}_{\mathbf{p}} \left\{ \frac{\mathbf{m}p'}{\mathbf{g}p' + \eta} \right\}$$

subject to

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

$$R_{i,\min} \leq \frac{1}{\lambda_i} \frac{m_i p_i}{\mathbf{g}p' + \eta} \;\leq\; R_{i,\max}$$

Fig. 17. Linear fractional programming for a single-cell system.

problem and is globally optimal. The E_b/I_0 and handoff constraints can be used to eliminate infeasible assignments and, hence, avoid unnecessary computations. In general, solving the optimization in its original max-max structure suffers computational complexity as the NLP subproblem or its equivalent LP problem needs to be solved for each of the base station assignments. To overcome the complexity problem, the optimization problem has been reformulated to a MINLP problem which combines the NLP subproblem with the base station assignment. The I-MINLP algorithm has been developed to achieve a high utilization of the network resources and the less complex S-MINLP algorithm has been developed to facilitate a centralized or partially decentralized implementation of the resource management. It has been shown that the S-MINLP algorithm outperforms the LSA algorithm and, with respect to the I-MINLP algorithm, runs faster with lower infeasibility at a cost of slightly reduced throughput. Computer simulation results demonstrate that, for a limited number of transmission rate levels, the I-MINLP algorithm has a reasonable tolerance to the path gain estimation error. A combination of the I-MINLP algorithm and the closed-loop power control can be applied to the resource management of the International Mobile Telecommunications in 2000 (IMT-2000) proposals [42].

APPENDIX

Proof of Lemma 1. For a large N, the approximation

$$\sum_{j=1, j\neq i}^{N} g_j p_j \approx \sum_{j=1}^{N} g_j p_j \qquad (19)$$

is valid, as the received signal power from one user (when $j = i$) is very small compared with the total received power from all the $N \, (\gg 1)$ users. By this approximation, the summation in the objective function becomes

$$\sum_{i=1}^{N} \frac{\lambda_i w_i g_i p_i}{\sum_{j=1, j\neq i}^{N} g_j p_j + \eta} \approx \frac{\sum_{i=1}^{N} \lambda_i w_i g_i p_i}{\sum_{j=1}^{N} g_j p_j + \eta} = \frac{\mathbf{m}p'}{\mathbf{g}p' + \eta}. \qquad (20)$$

Thus, the problem in Fig. 1 can be written in a new structure as shown in Fig. 17. It can be seen that the new structure is a linear fractional programming (LFP) problem. The theorem in the following states that for

any LFP problem, there exists an equivalent LP problem. Using this theorem, the LFP problem in Fig. 17 is equivalent to the LP problem in Fig. 2. Therefore, resource management in a single cell can be modeled as an LP problem and can be solved efficiently by LP methods.   ☐

Theorem A1 [43]. *The linear fractional programming*

$$\text{Maximize}_{\mathbf{p}} \left\{ \frac{\mathbf{m}p'}{\mathbf{g}p' + \eta} \right\} \text{ subject to } \mathbf{p} \geq 0, \; \mathbf{A}p = v, gp' + \eta > 0 \qquad (21)$$

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

$$\text{Maximize}_{\mathbf{y}, u} \{\mathbf{m}y'\} \text{ subject to } \mathbf{y} \geq 0, u > 0, \qquad (22)$$
$$\mathbf{g}y' + \eta u = 1, \mathbf{A}y - vu = 0,$$

where $\mathbf{p}$, $\mathbf{y}$, and $\mathbf{v}$ belong respectively to $\Re^N$, $A \in \Re^{N \times N}$, and $u \in \Re$, and it is assumed that no point $(\mathbf{y}, 0)$ with $\mathbf{y} \geq 0$ is feasible for (22).

Proof of Corollary 1. The throughput of a single cell is

$$R_c = \frac{1}{\bar{\gamma}} \sum_{i=1}^{N} \gamma_i r_i = \frac{1}{\bar{\gamma}} \sum_{i=1}^{N} \gamma_i \frac{W}{\gamma_i} \frac{g_i p_i}{\sum_{j=1, j\neq i}^{N} g_j p_j + \eta}, \qquad (23)$$

$$\approx \frac{W}{\bar{\gamma}} \frac{\sum_{i=1}^{N} g_i p_i}{\sum_{j=1}^{N} g_j p_j + \eta}, \qquad (24)$$

$$\approx \frac{W}{\bar{\gamma}}. \qquad (25)$$

In writing (23), the E_b/I_0 constraint for the BER requirement is used. Equation (24) is derived using the approximation (19) for a large N and the assumption that η is negligible as compared with the multiple access interference term.   ☐

Proof of Theorem 1. The variable u_{a_i} is positive and the vector $\mathbf{y}_{a_i}$ is nonnegative. Accordingly,

$$y_{ia_i} = p_i u_{a_i} \leq P_{i,\max} u_{a_i}, \qquad (26)$$

$$w_i g_{ia_i} y_{ia_i} = w_i g_{ia_i} p_i u_{a_i} = \frac{w_i g_{ia_i} p_i}{\sum_{j=1, j\neq i}^{N} g_{ja_i} p_j + \eta}, \qquad (27)$$

$$\sum_{j=1, j\neq i}^{N} g_{ja_i} y_{ja_i} + \eta u_{a_i} = u_{a_i} \left(\sum_{j=1, j\neq i}^{N} g_{ja_i} + \eta \right) = 1. \tag{28}$$

Thus, the point $(\mathbf{y}_{a_i}, u_{a_i})$ is feasible. Conversely, if $(\mathbf{y}_{a_i}, u_{a_i})$ is feasible and the point $(\mathbf{y}_{a_i}, 0)$ is infeasible, then $u_{a_i} > 0$ and $\mathbf{p} = \mathbf{y}_{a_i}/u_{a_i}$ satisfies the constraints. Therefore, (6) and (7) map the optimization problem one-by-one onto the equivalent problem as presented in Fig. 5. The first constraint in this figure is a combination of the following constraints.

$$0 \quad \leq y_{ia_i} \leq P_{i,\max} u_{a_i}, \tag{29}$$

$$\frac{R_{i,\min}}{w_i g_{ia_i}} \leq y_{ia_i} \leq \frac{R_{i,\max}}{w_i g_{ia_i}}. \tag{30}$$

The result is an LP problem. □

Proof of Corollary 2. The E_b/I_0 constraint for user i connected to base station k is given in Fig. 4. The lower bound on the path gain can be evaluated based on the best possible traffic condition in the network. This condition occurs when there are no interfering users in the network and user i transmits at its maximum power and minimum rate. Substituting these values in the constraint and evaluating g_{ik}, we get

$$g_{ik} \geq \frac{\eta \, \gamma_i \, R_{i,\min}}{W \, P_{i,\max}} \tag{31}$$

which gives the desired lower bound. When the path gain is smaller than this bound, under no circumstance can the E_b/I_0 at the receiver satisfy the target BER. □

Proof of Corollary 3. For the feasibility condition of a system of two users and two base stations, we use the first phase of the simplex method. Using the rate constraint in Fig. 1, the constraints of this system are given in the following six inequalities

$$R_{1,\min} \leq \frac{w_1 g_{1a_1} p_1}{g_{2a_1} p_2 + \eta} \leq R_{1,\max}, \tag{32}$$

$$R_{2,\min} \leq \frac{w_2 g_{2a_2} p_2}{g_{1a_2} p_1 + \eta} \leq R_{2,\max}, \tag{33}$$

$$p_1 \leq P_{1,\max}, \tag{34}$$

$$p_2 \leq P_{2,\max}. \tag{35}$$

To perform the simplex feasibility test, we need to express the constraints in the standard form [44]. That is, to alter the above inequalities into equalities. For this purpose, we add the slack variable κ_j, $j \in \{1, 2, \ldots, 6\}$, and the artificial variables ν_3 and ν_4 to the inequalities. Thus, the constraints become

$$\frac{w_1}{R_{1,\max}} g_{1a_1} p_1 - g_{2a_1} p_2 + \kappa_1 = \eta, \tag{36}$$

$$\frac{w_2}{R_{2,\max}} g_{2a_2} p_2 - g_{1a_2} p_1 + \kappa_2 = \eta, \tag{37}$$

$$\frac{w_1}{R_{1,\min}} g_{1a_1} p_1 - g_{2a_1} p_2 + \nu_3 - \kappa_3 = \eta, \tag{38}$$

$$\frac{w_2}{R_{2,\min}} g_{2a_2} p_2 - g_{1a_2} p_1 + \nu_4 - \kappa_4 = \eta, \tag{39}$$

$$p_1 - P_{1,\max} + \kappa_5 = 0, \tag{40}$$

$$p_2 - P_{2,\max} + \kappa_6 = 0. \tag{41}$$

We solve this system of linear equations symbolically for p_1, p_2, and different sets of four slack and artificial variables. The desired feasibility condition is derived by applying the nonnegativity property of p_1 and p_2 to the solution of the linear system with the variables p_1, p_2, ν_3, ν_4, κ_5, and κ_6, as given in (8). □

REFERENCES

[1] ARIB IMT-2000 Study Committee, "Japan's Proposal for Candidate Radio Transmission Technology on IMT-2000: W-CDMA," Assoc. Radio Industries and Businesses, Japan, June 1998.

[2] ETSI/UTRA, "The ETSI UMTS Terrestrial Radio Access (UTRA) ITU-R RTT Candidate Submission," *European Telecomm. Standard Inst.,* June 1998.

[3] TIA/TR45.5.4, "The cdma2000 ITU-R RTT Candidate Submission (0.18)," Telecomm. Industry Assoc., 1998.

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

[5] S.V. Hanly, "An Algorithm for Combined Cell-Site Selection and Power Control to Maximize Cellular Spread Spectrum Capacity," *IEEE J. Selected Areas Comm.,* vol. 13, no. 7, pp. 1332-1340, Sept. 1995.

[6] L. Papavassiliou and L. Tassiulas, "Joint Optimal Channel, Base Station and Power Assignment for Wireless Access," *IEEE/ACM Trans. Networking,* vol. 4, no. 6, pp. 857-872, Dec. 1996.

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

[8] R. Rezaiifar, A.M. Makowski, and S.P. Kumar, "Stochastic Control of Handoffs in Cellular Networks," *IEEE J. Selected Areas Comm.,* vol. 13, no. 7, pp. 1348-1362, Sept. 1995.

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

[10] C.I. Sabnani and K.K. Sabnani, "Variable Spreading Gain CDMA with Adaptive Control for True Packet Switching Wireless Network," *Proc. IEEE Int'l Conf. Comm.,* pp. 725-730, 1995.

[11] L.C. Yun and D.G. Messerschmitt, "Power Control for Variable QoS on a CDMA Channel," *Proc. IEEE Military Comm. Conf.,* pp. 178-182, 1994.

[12] A. Sampath, P.S. Kumar, and J.M. Holtzman, "Power Control and Resource Management for a Multimedia CDMA Wireless System," *Proc. IEEE Int'l Symp. Personal, Indoor, and Mobile Radio Comm.,* vol. 1, pp. 21-25, 1995.

[13] I.F. Akyildiz, D.J. Goodman, and L. Kleinrock, "Mobility and Resource Management in Next Generation Wireless Systems," *IEEE J. Selected Areas Comm.,* vol. 19, no. 10, pp. 1825-1830, Oct. 2001.

[14] F. Berggren, S.-L. Kim, R. Jäntti, and J. Zander, "Joint Power Control and Intracell Scheduling of DS-CDMA Nonreal Time Data," *IEEE J. Selected Areas Comm.,* vol. 19, no. 10, pp. 1860-1870, Oct. 2001.

[15] V. Huang and W. Zhuang, "Optimal Resource Management in Packet-Switching TDD CDMA Systems," *IEEE Personal Comm.,* vol. 7, no. 6, pp. 26-31, Dec. 2000.

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

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

[18] M. Soleimanipour, W. Zhuang, and G.H. Freeman, "Modeling and Resource Allocation in Wireless Multimedia CDMA Systems," *Proc. IEEE Vehicular Technology Conf.,* pp. 1279-1283, 1998.

[19] S. Lombardi and W. Zhuang, "Soft Handoff in a CDMA Wireless ATM Environment," *Computer Comm.,* special issue on Recent Advances in Mobile Comm. Networks, vol. 23, no. 5-6, pp. 525-532, Mar. 2000.

[20] H.P. Williams, *Model Building in Mathematical Programming.* New York: Wiley, 1990.

[21] M. Soleimanipour, "Modeling and Resource Management in Wireless Multimedia WCDMA Systems," PhD thesis, Electrical and Computer Eng., Univ. of Waterloo, 1999.

[22] W. Zhuang, "Integrated Error Control and Power Control for DS-CDMA Multimedia Wireless Communications," *IEE Proc.—Comm.,* vol. 146, no. 6, pp. 359-365, Dec. 1999.

[23] D.P. Bertsekas, *Nonlinear Programming.* Belmont, Mass.: Athena Scientific, 1995.

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

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

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

[27] J. Abadie and J. Carpentier, "Generalization of the Wolfe Reduced Gradient Method to the Case of Nonlinear Constraints," *Optimization,* R. Fletcher, ed., pp. 37-47, 1969.

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

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

[30] C.A. Floudas, *Nonlinear and Mixed-Integer Optimization.* Oxford Univ. Press, 1995.

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

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

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

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

[35] D. Kendrick and A. Meeraus, "GAMS, An Introduction," technical report, Development and Research Dept., World Bank, 1985.

[36] A.J. Viterbi, A.M. Viterbi, and E. Zehavi, "Other-Cell Interference in Cellular Power-Controlled CDMA," *IEEE Trans. Comm.,* vol. 42, no. 4, pp. 1501-1504, Apr. 1994.

[37] A.M. Viterbi and A.J. Viterbi, "Erlang Capacity of a Power Controlled CDMA System," *IEEE J. Selected Areas Comm.,* vol. 11, no. 6, pp. 892-900, Aug. 1993.

[38] M.C. Ferris, "MATLAB and GAMS: Interfacing Optimization and Visualization Software," Math. Programming Technical Report 98-19, Computer Science Dept., Univ. of Wisconsin, Nov. 1998. Also available at http://www.cs.wisc.edu/math-prog/matlab.html.

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

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

[41] G.L. Stüber, *Principles of Mobile Communication.* second printing, Boston: Kluwer, 1997.

[42] M. Soleimanipour, G.H. Freeman, and W. Zhuang, "A Partially Decentralized Resource-Management Scheme for IMT-2000," *Proc. IEEE Globecom'00,* pp. 1548-1553, 2000.

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

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

Majid Soleimanipour received the BSc degree in communications engineering from Sharif University of Technology, Tehran, in 1985, the MSc degree in communications engineering from K.N. Toosi University of Technology, Tehran, in 1991, and the MSc and PhD degrees in electrical engineering from the University of Waterloo, Waterloo, Canada, in 1995 and 1999, respectively. From 1985 to 1990, he was with Nasr Electronic R&D Center where he worked as a researcher and project manager. He directed the center from 1987 to 1990 as the general manager. From 1990 to 1993, he was engaged in several projects in mobile communications. Since 1999, Dr. Soleimanipour has been with Imam Hossein University, Tehran, as an assistant professor and has been the Dean of the Faculty of Engineering since 2000.
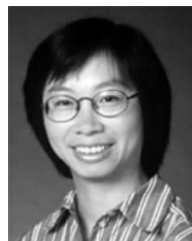
Weihua Zhuang (M'93-SM'01) received the BSc and MSc degrees from Dalian Marine University, China, in 1982 and 1985, respectively, and the PhD degree from the University of New Brunswick, Canada, in 1993, all in electrical engineering. Since October 1993, she has been with the Department of Electrical and Computer Engineering, University of Waterloo, Ontario, Canada, where she is a professor. She is a coauthor of the textbook *Wireless Communications and Networking* (Prentice Hall, 2002). Her current research interests include multimedia wireless communications, wireless networks, and radio positioning. Dr. Zhuang received the Premier's Research Excellence Award (PREA) in 2001 from the Ontario Government for demonstrated excellence of scientific and academic contributions. She is a licensed professional engineer in the Province of Ontario, Canada. She is a senior member of the IEEE.

George H. Freeman's bio and photo are not available.

▷ **For more information on this or any computing topic, please visit our Digital Library at** http://computer.org/publications/dlib.