An Approach to Connection Admission Control in Single-Hop Multiservice Wireless Networks With QoS Requirements

Tara Javidi, *Member, IEEE,* and Demosthenis Teneketzis, *Fellow, IEEE*

***Abstract*—We present an approach to connection admission control in single-hop multiservice wireless networks with quality of service (QoS) requirements. The approach consists of two steps: 1) the specification of an admission region that captures the QoS requirements and 2) the formulation of a generalized knapsack problem that captures the connection admission control. To illustrate approach 1), we determine an outage-based admission region; to illustrate approach 2), we investigate the performance of the greedy admission policy in a generalized knapsack problem.**

***Index Terms*—Connection admission, knapsack scheduling, outage, quality of service (QoS), wireless communication.**

I. Introduction

THE mobile wireless environment provides serious challenges such as limited bandwidth, low-capacity channels, and interference among users. As a result, an important network layer problem in the design of wireless systems is how to allocate the limited resources efficiently while providing quality of service (QoS) guarantees to the applications in terms of bit rate and loss. This problem becomes more acute for next-generation integrated-services networks, which aim to support heterogenous traffic. Next-generation networks will provide various services such as data, voice, video, etc., each with its own QoS requirements [which are expressed in terms of signal-to-interference ratio (SIR), outage probability, latency, loss rate, etc.], statistical description, activity factor, and generated revenues per unit of service. In such networks, the desirable resource allocation is achieved by the base station through connection admission decisions in the presence of new connection requests. A connection admission decision consists of granting or rejecting connection to a specific service. Granting an admission is equivalent to a contract where the newly connected service is guaranteed a set of desirable QoS measures for the full length of connection while generating revenue at a prespecified rate. As a result, an efficient allocation of resources is achieved by constructing a connection admission strategy that maximizes the average expected revenue and satisfies the QoS requirements of each connected service. Such a construction is the solution to a constrained stochastic dynamic optimization problem. Notice that quality of service for a connection is a dynamic variable whose statistics depend on the chosen admission strategy. Hence, a key feature of this optimization problem is that there is a two-way coupling between the constraints resulting from the QoS requirements and the admission policy. Such a two-way coupling results in a computationally challenging and analytically intractable optimization problem.

In this paper, we present an alternative approach to the above resource allocation problem. We propose a decomposition of this problem into two subproblems: 1) the specification of an admission region $\mathcal{A}$, which guarantees the QoS requirements for each connected user, independently of the admission policy; and 2) the determination of a connection admission policy that is optimal within the class $\Pi_{\mathcal{A}}$ of policies restricted to the admission region $\mathcal{A}$. This reduces the complexity of the problem to a great extent and allows us to understand the interaction among different layers of wireless communication systems.

The remainder of this paper consists of four parts. In Section II, we 1) formulate the CAC in single-hop multiservice wireless networks with QoS requirements; 2) discuss the nature of this problem and the need for alternative tractable methods to solve it; and 3) propose the aforementioned decomposition. In Section III, we present an approach to defining outage probability as a system-wide QoS measure for cellular systems. We describe how to use this approach to define an admission region where, independently of admission strategies, the requirements on probability of outage are satisfied. We illustrate the approach by constructing the aforementioned admission region for a few examples. In Section IV, we address the CAC problem in the presence of a predefined admission region $\mathcal{A}$. We show that this problem is equivalent to the scheduling of a stochastic "generalized knapsack." We discuss the nature of the optimal policy under different scenarios. Because of the difficulty of the implementation of such policies, we focus on the greedy policy that is easy to implement. We determine conditions on rates of revenue associated with each class of users—with a known average call-time and arrival rate—sufficient to guarantee the optimality of the greedy policy. In Section V, we present conclusions and reflections.

The contribution of this paper is threefold.

1) The decomposition of the problem of resource allocation in a wireless network with QoS requirements into two subproblems:
   a) the construction of an admission region;
   b) the generalized knapsack scheduling.

Manuscript received February 20, 2002; revised August 23, 2002 and November 7, 2002. This research was supported in part by ARO under Grant DAAH04-96-1-0377 and the National Science Foundation under Grant ECS-9979347.

T. Javidi is with the Electrical Engineering Department, University of Washington, Seattle, WA 98102 USA (e-mail: tara@ee.washington.edu).

D. Teneketzis is with the Electrical Engineering and Computer Science Department, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: teneket@eecs.umich.edu).

Digital Object Identifier 10.1109/TVT.2003.808802

Such a decomposition results in a tractable resource-allocation problem and addresses important issues regarding the interaction among various layers of the network.

2) The development of a statistical model for outage and the analytical construction of an outage-based admission region.
3) The determination of conditions on rates of revenue associated with each class of users with a known average call time and arrival rate, sufficient to guarantee the optimality of the greedy admission policy.

II. Decomposition of Connection Admission Control Problems in Cellular Systems With QoS Requirement

In a wireless system, the desirable resource allocation is achieved through two separate mechanisms of power-rate assignment (PRA) and connection admission control (CAC). In other words, the resource-allocation mechanisms should be designed such that the total generated revenue is maximized while guaranteeing an acceptable quality of service to the admitted connections. Let π and g be the CAC and PRA strategies. Let $q_k^i(\pi, g)$ be the vector of quality of service for user k of type i under CAC and PRA strategies π and g; each component of the vector determines the system performance in terms of one of the QoS requirements as observed by user k of type i. In general

$$q_k^i(\pi, g) := \theta_i(\{\mathbf{x}^\pi(t)\}_{t=1}^T, \{\mathbf{P}^g(t)\}_{t=1}^T, \{\mathbf{R}^g(t)\}_{t=1}^T \qquad (1)$$

where $\mathbf{x}^\pi(t) = (x_1^\pi(t), x_2^\pi(t), \ldots, x_L^\pi(t))$, $x_i^\pi(t)$ is the number of connections of type i present at the system at time t, $\mathbf{P}^g(t)$ represents the power assignment vector at time t, and $\mathbf{R}_k^g(t)$ is the rate assigned to user k at time t. Note that the superscript $\pi(g)$ in vector(s) $\mathbf{x}^\pi(t)$ ($\mathbf{P}^g(t)$ and $\mathbf{R}_k^g(t)$) indicates that the number of admitted connections (vector of assigned powers and rates to admitted connections) depends on policy $\pi(g)$. In addition, the form of θ_i depends on the nature of the desirable QoS criteria, the physical layer, and the power assignment rule. Total expected reward generated over a finite horizon T is as follows:

$$V(\pi, g) := E\left\{\sum_{t=0}^{T}\sum_{i=1}^{L}\sum_{k=1}^{x_i^\pi(t)} c_i(R_k^g(t))\right\} \qquad (2)$$

where $c_i(R_k^g(t))$ is the rate of revenue generated by user k of type i when it is assigned the rate $R_k^g(t)$. Here we assume that there exists a bisection between the assigned rate and the rate of information transmission.

In systems where the maximum rate of transmission of any user and the power-control mechanism is fixed and known, CAC is the only mechanism to guarantee a certain level of service while maximizing total revenue over horizon T. Mathematically, this problem can be formulated as

$$\max_\pi E\left\{\sum_{t=0}^{T}\sum_{i=1}^{L} c_i x_i^\pi(t)\right\} \qquad (3)$$

subject to

$$q_k^i(\pi) = \theta_i(\mathbf{x}^\pi(1), \ldots, \mathbf{x}^\pi(T)) \in Q_i \qquad (4)$$

where c_i is the rate of revenue generated by a user of type i, $q_k^i(\pi)$ is the same as above, and Q_i is the set of acceptable QoS for a connection of type i. In this paper, we focus our attention on the above CAC problem.

Notice that quality of service for a connection is a dynamic variable whose statistics depend on the chosen admission strategy. Hence, a key feature of this optimization problem is that there is a two-way coupling between the constraints resulting from the QoS requirements and the admission policy. Such a two-way coupling results in a computationally challenging and analytically intractable optimization problem.

In this paper, we propose the two-step decomposition described in the Introduction. Such a decomposition results in a one-way coupling between the constraints present in the resource-allocation problem and the determination of an optimal allocation policy. In other words, we propose the following problems.

P1) Find the largest coordinate convex set $\mathcal{A}$ such that for all $\pi \in \Pi_\mathcal{A}$

$$q_k^i(\pi) \in Q_i \qquad \forall k, \forall i. \qquad (5)$$

P2) Given the admission region $\mathcal{A}$ [constructed in P1)], solve the optimization problem

$$\max_{\pi\in\Pi_\mathcal{A}} E\left\{\sum_{t=0}^{T}\sum_{i=1}^{L} c_i x_i^\pi(t)\right\}. \qquad (6)$$

Though, in general, our approach results in a suboptimal solution for the original problem, it reduces the complexity of the problem. Furthermore, it creates a conceptual framework for understanding the interaction among different layers of wireless communication systems, such as physical layer concerns, QoS requirements, and network layer resource allocation. In other words, the admission region $\mathcal{A} = \{\mathbf{x} : q_k^i(\pi) \in Q_i \text{ for } \forall\pi \in \Pi_\mathcal{A}\}$ conceptualizes the physical channel and QoS requirements; and the constrained optimization problem is reduced to an optimization over the set of admissible policies $\Pi_\mathcal{A}$, which is regulated at the network layer.

Note that the coordinate convexity of set $\mathcal{A}$ is a natural requirements for any desirable admission region. Mathematically, set $B \subset \mathbb{R}^L$ is coordinate convex if and only if for any $\mathbf{x} \in B$ and any $l \in \{1, 2, \ldots, L\}$, $\mathbf{x} - \mathbf{e}_l \in B$, where $\mathbf{e}_l$ is the unity vector along dimension l. Coordinate convexity of set $\mathcal{A}$ captures the unacceptability of forced departures in an admission region.

In the wireless scenarios, the QoS for each class of users depends on the number and type of other users present in the system, and, in general, performance decreases as the number of users increases. As a result, QoS guarantees can create notions of capacity or regions of admission. In other words, in order to provide QoS guarantees, the possible combinations of the number of each type of users present in the system must be limited to a set called the admission region. Such an admission region summarizes, for the purpose of resource allocation and scheduling at the network layer, the physical layer characteristics of the network, the characteristics of the potential users, the

QoS requirements of each user, and the effects of interference caused by different applications (see, for example, [1]–[6]).

After specifying the admission region, our goal is to design a CAC strategy for the system. This leads to the following problem: if CAC decisions are restricted to the admission region $\mathcal{A}$ (so as to satisfy the QoS requirements), and if each class of users generates different rates of revenue, the objective is to determine CAC strategies that maximize a long-run average revenue. In general, the determination of optimal CAC strategies depends on the statistical characteristics of arrivals and service times, and the rate of revenue associated with each class of users.

In this paper, we illustrate the above-described approach as follows: we first construct an admission region based on QoS requirements expressed by the probability of outage. Afterwards, we determine an optimal admission strategy for a constrained resource-allocation problem where the constraint is described by a general admission region $\mathcal{A}$. Such a general admission region can be thought of as the intersection of an outage-based admission region (described in this paper) with those based on other reasonable QoS measures such as bit error rate, latency, and throughput (as proposed in [1]–[6]).

III. Outage-Based Admission Region

Outage probability is an important performance measure in cellular networks. In a cellular scenario, low signal-to-(noise plus interference) ratios (SNIR) can increase bit error rate (BER), but more importantly, if this ratio remains low for a long enough duration, it can cause an outage in an ongoing service (due to loss of synchronization, etc). This will result in disconnection of an admitted call. In most common scenarios, this is considered a more severe form of low performance than blocking (which occurs when a new call is denied admission to the cell, hence the network). As a result, our goal is to find a way to quantify and measure the occurrence of this undesirable event and its effect on system performance. An appropriate measure of outage is considered a main performance measure for traditional cellular networks. We believe that outage probability together with average bit error rate and throughput can form a reasonable set of performance criteria or QoS requirements for certain types of traffic [e.g., voice and data streams in pre-third-generation code-division multiple-access (CDMA) systems]. As the first step in our approach to the connection admission control in a multiservice wireless system with outage-based QoS requirements, the main goal of this section is to summarize the outage-based QoS into an admission region. In other words, we provide a procedure to determine the admission region of a cellular system where the probability of outage for each user cannot exceed a prespecified threshold (depending on the class of the user).

We describe an outage by two parameters: 1) the SNIR threshold γ and 2) a minimum duration τ. An outage occurs when the SNIR remains below the threshold γ for a period longer than or equal to τ. In most of the currently available literature (e.g., see [7] and [1]), an outage is assumed to occur when the SNIR falls below a threshold γ^*. We believe that this is not sufficient to capture the essence of an outage, since it ignores statistical correlation or burstiness in the incoming traffic stream. It is intuitively expected that traffic streams with a high level of burstiness are more likely to cause an outage than nonbursty or independent identically distributed streams with the same level of instantaneous instantaneous interference. Similarly, the memory present in fading channels directly affects how long the impairment will last, hence it affects the occurrence of an outage. In other words, the drop in SNIR below γ does not result in an outage instantaneously; an outage results when the SNIR is low for an extended period of time, i.e., a time period that exceeds a minimum duration τ. With this definition, the occurrence of outage events strictly depends on the second-order statistics of the interference and/or fading. A characterization of outage in terms of both the threshold γ and the time duration τ has appeared only in [8] and [9]. One key feature of [8] and [9] is that the effect of other users on the outage probability is not taken into account. That is, the effect of the (random) number of active users and the statistical variation of their channels on the outage probability is ignored. Attention in both [8] and [9] is restricted to one user and the effect of its physical channel on the outage probability. In general, the performance of a wireless system critically depends on two factors: 1) the condition of the physical channel and 2) the interference created by other users. The approach to outage that we propose captures both of the aforementioned factors.

In this paper, we propose two measures to quantify outage: 1) probability of outage and 2) frequency of outage. We will examine both measures as functions of the number and characteristics of users in the system, and will critique the merit of each measure. Since one of the components of our proposed approach to the connection admission control problem is to construct an admission region, where QoS requirements are satisfied, we believe that the outage probability is the appropriate measure for quantifying outage. Indeed, we show that by incorporating the effect of multiple-access interference into our approach, we are able to relate the probability of outage to the number and type of users present in the system and, therefore, to determine an admission region associated with the maximum acceptable outage probability for each type of user.

The salient features of our approach are the following.

1) We model the statistical variation of the physical channel by a Markov chain (as in [9]).
2) We consider several types of users in terms of their statistical activity, and QoS requirements.
3) We fix the total number of users admitted by the system, and we assume that the status of each user switches between “active” and “inactive” according to a Markov rule [independent of 1)]. The status of a particular user is not necessarily independent of that of another user.

As a result of the aforementioned features, we can construct a model that allows us to define, for any multiple access scheme, the SNIR ratio and, hence, determine for any parameters γ and τ the outage probability as a function of the fixed number of users present in the system. This in turn allows us to analytically determine the capacity of the system (described in terms of an admission region) associated with maximum acceptable probability of outage. Therefore, we achieve two main goals in this section:

1) the development of an approximate statistical model for outage and calculation of the appropriate measures: frequency and probability of outage;
2) the analytic determination of an admission region based the on desired performance of the system with regard to outage probability.

This section is organized as follows. In Section III-A, we construct a stochastic model and analytically study and calculate the probability of outage. In Section III-B, we present examples illustrating the modeling and results in Section II-A.

A. Outage-Based Admission Region for Multiuser Systems With Markov Channels

1) Philosophy of Our Approach: We address the issue of outage within the context of QoS requirements. A user in the system encounters an outage event when its received SNIR at the base station falls below a threshold for an extended period of time. Hence, an outage is experienced by each user individually. Therefore, the key conceptual issue is how to analytically describe an outage event as a system-wide QoS criterion. We address this issue by introducing a fictitious observer/user and by defining an outage incurring during this user's service time. To guarantee that the outage-based QoS requirements are satisfied for every type of user that may be admitted by the system, we proceed as follows. We consider a separate fictitious observer/user for each type of traffic. Such a user is always active and is identical to the actual users of the same type in terms of the statistics of the physical channel, SNIR threshold, and minimum outage duration. Each fictitious observer/user does not create any interference in the system, and hence has no effect on the performance of the system. The probability or frequency of outage for such a user is a conservative bound on the outage probability of each user of the same type. The system-wide QoS requirement in terms of probability or frequency of outage is met if and only if the probability or the frequency of outage for each of the aforementioned fictitious users is below a prespecified value (that depends on the type of user), which reflects the QoS requirement.

In this section, we construct the outage-based admission region following the above philosophy. In Section III-A2, we formulate the outage problem associated with a fictitious observer/user u_0 whose statistical variation of its channel, the SNIR threshold γ, and the minimum outage duration τ are given. We fix the number of admitted users and model the effect of channel variations and interference as a super Markov chain (SMC). Then, we identify the states of the SMC where the combination of channel variation and interference causes an SNIR below γ. In Section III-A3, we use the formulation of Section III-A2 to calculate the probability and frequency of outage associated with u_0 (when the number of admitted users is fixed). In Section III-A4, we present examples illustrating our results and critique the aforementioned formulation criteria. In Section III-A5, we construct an admission region where the system-wide QoS in terms of outage probability is guaranteed.

2) Outage Formulation for a Given Observer/User in the Presence of a Fixed Number of Users: In this section, we fix the number admitted users and then develop an approach to defining and computing the probability and the frequency of an outage for a fictitious observer/user u_0, whose channel statistics, SNIR threshold γ, and minimum outage duration τ are given.

In a wireless setting, the received SNIR of an observer/user u_0 depends on two decoupled factors:

1) the effect of physical channel in the absence of other users; this captures events like additive noise, fading, and/or shadowing (in the presence or absence of power-control mechanisms);
2) the effect of the presence, power, and channel statistics of the other active users admitted in the system.

Therefore, to determine the probability of outage, we need 1) to model the channel degradation, 2) to model the interference of other admitted users, and 3) to construct a "super Markov chain" combining 1) and 2) in order to describe the received SNIR of u_0.

It is very common to model the effect of the channel on SNIR in the absence of other users as a Markov chain. The validity of such model has been extensively studied and confirmed in the literature (see [10]). The most commonly used example of this kind is the Gilbert channel. In general, such an MC is defined by its state-space $\mathcal{H} = \{h_1, h_2, \ldots, h_I\}$, and its transition matrix

$$\mathbf{A} = [a_{kl}] = [\text{Prob}\{X(t+1) = h_l | X(t) = h_k\}]. \quad (7)$$

Note that in the case of an ideal power-control mechanism, the state-space $\mathcal{H}$ is reduced to a singleton $\{h\}$; hence $\mathbf{A} = 1$. In the case of power control with quantized error $\pm\delta$, we have $\mathcal{H} = \{h - \delta, h, h + \delta\}$. We assume that channel states of individual users are mutually independent.

To model the interference of other admitted users, we assume that there are L types of users in terms of QoS requirements, transmission power, and the activity factor [11], and there are $(M_1, M_2, \ldots, M_L)$ users admitted to the system (not including u_0). At any time slot, each admitted user can be active ("on") or inactive ("off"). Since only active users interfere with the received signal of u_0, we need to find an appropriate model to describe the evolution of the users' "on" periods. In this paper, we assume that active and inactive periods for a user of type l evolve according to a b_l-order Markov chain. Consequently, we can model the activity of user i of type l by a Markov chain of size $B_l = 2^{b_l}$; we denote the states of this Markov chain by an integer $n_l^i \in \{1, 2, \ldots, B_l\}$. In general, we assume that the activity of all users can be correlated.

Based on the above, we can express the state of users (in terms of being active or inactive) by the following random array:

$$(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_L) = ((n_1^1, \ldots, n_1^{M_1}), \ldots, (n_L^1, \ldots, n_L^{M_L})). \quad (8)$$

By construction, this array evolves according to a known Markov rule. Let $\mathbf{T}$ be the transition matrix for this Markov chain, i.e.,

$$\mathbf{T}_{ij} = \text{Prob}\{(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_L) \mid (\boldsymbol{m}_1, \ldots, \boldsymbol{m}_L)\}. \quad (9)$$

Note that T is a square matrix of dimension $\prod_{l=1}^{L} B_l^{M_l}$.

To describe the received SNIR of u_0, we construct an SMC, which represents the variation of the physical channel for u_0, the channel variation of other users, and the state of the admitted users. The states of this SMC are arrays of type

$$\mathbf{s} := (h^{u_0}, (h^{1,1}, \ldots, h^{1,M_1}), \ldots, (\boldsymbol{n}_1, \ldots, \boldsymbol{n}_L))$$

where $h^{u_0} \in \mathcal{H}_{u_0}$ is the state of the channel between user u_0 and the base station; $h^{l,k} \in \mathcal{H}_l$ for $l = 1, 2, \ldots, L$, and $k = 1, 2, \ldots, M_l$ is the state of the physical channel (in the absence of other users) between user k of type l and the base station; and $\boldsymbol{n}_l$, $l = 1, 2, \ldots, L$, is a vector of length M_l defined in (8). The state-space of this SMC is $\mathcal{S} = \mathcal{H}_{u_0} \times \prod_{l=1}^{L} {\mathcal{H}_l}^{M_l} \times \prod_{l=1}^{L} \{1, 2, \ldots, B_l\}^{M_l}$. Since, by assumption, the state $h^{i,l}$ of the physical channel for a user is independent of the number of the other users and their channel state, the transition probability for this SMC can be easily obtained, using (9). It is

$$\mathbf{P} = \mathbf{T} \otimes \overbrace{\mathbf{A}_L \otimes \cdots \otimes \mathbf{A}_L}^{M_L} \otimes \cdots \otimes \overbrace{\mathbf{A}_1 \otimes \cdots \otimes \mathbf{A}_1}^{M_1} \otimes \mathbf{A}_0 \quad (10)$$

where $\mathbf{A}_0$ is the transition matrix of the Markov physical channel between observer/user u_0 and the base station; $\mathbf{A}_l$ for $l = 1, 2, \ldots, L$ is the transition matrix of the Markov physical channel between a user of type l and the base station; and $A \otimes B$ denotes the Kronecker product of the matrices A and B.

To define an outage event mathematically, we must specify the received SNIR of observer/user u_0 at each state $\mathbf{s} \in \mathcal{S}$. This SNIR is a function $f_{u_0} : \mathcal{S} \longmapsto \mathbb{R}_+$. The exact form of $f_{u_0}(\cdot)$ depends on the dynamics of multiple-access interference, and possibly the power-control mechanism. For instance, for a CDMA system where users of the same class have a common transmitted power and there is no power control, the form of function f_{u_0} is

$$f_{u_0}(\mathbf{s}) = \frac{h^{u_0} P_0 G_0}{\eta + \sum_{l=1}^{L} P_l \sum_{k=1}^{M_l} \psi_{l,k} h^{l,k}} \quad (11)$$

where η is the noise power (that includes the expected total interference from the adjacent cells), G_0 is the spreading gain for user u_0, $h^{l,k}$ is the channel gain between the kth user of type l and the base station, $\psi_{l,k} \in \{0, 1\}$ is an indicator function that is equal to one when the kth type l user is active, and P_l is the common transmitted power for all type-l users. This form can extend to CDMA systems with power control where each class of users has a common targeted power and where $h^{l,k}$ represents the error of the power-control mechanism.

After specifying the SNIR of user u_0 at each state, we define the sets of "bad states" $\mathcal{B}$ and "good states" $\mathcal{G}$ as

$$\mathcal{B} := \{\mathbf{s} \in \mathcal{S} | f_{u_0}(\mathbf{s}) < \gamma\} \quad (12)$$

$$\mathcal{G} := \mathcal{S} - \mathcal{B}. \quad (13)$$

Based on the above classification of states, we can now formally define the following.

Definition 1: An outage is an event where the state of the SMC enters $\mathcal{B}$ and stays in $\mathcal{B}$ for at least τ units of time.

Definition 2: The probability of an outage is defined as the probability that a randomly selected time slot belongs to an outage event; it is denoted by P_{outage}.

Definition 3: The frequency of outage is defined as the frequency of entering outage events; it is denoted by f_{outage}.

Definition 4: The average outage duration is defined as the expected length of an outage event; it is denoted by $E[T_{\text{outage}}]$.

a) Worst case scenario: In general, the dimension of the matrix $\mathbf{P}$ can be very large. This creates a practical difficulty in the calculation of frequency and probability of outage. To deal with this difficulty, we can analyze the worst case condition for user/observer u_0, where all the other users are in their best physical channel realization. In other words, we replace $\mathcal{H}_l$ with a singleton $\{h_{\text{best}}^l\}$, where $h_{\text{best}}^l = \max \mathcal{H}_l$; hence $\mathbf{A}_l = 1$. In this situation, the SNIR can be expressed as

$$\begin{aligned} f_{u_0}^w(\mathbf{s}) &= \frac{h^{u_0} P_0 G_0}{\eta + \sum_{l=1}^{L} h_{\text{best}}^l P_l \sum_{k=1}^{M_l} \psi_{l,k}} \\ &= \frac{h_0^u P_0 G_0}{\eta + \sum_{l=1}^{L} h_{\text{best}}^l P_l r_l} \end{aligned} \quad (14)$$

where r_l is the the total number of active users of type l and, as before, η is the noise power, G_0 is the spreading gain for user u_0, h^{u_0} is the channel gain between u_0 and the base station, and P_l is the common transmitted power for all type-l users.

Equation (14) implies that, in this case, the effect of other users is like a noise term proportional to the number of active users. A state for the worst case scenario is a vector of the form $(h, r_1, r_2, \ldots, r_L)$. If we denote by I_l (I_{u_0}) the size of set $\mathcal{H}_l$ $(\mathcal{H}_{u_0})$, then the dimension of the state space in the worst case scenario is $I_{u_0} \times \prod_{l=1}^{L} (M_l + 1)$, whereas the state space in the original formulation is of dimension $I_{u_0} \times \prod_{l=1}^{L} {I_l}^{M_l} \times \prod_{l=1}^{L} B_l^{M_l} = I_{u_0} \times \prod_{l=1}^{L} (I_l \times B_l)^{M_l}$. In the worst case problem, the transition matrix of the SMC is

$$\mathbf{P} = \mathbf{T}' \otimes \mathbf{A}_0 \quad (15)$$

where $\mathbf{T}'$ is defined as

$$\mathbf{T}'_{ij} = \text{Prob}\{(r_1, r_2, \ldots, r_L) = j \mid (r_1, r_2, \ldots, r_L) = i\}. \quad (16)$$

$\mathbf{T}'$ can always be determined from $\mathbf{T}$, defined by (9), but in most cases $\mathbf{T}'$ can also be constructed using the traffic model of each user type directly.

As before, the sets of "bad" and "good" states are

$$\mathcal{B} := \{\mathbf{s} \in \mathcal{S} | f_{u_0}^w(\mathbf{s}) < \gamma\} \quad (17)$$

$$\mathcal{G} := \mathcal{S} - \mathcal{B}. \quad (18)$$

An outage occurs when the state of the Markov chain remains in set $\mathcal{B}$ for at least τ units of time.

Remark: In the worst case analysis, it is possible to upper bound SNIR by function $f_{u_0}^w$, which only depends on the number of active users, rather than each user's channel state. This property allows for a reduction in the size of the state space, and hence reduces the computational complexity in determining the admission region. We note that any approximation of the SNIR function that reduces the dependency on individual channel states results in a reduction in computational complexity of the problem. On the other hand, the cost of such reduction of complexity is the introduction of error in the calculation of frequency and probability of outage. In other words, there exists a tradeoff between the computational complexity of the method and the tightness and/or the precision of the constructed admission region.

3) Outage Analysis for a Given Observer/User in the Presence of a Fixed Number of Users: The probability and frequency of an outage event in the constructed SMC can be studied in the framework of [9].

Consider the constructed SMC and the associated transition matrix $\mathbf{P}$ with it. We follow [9] to establish the necessary equations and relations that describe the probability of outage. Note that the SMC is mathematically equivalent to the physical Markov Channel studied in [9], even though the SMC, in general, has a much larger state space, and it has a very specific structure due to its construction. Hence, after introducing the appropriate notation and definitions, we can use results provided by [9] for the analysis of the probability of outage.

a) Definitions: Let the row-vector $\boldsymbol{\pi}$ denote the stationary distribution of the SMC.

Define $\boldsymbol{\pi_G}$ and $\boldsymbol{\pi_B}$ as

$$\boldsymbol{\pi_G}(\mathbf{s}) = \begin{cases} \boldsymbol{\pi}(\mathbf{s}), & \text{if } \mathbf{s} \in \mathcal{G} \\ 0, & \text{otherwise} \end{cases} \tag{19}$$

$$\boldsymbol{\pi_B}(\mathbf{s}) = \begin{cases} \boldsymbol{\pi}(\mathbf{s}), & \text{if } \mathbf{s} \in \mathcal{B} \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

Define $\boldsymbol{P_B}$ as the matrix with entries

$$\boldsymbol{P_B}(i,j) = \begin{cases} \mathbf{P}(i,j), & \text{if } j \in \mathcal{B} \\ 0, & \text{if } j \in \mathcal{G}. \end{cases} \tag{21}$$

For any integer k, define $\sigma_{\mathcal{GB}}(k)$ to be the probability that the channel state is in $\mathcal{G}$ at time t and in $\mathcal{B}$ at times $t+1, t+2, \ldots, t+k$.

b) Results: We establish analytical expressions for probability and frequency of outage under the assumption that the system is operating in steady-state. For that matter, we need some results from [9], which we state in the form of facts.

Fact 1:

$$\sigma_{\mathcal{GB}}(k) = \boldsymbol{\pi_G} \boldsymbol{P_B}^k \mathbf{1} \tag{22}$$

where $\mathbf{1}$ is a column vector whose elements are all one.

Lemma 1:

$$P_{\text{outage}} = f_{\text{outage}} \times E[T_{\text{outage}}]. \tag{23}$$

Proof: We can write

$$P_{\text{outage}} = \lim_{N\to\infty} \frac{\sum_{j=1}^{N} {T_{\text{outage}}}^j}{L(N)} \tag{24}$$

where ${T_{\text{outage}}}^j$ is the time duration of the jth outage period and

$$L(N) = \inf\{t : \text{there are at least } N \text{ outage periods in the time interval } [0,t]\}.$$

On the other hand, the random sequence $\{{T_{\text{outage}}}^j\}$ is stationary. Hence

$$\begin{aligned} P_{\text{outage}} &= \lim_{N\to\infty} \frac{N E[T_{\text{outage}}]}{L(N)} \\ &= \lim_{N\to\infty} E[T_{\text{outage}}] \frac{N}{L(N)} \\ &= E[T_{\text{outage}}] \lim_{N\to\infty} \frac{N}{L(N)} \\ &= E[T_{\text{outage}}] \times f_{\text{outage}}. \end{aligned}$$

■

Lemma 1 can be used to establish the following.

Proposition 1:

$$f_{\text{outage}} = \sigma_{\mathcal{GB}}(\tau). \tag{25}$$

Proof: We first prove that

$$P_{\text{outage}} = E[T_{\text{outage}}] \times \sigma_{\mathcal{GB}}(\tau). \tag{26}$$

Then from (23) and (26), we establish (25). To prove (26), we consider the definition of P_{outage}. By the definition, P_{outage} can be calculated as the sum of the probability of a time slot belonging to a sequence of bad states weighted by the length of the period. In other words

$$\begin{aligned} &P_{\text{outage}} \\ &\quad= \sum_{t=\tau}^{\infty} t \, \text{Prob}\{\text{state in } \mathcal{B} \text{ for } t \text{ units of time}\} \\ &\quad= \sigma_{\mathcal{GB}}(\tau) \sum_{t=\tau}^{\infty} t \, \frac{\text{Prob}\{\text{state in } \mathcal{B} \text{ for } t \text{ unit of time}\}}{\sigma_{\mathcal{GB}}(\tau)} \\ &\quad= \sigma_{\mathcal{GB}}(\tau) \sum_{t=\tau}^{\infty} t \, \text{Prob}\{\text{an outage period of length } t\} \\ &\quad= \sigma_{\mathcal{GB}}(\tau) \times E[T_{\text{outage}}]. \end{aligned}$$

■

Proposition 1 and Fact 1 establish the following analytic expression for the frequency of outage.

Corollary 1:

$$f_{\text{outage}} = \boldsymbol{\pi_G} \boldsymbol{P_B}^{\tau} \mathbf{1}. \tag{27}$$

To obtain an analytic expression for P_{outage}, we use the following fact.

Fact. 2: The probability of outage is given as

$$P_{\text{outage}} = \tau \boldsymbol{\pi_G} \boldsymbol{P_B}^{\tau} \mathbf{1} + \boldsymbol{\pi_G} \boldsymbol{P_B}^{\tau+1} (I - \boldsymbol{P_B})^{-1} \mathbf{1}. \tag{28}$$

Based on Fact 2, we establish an alternative analytical expression for the probability of outage. The new expression is easier to compute, as it involves neither inversion of a matrix nor calculation of vector $\boldsymbol{\pi_G}$.

Proposition 2:

$$P_{\text{outage}} = \boldsymbol{\pi} \boldsymbol{P_B}^{\tau} \left(\tau I - (\tau - 1) \boldsymbol{P_B}\right) \mathbf{1}. \tag{29}$$

Proof: From (28), we have

$$\begin{aligned} P_{\text{outage}} =& \tau \boldsymbol{\pi_G} \boldsymbol{P_B}^{\tau} \mathbf{1} + \boldsymbol{\pi_G} \boldsymbol{P_B}^{\tau+1} (I - \boldsymbol{P_B})^{-1} \mathbf{1} \\ =& \boldsymbol{\pi_G} \boldsymbol{P_B}^{\tau} \big((\tau-1)I + (I - \boldsymbol{P_B})(I - \boldsymbol{P_B})^{-1} \\ & \quad + \boldsymbol{P_B}(I - \boldsymbol{P_B})^{-1}\big) \mathbf{1} \\ =& \boldsymbol{\pi_G} \boldsymbol{P_B}^{\tau} \left((\tau-1)I + (I - \boldsymbol{P_B})^{-1}\right) \mathbf{1} \\ =& (\boldsymbol{\pi} - \boldsymbol{\pi_B}) \boldsymbol{P_B}^{\tau} \left((\tau-1)I + (I - \boldsymbol{P_B})^{-1}\right) \mathbf{1} \\ =& (\boldsymbol{\pi} \boldsymbol{P_B}^{\tau} - \boldsymbol{\pi_B} \boldsymbol{P_B}^{\tau}) \left((\tau-1)I + (I - \boldsymbol{P_B})^{-1}\right) \mathbf{1} \\ =& (\boldsymbol{\pi_B} \boldsymbol{P_B}^{\tau-1} - \boldsymbol{\pi_B} \boldsymbol{P_B}^{\tau}) \left((\tau-1)I + (I - \boldsymbol{P_B})^{-1}\right) \mathbf{1} \\ =& \boldsymbol{\pi_B} \boldsymbol{P_B}^{\tau-1} (I - \boldsymbol{P_B}) \left((\tau-1)I + (I - \boldsymbol{P_B})^{-1}\right) \mathbf{1} \\ =& \boldsymbol{\pi_B} \boldsymbol{P_B}^{\tau-1} \left((\tau-1)(I - \boldsymbol{P_B}) + I\right) \mathbf{1} \\ =& \boldsymbol{\pi_B} \boldsymbol{P_B}^{\tau-1} \left(\tau I - (\tau-1) \boldsymbol{P_B}\right) \mathbf{1} \\ =& \boldsymbol{\pi} \boldsymbol{P_B}^{\tau} \left(\tau I - (\tau-1) \boldsymbol{P_B}\right) \mathbf{1} \end{aligned}$$

where the third equality holds since $\boldsymbol{\pi_B} = \boldsymbol{\pi} - \boldsymbol{\pi_G}$ and the fourth and sixth equalities result from the fact that $\boldsymbol{\pi_B} = \boldsymbol{\pi} \boldsymbol{P_B}$.

■

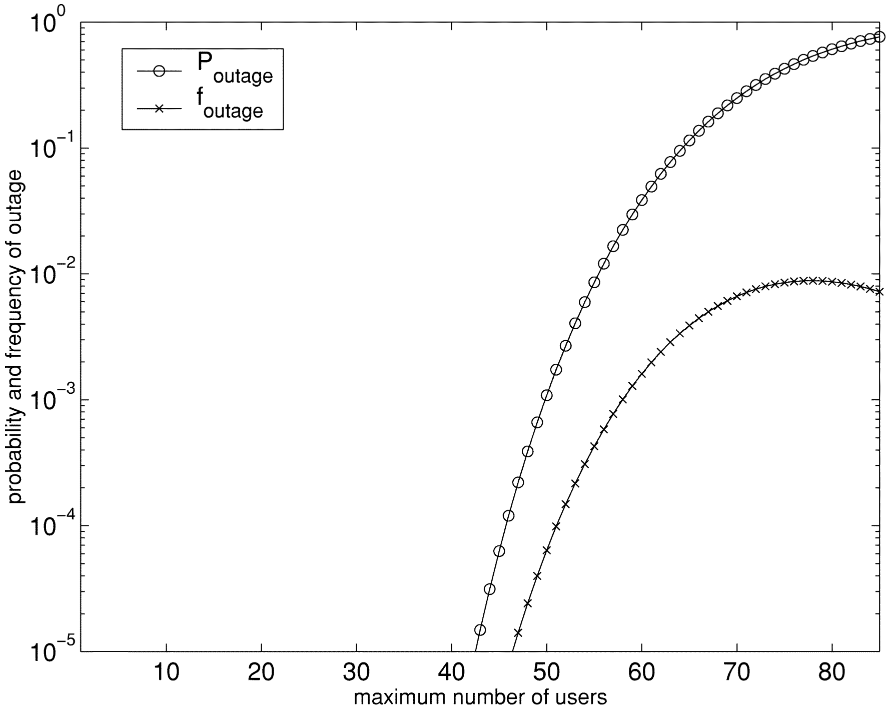

Fig. 1. P_{outage} and f_{outage} versus M for cases in Section III-A4.

4) Illustration and Critique of the Measures: The formulation of outage presented in Section III-A2 and the analysis of Section III-A3 provide an expression for the probability (frequency) of outage of a fictitious user of type l, $l = 1, 2, \ldots, L$, as a function $g_l : \mathbb{Z}_+^L \longmapsto \mathbb{R}_+$ ($\hat{g}_l : \mathbb{Z}_+^L \longmapsto \mathbb{R}_+$) of the vector of admitted users $\underline{M} = (M_1, M_2, \ldots, M_L)$.

For the case where there is only one class of users, we illustrate the behavior of g_1 ($\hat{g}_1$) via the following example. We consider a CDMA system with only one type of traffic, where there is a power-control mechanism with error of 5% where the maximum Doppler frequency of the channel is 100 Hz. We assume that $T = .002\tau_f$, where T is the time-slot duration and τ_f is the fading cycle, $\gamma = 3.1$ dB, spreading gain $G_0 = 64$, and $\tau = 7T$ (the value recommended by ITU-T [12]).

We assume that the traffic consists only of voice users, whose activity model follows a simple memoryless process—i.e., if M represents the fixed number of users admitted to the system, the transition probability of the number of active users at each time, denoted by $N(t)$, can be expressed as

$$P\{N(t+1) = n | N(t) = m, M\} = \begin{cases} (M-m)\lambda, & \text{if } n = m+1 \\ m\mu, & \text{if } n = m-1 \\ 1 - m\mu - (M-m)\lambda, & \text{if } n = m \\ 0, & \text{otherwise} \end{cases}$$

where $m \leq M$, λ is the activation rate of each inactive user, and μ is the probability that an active user becomes inactive. Note that $\lambda/(\lambda+\mu)$ is the voice activity factor and is around 0.4.

After identifying the transition matrix, determining the SNIR at each state, and finally labeling "bad" and "good" states according to (17) and (18), we calculate the probability and frequency of outage as functions $g_1(M)$ and $\hat{g}_1(M)$, respectively. Fig. 1 shows the result of such a calculation.

In Fig. 1, the frequency of an outage is not monotone in the number of admitted users. This is due to the fact that as the number of admitted users increases, the average number of time slots where the state is in the "bad" set $\mathcal{B}$ increases. This implies that as the number of admitted users increases, the probability that at any time slot the state of the channel is good or that the time slot does not belong to an outage decreases dramatically. However, because of Proposition 1, entering an outage at any time slot t requires the state of the system at time $t-1$ to be "good"; hence the probability of entering an outage event (frequency of outage) ultimately decreases, since the probability of being in a "good" state decreases.

The form of the function $\hat{g}_1$ indicates that in a dynamic system where there are connection arrivals and departures, the frequency of outage can increase in the case of a connection departure. On the other hand, the departure times are random and independent of control actions (which only include the admission decisions). This implies that even if at a particular instance of time the frequency of outage is within an acceptable range, it may not continue to be within that range no matter what the connection admission control policy is. Hence, the frequency of outage is not well suited to use in the construction of an outage-based admission region or to define an outage-based capacity.

The function g_1, which is the probability of outage for our example, is monotone in the number of admitted users. Other examples for one and two types of users, provided in Section III-B, as well as in [13], show that $g_1(M)$ (in case of examples with one type of user, discussed in Section III-B1) and $g_l(M_1, M_2)$, $l = 1, 2$ (in cases studied in Section III-B2), are all monotone increasing in M_1 and M_2. Based on these numerical examples, we propose Conjecture 1.

Conjecture 1: In any cellular system, $g_l(\underline{M})$ is component-wise increasing for all $l = 1, 2, \ldots, L$.

Based on this conjecture, we propose that the probability of outage is the appropriate measure to define an outage-based admission region. In Section III-A5, we present a procedure on how to construct an admission region based on probability of outage. The procedure does not depend on the validity of Conjecture 1. At the end of Section III-A5, we discuss how the validity of Conjecture 1 simplifies the procedure.

5) Construction of the Admission Region: We now discuss how to use the results obtained in Section III-A3 to construct an admission region when the probability of outage is the QoS requirement under consideration. An admission region is the set of all combinations of admitted users such that if connection admissions are restricted to a subset of its interior, the probability of an outage encountered by a fictitious observer/user of type l is less than a prespecified threshold $P_{\max}^l$ for all $l = 1, 2, \ldots, L$.

The formulation of probability of outage presented in Section III-A2 and the analysis of Section III-A3 provide an expression for the probability of outage of a fictitious user of type l ($l = 1, 2, \ldots, L$) as a function $g_l : \mathbb{Z}_+^L \longmapsto \mathbb{R}_+$ of the vector of admitted users $\underline{M} = (M_1, M_2, \ldots, M_L)$. Therefore, for a fixed type-l fictitious user (i.e., $\mathcal{H}_{u_0} = \mathcal{H}_l$, $\mathbf{A}_0 = \mathbf{A}_l$, $\tau_0 = \tau_l$, $\gamma_0 = \gamma_l$, and $f_{u_0}(\cdot) = f_l(\cdot)$), the region where QoS (expressed by the probability of outage) is guaranteed for that type of user is

$$\mathcal{R}_l := g_l^{-1}([0, P_{\max}^l]) = \{\underline{M} : g_l(\underline{M}) \le P_{\max}^l\}. \tag{30}$$

Consequently, the region where the QoS is guaranteed for all users is

$$\mathcal{R} := \bigcap_{l=1}^{L} \mathcal{R}_l = \bigcap_{l=1}^{L} \{\underline{M} : g_l(\underline{M}) \le P_{\max}^l\}. \tag{31}$$

Since it is not desirable for any admission strategy to terminate an unfinished service, we define the admission region $\mathcal{A}$ as the largest coordinate convex subset of $\mathcal{R}$, i.e.,

$$\mathcal{A} := \sup_{\mathcal{C} \subseteq \mathcal{R}} \{\mathcal{C} : \text{ if } \underline{M} \in \mathcal{C}, \underline{M}' \le \underline{M} \text{ then } \underline{M}' \in \mathcal{C}\}. \tag{32}$$

Recall that our analysis is valid for a fixed number of admitted users. In a wireless system, the number of users (active and inactive) present in the system varies with time. To establish the validity of our analysis for wireless systems, we prove the following theorem.

Theorem 1: The admission region $\mathcal{A}$ is a conservative bound on the number of admitted users for which the QoS expressed by the probability of outage is met.

Proof: Fix a fictitious observer/user, say, of type l. Restrict an admission policy to $\mathcal{A}$. Since $\mathcal{A}$ is coordinate convex, $\underline{M}(t) \in \mathcal{A}$ for all t, where $\underline{M}(t)$ is the vector indicating the number of users of each type present in the system at time t. Hence $\underline{M}(t) \in \mathcal{R}_l$ for all t. Pick $\underline{M}_0 = \arg\max_{\underline{M} \in \mathcal{R}_l} g_l(\underline{M})$. Since $\underline{M}(t) \in \mathcal{R}_l$ for all t, the probability of outage in the dynamic system for the aforementioned fictitious observer/user is less than or equal to $g_l(\underline{M}_0)$, and by the construction of $\mathcal{R}_l$

$$g_l(\underline{M}_0) \le P_{\max}^l. \tag{33}$$

Since l is arbitrary, the admission region $\mathcal{A}$, defined by (32), is a conservative bound on the number of users admitted by the dynamic system for which the QoS requirement expressed by the probability of outage is satisfied. ■

Now we discuss the implication of Conjecture 1 on the methodology used to construct the admission region based on outage probability. If Conjecture 1 is true, then $g_l(M_1, M_2, \ldots, M_L)$, $l = 1, 2, \ldots, L$, is increasing in each coordinate $M_1, M_2, \ldots, M_L$. This implies that region $\mathcal{R}_l$ $l = 1, 2, \ldots, L$, defined by (30) is coordinate convex. This, in turn, implies that region $\mathcal{R}$ defined by (31) is coordinate convex; hence $\mathcal{A} = \mathcal{R}$.

B. Special Cases, Examples, and Discussion

In this section, we present examples illustrating our approach. In all the examples, we consider the outage problem in a CDMA pre-third-generation wireless systems. In such systems, traffic mainly consists of voice, or data streams that are compressed and then treated as voice [14]. This kind of traffic, when the number of users is fixed, can be appropriately modeled by an Engseth birth–death chain (see [15]). Therefore, it is appropriate to follow the procedure given in Definition 4a.

In the remainder of the section, we formally introduce the Engseth traffic model. In Section III-B1 and B3, we compute the probability of outage and the resulting admission regions under a variety of power-control scenarios and system parameters.

We consider L types of traffic. Let the component M_k of the vector $\underline{M} = (M_1, \ldots, M_L)$ represent the fixed number of users of type k admitted to the system. Let $\underline{N}(t) = (N_1(t), \ldots, N_L(t))$ denote the vector of the number of active users of each type. Then the transition probability for the Engseth model is given as follows:

$$P\{\underline{N}(t+1) = \boldsymbol{n} | \underline{N}(t) = \boldsymbol{m}, \underline{M}\} = \begin{cases} (M_k - m_k)\lambda_k, & \text{if } n = m + e_k \\ m_k \mu_k, & \text{if } n = m - e_k \\ 1 - \sum\limits_{k \in K_{\text{on}}} m_k \mu_k - \sum\limits_{k \in K_{\text{off}}} (M_k - m_k)\lambda_k, & \text{if } n = m \\ 0, & \text{otherwise} \end{cases}$$

where e_k is a column vector whose elements are all zero except for the kth element, which is one, λ_k is the activation rate of each inactive user of type k, μ_k is the probability that an active user becomes inactive, $K_{\text{off}} = \{k : m_k < M_k\}$, and $K_{\text{on}} = \{k : 0 < m_k\}$. Note that $\lambda_k/(\lambda_k + \mu_k)$ is the activity factor of each stream. For voice users, this is around 0.4. For data users, it varies with the application, and it depends on the burstiness and information bandwidth of the stream, as well as the compression method employed.

1) Example: Homogeneous Traffic: We study the homogeneous traffic scenario. We first construct the SMC associated with this model. Let M denote the number of admitted users in the system. We construct $\mathbf{T}'_M = [\text{Prob}\{r = i/r = j, M\}]$. For the cases that follow, we use different channel models $(\mathcal{H}, \mathbf{A}_0)$. We construct the transition probability $\mathbf{P}$ associated with each $(\mathcal{H}, \mathbf{A}_0)$ pair using (15).

For all the scenarios under study, we assume that $\lambda/(\lambda + \mu) = 0.4$, $T = .002\tau_f = \tau_c$ (T is the time slot duration, τ_f is the fading cycle, τ_c is the shadowing cycle), $\gamma = 3.1$ dB, and the spreading gain is $G_0 = 64$.

After identifying the transition matrix, determining the SNIR at each state, and finally labeling "bad" and "good" states according to (17) and (18), we calculate the probability of an

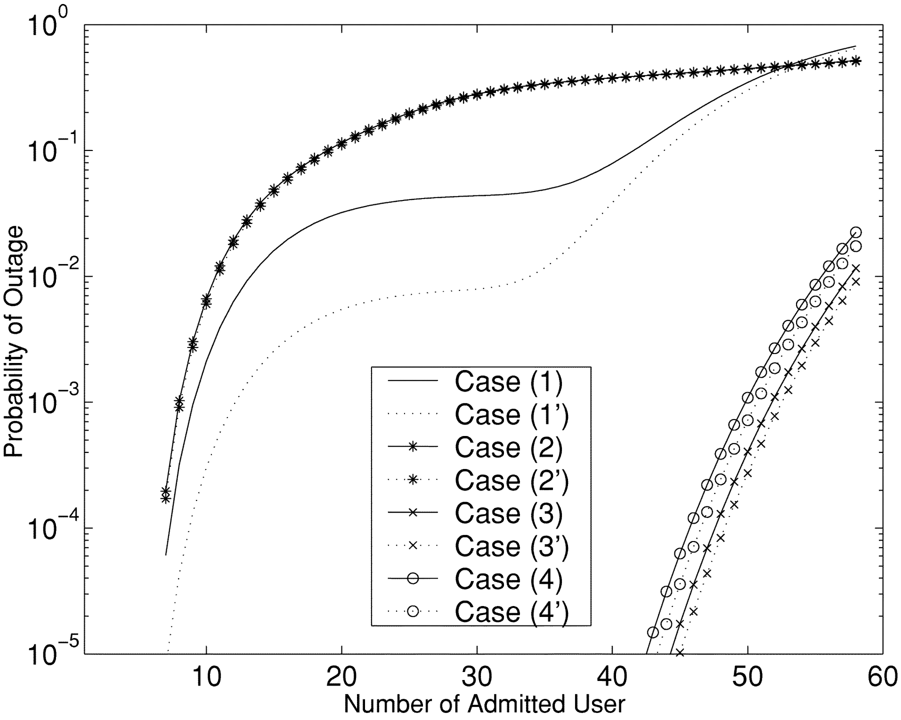

Fig. 2. P_{outage} versus number of admitted users M for cases in Section III-B1.

outage as a function $g_1(M)$. Fig. 2 shows the result of such a calculation for CDMA systems when:

1) the channel follows a Gilbert model with average burst lengths of four, with steady-state probability of the bad-channel-state equal to 0.1 and $\tau = 7T$ (the value recommended by ITU-T [12]);
1') channel is similar to 1) and $\tau = 15T$;
2) channel is an appropriate approximation to a log-normal shadowing channel with the maximum Doppler frequency of 100 Hz and correlation distance of 1 m, as given by [9], and $\tau = 7T$;
2') channel is similar to 2) and $\tau = 15T$;
3) an ideal power-control mechanism is implemented and $\tau = 7T$;
3') channel is similar to 3) and $\tau = 15T$;
4) power control is applied with error of 5% and $\tau = 7T$;
4') channel is similar to 4) and $\tau = 15T$.

2) Example: Two Types of Traffic: We study the outage problem for the same CDMA system as in Section III-B1 when the traffic consists of two classes of users with different activity factors, spreading gains, and outage parameters; these parameters are $\lambda_1/(\lambda_1+\mu_1) = 0.4$, $\lambda_2/(\lambda_2+\mu_2) = 0.6$, $G_1 = 64$, $G_2 = 32$, $P_2 = 2P_1$ $\tau_1 = \tau_2$, $\gamma_1 = 3.1$ dB, and $\gamma_2 = 3.3$ dB. We set the maximum acceptable probability of outage to be equal to 10^{-3}. Under this specification, Fig. 3 shows the admission region when

1) channel is described by a Gilbert model similar to that in Section III-B1 and $\tau = 7T$;
1') channel is described by a Gilbert model similar to that in Section III-B1 and $\tau = 15T$;
2) there is an ideal power-control mechanism and $\tau = 7T$;
2') there is an ideal power-control mechanism and $\tau = 15T$.

IV. Connection Admission Control

As discussed in the introduction, we propose to formulate the CAC problem in a single-hop multiservice network with QoS requirements as a constrained stochastic dynamic optimization problem, where the constraint describes the admission region. One standard approach to describing the admission region for such a problem is to define a total capacity for the network and associate an effective bandwidth to each class of users. This approach approximates the boundary of the admission region with a linear function of the number of each type of users (see [16], [4], and [5]). In cases where all the QoS requirements are summarized by an effective bandwidth-based admission region, the CAC problem is equivalent to a classical knapsack problem. Indeed, when the admission region is described by a linear inequality, the CAC problem is reduced to the search of allocation schemes that share a resource of finite and fixed capacity among several classes of traffic in a manner that is optimal with respect to the total expected generated revenue. This problem is equivalent to scheduling a stochastic knapsack, a well-studied subject in stochastic networks and operations research. The classical knapsack problem involves a knapsack of capacity W resource units and K classes of connections, with each connection of class-k occupying w_k units (w_k is the effective bandwidth of class-k users), and having random arrival and holding times with rates λ_k and μ_k, respectively. A revenue of c_k per unit of time is incurred while a class-k connection is placed into the knapsack. The new connections may be admitted to the knapsack as long as the sum of the occupied bandwidth does not exceed the knapsack capacity. The connections who are denied admission to the knapsack are lost. The objective is to determine an admission policy that maximizes the long-run average revenue (see [17,

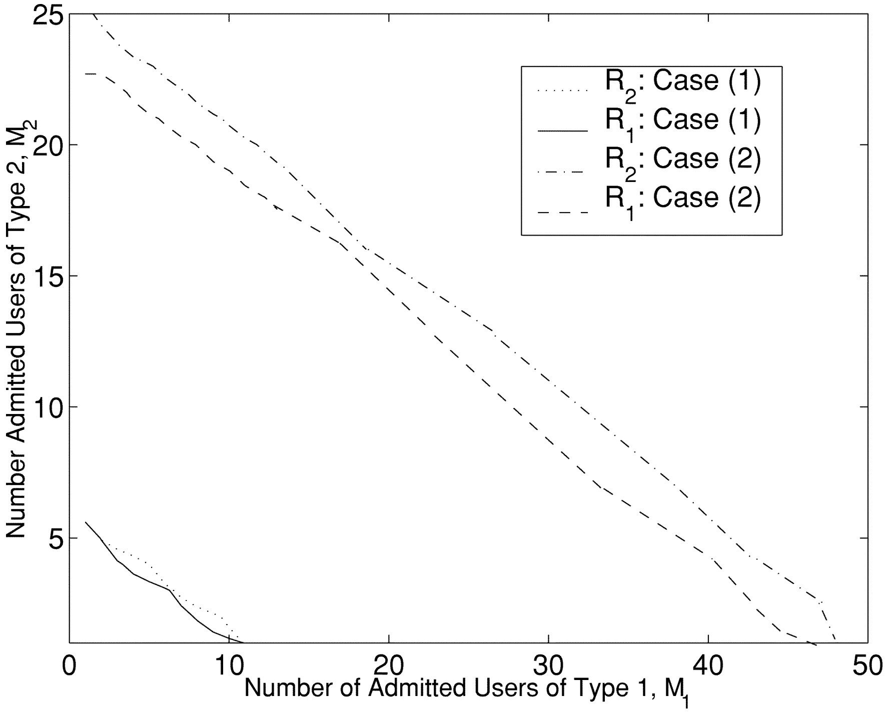

Fig. 3. Admission regions for cases in Section III-B2.

ch. 2–4] and the references therein for details on the classical stochastic knapsack problem).

In general, all QoS requirements considered simultaneously are summarized by an admission region, the boundary of which need not be a line. In such a situation, a reasonable assumption on the nature of the admission regions and their boundaries is coordinate convexity, which implies that no forced termination of service is required in order to meet QoS requirements. Based on this observation, in this paper we propose the formulation and investigation of a "generalized knapsack" whose scheduling is equivalent to the CAC problem with a coordinate convex admission region. In a generalized knapsack problem, there are K class of connections. The number and configuration of served connections are restricted to a coordinate convex set $\mathcal{A} \in \mathbb{R}_{+}^{K}$; we denote the boundary of the admission region by the set B. There are multiple identical parallel servers that serve the connections admitted in the system. The rate of service and arrivals for class-k connections is μ_k and λ_k, respectively. Each class-k connection generates a revenue of rate c_k while being served in the generalized knapsack. New connections can be potentially admitted if the resulting configuration and number of admitted connections are still in the admission region $\mathcal{A}$. The connections that are denied admission to the system are lost. The objective is to determine an admission strategy to maximize the long-run average revenue. A more detailed description of the generalized knapsack problem that is necessary for the analysis of the CAC problem will be given in Section IV-A.

To motivate the analysis of the CAC problem presented in Section IV-A, we first briefly discuss and critique the results available on the classical knapsack problem. Several variants of the classical knapsack problem have been carefully studied in the literature (for example, see [18]–[25], [17], [26]–[28]). In [23]–[25], there is a one-time reward that is fixed and known, and is obtained at the instance of admission. This feature makes the problem considered in [23]–[25] distinctly different from the problem we consider in this paper. In [22] and [28], it is assumed that no job admitted to the knapsack leaves the system, i.e., the problem changes to a packing problem. Such a problem is also distinctively different from ours. The model and the formulation of knapsack problem considered in [18], [20], [21], [17, ch. 4], [26], and [27] are similar to our problem. A Markov decision process (MDP) approach is used in this class of references for the analysis of the classical knapsack problem. It has been shown that solving the appropriate MDP, through standard numeric programming methods, can be analytically intractable and computationally complex. In [17] and [26], the authors propose a standard linear programming (LP) technique to solve the dynamic programming (DP) equation associated with the MDP describing the classical knapsack problem and compare the computational complexity of such a technique to the standard value or policy iteration techniques. Furthermore, it is known that the optimal solution to such MDP (which is the optimal admission policy for the classical knapsack), in general, lacks any specific structure or well-definable property (see [21]).

The general lack of structure in the optimal admission policy for the classical knapsack motivated the study of high-performance suboptimal policies that can be computed in reasonable amount of time. The result in [20] provides an approximation bound on the optimal performance and a heuristic in constructing efficient (suboptimal) policies. In [21] and [27], the authors restrict their attention to the class of coordinate convex (c-c) policies and attempt to characterize the optimal c-c policy. Coordinate convex policies are easier to study since under any c-c policy, the steady-state distribution is of a product form that makes the problem more tractable. Unfortunately, under special case scenarios, it can be shown that the optimal c-c policy is far from optimal (see [21]).

The complicated nature of the optimal connection admission control policies creates a practical difficulty for their implementation as viable CAC policies in high-speed networks. Consequently, in this paper, we consider the "greedy policy," which has a very simple implementation. The greedy policy admits any request for connection if there are sufficient resources available. We determine conditions on the rates of revenue generated by different classes of connections sufficient to guarantee the optimality of the greedy policy. The problem we address can be thought of as follows: how should each type of service provided by the network be charged so that it should be optimal to admit every request for connection provided that there are sufficient resources?

The remainder of this section is organized as follows. In Section IV-A, we formulate the CAC problem with two classes of connections as a generalized knapsack problem, called Problem (P). In Section IV-B, we define a new problem related to Problem (P), referred to as Problem (P'). We analyze Problem (P') and show that an optimal solution to Problem (P') is also an optimal solution to Problem (P). Section IV-C includes a brief discussion of future work and further extensions of the CAC problem.

A. The Generalized Stochastic Knapsack Problem With Two Classes of Connections

The generalized stochastic knapsack problem with two classes of users can be formulated as follows.

Problem (P): Consider a finite coordinate-convex set $\mathcal{A} \subset \mathbb{R}_+^2$ that contains the origin. A two-dimensional generalized knapsack associated with set $\mathcal{A}$ consists of a system of identical servers in parallel that can serve two classes of service within its support region $\mathcal{A}$. That is, the knapsack may serve x_1 number of class-1 connections and x_2 of class-2 connections only if $(x_1, x_2) \in \mathcal{A}$. Each connection of class k, $k = 1, 2$, is characterized by its arrival rate λ_k and the rate of its service time μ_k. We assume that the arrival and service statistics of each connection are independent of each other and independent of the arrival and service statistics of other connections; the service time for a connection of type k is a memoryless random variable with mean $1/\mu_k$; and at each unit of time there is at most one new connection arrival to the system. Each arriving connection can be admitted to the knapsack if the resulting number of connections is in $\mathcal{A}$. If a request for connection is rejected, the connection is lost. An admitted connection remains in the knapsack until its service is completed. Without any loss of generality and for clarity, we assume that arrivals and departures within the time slot from time t to $t{+}1$ occur in the open interval $(t, t{+}1)$; furthermore, departures occur at the end of a time slot, whereas arrivals occur at the beginning of a time slot. Thus, if we define t^+ and t^- as

$$\begin{aligned} t^+ = &\inf\{s \in (t, t+1) : s \text{ is any time} \\ &\text{after the arrival time of new} \\ &\text{connection requests and also} \\ &\text{admission decisions in slot } t\} \\ (t+1)^- = &\sup\{s \in (t, t+1) : s \text{ is any time} \\ &\text{before the completion time of} \\ &\text{any connection whose service} \\ &\text{ends in time slot } t\} \end{aligned}$$

we have $t^+ < (t+1)^-$. Each admitted connection of type $k = 1, 2$ generates a revenue of rate c_k while being served in the knapsack. The goal is to find an optimal admission strategy that maximizes the total expected revenue over a finite horizon T.

Remark: As a result of our formulation, a packet of type k may be admitted in the system at t^+, complete service at $(t+1)^-$, and result in a revenue c_k.

The main result of this section is summarized by the following theorem.

Theorem 2: If

$$\frac{\lambda_2}{\mu_2} \le \frac{c_1}{c_2} \le \frac{\mu_1}{\lambda_1} \tag{34}$$

then the policy that follows the greedy rule at all times is optimal for Problem (P).

B. Analysis of Problem (P)

We proceed to solve Problem (P) as follows. First, we formulate another problem, called Problem (P'), which includes the original knapsack described in Problem (P), and an auxiliary knapsack. Afterwards, we analyze Problem (P'). We determine conditions sufficient to guarantee the optimality of the greedy admission policy for Problem (P'). Finally, we show that conditions that are sufficient to guarantee the optimality of the greedy policy for Problem (P') are also sufficient to guarantee the optimality of the greedy policy for Problem (P).

Problem (P') includes an auxiliary (generalized) knapsack Q whose admission region is defined by

$$\mathcal{A}_q = \{(y_1, 0), (0, y_2) : y_1 \le b_1 \text{ and } y_2 \le b_2\} \tag{35}$$

where b_1 and b_2 are defined as

$$b_1 = \max_{(x_1,x_2)\in B} \max_{k\in\mathbb{Z}} \{k : (x_1 + k, x_2 - 1) \in \mathcal{A}\} \tag{36}$$

$$b_2 = \max_{(x_1,x_2)\in B} \max_{k\in\mathbb{Z}} \{k : (x_1 - 1, x_2 + k) \in \mathcal{A}\} \tag{37}$$

where B is the integer boundary of $\mathcal{A}$, i.e., $B := \{(x_1, x_2) \in \mathcal{A} | (x_1 + 1, x_2) \notin \mathcal{A} \text{ and } (x_1, x_2 + 1) \notin \mathcal{A}\}$. This auxiliary knapsack is used in the same manner as the original one, but the connections admitted to the auxiliary knapsack do not generate any revenue. Knapsack Q can be used to provide information about the service times of certain selected connections that were denied admission to the original knapsack. Obviously, any admissible connection control policy for Problem (P') that does not use the information provided by the auxiliary knapsack is equivalent to an admissible admission policy for Problem (P).

Problem (P'): We consider a system consisting of the original knapsack of Problem (P) and the auxiliary knapsack Q defined by (35)–(37). The state of this system at any time t is defined by (x_1, x_2, y_1, y_2), where $(x_1, x_2) \in \mathcal{A}$ and $(y_1, y_2) \in \mathcal{A}_q$. The original knapsack operates in exactly the same manner as in Problem (P). Connections admitted to the the auxiliary knapsack Q may be dropped before they complete their service. The forced departure times of connections to knapsack Q may depend on the admission policy. Forced departures from Q in time slot t, like admissions and rejections of arrived requests from knapsack Q in time slot t, occur at the beginning of the time slot, i.e., before time t^+. Each connection of type k, $k = 1, 2$, admitted to the original knapsack generates a revenue at rate c_k. Connections admitted to the auxiliary knapsack Q do not generate any revenue. An admissible policy π is a sequence $\pi_0, \pi_1, \ldots, \pi_t, \ldots, \pi_{T-1}$ of functions of the form

$$\pi_t(x_1, x_2, y_1, y_2, a_1(t), a_2(t)) = (x_1 + u_1, x_2 + u_2, y_1', y_2')$$

where for any t, $(a_1(t), a_2(t)) \in \{(0,0), (1,0), (0,1)\}$ denotes the number of arrivals of each type of connection requests at time t, u_i, $i = 1, 2$, represents the number of connection admissions of type i into the original knapsack, i.e., $u_i \in \{0, 1\}$ if $(x_1 + u_1, x_2 + u_2) \in \mathcal{A}$, and $u_i = 0$ if the original knapsack is full, $(y_1', y_2') \in \mathcal{A}_q$, and $u_i + (y_i' - y_i) \le a_i(t)$. The objective is to determine an admission policy that maximizes the expected revenue over a finite horizon T.

To proceed with the analysis of Problem (P'), we need the following.

Definition 5: A policy π is said to follow the *greedy* rule at time t if for any state (x_1, x_2, y_1, y_2)

$$\begin{aligned}
&\pi_t(x_1, x_2, y_1, y_2, 1, 0)\\
&= \begin{cases} (x_1 + 1, x_2, 0, 0), & \text{if } (x_1 + 1, x_2) \in \mathcal{A} \\ (x_1, x_2, 0, 0), & \text{otherwise} \end{cases}\\
&\pi_t(x_1, x_2, y_1, y_2, 0, 1)\\
&= \begin{cases} (x_1, x_2 + 1, 0, 0), & \text{if } (x_1, x_2 + 1) \in \mathcal{A} \\ (x_1, x_2, 0, 0), & \text{otherwise.} \end{cases}
\end{aligned}$$

Definition 6: A policy π is said to follow the *$(\mathbf{m}, \mathbf{n})$-tracking-greedy* rule at time t if for any state (x_1, x_2, y_1, y_2)

$$\begin{aligned}
&\pi_t(x_1, x_2, y_1, y_2, 1, 0)\\
&= \begin{cases} (x_1{+}1, x_2, y_1, y_2), & \text{if } (x_1{+}1, x_2) \in \mathcal{A} \text{ and} \\ & (x_1{+}y_1 - m{+}1, x_2{+}y_2 - n) \in \mathcal{A} \\ (x_1, x_2, y_1{+}1, y_2), & \text{if } (x_1{+}1, x_2) \notin \mathcal{A} \text{ and} \\ & (y_1{+}1, y_2) \in \mathcal{A}_q \text{ and} \\ & (x_1{+}y_1 - m{+}1, x_2{+}y_2 - n) \in \mathcal{A} \\ (x_1, x_2, y_1, y_2), & \text{otherwise} \end{cases}\\
&\pi_t(x_1, x_2, y_1, y_2, 0, 1)\\
&= \begin{cases} (x_1, x_2{+}1, y_1, y_2), & \text{if } (x_1, x_2{+}1) \in \mathcal{A} \text{ and} \\ & (x_1{+}y_1 - m, x_2{+}y_2 - n{+}1) \in \mathcal{A} \\ (x_1, x_2, y_1, y_2{+}1), & \text{if } (x_1, x_2{+}1) \notin \mathcal{A} \text{ and} \\ & (y_1, y_2{+}1) \in \mathcal{A}_q \text{ and} \\ & (x_1{+}y_1 - m, x_2{+}y_2 - n{+}1) \in \mathcal{A} \\ (x_1, x_2, y_1, y_2), & \text{otherwise.} \end{cases}
\end{aligned}$$

Definition 7: A policy π is said to follow the *emptying-greedy* rule at time t if for any state (x_1, x_2, y_1, y_2)

$$\begin{aligned}
&\pi_t(x_1, x_2, y_1, y_2, 1, 0)\\
&= \begin{cases} (x_1 + 1, x_2, y_1, y_2), & \text{if } (x_1 + y_1 + 1, x_2 + y_2) \in \mathcal{A} \\ (x_1, x_2, y_1, y_2), & \text{otherwise} \end{cases}\\
&\pi_t(x_1, x_2, y_1, y_2, 0, 1)\\
&= \begin{cases} (x_1, x_2 + 1, y_1, y_2), & \text{if } (x_1 + y_1, x_2 + y_2 + 1) \in \mathcal{A} \\ (x_1, x_2, y_1, y_2), & \text{otherwise.} \end{cases}
\end{aligned}$$

A policy that follows the greedy rule admits to the original knapsack any request for service at any time if the resulting state is admissible, and empties the auxiliary knapsack. Such a policy is a trivial extension of a complete sharing (greedy) policy (see [17]) for Problem (P). In fact, the policy that follows the greedy rule is equivalent to a complete sharing policy for Problem (P), since it always empties the auxiliary knapsack Q and does not use the information provided by it. A policy that follows the (m, n)-tracking-greedy rule and is applied to a system with initial state $(x_1, x_2, 0, 0)$ attempts to duplicate, as closely as possible, the admission decisions associated with a greedy policy when it is applied to another system whose initial state is $(x_1 - m, x_2 - n, y_1, y_2)$ and its arrival times and completion times are coupled with those of the original system. A policy that follows the emptying-greedy rule and is applied to a system with initial state (x_1, x_2, y_1, y_2) duplicates the admission decisions associated with a greedy policy when it is applied to another system whose initial state is $(x_1 + y_1, x_2 + y_2, 0, 0)$, and its arrival times and completion times are coupled with those of the original system.

We now prove the following result for Problem (P').

Theorem 3: Consider Problem (P'). If (34) holds, it is always optimal to follow the greedy rule at each time t.

Proof: We prove the assertion of the theorem by induction. We first establish the basis of induction by showing that at horizon $T-1$, it is optimal to follow the greedy rule. Then we assume it is optimal to follow the greedy rule from time $t+1$ on and show that it is optimal to follow the greedy rule at time t.

Denote by $J_t^\pi(x_1, x_2, y_1, y_2)$ the total revenue generated by policy π along a sequence of arrivals and departures from time t on, assuming that the system is in state (x_1, x_2, y_1, y_2) at time t. Denote by π^* the policy that follows the greedy rule at each time t. We need to show that under (34)

$$E\{J_t^{\pi^*}(x_1, x_2, y_1, y_2)\} \ge E\{J_t^{\pi}(x_1, x_2, y_1, y_2)\} \tag{38}$$

for any policy π and any initial state (x_1, x_2, y_1, y_2).

Basis of Induction: For $t = T - 1$, we have

$$J_{T-1}^{\pi^*}(x_1, x_2, y_1, y_2) \ge J_{T-1}^{\pi}(x_1, x_2, y_1, y_2)$$

for every policy π and for any state (x_1, x_2, y_1, y_2)).

Induction Step: Assume that for any state (x_1, x_2, y_1, y_2) and any policy π, we have

$$E\{J_{t+1}^{\pi^*}(x_1, x_2, y_1, y_2)\} \ge E\{J_{t+1}^{\pi}(x_1, x_2, y_1, y_2)\}. \tag{39}$$

To establish the optimality of the greedy policy, it is sufficient to consider a policy π that is different from the greedy rule at time t (assume $state \notin B$), then follows the greedy rule from time $t+1$ on, and show that for any (x_1, x_2, y_1, y_2), $E\{J_t^{\pi^*}(x_1, x_2, y_1, y_2)\} \ge E\{J_t^{\pi}(x_1, x_2, y_1, y_2)\}$.

Without any loss of generality, we can assume that there is an arrival at time t (if not, then the two policies π and π^* generate identical revenues). Let (x_1, x_2, y_1, y_2) be the state at time t. The greedy policy admits the new connection into the original knapsack. For policy π, we consider the following cases.

Case 1: (This case holds only when $y_1 \ne 0$ or $y_2 \ne 0$.) Policy π admits the new arrival into the original knapsack but does not empty the auxiliary knapsack. In this situation, we have

$$E\{J_t^{\pi^*}(x_1, x_2, y_1, y_2)\} - E\{J_t^{\pi}(x_1, x_2, y_1, y_2)\} = 0 \tag{40}$$

since connections served in the auxiliary knapsack do not generate any revenue.

Case 2: The arrival is of type 1, and policy π does not admit the new connection into the original knapsack. Then at time t^+, the knapsack state under policies π^* and π is $(x_1+1, x_2, 0, 0)$ and (x_1, x_2, y_1, y_2), respectively.

To prove the induction step, we construct a policy $\tilde{\pi}$ as follows: policy $\tilde{\pi}$ follows the greedy rule at time t and marks the accepted connection at this time slot as j_1; from time $t+1$ until the stopping time $t+\tau$ (defined below), $\tilde{\pi}$ follows the (1, 0)-tracking-greedy policy; from time $t+\tau$ until stopping time $t+\tau+\sigma$ (defined below), policy $\tilde{\pi}$ follows the emptying-greedy policy; and from time $t+\tau+\sigma$, policy $\tilde{\pi}$ follows the greedy policy. The stopping times above are defined as:

τ $\min\{\tau_1, \tau_2, T\}$;
τ_1 the occupation time of connection j_1;
$t+\tau_2$ the first time under policy $\tilde{\pi}$ the auxiliary knapsack is empty of type-2 users, and a connection of type-1 arrives, and cannot be admitted into the original knapsack;
$t+\tau+\sigma$ the first time after $t+\tau$ such that under policy $\tilde{\pi}$, the auxiliary knapsack is empty.

We compare the performance of policy $\tilde{\pi}$ with that of policies π^* and π. First we compare policies π^* and $\tilde{\pi}$.

Lemma 2: Policy π^* outperforms policy $\tilde{\pi}$.

Proof :

$$\begin{aligned}&E\{J_t^{\pi^*}(x_1,x_2,y_1,y_2)\} - E\{J_t^{\tilde{\pi}}(x_1,x_2,y_1,y_2)\}\\ &=E\{J_{t+1}^{\pi^*}(x_1',x_2',0,0)\} - E\{J_{t+1}^{\tilde{\pi}}(x_1',x_2',0,0)\}\\ &\geq 0.\end{aligned}\tag{41}$$

The equality in (41) holds because π^* and $\tilde{\pi}$ are identical at t. The inequality in (41) holds because of the induction hypothesis given by (39). ■

Next we relate the performance of policies $\tilde{\pi}$ and π along any sample path.

Lemma 3: For each realization of arrivals and departures from time t on, we have

$$J_t^{\tilde{\pi}}(x_1,x_2,y_1,y_2) - J_t^{\pi}(x_1,x_2,y_1,y_2) = c_1\tau - c_2\sum_{i=1}^{N_2(\tau)} s_i \tag{42}$$

where $N_2(\tau)$ denotes the number of type-2 connections admitted to the auxiliary knapsack under policy π; and $s_i, i = 1, 2, \ldots, N_2(\tau)$ denotes the service time for each type-2 connection to the auxiliary knapsack under policy $\tilde{\pi}$.

Proof: By construction, the state of the system under policies π and $\tilde{\pi}$ is the same after time $t+\tau+\sigma$, and π and $\tilde{\pi}$ are identical after $t+\tau+\sigma$. Therefore, to compare the performance of π and $\tilde{\pi}$, we must compute the difference in revenue generated by π and $\tilde{\pi}$ in $[t, t+\tau+\sigma)$. To compute this difference, we compare the admission decisions of π and $\tilde{\pi}$ in the intervals $[t, t+\tau]$ and $(t+\tau, t+\tau+\sigma)$ separately.

By construction, policy $\tilde{\pi}$ admits the type-1 arrival at time t, and until time τ, the number of type-1 connections under $\tilde{\pi}$ is one more than the corresponding number under π. During the interval $[t, t+\tau]$, policy $\tilde{\pi}$ imitates policy π and admits new arrivals into the original knapsack as long as the system's state (under $\tilde{\pi}$) is not at the boundary (i.e., $(x_1, x_2) \notin B$, where (x_1, x_2) is the original knapsack's state under $\tilde{\pi}$). When the system state under $\tilde{\pi}$ is at the boundary and a new arrival is admitted by π, $\tilde{\pi}$ admits the same arrival and places it at the auxiliary knapsack. This is possible because of the specification of the auxiliary knapsack, given by (35)–(37).

By construction, between $t+\tau$ and $t+\tau+\sigma$, policy $\tilde{\pi}$ follows the emptying-greedy rule. Consequently, in $(t+\tau, t+\tau+\sigma)\tilde{\pi}$ imitates π in every admission decision. This is possible because the extra connections of type-2 admitted into the original knapsack during the interval $(t, t+\tau]$ under policy π are placed in the auxiliary knapsack under policy $\tilde{\pi}$.

Based on the above comparison of the admission decisions of π and $\tilde{\pi}$ in $[t, t+\tau+\sigma)$, the difference in revenue between π and $\tilde{\pi}$ can be computed by considering the connection j_1 and all type-2 connections admitted by policy $\tilde{\pi}$ into the auxiliary knapsack in the time interval $(t, t+\tau]$; we denote by $N_2(\tau)$ the number of these connections. Hence, the difference in performance between π and $\tilde{\pi}$ along any sample path is $c_1\tau - c_2\sum_{i=1}^{N_2(\tau)} s_i$, where s_i, $i = 1, 2, \ldots, N_2(\tau)$, is the service time of connection i of type-2 admitted into the auxiliary knapsack under policy $\tilde{\pi}$. ■

Based on Lemma 3 and (34), we obtain the following result.

Lemma 4: Policy $\tilde{\pi}$ outperforms policy π.

Proof: Use (42), fix τ, and calculate the difference between the conditional expectations of total generated revenues. We obtain

$$\begin{aligned}R(\tau) =&E\{J_t^{\tilde{\pi}}(x_1,x_2,y_1,y_2)|\tau\} - E\{J_t^{\pi}(x_1,x_2,y_1,y_2)|\tau\}\\ &=c_1\tau - c_2 E\{\sum_{i=1}^{N_2(\tau)} s_i|\tau\}\\ &\geq c_1\tau - c_2 E\{\sum_{i=1}^{a_2(\tau)} s_i|\tau\}\\ &=c_1\tau - \frac{c_2}{\mu_2}E\{a_2(\tau)|\tau\}\\ &=c_1\tau - \frac{c_2}{\mu_2}\lambda_2\tau\\ &\geq 0\end{aligned}\tag{43}$$

where $a_2(\tau)$ is the number of arrivals of type-2 in the time window $(t, t+\tau)$. The first inequality holds since the number of type-2 connections admitted in the time interval $(t, t+\tau)$ is bounded by the total number of arrivals of type-2 connections in that interval. Since $a_2(\tau)$ is independent of the service time of connections in the system, we use Wald's lemma to get the third inequality. The fourth equality is a consequence of the assumption on the rate of arrivals, and the last inequality holds because of (34).

Using the smoothing property of conditional expectation, we obtain

$$E\{J_t^{\tilde{\pi}}(x_1,x_2,y_1,y_2)\} - E\{J_t^{\pi}(x_1,x_2,y_1,y_2)\} \geq 0 \tag{44}$$

which proves that policy $\tilde{\pi}$ outperforms policy π. ■

Combining Lemmas 2 and 4, we obtain

$$E\{J_t^{\pi^*}(x_1,x_2,y_1,y_2)\} \geq E\{J_t^{\pi}(x_1,x_2,y_1,y_2)\}. \tag{45}$$

Consequently, the assertion of the theorem is true in Case 2.

Case 3: The arrival is of type 2 and policy π does not admit the new connection into the original knapsack. Then at time t^+, the knapsack state under policies π^* and π is $(x_1, x_2+1, 0, 0)$ and (x_1, x_2, y_1, y_2), respectively.

We construct policy $\hat{\pi}$ such that $\tilde{\pi}$ follows the greedy policy at time t and marks the accepted connection at this time slot as j_2; from time $t+1$ until stopping time $t+\tau'\hat{\pi}$ follows the (0, 1)-tracking-greedy policy, and from time $t+\tau'$ until stopping time $t+\tau+\sigma'$ policy $\hat{\pi}$ follows the emptying-greedy policy; and from time $t+\tau'+\sigma'$ on, policy $\hat{\pi}$ follows the greedy policy. The stopping times above are defined as follows:

τ' $\min\{\tau_1', \tau_2', T'\}$;

τ_1' occupation time of connection j_1;

$t+\tau_2'$ first time such that under policy $\hat{\pi}$ the auxiliary knapsack is empty of type-1 users, and a connection of type-2 arrives and cannot be admitted to the original knapsack;

$t+\tau+\sigma'$ first time after $t+\tau'$ that under policy $\tilde{\pi}$ the auxiliary knapsack is empty.

The following lemmas prove the induction step in this case. The proofs of Lemmas 5 and 6 are similar to those of Lemmas 2 and 3, respectively.

Lemma 5: Policy π^* outperforms policy $\hat{\pi}$.

Lemma 6: For each realization of arrivals and departures from time t on, we have

$$J_t^{\hat{\pi}}(x_1, x_2, y_1, y_2) - J_t^{\pi}(x_1, x_2, y_1, y_2) = c_2\tau - c_1 \sum_{i=1}^{N_1(\tau')} s_i' \tag{46}$$

where $N_1(\tau')$ denotes the number of type-1 connections admitted to the auxiliary knapsack under policy π; and $s_i', i = 1, 2, \ldots, N_1(\tau')$ denotes the service time for each type-1 connection to the auxiliary knapsack under policy $\tilde{\pi}$.

Lemma 7: Policy $\tilde{\pi}$ outperforms policy π.

Proof: Use (46), fix τ', and calculate the difference between the conditional expectation of total generated revenues. We have

$$\begin{aligned} R(\tau') =& E\{J_t^{\hat{\pi}}(x_1, x_2, y_1, y_2)|\tau'\} - E\{J_t^{\pi}(x_1, x_2, y_1, y_2)|\tau'\} \\ =& c_2\tau' - c_1 E\{\sum_{i=1}^{N_1(\tau')} s_i'|\tau'\} \\ \geq& c_2\tau' - c_1 E\{\sum_{i=1}^{a_1(\tau')} s_i'|\tau'\} \\ =& c_2\tau' - \frac{c_1}{\mu_1} E\{a_1(\tau')|\tau'\} \\ =& c_2\tau' - \frac{c_1}{\mu_1}\lambda_1\tau' \\ \geq& 0 \end{aligned} \tag{47}$$

where $a_1(\tau')$ is the number of arrivals of type-1 in $(t, t+\tau')$. The first inequality holds because the number of type-1 connections admitted in the time interval $(t, t+\tau')$ is bounded by the total number of the arrivals of type-1 connections in that interval. Since $a_1(\tau')$ is independent of service time of the connections in the system, we use Wald's lemma to get the third equality. The fourth equality follows directly from the assumption on the rate of arrivals, and the last inequality holds because of (34).

Using the smoothing property of the conditional expectations, we obtain from (47)

$$E\{J_t^{\hat{\pi}}(x_1, x_2, y_1, y_2)\} - E\{J_t^{\pi}(x_1, x_2, y_1, y_2)\} \geq 0 \tag{48}$$

which proves that policy $\hat{\pi}$ outperforms policy π. ■

From Lemmas 5 and 7

$$E\{J_t^{\pi^*}(x_1, x_2, y_1, y_2)\} \geq E\{J_t^{\pi}(x_1, x_2, y_1, y_2)\}. \tag{49}$$

Therefore, the assertion of the theorem is true in Case 3. Because of (40), (45), and (49), the proof of induction step is now complete. Hence, the proof of Theorem 3 is complete. ■

We use the result of Theorem 3 to prove Theorem 2.

Proof: [Theorem 2] Since the policy that the greedy rule for Problem (P') does not use any information provided by the auxiliary knapsack, it is equivalent to the greedy rule for Problem (P). Furthermore, the set of admissible policies for Problem (P) is a superset of the set of admissible policies of Problem (P). Therefore, because of Theorem 3, under (34), the greedy admission policy is optimal for Problem (P). ■

C. *Extensions, Generalization, and Future Work*

In this section, we address extensions of Problem (P). Furthermore, we discuss the sufficient condition for the optimality of the greedy policy given by (34), and its equivalent in the extensions of Problem (P).

1) Infinite Horizon: The result of Theorem 2 is valid for the infinite horizon version of Problem (P) with the criterion of average cost per unit time for the following reason: if a stationary policy is optimal for a finite horizon (T) problem with the total expected revenue, then it is also optimal for its infinite horizon counterpart with the corresponding average-cost-per-unit-time criterion.

2) Generalized Knapsack Problem L ($L > 2$) Classes of Connections: It is possible to establish, by arguments similar to those used in the proof of Theorems 2 and 3, the following result for the generalized knapsack problem with L types of connections.

Theorem 4: Consider Problem (P) with L types of users. If

$$c_l \geq \sum_{k \neq l} c_k \frac{\lambda_k}{\mu_k}, \qquad \forall l \in \{1, 2, \ldots, L\} \tag{50}$$

then the policy that follows the greedy rule at all times is optimal.

We note that as L increases, the sufficient conditions, described by (50), for optimality of the greedy admission policy become increasingly weak.

V. Conclusion

In this paper, we presented an approach to connection admission control for a single-hop multiservice wireless network with QoS requirements. In general, a connection admission control strategy creates a complicated two-way coupling between the physical layer, i.e., QoS, and the network layer, i.e., the optimal resource allocation. Our approach proposes a decomposition of the problem into two subproblems: admission region construction and generalized knapsack scheduling. The result of such decomposition is reducing the interaction of the two layers into a one-way coupling between the physical layer (QoS) and the network layer (CAC). To demonstrate the methodology, we then constructed an outage-based admission region. Simultaneous consideration of QoS requirements such as outage probability, average bit error rate, delay, etc., can be incorporated into the admission control problem by taking the intersection of the corresponding admission regions resulting from the above QoS requirements. Such an intersection defines the admission region for a generalized knapsack problem. We investigated a generalized knapsack problem and established conditions sufficient to guarantee the optimality of the greedy admission policy.

REFERENCES

[1] S. Oh and K. M. Wasserman, "Dynamic spreading gain control in multiservice CDMA networks," *IEEE J. Select. Areas Commun.*, vol. 17, pp. 918–927, May 1999.

[2] Z. Zhang, I. Habib, and T. Saadawi, "Bandwidth reservation for multimedia traffic over micro-cellular networks," in *Proc. IEEE 6th Int. Conf. Universal Personal Communication (ICPC)*, vol. 2, 1997, pp. 761–765.

[3] D. Mitra, M. I. Reiman, and J. Wang, "Robust dynamic admission control for unified cell and call QoS in statistical multiplexers," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 692–707, June 1998.

[4] D. Tse and S. Hanly, "Effective bandwidths in wireless networks with multiuser receivers," in *Proc. 17th Annu. IEEE Conf. Computer Communications (INFOCOM)*, vol. 1, 1998, p. 3542.

[5] D. Tse and S. V. Hanly, "Linear multiuser receivers: Effective interference, effective bandwidth and user capacity," *IEEE Trans. Inform. Theory*, vol. 45, pp. 641–657, Mar. 1999.

[6] F. K. L. Lee and M. Hamdi, "Providing deterministic quality of service guarantee in a wireless environment," in *Proc. IEEE 49th Vehicular Technology Conf. Moving into a New Millenium*, vol. 2, 1999, pp. 1727–1731.

[7] J. Lin, W. Kao, Y. T. Su, and T. Lee, "Outage and coverage consideration for micro-cellular mobile radio systems in a shadowed-Rician/shadowed-Nakagami environment," *IEEE Trans. Veh. Technol.*, vol. 48, pp. 66–75, Jan. 1999.

[8] N. B. Mandayam, P. Chen, and J. M. Holtzman, "Minimum duration outage for CDMA cellular systems: A level crossing analysis," *Wireless Personal Commun.*, no. 7, pp. 135–146, 1998.

[9] M. Zorzi, "Outage and error events in bursty channels," *IEEE Trans. Commun.*, vol. 46, pp. 349–356, Mar. 1998.

[10] H. S. Wang and N. Moayeri, "Finite-state Markov chain: A useful model for radio communication channels," *IEEE Trans. Veh. Technol.*, vol. 44, pp. 163–171, Feb. 1995.

[11] T. S. Rappaport, *Wireless Communications: Principle & Practice*. Englewood Cliffs, NJ: Prentice-Hall, 1996.

[12] R. O. Onvural, *Assynchronous Transfer Mode Networks: Performance Issues*. Norwood, MA: Artech House, 1994.

[13] T. Javidi and D. Teneketzis, "Outage, QoS, and admission region in a single cell," Control Group, EECS Dept., Univ. Michigan, Ann Arbor, MI, 2001.

[14] J. Sullivan and A. Mendelson, "Personal communication services: Bringing new quality and clarity to the enterprise," Phillips, InfoTech: PCS Rep. 1, 1997.

[15] S. Asmussen, *Applied Probability and Queues*, New York ed: Wiley, 1987.

[16] J. S. Evans and D. Everitt, "Effective bandwidth-based admission control for multiservice CDMA cellular networks," *IEEE Trans. Veh. Technol.*, vol. 48, pp. 36–46, Jan. 1999.

[17] K. W. Ross, *Multiservice Loss Models for Broadband Telecommunication Networks*. Berlin, Germany: Springer, 1995.

[18] C. Barnhart, J. Wieselthier, and A. Ephremides, "Admission-control policies for multihop wireless networks," *Wireless Networks*, vol. 1, pp. 373–387, 1995.

[19] G. J. Foschini, B. Gopinath, and J. F. Hayes, "Optimum allocation of servers to two types of competing costumers," *IEEE Trans. Commun.*, vol. COM-29, pp. 1051–1055, July 1981.

[20] A. Gavious and Z. Rosberg, "A restricted complete sharing policy for a stochastic knapsack problem in B-ISDN," *IEEE Trans. Commun.*, vol. 41, pp. 2375–2379, July 1994.

[21] S. Jordan and P. P. Varaiya, "Control of multiple service, multiple resource communication networks," *IEEE Trans. Commun.*, vol. 42, pp. 2979–2988, Nov. 1994.

[22] T. E. Lee and G. T. Oh, "The assymptotic value-to-capacity ratio for the multi-class stochastic knapsack problem," *Eur. J. Oper. Res.*, vol. 103, pp. 584–594, 1997.

[23] S. A. Lippman, "Applying a new device in the optimization of exponential queueing systems," *Oper. Res.*, vol. 23, no. 4, pp. 687–710, July/Aug. 1975.

[24] S. Martello and P. Toth, *Knapsack Problems*. New York: Wiley, 1990.

[25] B. L. Miller, "A queueing reward system with several costumer classes," *Manage. Sci.*, vol. 16, no. 3, pp. 234–245, Nov. 1969.

[26] K. W. Ross and D. H. K. Tsang, "Optimal circuit access policies in an ISDN environment: A Markov decision approach," *IEEE Trans. Commun.*, vol. 37, pp. 934–939, Sept. 1989.

[27] ——, "The stochastic knapsack problem," *IEEE Trans. Commun.*, vol. 37, pp. 740–747, July 1989.

[28] R. Van Slyke and Y. Young, "Finite horizon stochastic knapsack with applications to yield management," *Oper. Res.*, vol. 48, no. 1, pp. 155–171, Jan./Feb. 2000.

Tara Javidi (M'02) studied electrical engineering at the Sharif University of Technology, Iran. She received the M.S. degrees in electrical engineering systems and in applied mathematics stochastics from the University of Michigan, Ann Arbor, in 1998 and 1999, respectively.

She is currently an Assistant Professor at the Electrical Engineering Department, University of Washington, Seattle. Her research interests are in communication networks, stochastic resource allocation, and wireless communication.

Demosthenis Teneketzis (F'00) received the diploma in Electrical Engineering from the University of Patras, Patras, Greece, and the M.S., E.E., and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge.

He is a Professor of electrical engineering and computer science at the University of Michigan, Ann Arbor. In winter and spring 1992, he was a Visiting Professor at the Swiss Federal Institute of Technology, (ETH), Zürich. Prior to joining the University of Michigan, he was with Systems Control Inc., Palo Alto, CA, and Alphatech Inc., Burlington, MA. His research interests are in stochastic control, decentralized systems, queueing and communication networks, stochastic scheduling and resource allocation problems, mathematical economics, and discrete event systems.