Predictive and Adaptive Bandwidth Reservation for Hand-Offs in QoS-Sensitive Cellular Networks *

Sunghyun Choi and Kang G. Shin

Real-Time Computing Laboratory

Department of Electrical Engineering and Computer Science

The University of Michigan

Ann Arbor, Michigan 48109–2122

E-mail: {shchoi,kgshin}@eecs.umich.edu

Abstract

How to control hand-off drops is a very important Quality-of-Service (QoS) issue in cellular networks. In order to keep the hand-off dropping probability below a pre-specified target value (thus providing a *probabilistic* QoS guarantee), we design and evaluate *predictive* and *adaptive* schemes for the bandwidth reservation for the existing connections' hand-offs and the admission control of new connections.

We first develop a method to estimate user mobility based on an aggregate history of hand-offs observed in each cell. This method is then used to predict (probabilistically) mobiles' directions and hand-off times in a cell. For each cell, the bandwidth to be reserved for hand-offs is calculated by estimating the total sum of fractional bandwidths of the expected hand-offs within a mobility-estimation time window. We also develop an algorithm that controls this window for efficient use of bandwidth and effective response to (1) time-varying traffic/mobility and (2) inaccuracy of mobility estimation. Three different admission-control schemes for new connection requests using this bandwidth reservation are proposed. Finally, we evaluate the performance of the proposed schemes to show that they meet our design goal and outperform the static reservation scheme under various scenarios.

1 Introduction

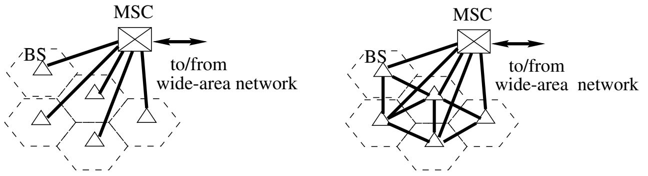
Recently, there has been a rapid growth of efforts in research and development to provide mobile users the means of "seamless" communications through wireless media. This has made it possible to implement and deploy the current cellular systems, PCS (Personal Communication Systems), and some commercial wireless LANs like WaveLAN [11]. There has also been a great demand for broadband multimedia communication involving digital audio and video. A number of researchers have been looking into communication services with guaranteed QoS such as delivery delay and link bandwidth in wired networks [1, 13, 16]. Limited efforts to support QoS guarantees in wireless/mobile networks have also been reported [2, 3, 9]. In addition to packet-level QoS issues (like packet-delay bound, throughput, and packet-error probability) considered in [2, 3, 9], connection-level QoS issues (related to connection establishment and management) need to be addressed in wireless/mobile networks, because users are expected to move around during communication sessions, causing hand-offs between cells. The current trend in cellular networks is to reduce cell size to accommodate more mobile users in a given area, and it will cause more frequent hand-offs, thus making connection-level QoS even more important.

One of the most important connection-level QoS issues is how to control (or reduce) hand-off drops due to lack of available channels in the new cell, since mobile users should be able to continue their on-going sessions. We will consider two connection-level QoS parameters: the probability P_{CB} of blocking new connection requests and the probability P_{HD} of dropping hand-offs. Ideally, we would like to have no hand-off drops so that on-going connections may be preserved as in a QoS-guaranteed wired network. However, this requires the network to reserve bandwidth in all cells a mobile might pass through; this is not possible in most cases, because the mobile's direction is not known *a priori*. Moreover, this per-connection/mobile reservation will severely under-utilize, and hence quickly deplete, bandwidth, which will, in turn, cause high P_{CB}.

Each cell can, instead, reserve fractional bandwidths of on-going connections in its adjacent cells, and this aggregate reserved bandwidth (of multiple on-going connections) can be used solely for hand-offs, not for new connection requests. The problem is then how much of bandwidth in each cell should be reserved for hand-offs. In this paper, we present a *predictive* and *adaptive* scheme for bandwidth reservation and admission control that keeps the hand-off dropping probability below a target value, $P_{HD,target}$. Since it is practically impossible to completely eliminate hand-off drops, the best one can do is to provide some form of *probabilistic QoS guarantees* by keeping P_{HD} below a pre-specified value. Our scheme is *predictive* as it estimates the directions and hand-off times of on-going connections in adjacent cells, and *adaptive* because it dynamically adjusts the amount of reserved bandwidth according to the estimation results and the observed hand-off dropping events.

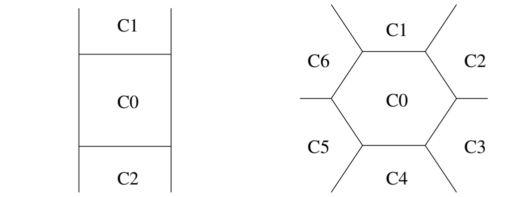
To reduce hand-off drops, researchers have also proposed adaptive QoS schemes in which a connection's QoS can be downgraded when there is an insufficient bandwidth avail-

*The work reported in this paper was supported in part by the US Department of Transportation under Grant No. DTFH61-93-X-00017. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the funding agency.



(a) Star topology case (b) Fully-connected case

Figure 1: Network topology among the MSC and BSs.



(a) 1-dimensional case (b) 2-dimensional case

Figure 2: Indexing of cells.

able in the new cell [6, 8].[1] In fact, QoS adaptation can be implemented independently of bandwidth reservation, and when both are used together, bandwidth reservation is made on the basis of the minimum QoS of each connection. In this way, our scheme can be integrated with any adaptive QoS scheme. The notion of bandwidth reservation for hand-offs was introduced in the mid-80s [5]. With this scheme, a portion of the link capacity is permanently reserved for hand-offs. This reserved bandwidth is not allowed to be used for new connections so that P_{HD} may be kept lower than P_{CB}. We will henceforth call this a *static* reservation scheme. As will be shown later, this static reservation scheme cannot effectively handle a variety of connection bandwidths, traffic loads, and users' mobility.

The rest of this paper is organized as follows. Section 2 describes the system specification and states the assumptions to be used. The users' mobility estimation based on an aggregate history of observations is presented in Section 3. Section 4 describes the proposed predictive, adaptive bandwidth reservation and three admission-control schemes. Section 5 presents and discusses the simulation results of the proposed and static-reservation schemes under various scenarios. Section 6 discusses related work, putting our scheme in a comparative perspective. Finally, the paper concludes with Section 7.

2 System Model

We consider a wireless/mobile network with a cellular infrastructure, comprising a wired backbone and a (possibly large) number of base stations (BSs). The geographical area covered by a BS is called a *cell*. A mobile,[2] while staying in a cell, communicates with another party, which may be a node connected to the wired network or another mobile, through the BS in the same cell. When a mobile moves into an adjacent cell in the middle of a communication session, a hand-off will enable the mobile to maintain connectivity to its communication partner, i.e., the mobile will start to communicate through the new BS, hopefully without noticing any difference.

A hand-off could fail due to insufficient bandwidth in the new cell, and in such a case, a *connection hand-off drop* occurs. Here, we preclude (1) delay-insensitive applications, which might tolerate long hand-off delays in case of insufficient bandwidth in the new cell at the time of hand-off; and (2) soft hand-off of the Code Division Multiple Access (CDMA) systems [15], in which a mobile can communicate via two adjacent BSs simultaneously for a while before the actual hand-off takes place. We propose to set aside some bandwidth in each cell for possible hand-offs from its adjacent cells. This reserved bandwidth can be used only for hand-offs from adjacent cells, but *not* by newly-requested connections in the cell. A connection is specified by its required bandwidth,[3] and a newly-requested connection in a cell requires a very simple admission test:

$$\sum_i b_i + b_{new} \leq C - B_r, \tag{1}$$

where C is the wireless link capacity, B_r is the *target reservation bandwidth*, i.e., the required bandwidth to be reserved for hand-offs, b_i is the bandwidth being used by an existing connection i, and b_{new} is the bandwidth required by the newly-requested connection. Upon arrival of a new connection request, B_r is updated predictively and adaptively — before performing the admission test Eq. (1) on the request — depending on the traffic status in adjacent cells. Note that B_r is a target, not the actual reserved bandwidth, since a cell may not be able to reserve the target bandwidth, i.e., $\sum_i b_i + B_r > C$. This can happen because a BS can control the admission of only newly-requested connections, not those connections handed off from adjacent cells.

Our bandwidth reservation is based on information from adjacent cells such as the number of existing connections and their bandwidth requirements. Thus, it is very important to maintain inter-BS communications. The underlying network topology for BSs can have mainly two possible configurations as shown in Figure 1. There is a node called "Mobile Switching Center" (MSC), which covers a number of BSs, and works as a gateway to and from the wide area network. Figure 1 (a) shows a star-topology interconnection among the MSC and BSs, in which there are no direct connections among BSs. This is a typical structure found in the currently-deployed cellular networks. In this environment, each BS delivers the information about existing connections in its cell to the MSC. The MSC will then determine the target reservation bandwidth in each cell, and accordingly, will perform the admission test for each newly-requested connection in a cell within its coverage. On the other hand, Figure 1 (b) shows the case where BSs are fully-connected. In this topology, BSs can communicate directly, not via the MSC, and each BS can determine the target reservation bandwidth, and hence, perform the admission test for each newly-requested connection in its cell.

All cells around each cell A are indexed:[4] A with 0, and the others with numbers beginning with 1 as shown in Figure 2. Let $C_{i,j}$ be connection j in cell i and $b(C_{i,j})$ be its required bandwidth. For simplicity, we assume that a mobile cannot have multiple connections simultaneously, so by an *active mobile*, we mean a mobile with one existing

[1] Reducing hand-off drops is one role of adaptive QoS. Other roles include reduction of P_{CB} and better utilization of network bandwidth by upgrading QoS if possible.

[2] We use the term "mobiles" to refer to mobile or portable devices, e.g., hand-held handsets or portable computers.

[3] A connection in QoS-sensitive networks might be specified by its required buffer space as well as bandwidth. However, in wireless networks, bandwidth (of wireless links) is the most precious resource, so we consider the bandwidth reservation only. Buffer space reservation can be treated similarly to the bandwidth reservation considered here, and admission control can be integrated with this buffer reservation.

[4] This is the cell A's (or its base station's) centric view.

connection.[5] The cellular system uses a fixed channel allocation (FCA) scheme, and cell i has a wireless link capacity $C(i)$. The unit of bandwidth is BU, which is the required bandwidth to support a voice connection. A connection runs through multiple wired and wireless links, and hence, we need to consider bandwidth reservation on both wireless and wired links for hand-offs. However, we will confine ourselves to reservation of wireless link bandwidth in each cell, because routing and/or re-routing upon hand-off of a connection is beyond the scope of this paper. Our scheme can be extended easily to include wired link bandwidth reservation by considering the routing and re-routing inside the wired network.

3 Mobility Estimation

We probabilistically model mobiles' hand-off behavior and estimate their mobility based on an aggregate history of hand-offs observed in each cell. In order to understand the rationale behind our mobility estimation, let's consider the usual road traffic as an example:

O1. There are speed limits in most roads, and mobiles' speeds usually are not much higher or lower than the speed limits.

O2. In local roads, traffic signals affect mobiles' movements significantly.

O3. During the rush hours, the speeds of all mobiles in a given geographical area are closely correlated.

O4. In many cases, the direction of a mobile can be predicted from the path the mobile has taken so far.

From the above observations, we expect that the hand-off behavior of a mobile will be probabilistically similar to the mobiles which came from the same previous cell and are now residing in the current cell. Hence, we can predict the next cell of a mobile and estimate its hand-off time by utilizing an aggregate history of observations in each cell. Even though the above observations were made from road traffic, the same method can be used for pedestrians because the speeds of pedestrians won't be that much different among themselves. In a typical outdoor cellular network, there will be both pedestrian and vehicular mobiles while in the indoor case, there are mostly pedestrians or non-moving objects.

Another possibility is to use mobile-specific histories as suggested in [8]. That is, each specific mobile's movement is observed over time, then the mobile's direction in a specific cell can be predicted by utilizing this observation. However, keeping track of each mobile's mobility over time is too costly, and in many cases, mobile-specific histories are not accurate enough to make good predictions. So, we preclude the availability of such information.

3.1 Hand-Off Estimation Functions

We now develop a scheme to estimate and predict mobility. This scheme will be executed by the BS of each cell in a distributed manner. For each mobile which moves into an adjacent cell from the current cell 0, the cell 0's BS caches the mobile's quadruplet, $(T_{event}, prev, next, T_{soj})$, called a *hand-off event quadruplet*, where T_{event} is the time when the mobile departed from the current cell, $prev$ is the index of the previous cell the mobile had resided in before entering the current cell, $next$ is the index of the cell the mobile entered after departing from the current cell, and T_{soj} is the sojourn time of the mobile in the current cell, i.e., the time span between the entry into and departure from the current cell. Note that $prev = 0$ means that the departed mobile started its connection in the current cell.

[5] Hence, we will use the terms "connection" and "mobile" interchangeably throughout this paper.

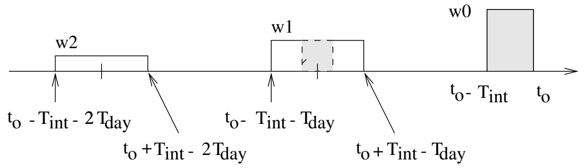


Figure 3: An example of periodic windows to obtain hand-off estimation functions with $N_{win\_days} = 2$.

From the cached quadruplets, the BS builds *hand-off estimation function*, which describes the estimated distribution of the next cell and sojourn time of a mobile, depending on the cell the mobile stayed before. One can also imagine that this probabilistic behavior of mobiles, especially in terms of sojourn time, will depend on the time of day, e.g., the sojourn time during rush hours will differ significantly from that during non-rush hours. We assume that the probabilistic behavior will mostly follow a cyclic pattern with the period of one day. A hand-off estimation function, at the current time t_o, is obtained as follows: for a quadruplet $(T_{event}, prev, next, T_{soj})$ such that

$$t_o - T_{int} - nT_{day} \le T_{event} < t_o + T_{int} - nT_{day}, \quad (2)$$

where T_{int} is the estimation interval of the function which is a design parameter, T_{day} is the duration of a day, i.e., 24 hours, and n (≥ 0) is an integer,

$$F_{HOE}(t_o, prev, next, T_{soj}) := w_n, \quad (3)$$

where $1 \ge w_n \ge w_{n+1}$, and $w_n = 0$ for all $n > N_{win\_days}$. The weight factor w_n is from the fact that the traffic condition in a cell during a specific period of days can vary over time. N_{win_days} is a design parameter so that the quadruplet observed more than $(N_{win\_days} \cdot T_{day} + T_{int})$ ago is determined out-of-date, and not used for the hand-off estimation function. One can easily see that the hand-off estimation functions are affected by the hand-off event quadruplets within the periodic windows of duration $2T_{int}$ as shown in Figure 3. Note that the duration $[t_o, t_o + T_{int}]$ is missing in the figure because it represents a future time, which is not meaningful in the definition of a hand-off event quadruplet.

In practice, it is desirable to limit the number of the quadruplets (1) used for the hand-off estimation function and (2) currently not used for the hand-off estimation function, but cached for future use, e.g., those with $t_o + T_{int} - T_{day} < T_{event} < t_o - T_{int}$ in Figure 3, in order to reduce the memory and computation complexity.[6] We define the *maximum hand-off estimation function size*, N_{quad}, as the maximum number of hand-off event quadruplets used for the hand-off estimation function for each pair of $(prev, next)$. This implies that we don't need the quadruplets from previous days if we observed enough during the last T_{int} interval. Up to N_{quad} cached quadruplets are used for the hand-off estimation with the following priority rule. First, the quadruplet which satisfies Eq. (2) with a smaller n gets higher priority. Second, among those satisfying Eq. (2) with

[6] The calculations for the mobility estimation will be dependent on the number of the quadruplets used for the hand-off estimation function as will be shown in the next section.

$$p_h(C_{0,j} \rightarrow next) := \begin{cases} \frac{\sum_{T_{ext\_soj}(C_{0,j})<t_{soj}\le T_{ext\_soj}(C_{0,j})+T_{est}} F_{HOE}(t_o, prev(C_{0,j}), next, t_{soj})}{\sum_{next'\in \mathbf{A}_0}\sum_{t_{soj}>T_{ext\_soj}(C_{0,j})} F_{HOE}(t_o, prev(C_{0,j}), next', t_{soj})}, \\ \quad \text{if } \sum_{next'\in \mathbf{A}_0}\sum_{t_{soj}>T_{ext\_soj}(C_{0,j})} F_{HOE}(t_o, prev(C_{0,j}), next', t_{soj}) \neq 0, \\ 0, \quad \text{otherwise.} \end{cases} \tag{4}$$

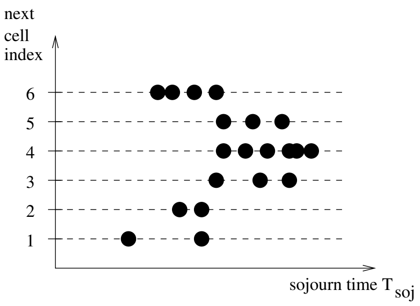


Figure 4: An example of the footprint of the hand-off estimation function for $prev = 1$.

the same n, the quadruplet with a smaller $|T_{event} - nT_{day}|$ gets higher priority. Figure 3 shows an example that only the quadruplets with the event times T_{event} within the shaded regions are used for the hand-off estimation function according to the priority rule, implying that the total number of quadruplets within the regions is N_{quad}. In order to reduce the cache memory size, those quadruplets observed at time t', i.e., $T_{event} = t'$, when the hand-off estimation function at time t' doesn't use any quadruplets observed previous days are not cached for future use, because they are unlikely to be used for the hand-off estimation function next day. Note that those quadruplets (1) with $T_{event} < t_o - T_{int} - N_{win\_days}T_{day}$ and (2) not used for the hand-off estimation function during the last $(T_{day} + T_{int})$ can be deleted from the cache memory.

There are other types of periodic and aperiodic patterns to consider for mobility estimation. These will be observed during weekends and holidays, and the mobility patterns will be significantly different from those during weekdays. So, another set of quadruplets will be cached for these special days, and the hand-off estimation functions for weekends, for example, will be built using Eqs. (2) and (3) by replacing T_{day} and N_{win_days} with $T_{week} = 7$ (days) and N_{win_weeks}, respectively. Figure 4 shows an example of footprint of the hand-off estimation function for $prev = 1$ without showing the values of w_n's. In the hand-off estimation function in a 3-dimensional space, the function is shown to have different heights, depending on the values of w_n's. The example is drawn from the same indexing as shown in Figure 2 (b). From the footprint, we observe that cell 4 is the farthest cell from cell 1 (i.e., the previous cell) through cell 0 (i.e., the current cell) among the adjacent cells of cell 0 since the sojourn times before entering cell 4 are generally shown to be among the largest. Note that the hand-off estimation function for a given $prev$ can generate a probability mass function for a two-dimensional random vector $(next, T_{soj})$, where $next$ is the predicted next cell and T_{soj} is the estimated sojourn time in the current cell. How this hand-off estimation function is used to estimate the user mobility is discussed next.

4 Predictive, Adaptive Bandwidth Reservation and Admission Control

We now describe predictive, adaptive bandwidth reservation and admission control to keep the hand-off dropping probability P_{HD} below $P_{HD,target}$ by utilizing the hand-off estimation functions described thus far.

4.1 Bandwidth Reservation

Our approach is based on the estimated mobility during the time window $[t_o, t_o + T_{est}]$, where t_o is the current time. We consider the behavior of a mobile in the current cell. The mobility of the active mobile with connection $C_{0,j}$ is estimated with $p_h(C_{0,j} \rightarrow i)$, the probability that $C_{0,j}$ hands off into cell i within T_{est}.

The hand-off probability can be computed using the hand-off estimation function as follows. The BS of a cell keeps track of each active mobile in its cell via the mobile's *extant sojourn time*. The extant sojourn time $T_{ext\_soj}(C_{0,j})$ of connection $C_{0,j}$ is the time elapsed since the active mobile with connection $C_{0,j}$ entered the current cell. Using Bayes' theorem [12], the hand-off probability $p_h(C_{0,j} \rightarrow next)$ at time t_o is calculated by Eq. (4), in which $prev(C_{0,j})$ is the cell which $C_{0,j}$ resided in before entering the current cell and $\mathbf{A}_i$ is the set of indices of cell i's adjacent cells. The equation represents the expected probability that $C_{0,j}$ hands off into cell $next$ with the sojourn time t_{soj} which is less than, or equal to, $T_{ext\_soj}(C_{0,j}) + T_{est}$ given the condition that $t_{soj} > T_{ext\_soj}(C_{0,j})$, which is the hand-off probability $p_h(C_{0,j} \rightarrow next)$.

Figure 5 shows an example of calculating $p_h(C_{0,j} \rightarrow 4)$, when $C_{0,j}$ entered cell 0 from cell 1, using the footprint of the hand-off estimation function for $prev(C_{0,j}) = 1$, shown in Figure 4. In the figure, the values of $F_{HOE}(t_o, 1, next', T_{soj})$ from all points at the right side of the vertical line at $T_{soj} = T_{ext\_soj}(C_{0,j})$ (i.e., in both dark and light shaded regions) are summed to obtain the denominator in Eq. (4). Because this value is not zero, the values of $F_{HOE}(t_o, 1, 4, T_{soj})$ from two points in the dark-shaded region are summed to obtain the numerator in Eq. (4). Then, we can complete the calculation of $p_h(C_{0,j} \rightarrow 4)$. Note that the mobile with connection $C_{0,j}$ is estimated to be stationary (i.e., non-moving) in cell 0 if there is no hand-off event in the hand-off estimation function with a sojourn time larger than the connection $C_{0,j}$'s extant sojourn time, i.e., the denominator in Eq. (4) is zero.

Now, using the probabilities of handing off connections into cell 0 from its adjacent cell i within T_{est} (i.e., hand-off probabilities $p_h(C_{i,j} \rightarrow 0)$), the required bandwidth $B^i_{r,0}$ to be reserved in cell 0 for the expected hand-offs from cell i is obtained as:

$$B^i_{r,0} = \sum_{j\in \mathbf{C}_i} b(C_{i,j})p_h(C_{i,j} \rightarrow 0), \tag{5}$$

where $\mathbf{C}_i$ is the set of indices of the connections in cell i and $b(C_{i,j})$ is connection $C_{i,j}$'s bandwidth. Finally, the target reservation bandwidth $B_{r,0}$ in cell 0, which is the aggregate

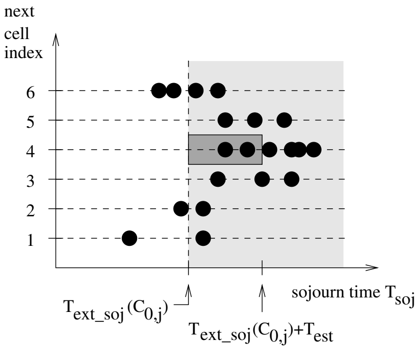


Figure 5: An example of calculating $p_h(C_{0,j} \rightarrow next)$ when $prev(C_{0,j}) = 1$ and $next = 4$.

bandwidth to be reserved in cell 0 for the expected hand-offs from adjacent cells within T_{est}, is calculated as:

$$B_{r,0} = \sum_{i \in \mathbf{A}_0} B^i_{r,0}, \tag{6}$$

where $\mathbf{A}_i$ is the set of indices of cell i's neighbors.

Note that the target reservation bandwidth is an increasing function of the estimation time T_{est} as $p_h(C_{i,j} \rightarrow 0)$ is an increasing function of T_{est}. There might be an optimal value of T_{est} for given traffic/mobility status in the sense of giving the smallest new connection blocking probability while keeping the hand-off dropping probability below the target. In our scheme, the estimation time will be adjusted adaptively in each cell independently of others, depending on the hand-off dropping events in the cell as described in the next section. Then, the estimation time T_{est} of cell $next$ (or $T_{est,next}$) will be used in Eq. (4). So, when the BS in cell 0 needs to update the value of $B_{r,0}$, the BS will inform the current value of $T_{est,0}$ to the adjacent cells, then the BS in each adjacent cell will calculate the required bandwidth for the expected hand-offs from that cell, i.e., $B^i_{r,0}$ for cell i, using Eq. (5), and will inform this value to cell 0's BS. Finally, cell 0's BS will calculate $B_{r,0}$ using Eq. (6).

4.2 Control of Mobility Estimation Time Window

Using our scheme, the bandwidth for hand-offs will be over-reserved (under-reserved) if T_{est} is too large (small). There might exist an optimal value of T_{est} for specific traffic load and user mobility, but these parameters in practice vary with time. Moreover, the mobility estimation functions used might not describe mobiles' behavior well, thus resulting in inaccurate mobility estimation even with the optimal T_{est}. We propose an adaptive algorithm for controlling the mobility estimation time window based on the hand-off dropping events in each cell so as to approximate the optimal T_{est} over time. Figure 6 shows the pseudo-coded algorithm executed by the BS in each cell to adjust the value of T_{est}.

Before running the algorithm, the reference window size w ($= \lceil 1/P_{HD,target} \rceil$) is determined and assigned to the observation window size w_{obs}. In addition, T_{est} is initialized to T_{start}, a design parameter, and the counts for hand-offs n_H and hand-off drops n_{HD} are reset to 0. As can be found in the pseudocode, w_{obs} is increased or decreased by w, and the constraint $P_{HD} < P_{HD,target}$ can be translated to that to keep the counted number n_{HD} of hand-off drops out of w_{obs} observed hand-offs below w_{obs}/w. During the runtime, whenever there is a hand-off drop after w_{obs}/w

```
if (w = ⌈1/P_HD,target⌉), then w_obs := w;
T_est := T_start;  n_H := 0;  n_HD := 0;
while (time increases) {
    if (hand-off into the current cell happens) then {
        n_H := n_H + 1;
        if (it is dropped) then {
            n_HD := n_HD + 1;
            if (n_HD > w_obs/w) then {
                w_obs := w_obs + w;
                if (T_est < T_soj,max) then T_est := T_est + 1;
            }
        }
        else if (n_H ≥ w_obs) then {
            if (n_HD ≤ w_obs/w and T_est > 1) then
                T_est := T_est − 1;
            w_obs := w;  n_H := 0;  n_HD := 0;
        }
    }
}
```

Figure 6: A pseudocode of the algorithm to adjust T_{est}.

drops, $T_{est} := T_{est} + 1$ and $w_{obs} := w_{obs} + w$. On the other hand, when there were less than, or equal to, w_{obs}/w hand-off drops out of w_{obs} observed hand-offs, $T_{est} := T_{est} - 1$ and $w_{obs} := w$. T_{est} is not greater than $T_{soj,max}$ in Figure 6, which is the maximum T_{soj} derived from the hand-off estimation functions in adjacent cells, because any value larger than that is meaningless. We also set the minimum value of T_{est} to 1 since if the value is too small, our scheme will reserve virtually no bandwidth irrespective of the existing connections in adjacent cells.

Given below are some considerations for the design of the estimation time window control algorithm.

- **C1.** When there were more hand-off drops than permitted, the algorithm should start to increase T_{est} quickly because of under-reserved bandwidth; otherwise, there will be continued hand-off drops.
- **C2.** The increment of T_{est} should not be too high. Otherwise, it might result in an over-reaction, hence over-reservation.
- **C3.** Due to over-reaction or decreased traffic load over time, there might be fewer hand-off drops than permitted, so the value of T_{est} should be decreased quickly. Otherwise, the bandwidth will continue to be over-reserved, hence under-utilizing the system.
- **C4.** T_{est} should not be decreased too much. Otherwise, it might result in an over-reaction, hence under-reservation.

There can be many candidate algorithms satisfying the above requirements. Especially, there might be many choices of increment and decrement step sizes, both of which were fixed at 1. We experimented with other choices like additive and multiplicative step sizes: the step size was increased additively $(1, 2, 3, \cdots)$ or multiplicatively $(1, 2, 4, \cdots)$ for consecutive increments and decrements. The main purpose of these choices is a prompt reaction to hand-off drops, i.e., **C1** and **C3**. However, these choices are found to cause over-reactions, and make the reserved bandwidth fluctuate severely between over-reservation and under-reservation. The algorithm presented here is the best one we have found so far.

Name	Description
AC1	Calculation of B_r in the current cell only.
AC2	Calculation of B_r in the current cell and every adjacent cell.
AC3	Calculation of B_r in the current cell and some adjacent cells only.

Table 1: Summary of the admission-control schemes.

4.3 Admission Control

The admission test after calculating the target reservation bandwidth can be as simple as given in Eq. (1). That is,

1. Check if $\sum_{j \in \mathbf{C}_0} b(C_{0,j}) + b_{new} \leq C(0) - B_{r,0}$.
2. If the above test is positive, the connection is admitted,

where $C(0)$ and b_{new} are the link capacity of cell 0 and the bandwidth of the newly-requested connection, respectively. This simple admission-control scheme will henceforth be referred to as **AC1**. However, when there is not enough bandwidth left unused by existing connections that can be reserved for hand-offs, it is meaningless to calculate the target reservation bandwidth. If this situation lasts for an extended period due to continued incoming hand-offs, the problem becomes more serious because some of the incoming hand-offs will be continuously dropped due to the unavailability of reserved bandwidth, triggering further increase of T_{est}. This, in turn, requires to reserve more bandwidth that doesn't exist. This situation can happen when adjacent cells accept new connections solely according to **AC1** and those admitted connections continue to be handed off into the current cell even though it doesn't have enough bandwidth.

To handle this problem, the admission test should check available bandwidths of adjacent cells as well as the current cell. Then, the admission test is given by

1. For all $i \in \mathbf{A}_0$, check if $\sum_{j \in \mathbf{C}_i} b(C_{i,j}) \leq C(i) - B_{r,i}$.
2. Check if $\sum_{j \in \mathbf{C}_0} b(C_{0,j}) + b_{new} \leq C(0) - B_{r,0}$.
3. If all of the above tests are positive, then the connection is admitted.

We call this scheme **AC2**. Note that using this admission test, the current cell and all of its adjacent cells must calculate $B_{r,i}$ for each new admission request, and this is costly. In fact, the undesirable situation described in the beginning of this subsection is expected to happen only in heavily-loaded networks. So, we present a hybrid scheme which requires only those adjacent cells which "appear" to be unable to reserve the target reservation bandwidth, to calculate the target bandwidth again and participate in the admission test. Note that $B_{r,i}$ is a time-varying function, and updated upon admission test. Upon arrival of a new connection request at cell 0, if the current target reservation bandwidth of an adjacent cell i, $B_{r,i}^{curr}$, which was calculated for a previous admission test, is not reserved fully, this cell will re-calculate $B_{r,i}$, and participate in the admission test.

1. For all $i \in \mathbf{A}_0$ such that $\sum_{j \in \mathbf{C}_i} b(C_{i,j}) + B_{r,i}^{curr} > C(i)$, calculate $B_{r,i}$ newly, set $B_{r,i}^{curr} := B_{r,i}$, and check if $\sum_{j \in \mathbf{C}_i} b(C_{i,j}) \leq C(i) - B_{r,i}$.
2. Check if $\sum_{j \in \mathbf{C}_0} b(C_{0,j}) + b_{new} \leq C(0) - B_{r,0}$.
3. If all the above tests are positive, then the connection is admitted.

We refer this scheme to **AC3**. Table 1 shows the summary of the admission-control schemes described thus far. These schemes will be comparatively evaluated in the next section.

5 Comparative Performance Evaluation

This section presents and discusses the evaluation results of the proposed schemes as well as the static reservation scheme for comparative purposes. We first describe the assumptions and specifications used for the simulation study.

5.1 Simulation Assumptions and Specifications

In our simulation environment, mobiles are traveling along a straight road (e.g., cars on a highway). This environment is the simplest in the real world, representing a one-dimensional cellular system as in Figure 2 (a). We make the following assumptions for our simulation study:

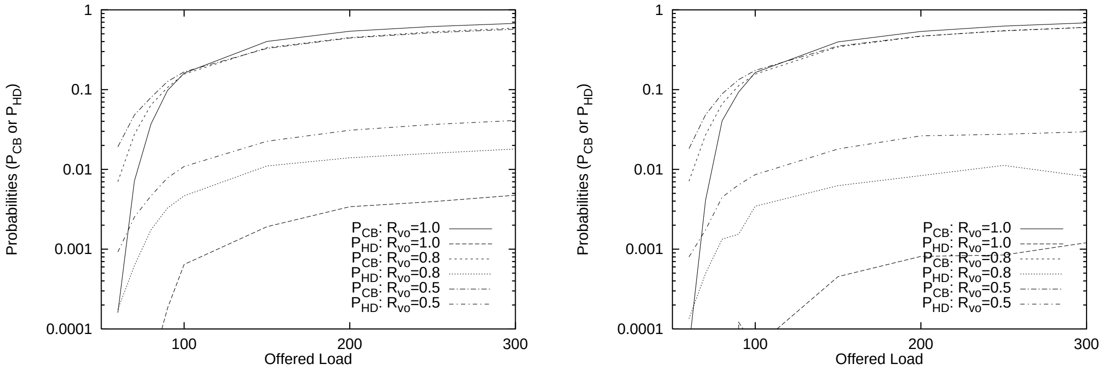
- **A1.** The whole cellular system is composed of 10 linearly-arranged cells, for which the diameter of each cell is 1 km. Cells are numbered from 1 to 10, i.e., cell $<i>$ represents the i-th cell.
- **A2.** Connection requests are generated according to a Poisson process with rate λ (connections/second/cell) in each cell. A newly-generated connection can appear anywhere in the cell with an equal probability.
- **A3.** A connection is either for voice (requiring 1 BU) or for video (requiring 4 BUs) with probabilities R_{vo} and $1 - R_{vo}$, respectively, where the *voice ratio* $R_{vo} \leq 1$.
- **A4.** Mobiles can travel in either of two directions with an equal probability with a speed chosen randomly between SP_{min} and SP_{max} (km/hour). Each mobile will run straight through the road with the chosen speed, i.e., mobiles will never turn around.
- **A5.** Each connection's lifetime is exponentially-distributed with mean 120 (seconds).
- **A6.** Each cell has a fixed link capacity 100 BUs, i.e., $C(i) = C = 100$ for all i.

Note that the fixed capacity assumption is not necessarily true in practice. For example, CDMA systems have a softer notion of capacity, in which the capacity depends on the target interference level. This target interference level is affected by the desired error performance of the system, which can be negotiable in some cases.

Each simulation run starts without any pre-memorized hand-off event quadruplets. As simulations are run, quadruplets will be collected, and will affect the hand-off estimation functions $F_{HOE}(t, prev, next, T_{soj})$. Under the above assumptions, the border cells (i.e., cells $<1>$ and $<10>$) will face fewer mobiles because there are no mobiles entering from the outside of the cellular system. Then, cells near the center (such as cells $<5>$ and $<6>$) will be more crowded by mobiles than those near the borders. This uneven traffic load can affect the performance evaluation of our proposed schemes, hence making it difficult to comprehend their operations correctly. So, we connected two border cells, i.e., cells $<1>$ to $<10>$, artificially so that the whole cellular system forms a ring architecture as was assumed in [10] (unless stated otherwise).

The parameters used include: $P_{HD,target} = 0.01$, $T_{start} = 1$ (second), $N_{quad} = 100$, $T_{int} = 1$ (hour), $N_{win\_days} = 1$, and $w_0 = w_1 = 1$. A frequently-used measure is the *offered load* per cell, L, which is defined as connection generation rate $\times$ connections' bandwidth $\times$ average connection lifetime, i.e.,

$$L = (1 \cdot R_{vo} + 4 \cdot (R_{vo} - 1)) \cdot \lambda \cdot 120, \quad (7)$$



(a) high user mobility (b) low user mobility

Figure 7: P_{CB} and P_{HD} vs. offered load: static reservation with $G = 10$ BUs.

with the above-described assumptions. The physical meaning of the offered load per cell is the total bandwidth required on average to support all existing connections in a cell.

We considered a range of the offered load from 60 to 300. Generally, the desirable range of the offered load is less than, or equal to, the link capacity, 100 BUs, of each cell. It is undesirable to keep a cell over-loaded (i.e., the offered load is > 100) for an extended period of time, and in such a case, the cell must be split into multiple cells to increase the total system capacity. However, cells can get over-loaded temporarily. Suppose a mobile user's connection request is blocked once. Then, s/he is expected in most cases to continue to request a connection establishment until it is successful or s/he gives up. This likely behavior of mobile users will affect the offered load. Near the offered load $= 100$, P_{CB} will be around, or larger than, 0.1 in most cases, due to some reserved bandwidth for hand-offs, and in such a situation, if each connection-blocked user attempts to make a connection about 5 times, then the offered load will increase to about 150 in a very short time. Likewise, there might be some cases with the offered load of 300. This possible situation can be interpreted as a positive-feedback effect for increase in the offered load. We consider the large values of offered load such as 300, since even for these large offered loads, our goal to keep P_{HD} below a target value should be achieved.

5.2 Stationary Traffic/Mobility

First, we simulated for stationary traffic/mobility with constant new connection generation rate λ and mobile speed range $[SP_{min}, SP_{max}]$. Two cases of user mobility are considered: high user mobility with $[SP_{min}, SP_{max}] = [80, 120]$, and low user mobility with $[SP_{min}, SP_{max}] = [40, 60]$. For the stationary case, $T_{int} = \infty$ is used since the speed range and the offered load do not vary during each simulation run; so, $N_{days\_win} = 1$ is meaningless.

5.2.1 Static Reservation

First, we consider the performance of static reservation as a reference (for comparison). Figure 7 plotted P_{CB} and P_{HD} as the offered load increases for (a) high user mobility and (b) low user mobility when $G = 10$, i.e., 10 BUs are reserved permanently for hand-offs in each cell. Three different values of the voice ratio R_{vo} are examined: $R_{vo} = 1.0, 0.8,$ and 0.5. The performance of this static scheme, in terms of both probabilities, is found to depend heavily on the voice ratio, user mobility, and offered load. Examples are:

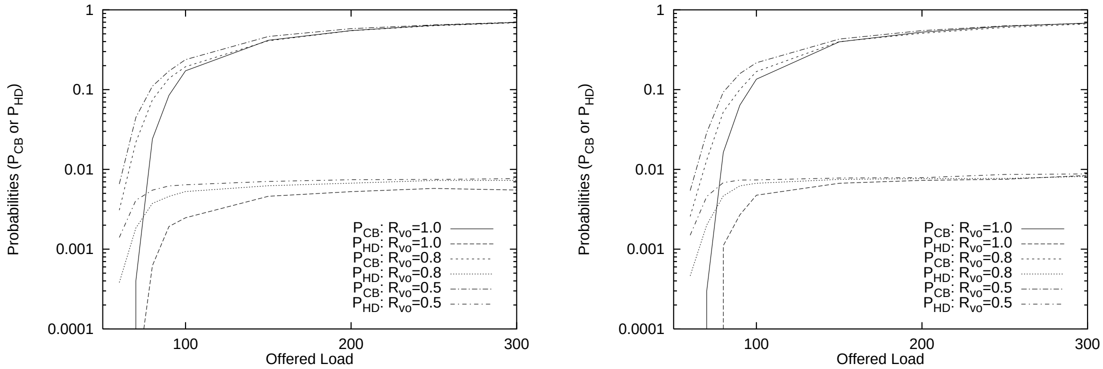
1. Static reservation of 10 BUs suffices to achieve our goal for $R_{vo} = 1.0$, but is not enough for $R_{vo} = 0.5$.
2. For $R_{vo} = 0.8$, 10-BU reservation seems enough for low user mobility as shown in Figure 7 (b), but not enough for high user mobility as shown in Figure 7 (a).
3. For $R_{vo} = 0.8$ and high user mobility, 10-BU reservation seems not enough for a highly over-loaded case (i.e., $L > 150$), but enough for the other case (i.e., $L < 150$). Moreover, for $R_{vo} = 1.0$, 10-BU reservation seems more than enough (i.e., over-reserved) for the under-loaded case (i.e., $L < 100$) since the observed P_{HD} value is too small (< 0.001 for high user mobility, and < 0.0001 for low user mobility), compared to $P_{HD,target} = 0.01$.

The voice ratio, mobile user speed, and offered load could in reality be any value and can even fluctuate. Hence, our goal cannot be achieved with static reservation, necessitating some form of adaptive reservation.

5.2.2 Performance of Admission Control AC3

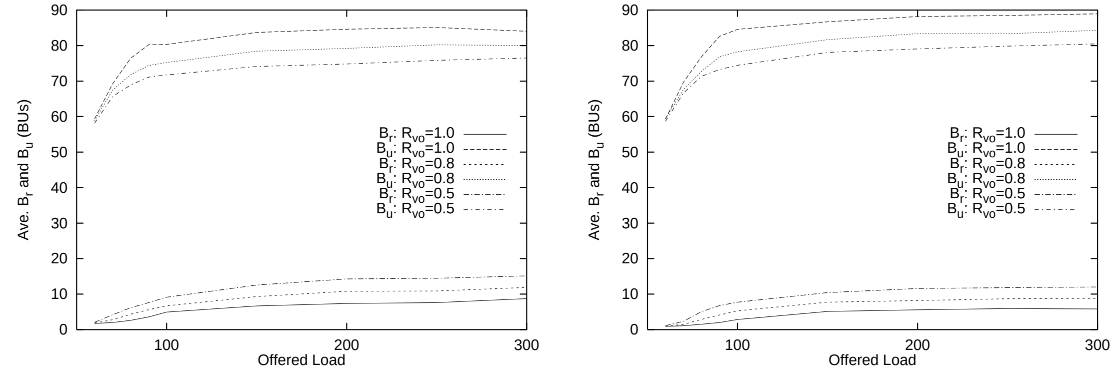
We first consider the performance of **AC3**, which is claimed to be the best among the three alternatives. Figure 8 shows P_{CB} and P_{HD} as the offered load increases for (a) high user mobility and (b) low user mobility. For the entire range of the offered load we examined, P_{HD} is observed to be less than, or equal to, our target $P_{HD,target}$ $(= 0.01)$ irrespective of user mobility and voice ratio. Moreover, for given user mobility and voice ratio, the difference between P_{CB} and P_{HD} in the plot (of log scale) is getting smaller as the offered load decreases. This means that, as the offered load decreases, the BSs reserve less bandwidth. This is desirable as long as P_{HD} stays below the target value as shown in the graphs.

Adaptive reservation patterns while varying the offered load are plotted in Figure 9 with the average target reservation bandwidth B_r in each cell and the average bandwidth B_u used by the existing connecitons in each cell. As the offered load increases, B_r in a cell increases monotonically, meaning that the target reservation bandwidth is controlled based on the offered load. The target reservation bandwidth gets saturated at the over-loaded region, because for the entire over-loaded region, regardless of the exact offered load



(a) high user mobility (b) low user mobility

Figure 8: P_{CB} and P_{HD} vs. offered load: **AC3**.



(a) high user mobility (b) low user mobility

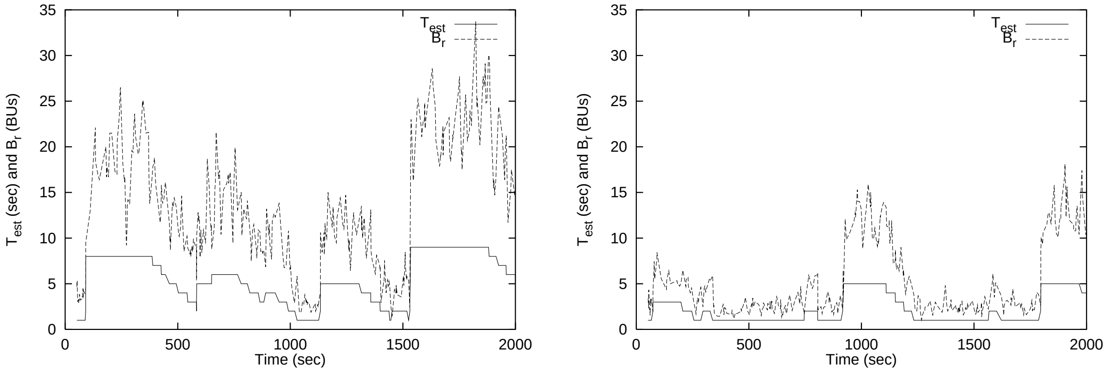
Figure 9: Average target reservation bandwidth B_r and average bandwidth used B_u vs. offered load: **AC3**.

value, the number of establishable connections will be limited by the link capacity. Our adaptive scheme reserves the bandwidth depending on the existing connections in adjacent cells, and hence the amount of the target reservation bandwidth will be almost the same for the entire over-loaded region.

We also observe that the target reservation bandwidth increases as the voice ratio R_{vo} decreases since the more video connections exist, the more bandwidth is needed. The average bandwidth used B_u is inversely proportional to the average target reservation bandwidth B_r since the reserved bandwidth can be used for handed-offs only. The reason why the sum of B_u and B_r is less than the capacity, 100, is that in **AC3**, the reserved bandwidths in adjacent cells are also checked for the admission test when these cells are suspected to have been over-loaded. By comparing two user-mobility cases, we observe that, for similar offered load and voice ratio, the high-mobility case reserves more bandwidth than the low-mobility case. For the low-mobility case, the chance of hand-offs would be smaller, and hence less bandwidth needs to be reserved.

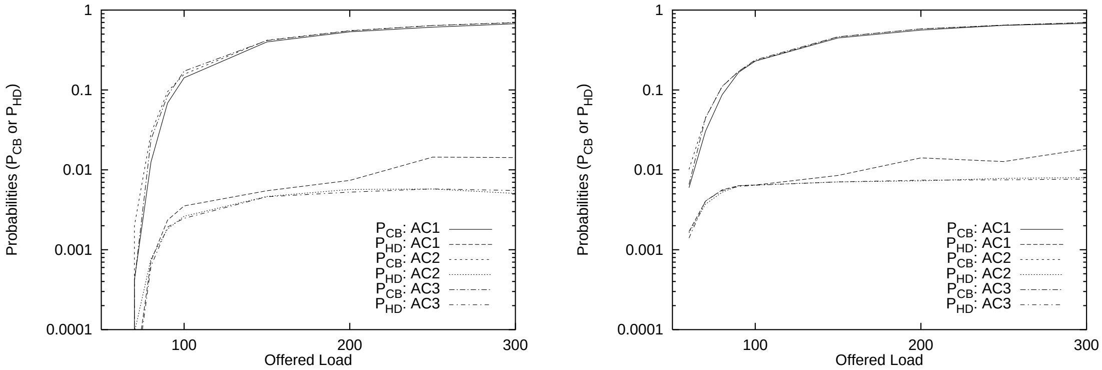
Next, let's consider the detailed operation of our scheme in each cell. Figure 10 shows T_{est} and B_r, starting from the beginning of a simulation run (i.e., $t = 0$) for the offered load = 300 and $R_{vo} = 1.0$ with high user mobility in (a) cell $<5>$ and (b) cell $<6>$. The values of T_{est} were observed to go up and down as time passes. Note that an increase of T_{est} by one corresponds to a connection's hand-off drop. The target reservation bandwidth fluctuates between over-reservation and under-reservation, depending on the value of T_{est}. The value of T_{est} seldom stays at a possible optimum value without fluctuation for the following two reasons: (1) hand-offs could be bursty, so when there are a number of hand-off drops, it is difficult to determine whether it is due to the insufficient reserved bandwidth or bursty hand-offs; and (2) the effectiveness of the reserved bandwidth is determined some time later; that is, whether the currently-reserved bandwidth is enough or not can be determined only after some mobiles enter the cell. We also observed the fluctuations of B_r even with a temporarily-constant T_{est} as the value of B_r depends on the number and type of connections in the adjacent cells and their extant sojourn times.

Figure 11 plotted P_{HD} for cells $<5>$ and $<6>$ while increasing time, obtained from the same simulation run used for Figure 10. Note that the increase of P_{HD} corresponds to hand-off drops. By comparing it with Figure 10, we can also observe that the increasing moments of P_{HD} and T_{est} coincide exactly as they should be. P_{HD} peaks over the target value $P_{HD,target}$ $(= 0.01)$ sometimes, but eventually goes below 0.01. Near the starting point, i.e., $t = 0$ (sec), our scheme seems not working well because the simulation starts without any pre-memorized hand-off event quadruplets $(T_{event}, prev, next, T_{soj})$ and with $T_{est} = T_{start} = 1$ (sec). As time goes on, the chance of peaks over the target value is low because (1) hand-off event quadruplets are observed and used for the hand-off estimation functions; (2) T_{est} is adapted; and (3) the effect of some more hand-off drops out of a large number of hand-offs is minor due to an averaging effect.



(a) cell <5> (b) cell <6>

Figure 10: T_{est} and B_r vs. time when the offered load is 300 and $R_{vo} = 1.0$ with high user mobility: **AC3**.



(a) $R_{vo} = 1.0$ (b) $R_{vo} = 0.5$

Figure 12: P_{CB} and P_{HD} vs. offered load for high user mobility.

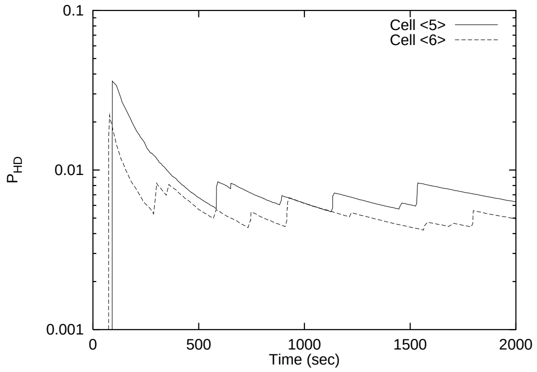


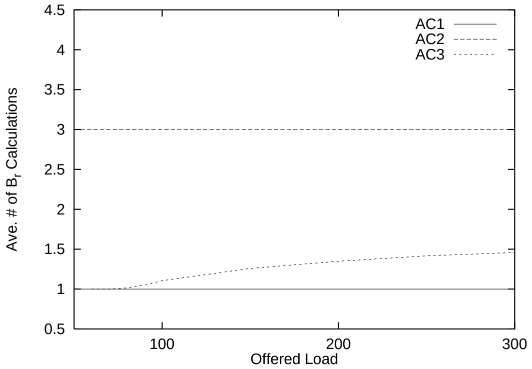
Figure 11: P_{HD} at cells $<5>$ and $<6>$ vs. time when the offered load is 300 and $R_{vo} = 1.0$ for high user mobility: **AC3**.

5.2.3 Comparison among Three Alternatives

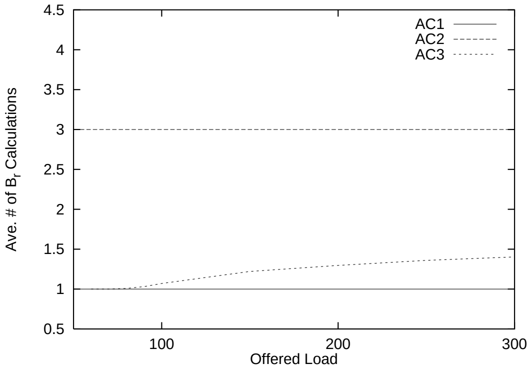
We now comparatively evaluate the performance of three difference schemes: **AC1**, **AC2**, and **AC3**. Figure 12 plots P_{CB} and P_{HD}. First, in terms of P_{CB}, three schemes work almost the same even though **AC1** has the smallest P_{CB} — with small differences – for the entire offered loads we examined. On the other hand, in terms of P_{HD}, **AC2** and **AC3** work almost the same, and **AC1** is worse. Our goal is not achieved in a highly over-loaded region (say, $L > 150$) for **AC1**. P_{HD} does not exceed 0.02 even at the offered load of 300, which is good because this small violation ratio might be tolerable in most practical applications.

Now, we consider the complexity of these schemes measured in average number of B_r calculations for the admission test of a new connection request ($= N_{calc}$). Note that, to calculate B_r in a cell, its BS needs to communicate with BSs in all adjacent cells. Figure 13 shows that N_{calc} for **AC1** is 1, irrespective of the offered load because only the BS of the cell in which the new connection was requested has to calculate B_r while $N_{calc} = 3$ for **AC2** because BSs in all adjacent cells are required to calculate B_r. For **AC3**, which is a hybrid of **AC1** and **AC2**, N_{calc} =1 for low offered load, but it starts to increase at about $L = 80$. However, the value is observed to be less than 1.5 in all of our simulations, i.e., less than a half of that of **AC2**. The complexity increase could be larger for two-dimensional cellular structures. Because **AC3** works almost the same as **AC2** in terms of P_{CB} while keeping P_{HD} below the target with a lower complexity according to our simulation results, we conclude that **AC3** is a better choice than **AC2**.

Now, we compare **AC1** with **AC3** by examining each cell when the system is over-loaded. Table 2 shows the state of each cell at the end of simulations when the offered load is 300 and $R_{vo} = 1.0$ for high user mobility with (a) **AC1** and (b) **AC3**. The first column represents the cell number, the second is P_{CB}, the third is P_{HD}, the fourth is the value of T_{est}, the fifth is the value of B_r, and the sixth is the value of B_u, all at the end of the simulations. From Table 2 (b), **AC3** is found to work similar throughout all cells in terms of P_{CB} while meeting the constraint $P_{HD} \leq P_{HD,target}$. B_r can



(a) high user mobility



(b) low user mobility

Figure 13: Average number of B_r calculations for an admission test vs. offered load:

Cell	P_{CB}	P_{HD}	T_{est}	B_r	B_u
1	1.28e-1	5.10e-3	1	1.04	89
2	9.59e-1	1.30e-2	45	102	89
3	9.57e-1	1.30e-2	45	93.6	97
4	2.41e-1	6.01e-3	1	2.13	93
5	9.79e-1	2.58e-2	45	89.9	99
6	3.83e-1	7.12e-3	1	1.97	87
7	9.80e-1	2.90e-2	45	85.8	100
8	2.75e-1	6.97e-3	2	3.22	94
9	9.57e-1	1.80e-2	45	10.4	88
10	9.26e-1	1.51e-2	45	102	81

(a) **AC1**

Cell	P_{CB}	P_{HD}	T_{est}	B_r	B_u
1	6.23e-1	6.53e-3	2	5.63	97
2	6.66e-1	6.57e-3	2	5.44	84
3	7.30e-1	7.54e-3	10	22.5	75
4	8.06e-1	7.80e-3	7	19.2	73
5	7.65e-1	7.31e-3	10	24.2	74
6	5.65e-1	6.74e-3	5	11.3	66
7	6.17e-1	6.42e-3	3	5.81	89
8	8.10e-1	6.87e-3	3	9.18	87
9	7.68e-1	7.42e-3	8	20.1	78
10	6.46e-1	5.41e-3	1	4.50	90

(b) **AC3**

Table 2: Status in each cell at the end of simulations when the offered load is 300 and $R_{vo} = 1.0$ with high user mobility.

	AC1		**AC3**	
Cell	P_{CB}	P_{HD}	P_{CB}	P_{HD}
1	0.	0.	5.61e-02	0.
2	5.24e-01	6.63e-03	5.38e-01	4.58e-03
3	9.66e-01	1.09e-02	7.83e-01	7.38e-03
4	2.84e-01	5.98e-03	7.06e-01	5.78e-03
5	9.45e-01	4.15e-02	6.35e-01	5.93e-03
6	4.17e-01	7.28e-03	7.44e-01	7.96e-03

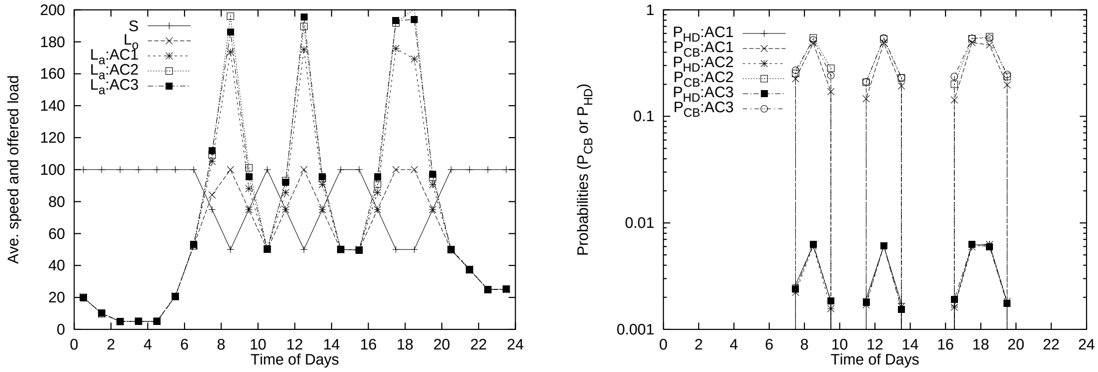
Table 3: Status in each cell at the end of simulations when the offered load is 300, $R_{vo} = 1.0$, and all mobiles follow one direction with high mobility.

change dramatically depending on the traffic condition in adjacent cells even with the same T_{est} as observed in Table 2. However, according to Table 2 (a) of **AC1**, the performance of each cell is found to fluctuate greatly, i.e., the performance in terms of P_{CB}, P_{HD}, T_{est}, and B_r drastically differ in roughly every two cells. This is not fair to those mobiles which want to establish new connections in cells with a very high P_{CB}, e.g., cells <2>, <3>, <5>, <7>, <9>, and <10> in the table. More importantly, P_{HD}'s of these cells are not bounded. This phenomenon was anticipated as explained in Section 4.3 when the admission test checks the current cell only as was done in **AC1**.

Table 3 shows the status of each cell at the end of simulations with a different mobility pattern when the offered load = 300 and $R_{vo} = 1.0$. For these simulations, the direction of mobiles are not chosen randomly. Instead, all mobiles follow the direction from cell <1> to cell <10>. Moreover, two border cells, i.e., cells <1> and <10>, are disconnected. Now, cell <1> won't have any incoming mobiles from adjacent cells. Naturally, P_{HD} will be zero at cell <1>. For **AC1**, we observe a behavior similar to that in Table 2 (a). Especially, because cell <1> doesn't care about the status of cell <2>, the BS of cell <1> accepted all new connection requests, hence $P_{CB} = 0$. Cell <2> also doesn't care about the status of cell <3>. These make cell <3> over-crowded, and eventually result in a very high P_{CB} (near 1) and over-target P_{HD} at cell <3>. This type of patterns appears every other cell as shown in the table. On the other hand, for **AC3**, cell <1> cares about cell <2>, and blocks some new connection requests. Every cell $<i>$ cares about the status of cell $<i+1>$. Eventually, balanced performance is observed over the entire system while every cell meeting the constraint on P_{HD}.

5.3 Time-Varying Traffic/Mobility

We now vary the connection generation rate λ and speed range $[SP_{min}, SP_{max}]$ over time. Each simulation is run for two days in simulation time. Figure 14 (a) shows time-varying averages of mobiles' speeds and offered loads. First, for a given value of the average speed (marked by S), the speed range is given by $[S-20,\ S+20]$ (km/h). Second, the original offered load (marked by L_o) is the traffic load from the new connections generated, which is the offered load L defined in Eq. (7). In this time-varying case, a blocked connection request will be re-requested with probability $1 - 0.1N_{ret}$ after waiting 5 seconds, where N_{ret} is the number of times a connection request has been made. So, depending on P_{CB}, the actual offered load L_a will vary, i.e., the larger P_{CB}, the larger L_a. From the figure, we observe that the values of L_a for different schemes are different when the system is highly-loaded even with the same L_o. Note that the fluctuations of the offered load and speed represent the reality, that is, the offered load peaks during rush hours (e.g., around 9 a.m., 1 p.m., and 5–6 p.m.) at low speeds.



(a) speed and offered load (b) probabilities

Figure 14: Time-varying case: (a) mobiles' average speed and offered load vs. time of days; and (b) P_{CB} and P_{HD} vs. time of days.

Figure 14 (b) shows P_{CB} and P_{HD} over time of days for three different schemes. The probability samples represent the average probability during the corresponding one-hour period, i.e., P_{CB} at $t = 8.5$ represents the average over the interval [8, 9]. First, we observe that outside the peak hour regions, both P_{CB} and P_{HD} are negligibly small. During the peak hours, P_{HD} is almost the same for different schemes, and bounded by $P_{HD,target}$ ($= 0.01$). On the other hand, P_{CB} of **AC1** is found to be lower than that of the other two schemes, and the differences between P_{CB}'s of **AC1** and **AC3** are larger compared to those from the stationary case in Figure 12. This is due to the positive feedback effect of the offered load increase; that is, from the original offered load, the difference between **AC1** and **AC3** could be small, but this small difference could be amplified through the retrials of each blocked connection request.

According to the results of the time-varying case, **AC1** is the best because it yields the lowest P_{CB} while meeting our goal. For this time-varying case, we considered only a regular traffic pattern: a high offered load for relatively short peak-hour periods (of 1 or 2 hours). However, **AC1** may have undesirable behaviors as previously observed in the time-invariant case, because there might be unexpected irregular traffic and mobility patterns in the real world. **AC3** was found to be robust in many different scenarios with relatively low complexity (up to 1.5 times that of **AC1** in our simulations). So, **AC3** is the most favorable among the schemes considered.

6 Related Work

We are not the first to attempt to design bandwidth-reservation and admission-control schemes to keep the connection hand-off dropping probability below a target value. The authors of [10] advocated the connection hand-off dropping probability as an important connection-level QoS parameter in wireless/mobile networks, and designed a distributed call admission-control scheme to keep the connection hand-off dropping probability below a specified limit. With their scheme, the BS obtains the required bandwidth for both the existing and hand-off connections after a certain time interval, then performs admission control so that the required bandwidth may not exceed the cell capacity. Their scheme was shown to be better than the static reservation scheme. The authors of [8] extended this scheme as a part of their proposal to accommodate heterogeneous connection bandwidths, and studied the effects of design parameters used in the scheme. The main problems of these schemes are: (1) they assumed the sojourn time of each mobile is exponentially-distributed, which is impractical. Moreover, it is not clear whether the scheme will still work when this assumption does not hold; and (2) there is no specified mechanism to predict which cells mobiles will move to.

The shadow cluster concept was suggested in [7] to estimate future resource requirements and perform admission control in order to limit the hand-off dropping probability, in which the shadow cluster is a set of cells around an active mobile. This scheme is based on the precise knowledge of each user mobility, depending on the location and time, which they assumed given. Our mobility estimation can provide the knowledge of mobility used in their scheme, but it is unclear how it will work if the knowledge is not accurate. (This may be the case if our cell-specific history-based mobility estimation is used.) How to determine the shadow cluster is also not defined clearly. Moreover, their scheme is computationally too expensive to be practical.

Our scheme is more realistic than the above-mentioned schemes, because (1) exponentially-distributed mobile sojourn times are not assumed, instead, mobiles' hand-off behaviors are estimated based on a history of observations in each cell; (2) our scheme is robust to the inaccuracy of mobility estimation and the time-variation of traffic/mobility, thanks to our mobility estimation time window control; and (3) due to the adaptability of our scheme, it is not required to determine the optimal value of parameters, which might depend on the traffic status, as in [8].

There were also limited efforts to estimate mobility. The authors of [8] explored mobility estimation for an indoor wireless system based on both mobile-specific and cell-specific observation histories. Mobile-specific observation of mobility is costly and not accurate in general. Our mobility estimation not only predicts the next cell to which a mobile will move, but also estimates the hand-off time (or sojourn time). This hand-off time estimation makes it possible for BSs to reserve bandwidth more efficiently.

There have also been research efforts for adaptive bandwidth reservation. The author of [6] suggested bandwidth reservation depending on the existing connections in adjacent cells. However, the scheme lacks such details as how much of bandwidth should be reserved. The bandwidth-reservation and admission-control schemes in [14] assume that the mobility of users is predictable, that is, mobility can be characterized by the set of cells the mobile is expected to visit during the lifetime of the mobile's connection. This assumption does not hold for most wireless/mobile networks. Moreover, the scheme reserves the required bandwidth at every cell and node in the mobility specification, which is usually excessive.

7 Conclusion and Future Work

In this paper, we designed and evaluated predictive, adaptive bandwidth reservation for hand-offs and admission control so as to keep the hand-off dropping probability below a pre-specified value. Our schemes utilize the following two components to reserve bandwidth for hand-offs: (1) *hand-off estimation functions* which are used to predict a mobile's next cell and estimate its sojourn time probabilistically based on its previously-resided cell and the observed history of hand-offs in each cell; and (2) *mobility estimation time window control scheme* in which, depending on the observed hand-off drops, the estimation time window size is controlled adaptively for efficient use of bandwidth and effective response to (1) time-varying traffic/mobility and (2) inaccuracy of mobility estimation.

We considered three different admission-control schemes depending on how many neighboring BSs participate in the admission decision of a new connection request. Through the performance and complexity comparisons, we concluded a hybrid one is superior to the others. Our best scheme is not optimal in the sense that there might be a better scheme resulting in a lower connection blocking probability while keeping the hand-off dropping probability below the target value. However, this scheme is not complex nor based on any impractical assumptions, and hence it is readily implementable. It is also shown to be robust and work well under a variety of traffic loads, connection bandwidths, and mobility.

We plan to evaluate our scheme in more realistic and general environments with two-dimensional cellular structures. This will include more realistic moving patterns of users (e.g., combined vehicular, pedestrian, and stationary mobiles) and their effects. Our scheme can be extended to utilize more information of user mobility. For example, mobiles' path/direction information — readily available from certain applications, such as the route guidance system of the Intelligent Transportation Systems (ITS) with the Global Positioning System (GPS) — can also be utilized in our scheme. Then, the mobility estimation function is used to estimate the sojourn time of a mobile only because the next cell of the mobile is known already. The modification of the proposed scheme to be used in the CDMA systems is also planned, where hand-off drops can be reduced due to (1) soft capacity notion and (2) soft hand-off support. We will integrate our work with routing and re-routing in the wired networks by considering bandwidth reservation in the wired links along the routes of hand-off connections.

Computational complexity of our scheme is reported in [4] as a part of the comparison study with other existing schemes in [10, 14]. In that paper, the three schemes are compared quantitatively through extensive simulations in terms of: (1) performance measures P_{CB}, P_{HD}, and B_u; (2) dependency on the design parameters; (3) dependency on the mobility estimation accuracy; and (4) complexity.

References

[1] A. Banerjea *et al.*, "The Tenet real-time protocol suite: design, implementation, and experience," *IEEE/ACM Trans. on Networking*, vol. 4, pp. 1–10, February 1996.

[2] S. Choi and K. G. Shin, "A cellular local area network with QoS guarantees for heterogeneous traffic," in *Proc. IEEE INFOCOM'97*, pp. 1032–1039, April 1997.

[3] S. Choi and K. G. Shin, "Uplink CDMA systems with diverse QoS guarantees for heterogeneous traffic," in *Proc. ACM/IEEE MobiCom'97*, pp. 120–130, September 1997.

[4] S. Choi and K. G. Shin, "Comparison of connection-admission control schemes in the presence of hand-offs in cellular networks," submitted for publication, April 1998.

[5] D. Hong and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized procedures," *IEEE Trans. on Vehicular Technology*, vol. 35, pp. 77–92, August 1986.

[6] K. Lee, "Supporting mobile multimedia in integrated service networks," *ACM Wireless Networks*, vol. 2, pp. 205–217, 1996.

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

[8] S. Lu and V. Bharghavan, "Adaptive resource management algorithms for indoor mobile computing environments," in *Proc. ACM SIGCOMM'96*, pp. 231–242, August 1996.

[9] S. Lu, V. Bharghavan, and R. Srikant, "Fair scheduling in wireless packet networks," in *Proc. ACM SIGCOMM'97*, pp. 63–74, September 1997.

[10] M. Naghshineh and M. Schwartz, "Distributed call admission control in mobile/wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 14, pp. 711–717, May 1996.

[11] K. Pahlavan and A. H. Levesque, *Wireless Information Networks*. New York, NY: Wiley-Interscience, 1995.

[12] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 3rd ed., 1991.

[13] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: the single-node case," *IEEE/ACM Trans. on Networking*, vol. 1, pp. 344–357, June 1993.

[14] A. K. Talukdar, B. R. Badrinath, and A. Acharya, "On accommodating mobile hosts in an integrated services packet network," in *Proc. IEEE INFOCOM'97*, pp. 1048–1055, April 1997.

[15] A. J. Viterbi, *CDMA: Principles of Spread Spectrum Communication*, Addison-Wesley, Reading, MA, 1995.

[16] Q. Zheng and K. G. Shin, "On the ability of establishing real-time channels in point-to-point packet-switched networks," *IEEE Trans. on Communications*, vol. 42, pp. 1096–1105, February/March/April 1994.