A Framework for Design and Evaluation of Admission Control Algorithms in Multi-Service Mobile Networks

Rahul Jain and Edward W. Knightly
Department of Electrical and Computer Engineering
Rice University

Abstract— **Supporting Quality of Service (QoS) guarantees in wireless networks requires that admission control algorithms incorporate user mobility, and limit the probability that sufficient resources are unavailable when a user must handoff. In this paper, we develop a framework for designing admission control algorithms in wireless networks that support guaranteed QoS. First, we devise a taxonomy to explore the mathematical structure and practical design tradeoffs encountered in developing admission control algorithms. We next introduce the Perfect Knowledge Admission Control Algorithm, which, while unrealizable in practice, serves as a benchmark for evaluating admission control algorithms by using future knowledge of handoff events to exactly control the admissible region. Finally, we perform an extensive set of simulations (including trace-driven simulations) and, applying the Perfect Knowledge Algorithm, we study several admission control algorithms from the literature, identify a number of key system parameters for algorithm design, and quantify the fundamental tradeoffs in complexity and accuracy as revealed by the taxonomy.**

I. Introduction

Resource allocation is an important component for future packet networks to support multimedia applications with Quality of Service (QoS) requirements. In wireless networks that support user mobility, client requirements are not limited to QoS parameters such as packet loss probability and minimum throughput, as users may also experience performance degradation due to properties of the wireless channel (e.g., from physical-layer channel errors) and due to user mobility from handoffs. While sophisticated service disciplines [9], [11] and medium access techniques [2], [4] may be used to mask the former problem, admission control must be used to address the latter problem [1], [3], [5], [7], [8], [10], [12], [13], [15], [17]. In particular, since a mobile user may handoff to neighboring cells during the lifetime of its call, network resources must be reserved even in cells other than the one in which the user was admitted. Otherwise, if sufficient resources are not available at a new cell when the mobile user must hand off, the call must either suffer prolonged periods of significantly reduced QoS or be dropped all together. Thus, we consider a key QoS metric provisioned via admission control to be P_{drop}, the probability that insufficient resources are available for handing off a mobile user. Moreover, we consider the critical factor for evaluating an admission control algorithm's effectiveness to be its ability to maximize resource utilization (and minimize call blocking) subject to the user's P_{drop} constraint.

This work was supported by NSF CAREER Award ANI-9733610, NSF Grant ANI-9730104, and the Texas Advanced Technology Program. The authors may be reached at http://www.ece.rice.edu/networks.

In this paper, we develop a new framework for designing and evaluating resource allocation and admission control algorithms in wireless and mobile networks that support quality of service, and in particular, algorithms that provision resources to control P_{drop}. Our contributions are three-fold.

First, we formulate a taxonomy of admission control algorithms for mobile multi-service networks which reveals both the structure and the fundamental design choices encountered in designing such algorithms. In particular, we first classify admission control algorithms according to whether they allocate resources via a *cell-occupancy* approach or a *spatial mobility* approach. In the former case, a cell's occupancy statistics are controlled with use of a flow model which characterizes the behavior of call arrivals, departures, and handoffs into and out of a cell, irrespective of a mobile user's previous or future locations. In contrast, the latter approach allocates network resources by exploiting the interdependence of cell-to-cell occupancies, i.e., the spatial mobility of a user or group of users. We next classify algorithms depending on whether or not they model user locations and mobility in a *spatially uniform* manner across the network's cells, thereby incorporating heterogeneity among cells. Finally, we distinguish between algorithms that control traffic on a per-user basis and those that manage resources on an aggregate per-cell basis. Using these dimensions, we describe a number of algorithms from the literature within the context of this taxonomy and illustrate several key tradeoffs in the design of admission control algorithms in terms of granularity of resource control, mathematical tractability, and efficacy of accurately controlling the admissible region while also provisioning the desired quality of service.

Second, we introduce the *Perfect Knowledge Algorithm* (PKA) to serve as a benchmark for performance evaluation of resource reservation algorithms in mobile networks. We show that PKA, while unrealizable for on-line admission control, serves its benchmarking purpose by achieving the maximal QoS-constrained admissible region of any algorithm that has complete knowledge of the future mobility behavior of all established (admitted) calls as well as the new call requesting admission, but as would be the case with an on-line algorithm,

without knowledge of calls that may request admission in the future. By comparing the admissible regions and QoS parameters obtained using a given admission control algorithm with those obtained by PKA, we can assess the error introduced by an admission control algorithm's approximations used for analytical tractability, and simplifications used to limit complexity and communication overhead.

Finally, we perform an extensive set of simulation and admission control experiments using a two-dimensional 64-cell network. Using implementations of several admission control algorithms and a suite of mobility models including the traces of [16], we use the Perfect Knowledge Algorithm to quantify the impact of the taxonomy's design tradeoffs in practical scenarios. Moreover, we use this framework to explore the importance of several design issues for admission control algorithms such as the mobility model, mobility speed, heterogeneity of bandwidth demands, and call arrival rate. Our results yield insights not only to the key issues for designing admission control algorithms, but also illustrate areas where further study is needed. For example, we find that while algorithms from the literature are successful in limiting network access to satisfy mobility QoS constraints, they can be quite conservative in certain environments such as high spatial correlation of user locations (e.g., as in a "downtown" mobility model) and low probability of handoff drop.

II. Taxonomy of Admission Control Algorithms

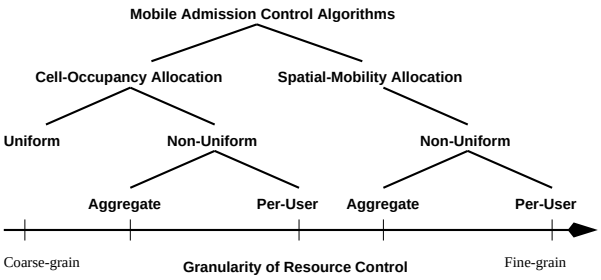


Fig. 1. Taxonomy of Admission Control Algorithms

In this section, we introduce a taxonomy of admission control algorithms for mobile and wireless cellular networks. We classify algorithms along three dimensions which, while having a simple mathematical representation, have significant implications regarding an algorithm's computational complexity, analytical tractability, and accuracy in properly controlling the admissible region. The taxonomy is illustrated in Figure 1 and as described in the Introduction leads to three design choices: (1) cell occupancy vs. spatial mobility, (2) spatially uniform vs. non-uniform, and (3) aggregate vs. per-user control.

Below, we show how this taxonomy results in five classes of admission control algorithms for mobile networks. Throughout the discussion, we denote $C(x, j, k, t)$ as the *cumulative* capacity handed off from cell j to cell k in the interval $[0, t]$ from user x. For simplicity of notation, denote $C(x, 0, k, t)$ as the capacity in cell k used by user x which originated its call in cell k. Hence, if user x's call request originates in a cell other than k then $C(x, 0, k, t) = 0$ for all t. Similarly, denote the cumulative capacity of calls that depart from the network from cell j by time t as $C(x, j, 0, t)$.

A. Spatially Uniform Occupancy Allocation

An admission control algorithm which we classify as "spatially uniform occupancy allocation" may be described as follows. First, under a spatially uniform model, users are equally likely to be located in any cell of the mobile network (or of a cell cluster). In other words, denoting the location of the mobile user x at time t by $L(x, t)$, and denoting the number of cells in a cluster by M,

$$P(L(x,t) = j) = \frac{1}{M} \tag{1}$$

for all users x and all cells $j = 1, \cdots, M$ in the cluster. Notice that with such uniformity of location, aggregate and per-user approaches are equivalent as indicated in Figure 1.

Mathematically, resource allocation algorithms in this class are concerned with a cell's occupancy behavior. We denote $\Omega(t)$ as a typical cell's occupancy at time t (expressed in bandwidth units for example), which may be expressed in terms of the aggregate flow model of Figure 2 as

$$\Omega(t) = \sum_{x} \sum_{k \in H_{in}} C(x, k, \cdot, t) - \sum_{x} \sum_{k \in H_{out}} C(x, \cdot, k, t) \tag{2}$$

where H_{in} denotes the set of cells with users handing off into the cell (including new calls) and H_{out} denotes the set of cells from users flowing out, which includes call departures.

As an example, if each of M cells has the capacity to support C users, each requiring capacity 1, then the probability that a cell is in an overload state (and hence hand-offs are being dropped) when N users are active in the system is given by

$$P(\Omega(t) \geq C) = \sum_{n=C}^{N} \binom{N}{n} \left(\frac{1}{M}\right)^{n} \left(1 - \frac{1}{M}\right)^{N-n} \tag{3}$$

A key advantage of a spatially uniform occupancy approach as taken in [1], [12] for example, is that it provides a framework for overcoming state-space explosion problems encountered in a multi-dimensional Markov model of a mobile network, as both per-user mobility and cell-to-cell interactions are not explicitly modeled. Moreover, using the flow model as in Figure 2, this framework can be applied to problems such as estimating the mean overflow duration, i.e., the mean time that $\Omega(t) \geq C$ (as in [1]), or determining an optimal capacity to set aside for guard channels [12], [13].

Roughly, an admission control algorithm's granularity of resource control is affected by the available information (whether measured, communicated by other cells, specified by users, etc.) so that the order of the algorithmic complexity can be expected to increase with the amount of information processed. The spatially uniform occupancy allocation approach can be characterized as in Equation (3), with a probability distribution vector of size C. Since this is the same for all cells, we can say that the information granularity of resource control is of order C.

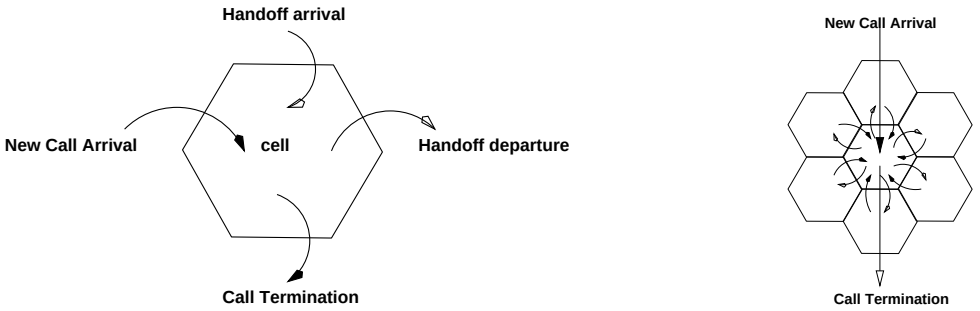


Fig. 2. Cell Occupancy vs. User Mobility Model

Hence, spatially uniform occupancy allocation is both computationally simple for on-line resource allocation as well as analytically tractable by treating mobile users on an aggregate basis and assuming spatially uniform user occupancy. In Section IV, we consider the impact of such a coarse-grained approach on an admission control algorithm's accuracy.

B. Spatially Non-Uniform Aggregate Occupancy Allocation

An admission control algorithm which allocates resources in a spatially non-uniform manner according to the aggregate behavior of the mobile users differs from the aforementioned approach in that different cells may have different occupancy characteristics. In particular, the occupancy of cell j at time t is given by

$$\Omega(j,t) = \sum_{x} \sum_{k \in H_{in,j}} C(x,k,j,t) - \sum_{x} \sum_{k \in H_{out,j}} C(x,j,k,t) \tag{4}$$

While cells need not have the same occupancy statistics, in this class of admission control algorithms, cell occupancies of neighboring cells are assumed to be uncorrelated. In particular, in a cell-occupancy approach, the spatial covariance function is approximated as zero such that

$$\psi_t(j,k) = E\Omega(j,t)\Omega(k,t) - E\Omega(j,t)E\Omega(k,t) \approx 0, j \neq k. \tag{5}$$

Here, we make the distinction between "occupancy" approaches versus "mobility" approaches according to whether the algorithm incorporates this *correlation* among occupancies in different cells, i.e., whether $\psi_t(j,k) \approx 0$, for $j \neq k$.

Thus, algorithms in this class have more fine grain resource control than in the uniform case, and can explicitly address the issue of spatial "hot spots", such as that which might arise in a cellular network with a downtown, for example. This class of algorithms therefore increases the order of the size of the information content to $M \cdot C$ as each cell maintains its own probability distribution vector for the occupancy. An example of an admission control algorithm in this class is found in [3] where a decoupling-of-states approach is introduced to address the state-space explosion problems incurred in Markov modeling of cellular networks. In Section IV, we further address the issue of spatial non-uniformity of user locality using a set of simulation experiments that include mobile users with non-uniformly distributed destinations.

C. Spatially Non-Uniform Per-User Occupancy Allocation

In contrast to the classes above, a spatially non-uniform per-user occupancy allocation scheme controls network resources according to individual user's occupancy characteristics.

Thus, in this class of admission control algorithms, each user has an associated steady-state per-cell occupancy density function given by

$$f(x,j,c) = \lim_{T \to \infty} \frac{\sum_{t=0}^{T} \left(\sum_{k \in H_{in}^{j}} C(x,k,j,t) - \sum_{k \in H_{out}^{j}} C(x,j,k,t) \right)}{\sum_{i=1}^{M} \sum_{t=0}^{T} \left(\sum_{k \in H_{in}^{i}} C(x,k,i,t) - \sum_{k \in H_{out}^{i}} C(x,i,k,t) \right)} \tag{6}$$

By incorporating each user's occupancy densities $f(x,j,c)$ into an admission control algorithm, resources may potentially be more accurately allocated than in an aggregate approach. The spatially non-uniform per-user occupancy approach associates a per-cell occupancy density with each user as in Equation (6). In particular, the size of information content in this case would be of order $N \cdot M$ for N active users in an M-cell network. However, such an approach also raises the question of how to obtain each user's $f(x,j,c)$ in practice as well as how to determine P_{drop} from these distributions.

In [17] an algorithm in this class is proposed in which this issue is addressed by having a user specify the set of cells that it expects to occupy during its connection lifetime as part of its traffic specification. In other words, at connection set-up time, users requiring QoS support must always specify their desired QoS parameters as well as their future bandwidth demands usually in the form of a maximum average rate and a maximum burst size. In [17], this specification is augmented with a set of cells Λ_x that user x will occupy such that if user x requires capacity c, the occupancy distributions are given by

$$F(x,j,r) = \begin{cases} 1(r > 0) & j \notin \Lambda_x \\ 1(r > c) & j \in \Lambda_x \end{cases} \tag{7}$$

where $1(\cdot)$ is an indicator function. Hence, with the use of *a priori* user mobility profiles, [17] performs admission control using spatially non-uniform per-user occupancy allocation.

D. Aggregate Mobility Allocation

The following two classes of admission control algorithms use *mobility allocation* and differ from the occupancy allocation schemes in that the temporal and spatial correlation of a user's locations or a group of users' locations over time is explicitly addressed. We refer to algorithms which account for such cell-to-cell interactions as employing "mobility allocation" such that the spatial occupancy covariance may be non zero, i.e., $\psi_t(j,k) \neq 0$ for some t, j, and k. More generally, a characterization of aggregate mobility behavior can be described by the distribution G as

$$P(\sum_{x \in j}\{C(x,j,k,t+s) - C(x,j,k,t)\} < C \mid \Omega(l,t), \\ l = 1, \cdots, M) = G(j,k,s,c) \quad (8)$$

so that G is the distribution of the capacity of all handoffs from cell j to k over intervals of length s given the network's current occupancy. Hence, an admission control algorithm in the aggregate mobility class requires a model or measurement scheme for determining $G(j,k,s,c)$ and a mechanism for estimating P_{drop} from the G's.

While this occupancy vs. mobility distinction is quite simple conceptually, it has significant implications in the design of admission control algorithms: for example, explicit modeling of cell-to-cell interactions using a multi-dimensional Markov chain results in a state space explosion [3]. Approaches to address this issue include [14], in which the asymptotic regions of very slow and very fast mobility are considered.

Thus, this class of admission control algorithms is the first under our taxonomy to explicitly address cell-to-cell mobility. Consequently, such algorithms must address how to model and characterize the cell-to-cell correlation via the $G(j,k,s,c)$ distribution, and require an increased granularity of information of $M^2 \cdot C \cdot T$, where T is the average call holding time in an M-cell network, expressed in time slots.

E. Per-User Mobility Allocation

As the name implies, per-user mobility allocation differs from the scheme above in that each user has a mobility profile H given by

$$P\left(C(x,j,k,t+s) - C(x,j,k,t) < C \mid L(x,t)\right) \\ = H(x,j,k,s,c) \quad (9)$$

which describes the distributions of the user's future locations and handoffs given its current location (or more generally, its past locations as well).

An admission control algorithm using such per-user, spatially non-uniform mobility characterizations employs the most fine grain resource control of the classes within the taxonomy, and has information content of order NM^2T. Such fine-grain control has potential benefits to an algorithm's accuracy, but encounters formidable problems of computational complexity and analytical tractability in specifying and manipulating such detailed user profiles, and in computing the relevant QoS parameters given these profiles.

We classify the "shadow cluster" approach of [7] as employing per-user mobility allocation. In [7], the complexity issues encountered in this class of admission control algorithms are addressed with a measurement-based approach. In particular, each user is characterized by "active mobile probabilities" which are adaptive and measurement-based versions of $H(x,j,k,s,c)$ so that a new mobile user is admitted only if a set of tests on H indicate that a new mobile call has sufficient probability of surviving for the duration of its call without encountering a hand-off drop. While state-space explosion issues are largely avoided with this approach, additional computational and communication overheads are encountered, as neighboring base-stations must dynamically adjust and communicate the active mobile probabilities of each user.

F. Discussion

In summary, with design decisions trading coarse to fine grain resource control for algorithmic scalability and simplicity, the above taxonomy can be viewed largely in terms of aggregation and heterogeneity. In particular, the above taxonomy distinguishes among algorithms according to their aggregation properties, namely (1) whether handoffs into a cell from neighboring cell are aggregated into an occupancy model (irrespective of the handing-off cell) or treated individually according to a mobility model which accounts for their previous locations; and (2) whether users (handing off or occupying a cell) are treated individually or on an aggregate per cell basis. Second heterogeneity also plays a key role in that the taxonomy distinguishes among algorithms according to whether they exploit spatial heterogeneity (as in a non-uniform approach) and whether they exploit user heterogeneity (as in a per-user approach).

In general, with a more fine grain approach, state-space explosion and analytical complexity issues become more pressing, and we have described a number of techniques that have been proposed to address these issues including decoupling of states, mobility traffic descriptors, asymptotic approximations, and measurement-based approaches. In Figure 1's illustration of the taxonomy, algorithm granularity is increasing from left to right as

$$C < MC < NM < M^2CT < NM^2T$$

in an M-cell network for N active users with mean call holding time T, cell capacity C, and assuming $N \sim MC$. To illustrate this granularity tradeoff using a scenario from Section IV with $C \sim 40, M \sim 64, N \sim 2560, T \sim 10$, we have

$$40 < 2,560 < 163,840 < 1,638,400 < 104,857,600$$

While aggregation simplifies resource management tasks, it may introduce *costs* in terms of accuracy, i.e., the ability of an algorithm to maximally utilize resources while also satisfying user's QoS constraints. We attempt to quantify such costs using the "Perfect Knowledge Algorithm" which we develop next.

III. PERFECT-KNOWLEDGE ALGORITHM

In this section, we introduce the *Perfect-Knowledge Algorithm* (PKA) to serve as a performance benchmark for admission control algorithms in mobile multi-service networks. In particular, one may view an admission control algorithm as making a sequence of admission decisions upon the arrival of each call request, and, as described in Section II, admitting or rejecting each call according to some resource reservation scheme. In evaluating the performance of a particular admission control algorithm, one must assess its effectiveness in making correct admission decisions, that is, whether the algorithm properly limited the handoff dropping probability to below the target P_{drop} and whether it did so while maximally utilizing network resources, admitting as many calls as possible subject to the QoS constraint.

Towards this end, we devise PKA, which, while unrealizable in practice, serves its benchmarking purpose by exploiting knowledge of future handoff events to assure that the maximal admissible region is obtained while satisfying the P_{drop} constraint. Thus, we can evaluate the performance and effectiveness of a practical on-line admission control algorithm by comparing utilization and QoS values obtained by a certain algorithm with those obtained using the idealized PKA.

The Perfect Knowledge Algorithm operates under the following three assumptions when performing an admission control test for a call arriving at time t: (A1) Characteristics of calls arriving at times $u > t$ are *not* known; (A2) If the call is deemed admissible at time t, it must be admitted and the decision cannot be reversed; (A3) Future handoffs of the new call and all established calls *are* known. Assumption (A1) and (A2) make PKA analogous to an on-line admission control algorithm since in both cases, future call requests are not known. With assumption (A3), PKA differs from on-line algorithms in that (A3) allows PKA to obtain an idealized admissible region using knowledge of future handoff events.

The goal of PKA is to maximize a cellular network's average utilization, U, while satisfying the empirical QoS constraint $\widehat{P}_{drop}$. Specifically, U is defined as the fraction of available capacity used over time, averaged over all cells

$$U = \frac{\sum_{j=1}^{M}\sum_{t=1}^{T} C_u(j,t)}{T\sum_{j=1}^{M} C(j)} \tag{10}$$

where $C_u(j,t)$ denotes the capacity utilized in cell j at time t and $C(j)$ is the available capacity in cell j. Notice that $C_u(j,t)$ is determined by the set of users which have been admitted to the network. The empirical dropping probability is defined as the fraction of handoffs dropped through time t, and is given by

$$\widehat{P}_{drop}(t) = \frac{N_{drops}(t)}{N_{handoffs}(t)} \tag{11}$$

where $N_{drops}(t)$ is the number of failed handoff attempts, and $N_{handoffs}(t)$ is the total number of handoff attempts by time t.

PKA, presented in pseudo-code in Figure 3, is invoked when a new user x requests admission to the cellular network at time t_0. The user has QoS requirement P_{drop} and mobility profile $P[x]$ which specifies user x's hand-off pattern and cell residence times through the lifetime of the call. As indicated in Line 1, PKA first checks that sufficient spare capacity is available in the cell of call initiation at t_0, and if not, user x is blocked (Line 2). Otherwise, subsequent time slots up to the call termination time are tested for overload (Lines 6-21). If at any time $t > t_0$ in which user x makes a handoff attempt sufficient capacity is not available (Lines 10-11), user x's call will be dropped (Line 12) and *new_drops* is incremented by one (Line 13).

If admitted, user x may induce handoff drops on other users in addition to possibly being dropped itself. Such a scenario is considered in Lines 16-19 and all such drops are tallied in Line 18. Note that *new_drops* and *new_handoffs* may not only be incremented when user x induces a drop, but also *decremented* when user x's admission prevents a drop as a consequence of an earlier induced drop on another user. In Line 22, the empirical

```
Procedure PKA (user x, profile P, time t0, QoS P_drop) {
1.  if (Cu[P[x].cell[t0]][t0]+P[x].bw > C[P[x].cell[t0]]) {
2.    BLOCK(x);
3.    return;
4.  }
5.  new_drops=0; new_handoffs=0;
6.  for (n=0; n ≤ P[x].numhandoffs; n++) {
7.    new_handoffs++;
8.    for (t=P[x].handofftime[n];
            t < P[x].handofftime[n]+P[x].tresid; t++) {
9.      zone=P[x].cell[t];
10.     if (Cu[zone][t]+P[x].bw > C[zone])
11.       if (t==P[x].handofftime[n]) {
12.         DROP(x);
13.         new_drops++;
14.         break;      /*Goto line 23*/
15.       }
16.       else if (Num_Handoffs[P[x].cell[t]][t] > 0) {
17.         DROP(HandOffUser[zone][t]
                [Num_HandOff[zone][t]]);
18.         Update(new_drops,new_handoffs);
19.       }
20.   }
21. }
22. frac_dropped = (total_drops + new_drops)/
          (total_handoffs + new_handoffs);
23. Compute(BLoss,BGain);
24. if ((frac_dropped ≤ P_drop) AND (BLoss < BGain)) {
25.   ADMIT(x);
26.   total_drops += new_drops;
27.   total_handoffs += new_handoffs;
28.   Update(Cu);
29. }
30. else BLOCK(x);
31. return;
}
```

Fig. 3. Perfect Knowledge Algorithm

drop probability, denoted *frac_dropped*, is computed as the fraction of handoff requests dropped (possibly including x itself) if user x is admitted. Next, in Line 23 PKA computes *BGain*, the gain in bandwidth utilization if user x is admitted, and *BLoss*, the loss in bandwidth utilization due to drops induced by admitting user x. Note that admitting an additional user does not necessarily increase average utilization, as $C_u(j,t)$ may be adversely affected from induced handoff drops. Finally, user x is admitted (Line 25) if *frac_dropped* will remain within the QoS requirement P_{drop}, and if there is a net gain in utilization from admitting the user, i.e., *BGain* $>$ *BLoss* (Line 24). Otherwise, user x is blocked (Line 30).

Result: PKA maximizes average utilization U subject to assumptions (A1)-(A3), while satisfying

$$\widehat{P}_{drop}(t) \leq P_{drop} \; \forall \; t.$$

Proof: Let t_0 denote the time that user x initiates a call, t_h the time that it hands off for the hth time, and b_x its bandwidth demand. Let t'_y denote the time that user x comes in conflict with user y, and $t''_y \geq t'_y$ denote the time at which user y's call terminates or is dropped, which ever occurs first. Denote the set of such conflicting users by $\mathcal{Y}$. It is possible that because of user y being dropped at time t'_y, a user z that would have been dropped at time t'_z, would be able to avoid being dropped up to time $t''_z > t'_z$. Denote the set of such users by $\mathcal{Z}$.

Let $H(x)$ denote the number of handoff attempts of user x during its call holding time, and H_0 the number of successful handoffs such that if user x is not dropped, H_0 is equal to $H(x)$. Let t_{H_0+1} denote the time when user x is dropped or when its call terminates. Then PKA admits user x if:

$$\text{(C1)} \;\; (t_{H_0+1} - t_0) \cdot b_x + \sum_{z \in \mathcal{Z}} (t''_z - t'_z) \cdot b_z > \sum_{y \in \mathcal{Y}} (t''_y - t'_y) \cdot b_y$$

$$\text{(C2)} \;\; \frac{N_{drops}(t) + new\_drops(t)}{N_{handoffs}(t) + new\_handoffs(t)} \leq P_{drop}, \forall t$$

Therefore, at time t_0, when an admission decision about user x is made, condition (C1) ensures that bandwidth utilization is maximized subject to the QoS constraint (C2). □

Thus, PKA can serve as a benchmark for evaluating the performance of on-line admission control algorithms by comparing their admissible regions with the maximal region obtained by the idealized Perfect-Knowledge Algorithm.

For a given set of admission control decisions, we define an on-line admission control algorithm's PKA *Error Index* as

$$\text{PEI} = \frac{U_{\text{AC}} - U_{\text{PKA}}}{U_{\text{PKA}}} \tag{12}$$

to reflect the utilization error of the on-line algorithm, i.e., PEI represents the difference in the admission decisions between PKA and the on-line algorithm scaled to utilization. In Section IV, we study a number of on-line admission control algorithms under a diverse set of mobility models using PKA and the PKA Error Index.

IV. Performance Study of Mobile Admission Control

In this section, we use an extensive set of simulation experiments to study admission control and resource allocation in mobile multi-service networks. We first illustrate the use of the Perfect-Knowledge Algorithm as a benchmark for the design and evaluation of admission control algorithms: we study a number of algorithms from the literature by exploring their ability to control the admissible region relative to PKA. We also use this study to illustrate key performance tradeoffs encountered relative to the algorithm taxonomy presented in Section II. Finally, we consider several system parameters such as the user mobility model, speed of the mobile units, and traffic heterogeneity. We study the impact of such parameters on both the idealized PKA as well as on the performance of on-line admission control algorithms. The results of our study provide insights into the admission control design considerations of greatest impact and point to aspects of admission control algorithms requiring further study.

Our simulation scenario as described below uses a two-dimensional cellular network with heterogeneous traffic sources and a diverse set of mobility models, including real-time mobility traces [16].

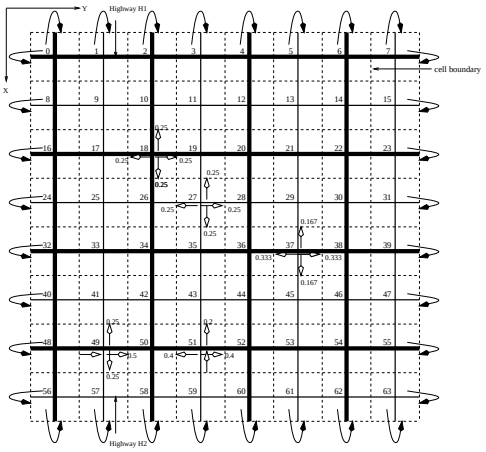
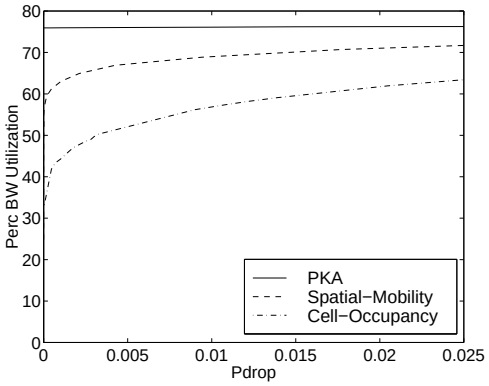


Fig. 4. Cellular topology

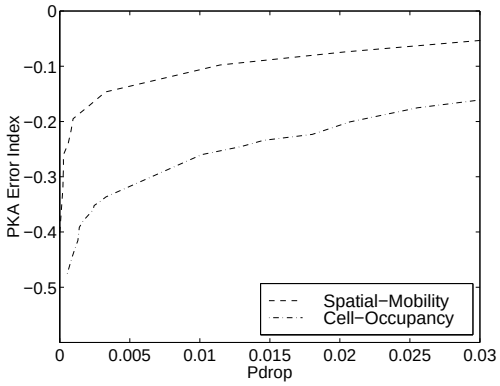
A. Network Topology and Traffic Model

Our simulated network consists of 64 cells located in a rectangular grid as depicted in Figure 4. Each cell has four neighbors, so that handoffs take place only between cells sharing an edge, and not just a vertex. The network wraps around such that, for example, a hand-off to the east of cell 63 wraps to cell 56. Each cell has capacity to support 40 Bandwidth Units (BUs).

Mobile calls originate from a uniformly random location at a Poisson rate of 1 call per minute per cell. Traffic demands consist of four classes requiring 1, 2, 4, and 8 BUs. The respective probability that a new call belongs to one of these classes is 0.5, 0.3, 0.1 and 0.1. Unless otherwise noted, time is slotted to 1 minute and calls have a geometrically distributed duration



(a) Average bandwidth utilization



(b) PKA Error Index

Fig. 5. Cell-occupancy vs. Spatial-mobility distribution

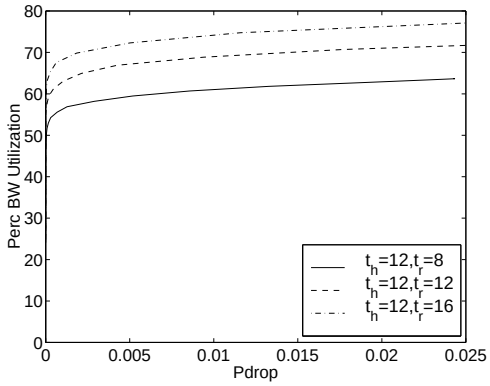


Fig. 6. Impact of mobility speed (or cell residence time)

with mean 12 minutes; cell residence time is geometrically distributed with mean 12 minutes. Simulations are performed for 6 hours of simulation time with a large number of call arrivals, departures, and handoff attempts. In each plot, 95% confidence intervals are within 2 bandwidth units or 5% utilization, and are not shown.

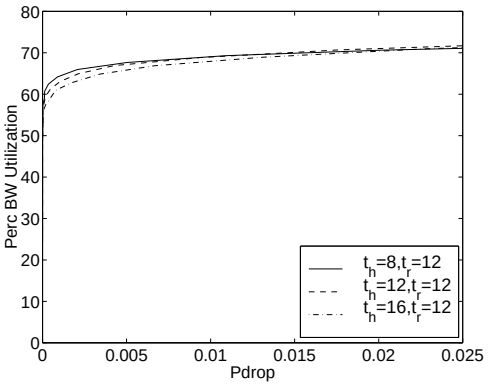
In order to study the impact of user mobility patterns on capacity allocation, we use the following five mobility models in our simulations: (1) *Multihop random*: at each handoff the user is equally likely to move in each of four possible directions. (2) *Hierarchical highway*: the cellular network has an overlay of "highways", or a set of mobility patterns that users are more likely to follow, with H1 indicating major highways and H2 indicating less popular roads. (3) *Destination model*: users move toward a uniformly random location chosen when the call is initiated; the probability of handing off in a particular direction is weighted according to the shortest path to the destination. (4) *Downtown model*: the four contiguous cells 0, 7, 56, and 63 are regarded as a downtown area, and mobile users are highly likely to have a destination within this area. (5) *Real-time mobility traces*: users move according to traces of the San Francisco Bay Area (voice) cellular network [16]. In this case, we use the actual Bay Area cellular network topology rather than that of Figure 4, and in the experiments, modify the traces to include calls with higher bandwidth demands than a voice call. Further details of these models may be found in [6].

B. Illustration of PKA Benchmark and Taxonomy

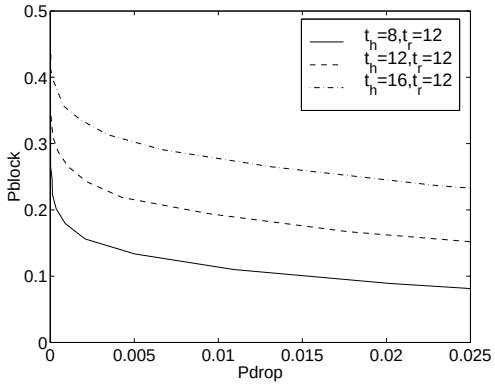
The goal of a call admission control algorithm is to utilize available resources as efficiently as possible such that the QoS demand is satisfied. To evaluate the performance of admission control algorithms, we compare their performance with that of PKA, which achieves a maximal admissible region while maintaining the required P_{drop}. The PKA Error Index (PEI), as proposed in Section III, can be used as a measure of accuracy of an admission control algorithm. We consider two admission control algorithms representative of two classes of our taxonomy: cell-occupancy ([1]), and spatial mobility ([7]). Figure 5(a) shows the average bandwidth utilization of the two algorithms, and also of PKA. Notice that the two online admission control algorithms, representing two fundamentally different classes of the taxonomy, are conservative, particularly for low P_{drop}. It may be noted that while PKA assumes knowledge of the complete profile of a user, the two online algorithms use only the mobility/occupancy distributions. The PEI vs. P_{drop} plot in Figure 5(b) suggests that for $P_{drop} < 0.01$, the two online algorithms are almost 10% to 25% more conservative as compared to the PKA. The spatial-mobility allocation converges to the PKA performance for $P_{drop} > 0.0025$, and outperforms the cell-occupancy allocation by 20% on the PKA Error Index, illustrating the potential accuracy gains of finer granularity of resource control.

C. Design Issues for Admission Control

Among the system parameters to be considered for designing mobile admission control algorithms are the mean call holding duration, the mean cell residence time (which is related to the speed of mobility of the users), the new call arrival rate, the mobility pattern (which determines the correlation in the occupancy levels of neighboring cells), and the heterogeneity of the users in demanding different classes of services requiring different bandwidths. Below, we use the simulation set-up described in Section IV-A, and, unless otherwise noted, we assume a destination

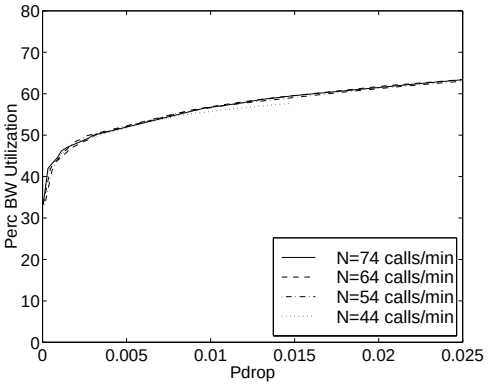


(a) Average bandwidth utilization

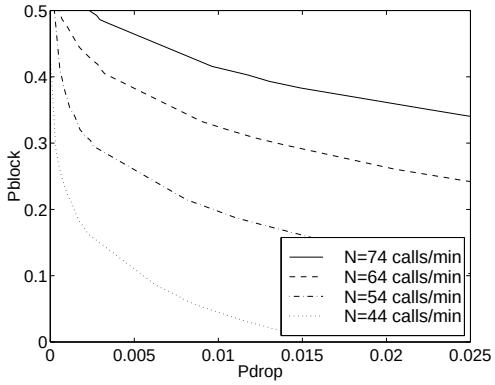


(b) P_{block}

Fig. 7. Impact of call holding time



(a) Average Bandwidth utilization



(b) P_{block}

Fig. 8. Impact of call arrival rate

model with geometrically distributed call holding duration and cell residence time, with a mean of 12 minutes for both.

To investigate the impact of the *mobility speed*, we generate several traces varying the mean cell residence time t_r from 8 to 16 minutes (keeping the mean call holding time t_h the same at 12 minutes). Figure 6 shows the average bandwidth utilization versus P_{drop} for the spatial mobility allocation. Observe that with increasing t_r, the utilization is higher for the same P_{drop}. Thus, for higher speed of mobile users, the utilization decreases.

We vary the mean *call holding time* t_h from 8 to 16 minutes (keeping the mean cell residence time t_r at 12 minutes). From Figure 7(a), observe that the average bandwidth utilization is almost the same for the range of the call holding time considered. Thus, it depends only on the mean cell residence time, and not the call holding time. However, Figure 7(b) indicates that with longer duration of the calls, the number of calls blocked increases without affecting the utilization, implying that the rate of handoff is far more important than the cumulative number of handoffs.

To investigate the impact of the *call arrival rate*, we perform simulations with cell-occupancy allocation by varying the call arrival rate from 0.68 calls/cell/minute to 1.16 calls/cell/minute. From Figure 8, observe that an increased call arrival rate increases the blocking probability, while the average bandwidth utilization remains nearly the same. Spatial mobility allocation and PKA also exhibit similar trends, indicating that for a new call arrival rate above a certain threshold, the average bandwidth utilization does not increase, while the blocking probability does.

In the experiments described in Figure 9, we investigate the impact of the *mobility pattern* using the traffic models described in Section IV-A. Figure 9 illustrates how the average bandwidth utilization depends on the mobility pattern. It shows that the average bandwidth utilization for the downtown model is almost 20% higher than for other models. The destination model mobility pattern also decreases utilization by 4% to 5% as compared to the highway and the random hop mobility models. Thus, we can conclude that the mobility pattern is an important issue in the design of admission control algorithms as it reflects the importance of the systems' spatial correlation structure.

To investigate the impact of the *heterogeneity* of the traffic, we vary the variance of the bandwidth demanded per call while keeping the mean bandwidth per call the same. From Figure 10, observe that with increased heterogeneity of the traffic, the average bandwidth utilization decreases significantly.

Finally, Figure 11 shows the results for trace-driven simulations using real-time mobility traces. We observe that both algorithms are again conservative for small P_{drop}, although the

spatial-mobility allocation algorithm converges to PKA's utilization. Moreover, the spatial mobility allocation outperforms the cell-occupancy allocation.

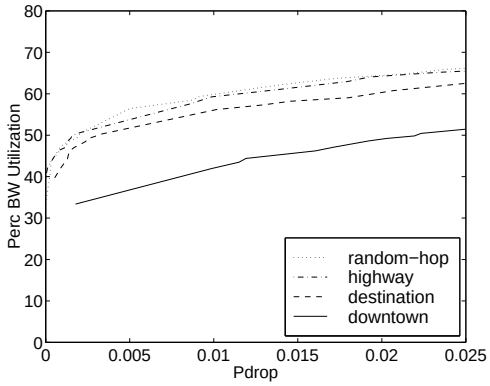


Fig. 9. Impact of the mobility pattern

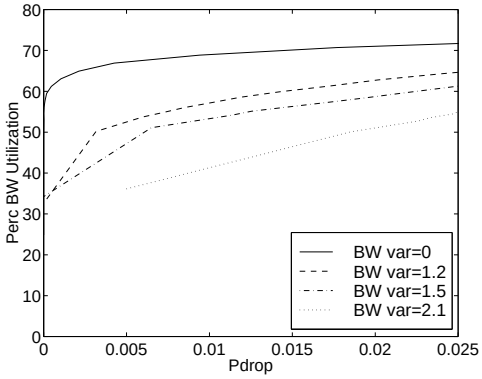


Fig. 10. Impact of heterogeneity

V. Conclusions

In this paper, we first devised a taxonomy of admission control algorithms to explore the structure and design issues encountered in algorithm design, and together with simulation experiments, we quantified the fundamental tradeoffs between an admission control algorithm's accuracy and granularity of resource control. We next designed a Perfect Knowledge Admission Control Algorithm and showed how it exactly controls the admissible region to serve as an ideal benchmark for evaluating practical on-line algorithms. Finally, we performed an extensive simulation study using a suite of mobility models and traces. We applied the taxonomy and PKA and explored a number of admission control design issues. We found for example, that algorithms from the literature can be quite conservative in certain environments such as high spatial correlation of user locations (such as in a "downtown" mobility model) and stringent QoS constraints on the probability of handoff drop. Our study thus serves as a framework for designing admission control algorithms that support guaranteed quality of service in wireless and mobile networks.

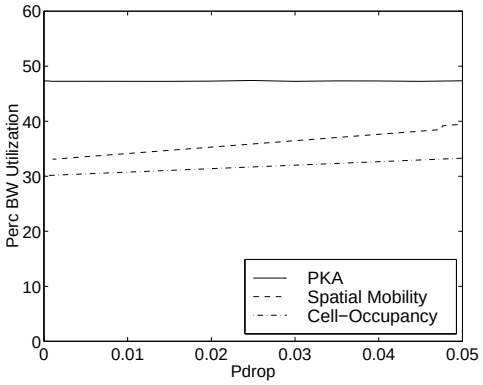


Fig. 11. Real-time mobility trace simulation

References

[1] A. Acampora and M. Naghshineh. An architecture and methodology for mobile-executed handoff in cellular ATM networks. *IEEE Journal on Selected Areas in Communications*, 12(8):1365–1375, October 1994.

[2] M. Arad and A. Leon-Garcia. A generalized processor sharing approach to time scheduling in hybrid CDMA/TDMA. In *Proceedings of IEEE INFOCOM '98*, San Francisco, CA, March 1998.

[3] C. Chao and W. Chen. Connection admission control for mobile multiple-class personal communications networks. *IEEE Journal on Selected Areas in Communications*, 15(8):1618–1626, 1997.

[4] S. Choi and K. Shin. A cellular wireless local area network with QoS guarantees for heterogeneous traffic. In *Proceedings of IEEE INFOCOM '97*, pages 1030–1037, Kobe, Japan, April 1997.

[5] S. Choi and K. Shin. Predictive and adaptive bandwidth reservation for hand-offs in QoS-sensitive cellular networks. In *Proceedings of ACM SIGCOMM'98*, Vancouver, British Columbia, August 1998.

[6] R. Jain and E. Knightly. Design and evaluation of admission control algorithms for mobile networks. Technical Report #9804, Rice University, Houston, TX, July 1998.

[7] D. Levine, I. Akyildiz, and M. Naghshineh. A resource estimation and call admission algorithm for wireless multimedia networks using the shadow cluster concept. *IEEE/ACM Transactions on Networking*, 5(1):1–12, February 1997.

[8] S. Lu and V. Bharghavan. Adaptive resource management for indoor mobile computing environments. In *Proceedings of ACM SIGCOMM'96*, Stanford, CA, August 1996.

[9] S. Lu, V. Bharghavan, and R. Srikant. Fair queueing in wireless packet networks. In *Proceedings of ACM SIGCOMM '97*, Cannes, France, September 1997.

[10] M. Naghshineh and M. Schwartz. Distributed call admission control in mobile/wireless networks. *IEEE Journal for Selected Areas in Communications*, 14(4):711–717, 1996.

[11] T. Ng, I. Stoica, and H. Zhang. Packet fair queueing algorithms for wireless networks with location-dependent errors. In *Proceedings of IEEE INFOCOM '98*, San Francisco, CA, March 1998.

[12] R. Ramjee, R. Nagarajan, and D. Towsley. On optimal call admission control in cellular networks. In *Proceedings of IEEE INFOCOM '96*, pages 43–50, San Francisco, CA, March 1996.

[13] M. Saquib and R. Yates. Optimal call admission to a mobile cellular network. In *IEEE Vehicular Technology Conference*, pages 190–194, Chicago, IL, July 1995.

[14] M. Sidi and D. Starobinski. New call blocking versus handoff blocking in cellular networks. *Wireless Networks*, 3(1):15–27, 1997.

[15] S. Singh. Quality of service guarantees in mobile computing. *Computer Communications*, 19(4):359–371, April 1996.

[16] Stanford University SUMATRA Group. Sumatra real-time mobiliy traces. *http://www-db.stanford.edu/sumatra/*, 1996.

[17] A. Talukdar, B. Badrinath, and A. Acharya. On accommodating mobile hosts in an integrated services packet network. In *Proceedings of IEEE INFOCOM '97*, Kobe, Japan, April 1997.