Jointly Optimal Power and Admission Control for Delay Sensitive Traffic in CDMA Networks With LMMSE Receivers

Cristina Comaniciu and H. Vincent Poor, *Fellow, IEEE*

***Abstract*—In this paper, quality-of-service (QoS) guarantees for multiclass code division multiple access networks are provided by means of cross-layer optimization across the physical and network layers. At the physical layer, the QoS requirements are specified in terms of a target signal-to-interference ratio (SIR) requirement, and optimal target powers are dynamically adjusted according to the current number of users in the system. At the network layer, the QoS requirements are the blocking probabilities and the call connection delays. The network layer guarantees that both physical layer and network layer QoS are met by employing admission control. An optimal admission control policy is proposed based on a semi-Markov decision process formulation. The tradeoff between blocking and delay is discussed for various buffer configurations. The advantage of advanced signal processing receivers is established using a comparative capacity analysis and simulation with the classical scenario in which the system uses matched filter receivers.**

***Index Terms*—Admission control, capacity, CDMA, cross-layer, multiuser detection, power control, QoS.**

I. Introduction

OVER the past decade, both wireless communications and wireline multimedia have experienced tremendous commercial success and growth. This increase in popularity for both services has opened new challenges for the next generation of wireless networks, which are expected to provide ubiquitous multimedia coverage with quality-of-service (QoS) guarantees. A promising technology for such networks is code division multiple access (CDMA) due to its soft capacity characterization, which allows a graceful degradation of the system performance as well as statistical multiplexing of streams with varied bit error rates and delay requirements. In CDMA systems, QoS delivery relies on interference management techniques such as power control, multiuser detection, and access/admission control. While there is a substantial amount of literature concerned with each of these topics (e.g., [1], [2], [5], [9]–[11], [15], [16], [18], and the references therein), very few papers [6], [12] address the problem of interconnecting physical layer design (e.g., multiuser detection) with network layer optimization. In [6] and [12], admission/access control algorithms for CDMA networks employing multiuser receivers are proposed. The approaches in [6] and [12] are fundamentally different. While [6] focuses on potential gains that can be achieved through access control by taking advantage of real-time traffic burst activity detection, in [12], the emphasis is on optimal call admission design for maximizing the network capacity (minimizing the blocking probability). The optimal call admission problem is formulated as a semi-Markov decision process (SDMP) with constraints on blocking probabilities and signal-to-interference ratio (SIR). It is shown that the optimal admission policy can be determined via a linear programming-based algorithm, and the network capacity can be further increased if users requesting connections are queued when resources are not available. However, the approach in [12] optimizes only the network layer performance, whereas the powers at the physical layer are not optimally chosen. Furthermore, only blocking probability constraints are considered in [12] since the proposed framework does not allow the control of call connection delays generated by buffering.

In this paper, QoS guarantees for multiclass CDMA networks are provided by means of cross-layer optimization across the physical and network layers. At the physical layer, the QoS requirements are specified in terms of a target SIR requirement, and optimal target powers are dynamically adjusted according to the current number of users in the system. A closed-form capacity expression is derived for large CDMA systems using linear minimum mean square error (LMMSE) receivers in multipath fading channels and under imperfect channel estimation conditions. Our physical capacity analysis shows that the maximum SIR that can be achieved for a particular class of users is limited by the estimation accuracy of their channel link gain.

At the network layer, we consider both blocking probabilities as well as call connection delay constraints. The network layer guarantees that both physical layer and network layer QoS are met by employing admission control. An optimal admission control policy is proposed based on an SDMP formulation.

The proposed framework for admission control has several advantages over the approach in [12]. Most importantly, it is a genuine cross-layer design that adapts to interference level changes at both the physical layer (power control) and at the network layer (admission control) while exchanging information between layers in both directions. Figs. 1 and 2 illustrate the conceptual design difference between the approach we propose in this paper and the one proposed in [12]. Moreover, our

Manuscript received August 29, 2002; revised April 8, 2003. This work was supported by the National Science Foundation under Grant CCR-99-79361 and by the New Jersey Center for Pervasive Information Technology. This paper was presented in part at the 40th Annual Allerton Conference on Communication, Computing, and Control, October 2002. The associate editor coordinating the review of this paper and approving it for publication was Dr. Alfred O. Hero, III.

The authors are with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: ccomanic@ee.princeton.edu; poor@ee.princeton.edu).

Digital Object Identifier 10.1109/TSP.2003.814467

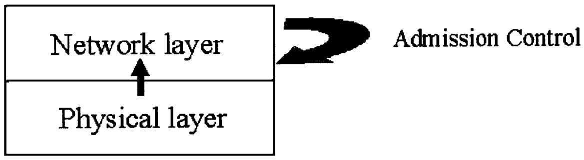


Fig. 1. Optimization of network layer performance.

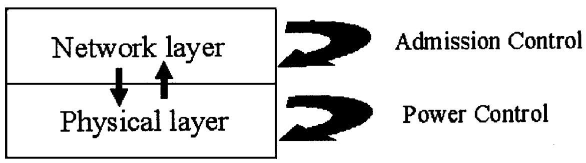


Fig. 2. Joint optimization across physical and network layer.

proposed framework can be extended to support multiple rate users and average connection delay requirements. The admission policy is based on the SMDP formulation and considers the interplay between blocking probabilities and delay requirements in a multiclass system. We show in this paper that if delay is not of concern, blocking probability constraints can be satisfied for all classes of users, using a simple threshold policy and appropriately selecting the buffer lengths. This suggests that the threshold policy may be a better option than the SMDP approach in [12]. The real advantage of an SMDP-based solution is that it provides more flexibility in achieving blocking probability/delay tradeoffs.

While implementing an optimal policy yields some performance gains over other classical admission control approaches, it can be seen that the most significant network capacity gain is a result of advanced signal processing at the physical layer. It is shown experimentally that using MMSE receivers roughly doubles the network capacity for given network QoS requirements. This is especially why it is so important to optimize the physical layer capacity by selecting appropriate target powers, according to the current level of interference, which is in turn controlled by the admission control.

The paper is organized as follows. In Section II, we introduce the system model, whereas Section III presents the asymptotic capacity derivation, and Section IV discusses the design for the optimal admission control. Section V illustrates the admission control performance using numerical examples and compares the results with classical approaches in call admission control. Finally, some concluding remarks are presented in Section VI.

II. System Model and Assumptions

We consider a single-cell, power-controlled synchronous CDMA system, which supports J classes of users, characterized by different target SIRs ϵ_j, different blocking probability requirements P_b^j, and different connection delay constraints Ξ_j, $j = 1, 2, \ldots, J$. All signature sequences are independent, randomly chosen, and normalized, having a fixed spreading gain N. Multiple rates can be achieved using multicodes [17], and the admission control design can easily be extended to account for multiple rate systems. Requests for connections occur with rates λ_j, $j = 1, 2, \ldots, J$ and are Poisson distributed. The call durations are exponentially distributed, and the mean duration is μ_j, $j = 1, 2, \ldots, J$. It is assumed that due to fast fading, the channel cannot be perfectly known, and it is characterized by its estimated average link gain $\overline{h}_j$, and its estimation variance ξ_j^2, both of which we assume to be the same for all users from an arbitrary class j. The implicit assumption for the channel model is that it is conditioned on the slower fading (free space path loss and shadow fading), which does not affect the received power over the time scale of interest. The effects of slow fading are absorbed into the attenuated transmitted power, which is defined as

$$P_k = z_k P_k^t, \qquad k = 1, 2, \ldots, K \tag{1}$$

where z_k is the path loss due to free space loss and shadow fading, P_k^t is the transmitted power, and K is the total number of users in the system.

For a multipath fading channel with L resolvable paths, we denote by $|\overline{h}_j|^2$ the equivalent estimated average power gain, which is defined as

$$|\overline{h}_j|^2 = \sum_{l=1}^{L} |\overline{h}_{jl}|^2, \qquad j = 1, 2, \ldots, J \tag{2}$$

where $\overline{h}_{jl}$ represents the average link gain for the lth path for a user in class j.

The number of users in each class is limited by the admission control according to the current system state and the admission policy. Users that cannot be immediately admitted into the system are queued using buffers of finite lengths $B(j)$, $j = 1, 2, \ldots, J$.

III. Physical Layer Constraints: Asymptotic Capacity

In this section, we derive the asymptotic system capacity for a multiclass CDMA system using LMMSE receivers under imperfect channel gain estimation. The analysis is based on large random matrix results for CDMA systems in fading environments presented in [8]. We start with the general expression for the SIR presented in [8] for a large system with random spreading. Considering a multipath fading environment with L resolvable paths, the SIR achieved by an arbitrary user (say the kth one) can be expressed as

$$\text{SIR} = \frac{P_k \sum_{l=1}^{L} |\overline{h}_{kl}|^2 \beta}{1 + P_k \xi_k^2 \beta} = \frac{P_k |\overline{h}_k|^2 \beta}{1 + P_k \xi_k^2 \beta} \tag{3}$$

where β is the unique fixed point in $(0, \infty)$ that satisfies

$$\beta = \left[\sigma^2 + \frac{1}{N} \sum_{k=2}^{K} I_k^*(L, \beta, \xi_k^2, |\overline{h}_k|^2)\right]^{-1} \tag{4}$$

with $I_k^*(L, \beta, \xi_k^2, |\overline{h}_k|^2) = (L-1) I(\xi_k^2 P_k, \beta) + I(P_k(\xi_k^2 + |\overline{h}_k|^2), \beta)$, and $I(p, \beta) = p/(1 + p\beta)$.

As $K, N \to \infty$, β approaches a constant for all users. If we impose the condition that the achieved SIR should be ϵ_j for all class j users, $j = 1, 2, \ldots, J$ (i.e., the minimum power solution is achieved with equality), and using the assumption that all users in the same class have the same channel characteristics, it follows from (3) that all users in the same class j must have equal attenuated transmitted powers:

$$P_j = \frac{\epsilon_j}{\beta(|\overline{h}_j|^2 - \xi_j^2 \epsilon_j)} = \frac{\epsilon_j}{\beta|\overline{h}_j|^2(1-\nu_j\epsilon_j)} \quad j = 1, 2, \ldots, J \tag{5}$$

where $\nu_j = (\xi_j^2/|\overline{h}_j|^2)$.

Since the powers must be positive and $\beta \geq 0$, an immediate result is that $(1-\nu_j\epsilon_j) \geq 0$, and the maximum achievable SIR is

$$\epsilon_j < 1/\nu_j, \qquad j = 1, 2, \ldots, J. \tag{6}$$

Denoting $Q_j = |\overline{h}_j|^2 P_j$ for the class j of users and imposing the constraint that the SIR condition should hold with equality ($\text{SIR}_j = \epsilon_j$), from (3), we can express β as

$$\beta = \frac{\epsilon_j}{Q_j(1-\nu_j\epsilon_j)}, \qquad j = 1, 2, \ldots, J. \tag{7}$$

Thus, for multiple classes of users, the following equality holds:

$$\frac{\epsilon_i}{Q_i(1-\nu_i\epsilon_i)} = \frac{\epsilon_j}{Q_j(1-\nu_j\epsilon_j)}, \qquad \forall\, i, j = 1, 2, \ldots, J. \tag{8}$$

Expressing the SIR condition for an arbitrary user in class 1, we have

$$\left[\sigma^2 + \frac{1}{N}\sum_{j=1}^{J}\sum_{k=2}^{K_J}\left((L-1)\frac{Q_j\nu_j}{1+Q_j\nu_j\frac{\epsilon_j}{Q_j(1-\epsilon_j\nu_j)}} + \frac{Q_j(1+\nu_j)}{1+Q_j(1+\nu_j)\frac{\epsilon_j}{Q_j(1-\epsilon_j\nu_j)}}\right)\right]^{-1} = \frac{\epsilon_1}{Q_1(1-\nu_1\epsilon_1)}. \tag{9}$$

Using the notation $\alpha_j = K_j/N$, i.e., α_j is the number of users per dimension, the equality in (9) becomes

$$\left[\sigma^2 + \sum_{j=1}^{J}\alpha_j\left((L-1)\frac{Q_j\nu_j}{1+\nu_j\frac{\epsilon_j}{(1-\epsilon_j\nu_j)}} + \frac{Q_j(1+\nu_j)}{1+(1+\nu_j)\frac{\epsilon_j}{(1-\epsilon_j\nu_j)}}\right)\right]^{-1} = \frac{\epsilon_1}{Q_1(1-\nu_1\epsilon_1)}. \tag{10}$$

Using (8), we have

$$\left[\sigma^2 + \sum_{j=1}^{J}\alpha_j\left((L-1)\nu_j\epsilon_j(1-\nu_1\epsilon_1)Q_1/\epsilon_1 + \frac{Q_1(1+\nu_j)\frac{\epsilon_j}{\epsilon_1}\frac{1-\nu_1\epsilon_1}{1-\nu_j\epsilon_j}}{1+(1+\nu_j)\frac{\epsilon_j}{(1-\epsilon_j\nu_j)}}\right)\right]^{-1} = \frac{\epsilon_1}{Q_1(1-\nu_1\epsilon_1)}. \tag{11}$$

Hence, the transmitted power for user 1 can be determined as

$$P_1^t = \frac{Q_1}{z_1|\overline{h}|_1^2} \tag{12}$$

where

$$Q_1 = \frac{\epsilon_1\sigma^2}{(1-\nu_1\epsilon_1)\left(1-\sum_{j=1}^{J}\alpha_j\left[(L-1)\nu_j\epsilon_j + \frac{\epsilon_j(1+\nu_j)}{1+\epsilon_j}\right]\right)}. \tag{13}$$

Since the transmitted power needs to be positive, and considering (6) and (13), the system capacity is restricted by the power control feasibility condition

$$\sum_{j=1}^{J}\alpha_j\left[(L-1)\nu_j\epsilon_j + \frac{\epsilon_j(1+\nu_j)}{1+\epsilon_j}\right] < 1. \tag{14}$$

The flat fading channel case can be obtained by setting $L = 1$.

The above derivation can be summarized in the following theorem.

Theorem 1: In an asymptotically large CDMA system ($K_j \to \infty$, $N \to \infty$ and $\alpha_j = K_j/N$ is constant) operating with linear MMSE receivers in a multipath fading environment with imperfect channel estimation, a minimum received power solution exists such that all users achieve their target SIRs if and only if

$$\epsilon_j < \frac{1}{\nu_j}$$

and

$$\sum_{j=1}^{J}\alpha_j\left[(L-1)\nu_j\epsilon_j + \frac{\epsilon_j(1+\nu_j)}{1+\epsilon_j}\right] < 1.$$

The minimum transmit power solution for a user in class i is given by

$$P_i^t = \frac{\epsilon_i\sigma^2}{z_i|\overline{h}_i|^2(1-\nu_i\epsilon_i)\left(1-\sum_{j=1}^{J}\alpha_j\left[(L-1)\nu_j\epsilon_j + \frac{\epsilon_j(1+\nu_j)}{1+\epsilon_j}\right]\right)}.$$

The above results can be straightforwardly extended to account for multiple rates implemented using multiple codes. Each high rate user in class j is assigned M_j signature sequences and transmits at M_j times the basic rate (obtained using the lowest spreading gain N). A high rate user is equivalent to M virtual low rate users. Hence, the power control feasibility condition can be easily adapted from (14) to be

$$\sum_{j=1}^{J}M_j\alpha_j\left[(L-1)\nu_j\epsilon_j + \frac{\epsilon_j(1+\nu_j)}{1+\epsilon_j}\right] < 1 \tag{15}$$

where $\alpha_j = K_j/N$, and K_j is the number of the class j users physically present in the system.

Theorem 1 illustrates the interdependence between the physical and the network layer performance. Referring to Fig. 2, we can see that the capacity conditions are passed to the network layer, and we will see shortly that they will determine the set of feasible system states used to select the optimal admission policy. On the other hand, the optimal power allocation (13) depends on the number of the users (in different classes) currently admitted in the network. This information should be passed along from the network layer to the physical layer such that appropriate target powers can be selected to optimize the physical layer performance.

IV. Network QoS: Optimal Admission Control

While the previous section characterizes the system capacity from the physical layer perspective, the admission policy must also take into account network QoS requirements such as blocking probability and average connection delay. We consider the interplay between physical and network layer constraints using an equivalent queueing system, as illustrated in Fig. 3. The average connection delays and the blocking probabilities can be derived using a queueing analysis. The service rate for each queue is varied by the admission control such that the power control feasibility holds, and hence, all users can meet their target SIRs. The equivalent queueing problem consists of J $M/M/K_j/B(j)$ queues [$B(j)$ represents the length of buffer for queue j], $j = 1, 2, \ldots, J$, for which the service rate depends on the current number of connections [such that $\alpha_j = K_j/N$ satisfies (14)], as well as on the current number of calls waiting for connection (according to delay and blocking constraints). We can formulate the admission control problem as an SMDP under the assumption that the following Markovian properties hold [13].

• Given the current decision time, when action $\mathbf{a}$ is chosen in state i, the time until the next decision epoch and the state at the next decision epoch depend only on the present state and on the current chosen action $\mathbf{a}$ and are independent of the past history of the system.

• The costs incurred until the next decision epoch also depend only on the current state and chosen action $\mathbf{a}$.

The optimality criterion for an SMDP is the long-run average cost per unit time.

SMDPs can essentially be solved by considering an equivalent discrete time average cost Markov decision process using a process called uniformization [4]. As a consequence, algorithms such as policy iteration, value iteration, and linear programming (LP) can be used to provide solutions for the SMDP problem. We are particularly interested in the LP approach, which allows us to introduce easily optimization constraints such as maximum allowed connection delays and/or blocking probabilities.

An SMDP is completely characterized [13] by the following.

- $p_{\mathbf{xy}}(\mathbf{a}) =$ the probability that at the next decision epoch the system will be in state $\mathbf{y}$, if action $\mathbf{a}$ is selected at the current state $\mathbf{x}$.
- $\tau_{\mathbf{x}}(\mathbf{a})$ (sojourn time) = the expected time until the next decision epoch after action $\mathbf{a}$ is chosen in the present state $\mathbf{x}$.

$$\tau_{\mathbf{x}}(\mathbf{a}) > 0, \qquad \forall \mathbf{x} \in \boldsymbol{X}, \quad \mathbf{a} \in \boldsymbol{A}_{\mathbf{x}}$$

  where $\boldsymbol{X}$ represents the state space, and $\boldsymbol{A}_{\mathbf{x}}$ represents the admissible action space.
- $c(\mathbf{x}, \mathbf{a}) =$ the expected costs incurred until the next decision epoch after action $\mathbf{a}$ is chosen in the current state $\mathbf{x}$.

In order to model the admission problem as an SMDP, we proceed to define the main constituents that characterize it: the state space, the decision epochs, the action space, the state dynamics, and the performance criteria (cost and probabilistic constraints).

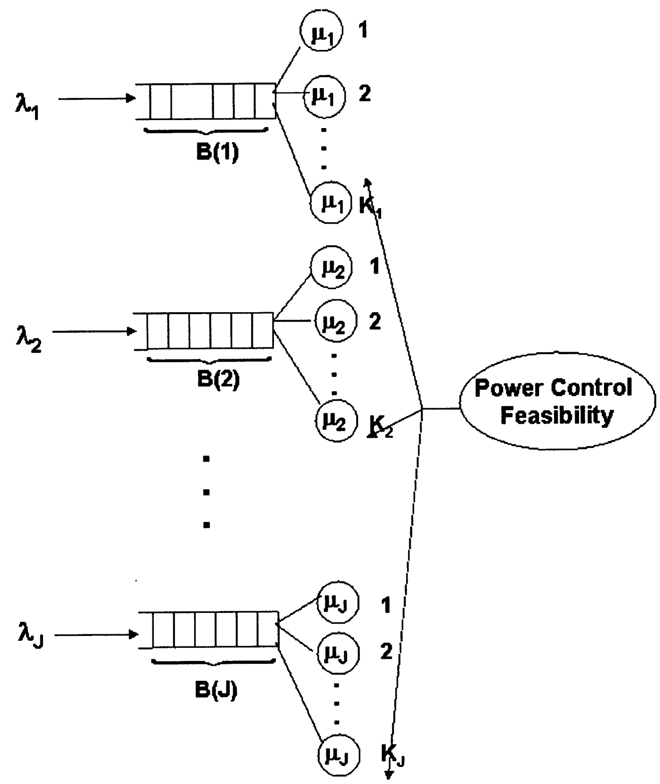


Fig. 3. Equivalent queueing system.

The State Space

The state of each queue j is characterized by the number of users $n_q^j(t)$ in the queue at time t, $t > 0$, and the number of servers $n_s^j(t)$ at time t (which is equivalent to the number of connections admitted for class j). Thus, we define the state of the system at decision epoch t as

$$\mathbf{x}(t) = [n_q^1(t), n_s^1(t), n_q^2(t), n_s^2(t), \ldots, n_q^J(t), n_s^J(t)]. \quad (16)$$

Since the arrivals and departures of users are random, $\{\mathbf{x}(t), t > 0\}$ represents a finite state stochastic process. The state space $\boldsymbol{X}$ is comprised of any state vector $\mathbf{x}$, such that SIR constraints can be met:

$$\boldsymbol{X} = \left\{ \mathbf{x}: n_q^i \le B(i), i = 1, 2, \ldots, J; \sum_{j=1}^{J} \frac{n_s^j}{N} \Lambda_j \le 1 \right\} \quad (17)$$

where we use the notation $\Lambda_j = (L-1)\nu_j\epsilon_j + (\epsilon_j(1+\nu_j)/(1+\epsilon_j))$.

For a multirate, multicode system, the state space is alternatively defined as

$$\boldsymbol{X} = \left\{ \mathbf{x}: n_q^i \le B(i), i = 1, 2, \ldots, J; \sum_{j=1}^{J} \frac{M_j n_s^j}{N} \Lambda_j \le 1 \right\} \quad (18)$$

where Λ_j has the same expression as before.

Decision Epochs

Every time a new user arrives and requests a new connection, and any time a departure occurs, the state of the system changes. Since these changes in the system state should affect the admission process, we choose the decision epochs to be the set of all arrival and departure instances.

Action Space

At each decision epoch, an action $\mathbf{a}$ is chosen that determines how the admission control will perform at the next decision moment. The action vector $\mathbf{a}$ is state dependent, and its components depend on the type of event: arrival or departure. In general, action $\mathbf{a}$ at decision epoch t is defined as

$$\mathbf{a}(t) = [a_1^a(t),\, a_1^d(t),\, a_2^a(t),\, a_2^d(t),\, \ldots,\, a_J^a(t),\, a_J^d(t)] \quad (19)$$

where a_i^a denotes the action for queue i if an arrival occurs, and a_i^d denotes the action for queue i if a departure occurs, which are defined as follows:

$$a_i^a = \begin{cases} 0; & \text{maintain the number of servers for queue } i \\ 1; & \text{increase (by 1) the number of servers for queue } i \end{cases}$$

$$a_i^d = \begin{cases} 0; & \text{decrease (by 1) the number of servers for queue } i \\ 1; & \text{maintain the number of servers for queue } i. \end{cases}$$

The action space can be defined as the set of all possible actions:

$$A = \{\mathbf{a} \colon \mathbf{a} \in \{0, 1\}^{2J},\ i = 1, 2, \ldots, J\}. \quad (20)$$

The action space must be restricted for a given state $\mathbf{x} \in X$ such that the selected action will not result in a transition into a state that it is not allowed (not in X). In addition, we restrict the admissible action space $(A_{\mathbf{x}})$ such that $1/\tau_{\mathbf{x}}(\mathbf{a}) > 0, \forall \mathbf{x} \in X$, $\mathbf{a} \in A_{\mathbf{x}}$. More specifically, we impose the condition that $(a_1^a, a_2^a, \ldots, a_J^a) \neq (0, 0, \ldots, 0)$ if the system is in state $\mathbf{x} = (0, 0, 0, 0, \ldots, 0, 0)$. If this is not true, $\tau_{\mathbf{x}}(\mathbf{a}) = \infty$ for this particular configuration, which would result in the system being forever in state $\mathbf{x} = (0, 0, 0, 0, \ldots, 0, 0)$. Thus, the admissible action space $A_{\mathbf{x}}$ can be defined as

$$\begin{aligned} A_{\mathbf{x}} = \{\mathbf{a} \in A \colon a_j^a = 0 \text{ if } \mathbf{x} + (0, 0, \ldots, \underbrace{0, 1}, \ldots, 0, 0) \notin X \\ \text{and } (a_1^a, a_2^a, \ldots, a_J^a) \neq (0, 0, \ldots, 0) \\ \text{if } \mathbf{x} = (0, 0, 0, 0, \ldots, 0, 0)\}. \end{aligned} \quad (21)$$

State Dynamics

The state dynamics of an SMDP can be characterized by the transition probabilities of its embedded chain and the expected sojourn time for each state-action pair [4].

We define the following notation:

- $\delta(x) = \begin{cases} 0; & x = 0 \\ 1; & x > 0. \end{cases}$
- $\mathbf{x_i} = [n_q^i, n_s^i]$ represents the state vector for class i users, such that the state vector for the system can be expressed as $\mathbf{x} = [\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_J}]$.
- $\mathbf{e_i^q}$ represents a vector of dimension $2J$ containing only zeros except for the position $2(i-1)+1$, which contains a 1; $\mathbf{x} + \mathbf{e_i^q}$ is equivalent to $\mathbf{x_i} + [1, 0]$ and maps an increase in the queue of class i users by 1.
- $\mathbf{e_i^s}$ represents a vector of dimension $2J$, containing only zeros except for the position $2(i-1)+2$, which contains a 1; $\mathbf{x} + \mathbf{e_i^s}$ is equivalent to $\mathbf{x_i} + [0, 1]$ and maps an increase in the number of servers for class i users (number of admitted users) by 1.

Derivations of $p_{\mathbf{xy}}(\mathbf{a})$ and $\tau_{\mathbf{x}}(\mathbf{a})$ rely on the statistical properties of the arrival and departure processes, which are Poisson distributed and mutually independent. It follows that the cumulative process is also Poisson, and thus, the cumulative event rate is the sum of the rates for all constituent processes. It should be mentioned that arrivals that are blocked do not constitute an event such that the cumulative process includes only the unblocked arrivals, which are also Poisson distributed with rate $\lambda_j(1 - P_b^j)$. Hence, the interevent time $\tau_x(\mathbf{a})$ (the expected sojourn time) can be defined as the inverse of the event rate

$$\tau_{\mathbf{x}}(\mathbf{a}) = \left[\sum_{i=1}^{J} \lambda_i a_i^a + \sum_{i=1}^{J} \lambda_i (1 - a_i^a) \cdot \delta(B(i) - n_q^i) + \sum_{i=1}^{J} \mu_i n_s^i \right]^{-1}. \quad (22)$$

Equation (22) can be interpreted as follows: The embedded chain always changes state when an arrival occurs unless the arrival is blocked (the queue is full, and no new servers are allocated for that particular queue); in addition, it always changes state when a departure occurs.

To derive the transition probabilities, we use the decomposition property of a Poisson process: An event of certain type occurs (e.g., arrival class j, departure class i) with a probability equal to the ratio between the rate of that particular type of event and the total cumulative event rate $1/\tau_{\mathbf{x}}(\mathbf{a})$. Hence, the transition probabilities for the embedded Markov chain are determined to be

$$p_{\mathbf{xy}}(\mathbf{a}) = \begin{cases} \lambda_i a_i^a \tau_{\mathbf{x}}(\mathbf{a}); & \text{if } \mathbf{y} = \mathbf{x} + \mathbf{e_i^s} \\ \lambda_i (1 - a_i^a) \delta(B(i) - n_q^i) \tau_{\mathbf{x}}(\mathbf{a}); & \text{if } \mathbf{y} = \mathbf{x} + \mathbf{e_i^q} \\ \mu_i n_s^i a_i^d \tau_{\mathbf{x}}(\mathbf{a}); & \text{if } \mathbf{y} = \mathbf{x} - \mathbf{e_i^q} \\ \mu_i n_s^i (1 - a_i^d) \tau_{\mathbf{x}}(\mathbf{a}) \\ \quad + \mu_i n_s^i a_i^d (1 - \delta(n_q^i)) \tau_{\mathbf{x}}(\mathbf{a}); & \text{if } \mathbf{y} = \mathbf{x} - \mathbf{e_i^s} \\ 0; & \text{otherwise.} \end{cases} \quad (23)$$

Optimal Policy: Linear Programming Approach

For any given state $\mathbf{x} \in X$, an action $\mathbf{a}$ is selected according to a specified policy R. A stationary policy R is a function that maps the state space into the admissible action space, where the class of admissible policies can be defined as

$$R_{\mathbf{x},\mathbf{a}} = \{R | R \colon X \to A_{\mathbf{x}},\ 1/\tau_{\mathbf{x}}(R) > 0\}. \quad (24)$$

According to [4], an average cost criterion for a given policy R and an initial state $\mathbf{x}_0$ can be associated with the SMDP:

$$J_R(\mathbf{x}_0) = \lim_{T\to\infty} \frac{1}{T} E\left\{\int_0^T c(\mathbf{x}(t),\, \mathbf{a}(t))\, dt\right\}. \tag{25}$$

An optimal policy R^* that minimizes an average cost criterion $J_R(\mathbf{x}_0)$ for any initial state $\mathbf{x}_0$ exists under the weak unichain assumption. The weak unichain assumption requires that the average cost optimal policy has no disjoint closed sets, but the class of transient states can vary from policy to policy, and nonoptimal policies may have multiple disjoint sets [13].

In (25), $c(\mathbf{x}(t),\, \mathbf{a}(t))$ can be interpreted as the expected cost until the next decision epoch and will be selected to meet the network layer performance criteria, as will be discussed in the next section. An optimal policy for the above defined SMDP process can be determined using a linear programming approach. The optimal policy $R^*(\mathbf{x}) \in R_{\mathbf{x},\,\mathbf{a}},\, \mathbf{x} \in X$ can be obtained using the decision variables $u_{\mathbf{xa}},\, \mathbf{x} \in X,\, \mathbf{a} \in A_{\mathbf{x}}$, which are obtained by solving the linear program associated with the SMDP [13]:

$$\min_{\substack{u_{\mathbf{x},\,\mathbf{a}}\geq 0\\ \mathbf{x}\in X\\ \mathbf{a}\in A_{\mathbf{x}}}} \sum_{\mathbf{x}\in X}\sum_{\mathbf{a}\in A_{\mathbf{x}}}\sum_{i=1}^{J} \theta_i c_i(\mathbf{x},\, \mathbf{a}) u_{\mathbf{x},\,\mathbf{a}} \tau_{\mathbf{x}}(\mathbf{a})$$

subject to the constraints

$$\sum_{\mathbf{a}\in A_{\mathbf{y}}} u_{\mathbf{ya}} - \sum_{\mathbf{x}\in X}\sum_{\mathbf{a}\in A_{\mathbf{x}}} p_{\mathbf{xy}}(\mathbf{a}) u_{\mathbf{xa}} = 0, \qquad \mathbf{y} \in X$$

and

$$\sum_{\mathbf{x}\in X}\sum_{\mathbf{a}\in A_x} u_{\mathbf{xa}} \tau_{\mathbf{x}}(\mathbf{a}) = 1. \tag{26}$$

A heuristic explanation [13] for (26) is to interpret $u_{\mathbf{xa}}\tau_{\mathbf{x}}(\mathbf{a})$ as the steady-state probability of being in state $\mathbf{x}$ and choosing action $\mathbf{a}$ (for an aperiodic Markov chain). Hence, the objective function is to minimize the long run average of the cost function per unit time. The first constraint in (26) can be interpreted as a balance equation, and the second constraint requires that the sum of the steady-state probabilities should be equal to 1.

The coefficient θ_i represents the weighting of the cost function for a particular class i.

The optimal policy for the admission control can be constructed as follows [13].

- $\forall \mathbf{x}$, choose any $\mathbf{a}^*$ such that $u_{\mathbf{xa}^*} > 0$; then set the optimal policy for state $\mathbf{x}$ to be $R^*(\mathbf{x}) = \mathbf{a}^*$.
- If $u_{\mathbf{xa}} > 0\, \forall \mathbf{a}^* \in \boldsymbol{A}_{\mathbf{x}}$, arbitrarily choose $\mathbf{a}^*$, and then, set the optimal policy for state $\mathbf{x}$ to be $R^*(\mathbf{x}) = \mathbf{a}^*$.

As we will see shortly, for our purpose of meeting QoS requirements in terms of blocking probabilities and average delay constraints, it is very important to be able to solve a constrained optimization. The linear programming approach allows us to introduce very easily probabilistic constraints related to some expected cost functions $c'(\mathbf{x},\, \mathbf{a})$ [13]:

$$\sum_{\mathbf{x}\in X,\, \mathbf{a}\in A_{\mathbf{x}}} c'(\mathbf{x},\, \mathbf{a}) u_{\mathbf{xa}} \tau_{\mathbf{x}}(\mathbf{a}) \leq C \tag{27}$$

where C is a fixed value constraint.

When probabilistic constraints are imposed, the optimal policy becomes a randomized policy: In each state, an action $\mathbf{a}$ is chosen randomly according to a probability $\pi^*_{\mathbf{a}}(\mathbf{x}) = u_{\mathbf{x},\,\mathbf{a}}/\sum_{\mathbf{a}\in A_{\mathbf{x}}} u_{\mathbf{x},\,\mathbf{a}}$. The randomized policy can be specified as a matrix $R^*_{(dim(X),\, dim(A_{\mathbf{x}}))}$, with each entry given as $R^*(i,\, j) = \pi^*_j(i)$. The $(i,\, j)$th entry for matrix R^* represents the probability that action j is selected when the system is in state i. The matrix R^* is determined offline, and the admission control randomly chooses actions at each decisions epoch, according to the corresponding probabilities from matrix R^*.

Cost Functions and Network QoS

The network layer performance measures are the blocking probabilities, $P_b^j,\, j = 1,\, \ldots,\, J$ (which reflect the system capacity) and the average connection delays $W_j,\, j = 1,\, \ldots,\, J$. More specifically, we determine the admission control policy such that network QoS requirements for all admitted users can be met. The network QoS requirements are given as

$$\begin{cases} P_b^j \leq \Psi_j, & j = 1,\, \ldots,\, J\\ W_j \leq \Xi_j, & j = 1,\, \ldots,\, J. \end{cases} \tag{28}$$

We now discuss how to select the cost functions $c(\mathbf{x},\, \mathbf{a})$ and $c'(\mathbf{x},\, \mathbf{a})$, such that the optimal policy can be obtained using the linear programming approach previously outlined.

Blocking Probabilility

In [12], the authors proved that the blocking probability can be expressed as an average cost criterion for a similar system setting. Therefore, in this paper, we just give an interpretation of the cost expression for the blocking probability for our particular framework. The expression obtained for the blocking probability is similar to that in [12], and the proof in [12] applies directly.

A call from class j is blocked if the queue is full and the action selected a_j^a is zero. We can thus define the blocking probability for class j as the fraction of time the system is in some subset of states $X^b = \{\mathbf{x} \in X\colon n_j^q = B(j)\}$, and the chosen action is in some subset of the admissible action space $A^b_{\mathbf{x}} = \{\mathbf{a} \in A_{\mathbf{x}}\colon a_j^a = 0\}$ [13]

$$\begin{aligned} P_b^j &= \sum_{\mathbf{x}\in X^b,\, \mathbf{a}\in A^b_{\mathbf{x}}} \tau_{\mathbf{x}}(\mathbf{a})\\ &= \lim_{T\to\infty} \frac{1}{T} \mathbf{E}\left\{\int_0^T (1 - a_j^a(t))\right.\\ &\qquad \left. \cdot\, (1 - \delta(B(j) - n_j^q(t)))\tau_{\mathbf{x}(t)}(\mathbf{a}(t))\, dt\right\}. \end{aligned} \tag{29}$$

Expression (29) represents the cumulative average blocking probability. We can then obtain an expression for the expected blocking probability until the next decision epoch, when action $\mathbf{a}$ is chosen in current state $\mathbf{x}$:

$$c(\mathbf{x},\, \mathbf{a}) = (1 - a_j^a)(1 - \delta(B(j) - n_j^q)). \tag{30}$$

To minimize a weighted sum of blocking probabilities for all users in the system, (30) gives the expected cost in (26). Furthermore, blocking probability constraints can be met by selecting $c'(\mathbf{x}, \mathbf{a}) = (1 - a_j^a)(1 - \delta(B(j) - n_j^q))$ and $C = \Psi_j$ in (27).

Average Delay

As opposed to the blocking probability, the average connection delay cannot be expressed as an average cost criterion. However, we will show that a combination of cost functions can be used to ensure that the QoS requirements in (28) are met for all users if the requirements are feasible.

The average connection delay can be expressed using a queueing analysis for the equivalent system in Fig. 3. The delay expression for a particular class j is given by Little's theorem as a function of the average number of calls in the jth queue N_q^j, the arrival rate λ_j, and the blocking probability for class j, P_b^j:

$$W_j = \frac{N_q^j}{\lambda_j(1 - P_b^j)}. \tag{31}$$

The delay restrictions imposed in (28) for a class j of users can be rewritten as

$$N_q^j \le \Xi_j \lambda_j (1 - P_b^j). \tag{32}$$

Since $P_b^j \le \Psi_j$ is also required, (32) is guaranteed to be met if

$$N_q^j \le \Xi_j \lambda_j (1 - \Psi_j). \tag{33}$$

Therefore, the network QoS requirements can be reformulated to be

$$\begin{cases} P_b^j \le \Psi_j, & j = 1, \ldots, J \\ N_q^j \le \Xi_j \lambda_j (1 - \Psi_j), & j = 1, \ldots, J. \end{cases} \tag{34}$$

The average number of calls in the queue can be expressed as an average cost function by selecting the expected cost until the next decision epoch to be

$$c(\mathbf{x}, \mathbf{a}) = n_q^j. \tag{35}$$

Hence, to determine an admission policy which satisfies the restrictions in (34), constraints on both the blocking probability and on the average number of queued call requests must be imposed. The last ones can be obtained by selecting $c'(\mathbf{x}, \mathbf{a}) = n_q^j$ and $C = \Xi_j \lambda_j (1 - \Psi_j)$ in (27).

As a consequence, we can formulate the following proposition.

Proposition 1: An optimal admission control policy can be determined as a solution of a constrained linear programming optimization, such that the network QoS requirements in (28) can be guaranteed for all users if the system is feasible. The linear program can be formulated as follows:

$$\min_{\substack{u_{\mathbf{x},\mathbf{a}} \ge 0 \\ \mathbf{x} \in X \\ \mathbf{a} \in A_{\mathbf{x}}}} \sum_{\mathbf{x} \in X} \sum_{\mathbf{a} \in A_{\mathbf{x}}} \sum_{i=1}^{J} \theta_i (1 - a_i^a) \cdot (1 - \delta(B(i) - n_i^q)) u_{\mathbf{x},\mathbf{a}} \tau_{\mathbf{x}}(\mathbf{a})$$

subject to the constraints

$$\begin{aligned} &\sum_{\mathbf{a} \in A_{\mathbf{y}}} u_{\mathbf{ya}} - \sum_{\mathbf{x} \in X} \sum_{\mathbf{a} \in A_{\mathbf{x}}} p_{\mathbf{xy}}(\mathbf{a}) u_{\mathbf{xa}} = 0, \qquad \mathbf{y} \in X \\ &\sum_{\mathbf{x} \in X} \sum_{\mathbf{a} \in A_x} u_{\mathbf{xa}} \tau_{\mathbf{x}}(\mathbf{a}) = 1. \\ &\sum_{\mathbf{x} \in X,\, \mathbf{a} \in A_{\mathbf{x}}} (1 - a_j^a)(1 - \delta(B(j) - n_j^q)) \\ &\qquad \cdot u_{\mathbf{xa}} \tau_{\mathbf{x}}(\mathbf{a}) \le \Psi_j, \qquad j = 1, \ldots, J, \\ &\sum_{\mathbf{x} \in X,\, \mathbf{a} \in A_{\mathbf{x}}} n_q^j u_{\mathbf{xa}} \tau_{\mathbf{x}}(\mathbf{a}) \le \Xi_j \lambda_j (1 - \Psi_j), \\ &\qquad j = 1, \ldots, J. \end{aligned} \tag{36}$$

The optimal solution obtained by solving (36) minimizes both the blocking probabilities (a weighted sum for all classes) as well as the average delays, subject to the network QoS constraints in (28).

Proof: The first three equations in (36) represent the standard linear programming solution for an SMDP formulation to obtain an optimal admission control policy that minimizes a certain average cost function. This cost function is selected according to our previous discussion such that the weighted sum of blocking probabilities for all classes of users is minimized.

Moreover, we already showed that the last two constraints in (36) reflect the network QoS requirements in (28), which are expressed in the equivalent form (34).

We now show that minimizing the blocking probabilities P_b^j minimizes the average delay W_j over all policies that satisfy the network QoS constraints. Without loss of generality, we consider an arbitrary class j for which the network QoS constraints are satisfied, and we denote $\phi_j = \Xi_j \lambda_j (1 - \Psi_j)$, such that

$$W_j \le \frac{\phi_j}{\lambda_j (1 - P_b^j)}. \tag{37}$$

The optimal admission policy R^*, determined as a solution to (36), yields a minimum blocking probability $P_b^{j^*}$ for class j over the set of all possible policies that satisfy the QoS constraints R_c:

$$P_b^{j^*} = \min_{R \in R_c} P_b^j(R). \tag{38}$$

Hence, $P_b^{j^*} \le P_b^j(R), \forall R \in R_c$, and given (37), we have that

$$W_{j^*}(R^*) \le \frac{\phi_j}{\lambda_j (1 - P_b^{j^*})} \le \frac{\phi_j}{\lambda_j (1 - P_b^j(R))}, \qquad \forall R \in R_c. \tag{39}$$

□

Proposition 1 gives the optimal admission policy that minimizes blocking probabilities and average delays under certain network QoS requirement constraints. However, the admission policy exists if the system is feasible, that is, if for the given arrival rate for each class, the network QoS requirements can be met for the given buffers' dimension. As we will see also in the numerical results section, not all buffers' configurations result in a feasible solution. The buffers' dimension is thus a parameter of the optimization being closely related to the blocking probability. In case of infeasibility, the linear programming can be reformulated for different buffers' configurations. We will present a numerical example in the next section.

There are also cases in which the arrival rate is simply too high, and no buffers' configuration can be found to accommodate the network QoS requirements. Theoretically, we can define a maximum arrival rate per class that can be supported by the network such that QoS requirements are met. This represents the network capacity. However, due to the dynamic programming approach that we used for designing admission control, it is hard to determine analytically what the network capacity is. The solution is to try to solve the LP optimization with increasingly lower arrival rates until the system becomes feasible. Let us denote by λ_j^* the network capacity determined using this trial and error procedure for an arbitrary class j. Since $\lambda_j > \lambda_j^*$, the final blocking probability requirements have to be relaxed, and the resulting blocking probability for class j will be

$$P_{bf}^j = P_b^j p_b^j \tag{40}$$

where $p_b^j = 1 - (\lambda_j^* / \lambda_j)$.

In other words, the admission control policy will be selected as a solution to an LP formulation with a lower arrival rate λ_j^* such that it will meet the specified QoS requirements for this arrival rate. The above discussion on the design of such optimal policy directly applies, with the only difference being that further action is needed to reduce the arrival rate, and the final admission control will be implemented in two steps. To reduce the arrival rate from initial rate λ_j to λ_j^*, a higher level admission control will be implemented as follows.

Before requesting a new call connection, each user in class j will run a Bernoulli trial experiment with a probability of success $p_{adm}^j = 1 - p_b^j$. In case of success, the request for connection is made; otherwise, the call is automatically rejected. This ensures that the rate of call connection requests is reduced to $\lambda_j^* = p_{adm}^j \lambda_j$, and the Poisson distribution of the call connection requests is preserved.

After the call connection request has been made, the call is admitted or rejected according to the previously discussed optimal admission policy, based on the SMDP formulation.

V. Numerical Examples and Comparisons

The performance of the proposed call admission control is illustrated for a two-class system having equal high transmission rates corresponding to an equivalent spreading gain $N_e = 8$. If $N_e = 8$ is the actual spreading gain (N), we expect that the asymptotic approximation used for deriving the system capacity will be loose. In [8], it was shown experimentally that the asymptotic approximation is fairly accurate for systems employing a spreading gain on the order of $N = 128$. As the spreading gain decreases, it was shown that while the achieved average SIR remains close to the theoretical values, the variance increases, thus increasing the uncertainty of obtaining a specified target SIR.

Alternatively, the equivalent spreading gain $N_e = 8$ can be obtained using $M = N/N_e$ multiple code transmission, and a higher spreading gain $N = 128$. Using a higher spreading gain will ensure the accuracy of the asymptotic approximation, and it is therefore more suitable for higher rate users.

The implementation of the admission control is transparent to the multicode implementation. The only difference resides in defining the state space according to (17) or (18), which are, in fact, equivalent for $M = N/N_e$.

We now discuss experimental results obtained for a multicode system with $N = 128$, where each user transmits using $M = 16$ codes. The equivalent spreading gain is $N_e = 8$.

The arrival and departure rates for calls in classes 1 and 2 are selected to be $\lambda_1 = 1$, $\lambda_2 = 0.5$, $\mu_1 = 0.25$, and $\mu_2 = 0.1375$, respectively. The target SIRs for both classes are $\epsilon_1 = \epsilon_2 = 10$.

For the simulations, the channel parameters are chosen to be $|\overline{h}|^2 = 1$ for the estimated average link gain, $\xi^2 = 0.05$ for the link gain estimation variance, and $L = 1$ (flat fading). According to the considered channel model, and given (6), the achieved SIR can be at most $\epsilon_j < 1/\nu_j = |\overline{h}|^2/\xi^2 = 20$. To enhance the system capacity, calls that cannot be accepted in the system immediately are queued using buffers of finite lengths $B(1)$ and $B(2)$, respectively.

The optimal policy for each experiment is obtained using a linear programming (LP) optimization. For each numerical example, a randomized optimal stationary policy R^* is obtained, which is then used for simulations. The average call connection delays are obtained from simulations, averaged over 10 000 call requests. The blocking probabilities can be obtained both as a result of the LP optimization, as well as from simulations implemented using the obtained optimal policy.

In the presented experiment, we determine an optimal admission control policy such that network QoS constraints are met: blocking probabilities for class 1 and 2, respectively, $\mathbf{\Psi} = [0.2, 0.1]$, and average delays for class 1 and 2, respectively, $\mathbf{\Xi} = [2.5, 0.67]$. As in Section IV, the delay constraints are equivalently expressed as constraints on the average number of users in the queues $\overline{\mathbf{n}} = [2, 0.3]$. We run the LP optimization for different buffer configurations. As discussed in Section IV, some configurations are infeasible: $\mathbf{B} = [1, 1]$, $\mathbf{B} = [1, 2]$, $\mathbf{B} = [1, 3]$, $\mathbf{B} = [2, 3]$, and $\mathbf{B} = [3, 3]$. The network performance is summarized in Table I. The first four columns represent the blocking probabilities and average number of queued calls obtained from the LP optimization. The last four columns represent simulation results. We can see that all four buffer configurations result in admission policies for which the imposed QoS requirements are met.

An interesting observation is that, as opposed to the case in which only blocking probability constraints are considered (as in [12]), buffer dimensioning cannot be implemented independently for each class. For the case in which no delay requirements are specified, lower blocking probabilities for both classes can be obtained if the buffer lengths are increased at the expense of an increased call connection delay. This can be illustrated by a simple example (see Fig. 4) in which only the minimization of blocking probability is considered, and no delay constraints are enforced. The cost function for the LP is the sum of the blocking probabilities for both classes (equally weighted).

We can see in Table I that when delay constraints are imposed, increasing the length of the buffer for class 2 (the most delay sensitive class) lowers P_b^2 but increases P_b^1. This is a consequence of the fact that class 2 users are more delay sensi-

TABLE I
NUMERICAL RESULTS FOR THE CASE WITH DELAY AND BLOCKING PROBABILITY CONSTRAINTS

B	P_b^1	$\bar{n}1$	P_b^2	$\bar{n}2$	P_b^1 -sim.	P_b^2 -sim.	delay 1	delay 2
[2,1]	0.1865	0.6865	0.1	0.1880	0.1797	0.1040	0.8349	0.4037
[2,2]	0.2	0.8283	0.0598	0.3	0.1881	0.0585	0.9706	0.6177
[3,1]	0.1645	1.1655	0.1	0.1847	0.1656	0.1022	1.3718	0.4027
[3,2]	0.1855	1.3702	0.0533	0.3	0.1907	0.0573	1.6654	0.6383

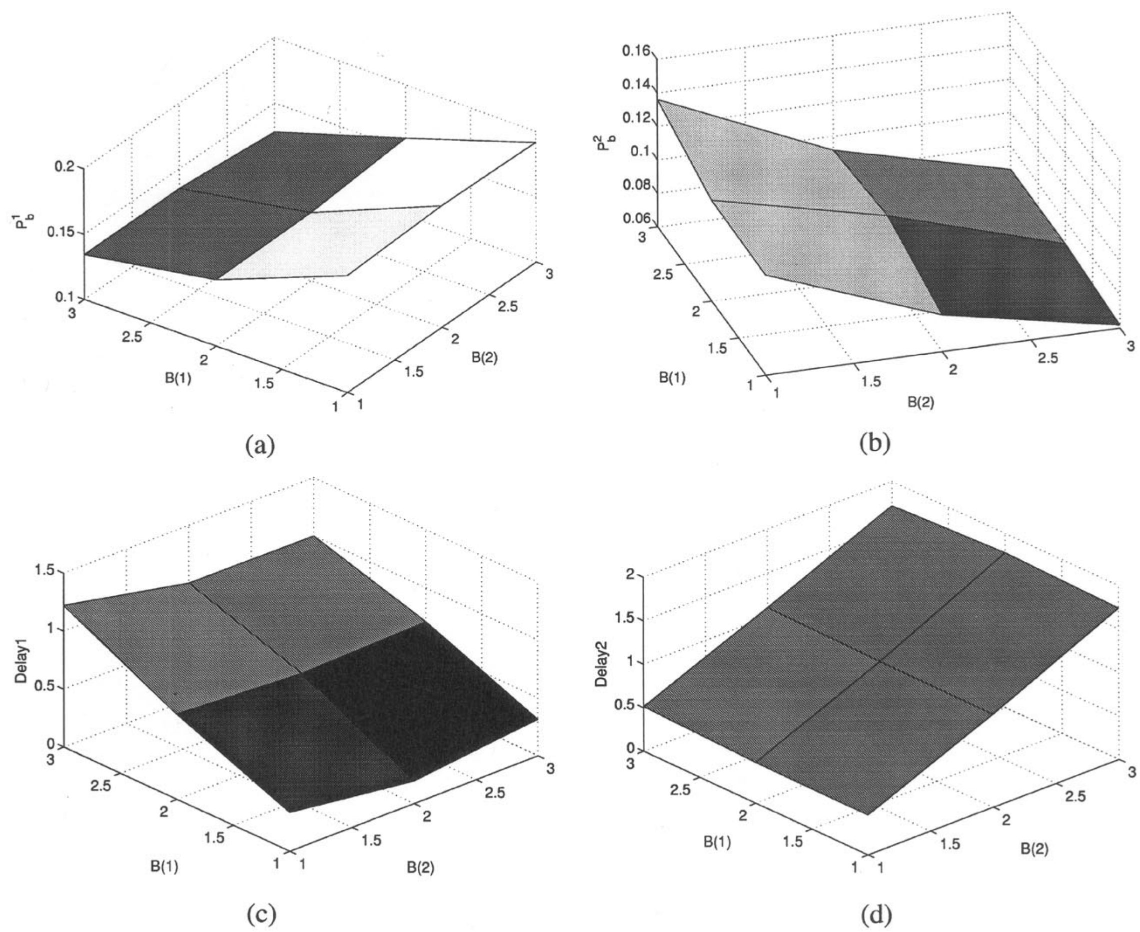


Fig. 4. System performance when no delay constraints are enforced. (a) and (b) Blocking probabilities for class 1 and 2, respectively. (c) and (d) Average connection delay for class 1 and 2, respectively.

tive, and by increasing their buffer length (and correspondingly decreasing their blocking probability), their service has to be sped up as well, so that the delay constraints can be met. The most delay-sensitive class (class 2) is an expensive class since increasing its share of capacity (lower blocking probability obtained using more buffering) affects the performance of all other classes in the system.

A. Multiuser Detection Performance Gain

We compare the capacity of a network using linear multiuser receivers with the classical scenario in which all users have conventional matched filter receivers. To derive the asymptotic capacity for the matched filter system, we rely on the SIR formula presented in [8]:

$$\text{SIR}_k = \frac{P_k \sum_{l=1}^{L} |\overline{h}_{kl}|^2 \beta}{1 + P_k \xi_k^2 \beta} = \frac{P_k |\overline{h}_k|^2 \beta}{1 + P_k \xi_k^2 \beta} \tag{41}$$

where β is the unique fixed point in $(0, \infty)$ that satisfies

$$\beta = \left[\sigma^2 + \frac{1}{N} \sum_{k=2}^{K} \left((L-1) I(\xi_k^2 P_k) + I(P_k(\xi_k^2 + |\overline{h}_k|^2))\right)\right]^{-1} \tag{42}$$

with $I(p) = p$.

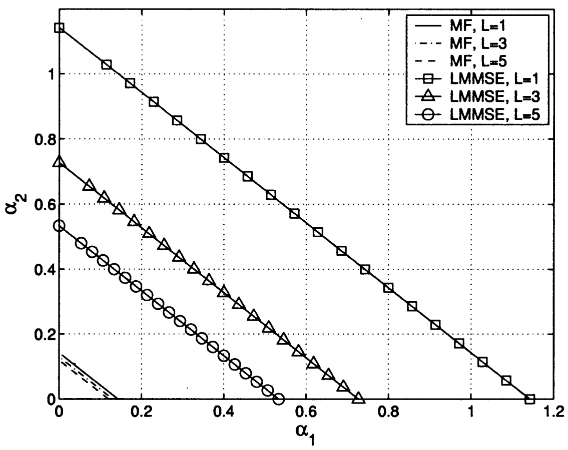


Fig. 5. Physical layer capacity comparisons: MF versus LMMSE.

Using a similar derivation as in Section III, the asymptotic system capacity for the matched filter case can be expressed as

$$\sum_{j=1}^{J} \alpha_j \epsilon_j \frac{L\nu_j + 1}{1 - \nu_j \epsilon_j} < 1. \tag{43}$$

In addition, the minimum transmit power solution for a user in class i is given by

$$P_i^t = \frac{1}{z_i |\overline{h}_i|^2} \frac{\epsilon_i \sigma^2}{(1 - \nu_i \epsilon_i)\left(1 - \sum\limits_{j=1,\, j \neq i}^{J} \alpha_j \epsilon_j \frac{L\nu_j + 1}{1 - \nu_j \epsilon_j}\right)}.$$

In Fig. 5, we present asymptotic physical layer capacity comparisons for LMMSE and matched filter systems for two classes of users when $L = \{1, 3, 5\}$. We can notice a very significant gain in capacity when LMMSE receivers are used. This has a great impact on the network capacity as well. We illustrate this by simulating the simplest example when no buffering is allowed (the delay is zero). The spreading gain is chosen to be $N_e = 32$. All other numerical values are chosen to be the same as in the previous experiments. The cost criterion is set to maximize the weighted sum of probabilities for the two classes (equal weights are considered). After optimization, $P_b^1 = P_b^2 = 0.5585$ is obtained for the matched filter case, whereas $P_b^1 = 2.9647 \times 10^{-4}$, and $P_b^2 = 2.7601 \times 10^{-4}$ is achieved for the LMMSE case. Thus, the achieved LMMSE network capacity is $\{\lambda_1, \lambda_2\} = \{0.9997, 0.4998\}$, compared with $\{\lambda_1, \lambda_2\} = \{0.4415, 0.220\,75\}$ for the matched filter case.

B. Comparisons With Classical Approaches in Admission Control

On a final note, we compare the performance of the proposed optimal admission policy with two classical approaches in call admission control: the complete sharing policy (see, for example, [6]) and the threshold policy (see, for example, [10]). The complete sharing policy accepts users in the system whenever the SIR condition can be met for all users, including the new call requesting connection. This policy cannot control the blocking probability or average connection delay. The obtained performance is simply characterized by the statistical properties of the traffic. Results for the complete sharing policy are presented in Table II.

TABLE II
NUMERICAL RESULTS FOR THE COMPLETE SHARING POLICY

B	P_b^1	P_b^2	delay 1	delay 2
[2,1]	0.1109	0.1892	0.4595	0.6017
[2,2]	0.1352	0.1118	0.5414	1.1026
[3,1]	0.0758	0.1857	0.6289	0.6144
[3,2]	0.0782	0.2019	0.6297	0.6385
[3,3]	0.1146	0.0891	0.9051	1.7118
[5,3]	0.0844	0.1059	1.5627	2.0285
[5,1]	0.0395	0.2234	0.9824	0.7031

On the other hand, a threshold policy may be designed to somewhat accommodate performance constraints for different classes of users. For this particular example with two classes of users, we partition the resources between the two classes such that blocking probability constraints can be met. To this extent, we first consider the most demanding class: class 2. In order to fairly compare the results, we impose the same blocking probability constraints as the ones considered for the optimal policy: [0.2, 0.1].

According to (14), the total number of users that can be accepted in the system for the considered numerical values is $K = K_1 + K_2 = 8$. We wish to find K_1 and K_2 such that $P_b^2 \leq 0.1$. For fixed K_2 and $B(2)$, we have an $M/M/K_2/B(2)$ queue, and the blocking probability can be computed as [3]

$$P_b^2 = p_0 \frac{K_2^{K_2} \rho_2{}^{B(2)}}{K_2!} \tag{44}$$

where $\rho_2 = \lambda_2 / K_2 \mu_2$, and p_0 is the probability of an empty queue and no one in service:

$$p_0 = \left[1 + \sum_{n=1}^{K_2 - 1} \frac{(K_2 \rho_2)^n}{n!} + \sum_{n=K_2}^{K_2 + B(2)} \frac{(K_2 \rho_2)^n}{K_2!} \frac{1}{K_2{}^{(n-m)}}\right]^{-1}. \tag{45}$$

The average connection delay experienced by calls in class 2 can be expressed as

$$W_2 = \frac{\rho_2 P_Q}{\lambda_2 (1 - \rho_2)} \tag{46}$$

where P_Q is the probability of queueing, which is defined as

$$P_Q = p_0 \frac{(K_2 \rho)^{K_2}}{K_2!} \sum_{n=K_2}^{K_2 + B(2)} \rho^{n-m}. \tag{47}$$

We note that both the blocking probability and the delay depend on K_2 and $B(2)$; therefore, they cannot be optimized independently. If we fix K_2 and $B(2)$ for a given constraint for the blocking probability, the delay is also automatically fixed to the value computed from (46). However, if delay is not of concern, a specified blocking probability can be obtained for class 2

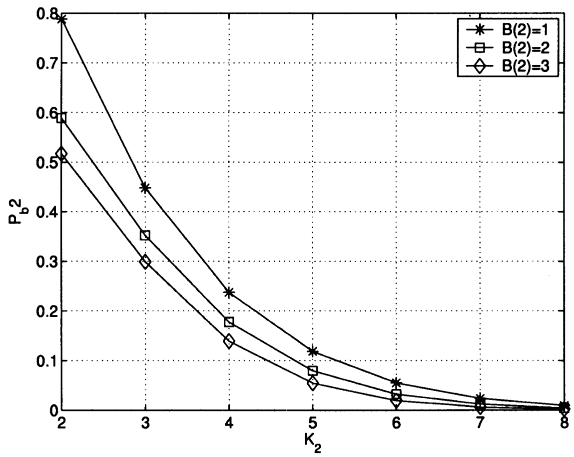


Fig. 6. Threshold policy. Blocking probability for class 2.

TABLE III
NUMERICAL RESULTS FOR THE THRESHOLD POLICY

B	P_b^1	P_b^2	delay 1	delay 2
[2,1]	0.3325	0.1132	1.3461	0.2340
[2,2]	0.3404	0.0654	1.4325	0.4983
[3,1]	0.3053	0.1154	2.2313	0.2342
[3,2]	0.3040	0.0742	2.1936	0.5359
[3,3]	0.2999	0.0505	2.1757	0.8041
[5,3]	0.2748	0.0540	4.1355	0.8135
[5,1]	0.2722	0.1046	4.0970	0.2282

as well by appropriately selecting the buffer length $B(2)$. This suggests that the threshold policy may be a better option than the SMDP approach in [12], if only blocking probability constraints are considered, due to its implementation simplicity.

We represent P_b^2 in Fig. 6 as a function of K_2 for three different values of the buffer length. It can be seen that the partition $[K_1, K_2] = [3, 5]$ gives $P_b^2 \approx 0.1$, depending on the designed buffer length. Therefore, we use this partition to obtain simulation results, which are summarized in Table III. We compare the results with the ones obtained using the optimal policy (see Table I). It can be seen that the QoS constraints $\mathbf{\Psi} = [0.2, 0.1]$ (for blocking probability) and $\mathbf{\Xi} = [2.5, 0.67]$ (for average delay) cannot be met. Lower blocking probabilities can be obtained if the buffer lengths are increased for both classes, but this comes at the expense of increased delay. We observe that the threshold policy is clearly suboptimal and lacks flexibility in guaranteeing the desired QoS. On the other hand, if the traffic statistics change frequently and must be estimated, the optimal policy based on linear programming may be too complex to implement, and suboptimal policies may be preferred. As an alternative, measurement-based admission control may also be considered in such cases (see, for example, [2] and [7]).

VI. CONCLUSIONS

In this paper, we have proposed joint optimization across the physical and the network layer to provide QoS in a multiclass CDMA network. Our proposed approach integrates admission control with power control and multiuser detection such that an optimal admission policy can be obtained, subject to physical layer QoS constraints (SIR targets) as well as network layer QoS constraints (admission delays and blocking probabilities). We have characterized the physical layer QoS by deriving the power control feasibility condition for an asymptotically large system under imperfect channel estimation for a multipath fading environment. We have shown that the maximal SIR that can be achieved for a particular class of users is limited by the estimation accuracy of their channel link gain. Optimal transmission powers for all classes of traffic have been jointly determined. Network layer constraints have been incorporated using an equivalent queueing system. The optimal admission policy has been obtained using linear programming as a solution to a semi-Markov decision process formalism. We have validated the analytical results by simulations, and the tradeoff between delay and blocking probability has been discussed for various buffer configurations. We have also emphasized the performance gains achieved using LMMSE receivers by providing comparisons with a classical scenario in which all users have conventional matched filter receivers. Further, the performance results (delays and blocking probabilities) for the proposed admission policy have been compared against two classical approaches for admission control design: the complete sharing policy and the threshold policy. It has been shown that, as expected, both the complete sharing policy and the threshold policy are suboptimal and lack flexibility in guaranteeing the desired QoS.

REFERENCES

[1] M. Andersin, Z. Rosberg, and J. Zander, "Soft and safe admission control in power-controlled mobile systems," *IEEE/ACM Trans. Networking*, vol. 5, pp. 255–265, Apr. 1997.

[2] Y. Bao and A. Sethi, "Performance-driven adaptive admission control for multimedia applications," in *Proc. IEEE Int. Conf. Commun.*, June 1999, pp. 199–203.

[3] D. Bertsekas and R. Gallager, *Data Networks*. Upper Saddle River, NJ: Prentice-Hall, 1992.

[4] D. P. Bertsekas, *Dynamic Programming and Optimal Control*. Belmont, MA: Athena Scientific, 1995.

[5] B. Chen and L. Tong, "Traffic aided multiuser detection for packet switching random access CDMA networks," in *Proc. 37th Annu. Allerton Conf. Commun., Contr., Comput.*, Monticello, IL, Sept. 1999, pp. 881–889.

[6] C. Comaniciu and N. B. Mandayam, "Integrated access control and multiuser detection for multimedia CDMA systems," in *Proc. Conf. Inform. Sci. Syst.*. Princeton, NJ, Mar. 2002.

[7] ——, "Wireless access to the World Wide Web in an integrated CDMA system," *IEEE Trans. Wireless Commun.*, to be published.

[8] J. S. Evans and D. Tse, "Large system performance of linear multiuser receivers in multipath fading channels," *IEEE Trans. Inform. Theory*, vol. 46, pp. 2059–2078, Sept. 2000.

[9] S. V. Hanly and D. N. Tse, "Power control and capacity of spread spectrum wireless networks," *Automatica*, vol. 35, no. 12, Dec. 1999.

[10] T. Liu and J. A. Silvester, "Joint admission/congestion control for wireless CDMA systems supporting integrated services," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 845–857, Aug. 1998.

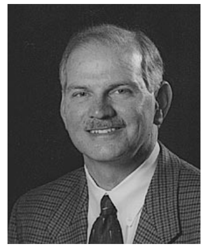
[11] U. Madhow and M. L. Honig, "MMSE interference suppression for direct-sequence spread-spectrum CDMA," *IEEE Trans. Commun.*, vol. 42, pp. 3178–3188, Dec. 1994.

[12] S. Singh, V. Krishnamurthy, and H. V. Poor, "Integrated voice/data call admission control for wireless DS-CDMA systems with fading," *IEEE Trans. Signal Processing*, vol. 50, pp. 1483–1495, June 2002.

[13] H. C. Tijms, *Stochastic Modeling and Analysis: A Computational Approach*. Chichester, U.K.: Wiley, 1986.
[14] D. Tse and O. Zeitouni, "Performance of linear multiuser receivers in random environments," *IEEE Trans. Inform. Theory*, vol. 46, pp. 171–178, Jan. 2002.
[15] S. Ulukus and R. D. Yates, "Adaptive power control and MMSE interference suppression," *ACM Wireless Networks, Special Issue on Multiuser Detection in Wireless Communications*, vol. 4, pp. 489–496, Nov. 1998.
[16] X. Wang and H. V. Poor, "Blind multiuser detection: A subspace approach," *IEEE Trans. Inform. Theory*, vol. 44, pp. 677–690, Mar. 1998.
[17] Y. Yao, H. V. Poor, and F. Sun, "User capacity for synchronous multirate CDMA systems with linear MMSE receivers," *IEEE Trans. Inform. Theory*, to be published.
[18] J. Zander, "Transmitter power control for co-channel interference management in cellular radio systems," in *Fourth WINLAB Workshop on Third Generation Wireless Information Networks*. Piscataway, NJ: Rutgers Univ. Press, Oct. 1993, pp. 241–248.

Cristina Comaniciu received the M.S. degree in electronics from the Polytechnic University of Bucharest, Bucharest, Romania, in 1993 and the Ph.D. degree in electrical and computer engineering from Rutgers University, New Brunswick, NJ, in December 2001.

From 1998 to 2001, she was with the Wireless Information Network Laboratory (WINLAB) at Rutgers University, working on integrated access control and detection algorithms for multimedia CDMA systems. She is currently a postdoctoral fellow in the Electrical Engineering Department, Princeton University, Pricneton, NJ. Her research interests include QoS for multimedia wireless networks, multiaccess protocols, radio resource management multiuser detection, and cross-layer design for wireless networks.



H. Vincent Poor (S'72–M'77–SM'82–F'87) received the Ph.D. degree in electrical engineering and computer science in 1977 from Princeton University, Princeton, NJ.

He is currently a Professor of electrical engineering at Princeton. He is also affiliated with Princeton's Department of Operations Research and Financial Engineering and with its Program in Applied and Computational Mathematics. From 1977 until he joined the Princeton faculty in 1990, he was a faculty member at the University of Illinois at Urbana-Champaign. He has also held visiting and summer appointments at several universities and research organizations in the United States, Britain, and Australia. His research interests are primarily in the area of statistical signal processing, with applications in wireless communications and related areas. Among his publications in this area is the forthcoming book *Wireless Communication Systems: Advanced Techniques for Signal Reception*.

Dr. Poor is a member of the National Academy of Engineering and is a Fellow of the Acoustical Society of America, the American Association for the Advancement of Science, the Institute of Mathematical Statistics, and the Optical Society of America. His IEEE activities include serving as the President of the IEEE Information Theory Society in 1990 and as a member of the IEEE Board of Directors from 1991 to 1992. Among his recent honors are an IEEE Third Millennium Medal in 2000, the IEEE Graduate Teaching Award in 2001, the Joint Paper Award of the IEEE Communications and Information Theory Societies in 2001, the NSF Director's Award for Distinguished Teaching Scholars in 2002, and a Guggenheim Fellowship for the period 2002 to 2003.