Wireless Networks 6 (2000) 289–305

A comparative study of bandwidth reservation and admission control schemes in QoS-sensitive cellular networks *

Sunghyun Choi [a,**] and Kang G. Shin [b]

[a] *Philips Research-USA, Briarcliff Manor, New York 10510, USA*
[b] *Real-Time Computing Laboratory, Department of Electrical Engineering and Computer Science, The University of Michigan, Ann Arbor, Michigan 48109-2122, USA*

This paper compares five different schemes – called CHOI, NAG, AG, BHARG, and NCBF – for reserving bandwidths for handoffs and admission control for new connection requests in QoS-sensitive cellular networks. CHOI and NAG are to keep the handoff dropping probability below a target value, AG is to guarantee no handoff drops through per-connection bandwidth reservation, and BHARG and NCBF use another type of per-connection bandwidth reservation. CHOI predicts the bandwidth required to handle handoffs by estimating possible handoffs from adjacent cells, then performs admission control for each newly-requested connection. On the other hand, NAG predicts the total required bandwidth in the current cell by estimating both incoming and outgoing handoffs at each cell. AG requires the set of cells to be traversed by the mobile with a newly-requested connection, and reserves bandwidth for each connection in each of these cells. The last two schemes reserve bandwidth for each connection in the predicted next cell of a mobile where the two schemes use different admission control policies. We adopt the history-based mobility estimation for the first two schemes. Using extensive simulations, the five schemes are compared quantitatively in terms of (1) handoff dropping probability, connection-blocking probability, and bandwidth utilization; (2) dependence on the design parameters; (3) dependence on the accuracy of mobility estimation; and (4) complexity. The simulation results indicate that CHOI is the most desirable in that it achieves good performance while requiring much less memory and computation than the other four schemes.

1. Introduction

Establishment and management of connections are crucial issues in QoS-sensitive cellular networks because users are expected to move around during communication sessions experiencing handoffs between cells. The current trend in cellular networks is to shrink cell size in order to accommodate more mobile users in a given geographical area. This results in more frequent handoffs, and makes connection-level QoS more difficult to achieve. Two important connection-level QoS parameters are the probability P_{CB} of blocking newly-requested connections and the probability P_{HD} of dropping handoffs due to the unavailability of channels in the new cell. As in a wired network with QoS guarantees, mobile users, once their connections are set up, should be able to continue communication as long as they want.

Since it is practically impossible to completely eliminate handoff drops, the best one can do is to provide some form of *probabilistic* QoS guarantees. Recently, two connection-admission schemes have been proposed to keep the handoff dropping probability below a target value $P_{\mathrm{HD,target}}$. Limiting P_{HD} below $P_{\mathrm{HD,target}}$ will henceforth be called the *design goal*. Both schemes are based on the estimation of handoffs that may occur during a specific time window. First, using the scheme proposed in [1] (referred to as CHOI), the base station (BS) of a cell calculates the required bandwidth to be reserved for anticipated handoffs from adjacent cells upon arrival of a new connection request.[1] The mobility (i.e. handoff behavior) of each user is estimated using a history of handoffs observed in each cell. Using this estimation, one can compute the bandwidth required to handle the handoffs that are predicted to occur within a specific time window. It also adaptively controls the window size depending on the observed handoff dropping events.

In the second scheme proposed in [6] (referred to as NAG), the BS considers not only incoming handoffs from adjacent cells, but also outgoing handoffs to adjacent cells. The BS then calculates the total required bandwidth in its cell for both handed-off and existing connections. Originally, this scheme was evaluated based on: (1) an exponentially-distributed time each mobile stays in a cell; and (2) the perfect knowledge about the mobility and lifetime of each user connection, i.e. known handoff and connection-termination rates. Under these assumptions, NAG was shown to meet the design goal of keeping P_{HD} below a target value. However, these two assumptions do not usually hold in reality, and hence, we adopt the history-based mobility estimation scheme developed for CHOI under more realistic assumptions.

NAG may appear to be superior to CHOI because it considers more states on the mobility in each cell. How-

* 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.

** This work was done while the author was with the University of Michigan.

[1] Three alternative admission-control schemes were considered in the paper. We refer to the best scheme AC3 of these three as CHOI here.

ever, as we shall show later, CHOI performs as good as, and requires much less resources than, NAG. The former requires much more computation and memory to keep P_{HD} below $P_{\text{HD,target}}$ over a variety of traffic loads, and it is very sensitive to the choice of a design parameter. In contrast, CHOI is found to be insensitive to inaccuracies in mobility estimation and achieve the design goal with much less computation and memory than NAG.

Also considered is an admission-control scheme (referred to as AG) which guarantees no handoff drop for any existing connection. Using the first two schemes, it is not possible to completely eliminate handoff drops. No handoff drops can be achieved only by checking bandwidth availability and reserving each connection's bandwidth in *all* cells the mobile (which is requesting a new connection) is to traverse in future. It is practically impossible to know these cells in advance during the admission-control phase. The basic concept of this scheme was proposed in [9] assuming the availability of such information. We will show how costly it is to make the handoff dropping probability zero even under this impractical assumption. The fourth admission-control scheme (referred to as BHARG) is based on per-connection bandwidth reservation [4,5]. This scheme, unlike the first three schemes described above, does not have any specific design goal. The next cell each mobile will move to is predicted, and the mobile's per-connection bandwidth is reserved in the cell. By doing this, it is possible to reduce P_{CB} to almost zero. In fact, the authors of [4] proposed to use this per-connection bandwidth and admission control when the next cell of a mobile can be predicted, and to use a variant of NAG when it is not. The last admission-control scheme (referred to as NCBF) is a slight twist of BHARG. In this scheme, both the current cell and the predicted next cell of the new connection-requesting mobile should have enough bandwidth to admit the request while BHARG requires only the current cell to have enough bandwidth. It will be shown that both BHARG and NCBF are still costly as compared to the first two schemes due to their per-connection bandwidth reservation requirement.

There is one more scheme that limits P_{HD} below a target [3]. It uses the "shadow cluster" concept to estimate future resource requirements and perform admission control to limit P_{HD}, in which the shadow cluster is a set of cells around a mobile. This scheme is based on availability of the accurate knowledge of each user's mobility, depending on his/her location and time. The mobility estimation used here may provide this scheme with the needed knowledge of mobility, but it is unclear how it will work if the knowledge is not accurate, as is usually the case when the history-based mobility estimation is used. Also, the scheme did not address clearly how to determine the shadow cluster. Moreover, the scheme is computationally too expensive to be practical, as compared to the five schemes considered here.

The notion of bandwidth reservation for handoffs and admission control for new connections was introduced in the mid-eighties [2]. In this scheme, a portion of the link capacity is permanently reserved for handoffs. It was shown that this static reservation scheme is optimal in minimizing a linear objective function of the connection blocking probability and the handoff dropping probability when both new and handoff connection arrivals are Poisson, and connection durations are exponentially distributed [8]. As shown in [1], this is not effective enough to handle a variety of connection bandwidths, traffic loads, and user's mobility. Basically, any form of QoS cannot be guaranteed with this scheme. CHOI and NAG were claimed – in [1] and [6], respectively – to be superior to the conventional static bandwidth reservation scheme.

The paper is organized as follows. Section 2 states the system specifications and assumptions. The users' mobility estimation based on an aggregate history of observations is presented in section 3. Section 4 describes CHOI, section 5 describes NAG utilizing mobility estimation, and section 6 presents three per-connection bandwidth reservation schemes, AG, BHARG, and NCBF. Section 7 quantitatively compares these five bandwidth reservation and admission-control schemes. Finally, the paper concludes with section 8.

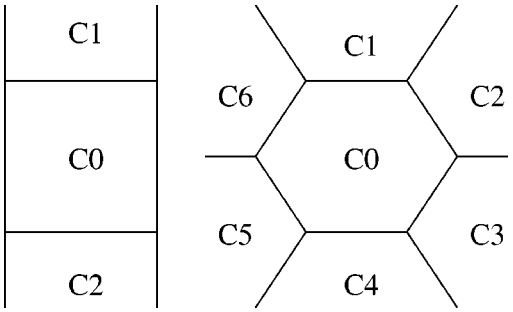
2. System model

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 residing in a cell, communicates through its current BS with another party, which may be a node connected to the wired network or another mobile. When a mobile moves into an adjacent cell in the middle of a communication session, a handoff will enable the mobile to maintain seamless connectivity to its communication partner, i.e. the mobile will continue to communicate through the new BS, preferably without noticing any difference. A handoff could fail due to insufficient bandwidth available in the new cell, and in such a case, a *connection handoff drop* occurs. Here, we preclude delay-insensitive applications, which can tolerate long handoff delays in case of insufficient bandwidth in the new cell at the time of handoff.

For simplicity, BSs are assumed to be fully-connected so that they communicate with each other through the wired links. However, this assumption is not always required as discussed in [1], and will not affect the results in this paper. Under this assumption, the admission control considered in this paper can be performed by each BS, which receives a new connection request from a mobile in its cell. All cells around a cell A are indexed:[3] A is labeled with 0, and the others with numbers beginning 1 as shown in figure 1. Let $C_{i,j}$ be the jth connection in cell i and $b(C_{i,j})$ be its re-

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

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



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

Figure 1. Indexing of cells.

quired bandwidth. For simplicity, we assume that a mobile does not have multiple simultaneous connections, so that by an *active mobile*, we mean a mobile with one existing connection.[4]

The cellular system uses a fixed channel allocation (FCA) scheme, and each cell has a wireless link capacity C. The unit of bandwidth is BU, which is the required bandwidth to support one voice connection. A connection runs through multiple wired and wireless links, and hence, we need to consider the admission control on both wireless and wired links. For a new connection to be admitted, the admission tests on all the nodes along the route of the connection (traversing both wireless and wired links) should be positive. However, we will confine ourselves to the admission control on the wireless link in each cell, because routing and/or rerouting upon handoff of a connection is beyond the scope of this paper. The schemes considered here can be easily extended to include the admission control on wired links by considering the routing and rerouting inside the wired network.

3. Mobility estimation

The direction and speed of active mobiles are, in general, unknown to the underlying wired network (or BSs). However, for effective admission control with our design goal, it is necessary to have a good mobility-estimation scheme. We introduce here the mobility-estimation scheme [1] that is based on a history of handoffs observed in each cell. This scheme is motivated by road traffic: the mobility in terms of a mobile's speed and direction in a cell is probabilistically similar to that of those mobiles that came from the same previous cell and are now residing in the same cell. The rationale behind this scheme is the existence of the traffic signals and/or signs (e.g., speed limits) and the possible correlation between mobiles' previous and future paths. This scheme might not produce very accurate mobility estimation due to its dependency on the observation, but is feasible in practice, and was found to work well with the scheme CHOI [1].

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

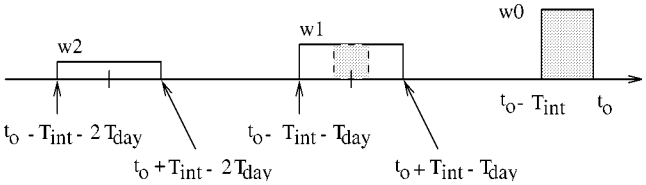


Figure 2. An example of periodic windows to obtain handoff estimation functions with $N_{\text{win\_days}} = 2$.

3.1. Handoff estimation functions

We now describe how to estimate and predict mobility. This function will be executed by the BS of each cell in a distributed manner. For each mobile which moves into a neighbor cell from the current cell 0, the current cell's BS caches the mobile's quadruplet, (T_{event}, *prev*, *next*, T_{soj}), called a *handoff event quadruplet*, where T_{event} is the time 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.

From the cached quadruplets, the BS builds *handoff estimation function*, which describes the estimated distribution of the next cell and sojourn time of a mobile, depending on the cell the mobile previously resided in. 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 handoff estimation function at the current time t_o is obtained as follows: for a quadruplet (T_{event}, *prev*, *next*, T_{soj}) such that

$$t_o - T_{\text{int}} - nT_{\text{day}} \leqslant T_{\text{event}} < t_o + T_{\text{int}} - nT_{\text{day}}, \quad (1)$$

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 ($\geqslant 0$) is an integer,

$$F_{\text{HOE}}(t_o, \textit{prev}, \textit{next}, T_{\text{soj}}) := w_n, \quad (2)$$

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

In practice, it is desirable to limit the number of the quadruplets (1) used for handoff estimation and (2) currently not used for handoff estimation, but cached for future use, e.g., those with $t_o + T_{int} - T_{day} < T_{event} < t_o - T_{int}$ in figure 2, in order to reduce the memory and computation complexity.[5] We define the *maximum handoff estimation function size*, N_{quad}, as the maximum number of handoff event quadruplets used for handoff estimation for each *prev*. This implies that we do not 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 handoff estimation with the following priority rule. First, the quadruplet that satisfies equation (1) with a smaller n gets higher priority. Second, among those satisfying equation (1) with the same n, the quadruplet with a smaller $|T_{event} - nT_{day}|$ gets higher priority. Figure 2 shows an example that only the quadruplets with the event times T_{event} within the shaded regions are used for handoff estimation according to the priority rule, implying that the total number of quadruplets within the regions be N_{quad}. In order to reduce the caching memory size, those quadruplets observed at time t', (i.e. $T_{event} = t'$), when the handoff estimation function at time t' does not use any quadruplets observed during previous days, are not cached for future use, because they are unlikely to be used for handoff estimation 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 handoff estimation function during the last $(T_{day} + T_{int})$ can be deleted from the cache entries.

Figure 3 shows an example of footprint of the handoff estimation function for *prev* $= 1$ without showing the values of w_n's. The handoff estimation function in a three-dimensional space will have different heights, depending on the values of w_n's. The example is drawn from the same indexing as shown in figure 1(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 neighbors of cell 0 since the sojourn times before entering cell 4 are generally shown to be the largest. Note that the handoff estimation function for 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. Then, the probability that a connection which arrives from cell *prev* at time t_o, will reside in the current cell for t_{soj}, where $T_{min} < t_{soj} \leqslant T_{max}$, and depart to cell *next* can be estimated by

$$\Pr(T_{min} < t_{soj} \leqslant T_{max} \ \& \text{ departure to cell } next) = \frac{\sum_{T_{min} < t_{soj} \leqslant T_{max}} F_{HOE}(t_o, prev, next, t_{soj})}{\sum_{next' \in \mathbf{A}_0} \sum_{0 < t_{soj} < \infty} F_{HOE}(t_o, prev, next', t_{soj})}, \quad (3)$$

where $\mathbf{A}_0$ is the set of indices of cell 0's neighbors.

[5] The calculations required for mobility estimation will be dependent on the number of the quadruplets used for the handoff estimation function as will be shown in the next section.

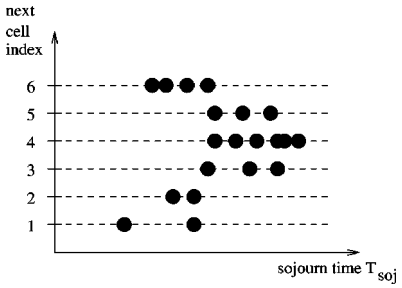


Figure 3. An example of the footprint of handoff estimation function for *prev* $= 1$.

4. Admission control with estimation of incoming handoffs only

We first introduce the admission control scheme CHOI in [1] to keep P_{HD} below $P_{HD,target}$ by utilizing the handoff estimation function described thus far.

4.1. Target reservation bandwidth

This 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 an active mobile with connection $C_{0,j}$ is estimated with the probability, $p_h(C_{0,j} \to i)$, that $C_{0,j}$ hands off into cell i within T_{est}.

The handoff probability can be computed using the handoff 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*. Connection $C_{0,j}$'s extant sojourn time, $T_{ext\_soj}(C_{0,j})$, is the time elapsed since the active mobile with connection $C_{0,j}$ entered the current cell. Using Bayes' theorem [7], the handoff probability $p_h(C_{0,j} \to next)$ at time t_o is calculated by

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

where *prev*$(C_{0,j})$ is the cell in which $C_{0,j}$ resided before entering the current cell and $\mathbf{A}_i$ is the set of indices of cell i's neighboring 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 $t_{soj} > T_{ext\_soj}(C_{0,j})$. This is the handoff probability $p_h(C_{0,j} \to next)$.

Figure 4 shows an example of calculating $p_h(C_{0,j} \to 4)$, when $C_{0,j}$ entered cell 0 from cell 1, using the footprint of the handoff estimation function for *prev*$(C_{0,j}) = 1$, shown in figure 3. 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

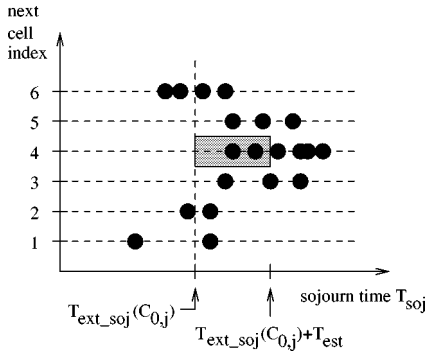


Figure 4. An example of calculating $p_h(C_{0,j} \rightarrow next)$ when $prev(C_{0,j}) = 1$ and $next = 4$ using the footprint of $F_{HOE}(t_o, 1, next', T_{soj})$.

light shaded regions) are summed to obtain the denominator in equation (4). Because this value is nonzero, 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 equation (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. nonmoving) in cell 0 if there is no handoff event in the handoff estimation function with a sojourn time larger than the connection $C_{0,j}$'s extant sojourn time, i.e. the denominator in equation (4) is zero.

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

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

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 bandwidth to be reserved in cell 0 for the expected handoffs from adjacent cells within the estimation time T_{est}, is calculated as

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

Note that $B_{r,0}$ is a target, not the actual reserved bandwidth, since a cell may not be able to reserve the target bandwidth. This can happen because a BS can control the admission of only newly-requested connections, not those connections handed off from adjacent cells.

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 yielding the least connection-blocking probability while keeping the handoff dropping probability below the target. In this scheme, the estimation time will be adjusted adaptively in each cell independently of others, depending on the handoff dropping events in the cell as described in the next subsection. Then, the estimation time T_{est} of cell *next* (or $T_{est,next}$) will be used in equation (4). So, when

```
01. if (w = ⌈1/P_HD,target⌉), then w_obs := w;
02. T_est := T_start;  n_H := 0;  n_HD := 0;
03. while (time increases) {
04.     if (handoff into the current cell happens) then {
05.         n_H := n_H + 1;
06.         if (it is dropped) then {
07.             n_HD := n_HD + 1;
08.             if (n_HD > w_obs/w) then {
09.                 w_obs := w_obs + w;
10.                 if (T_est < T_soj,max) then T_est := T_est + 1;
11.             }
12.         }
13.         else if (n_H ⩾ w_obs) then {
14.             if (n_HD ⩽ w_obs/w and T_est > 1) then
15.                 T_est := T_est − 1;
16.             w_obs := w;  n_H := 0;  n_HD := 0;
17.         }
18.     }
19. }
```

Figure 5. A pseudo-code of the algorithm to adjust T_{est} in each BS.

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 use equation (5) to calculate the required bandwidth for the expected handoffs from that cell, (i.e. $B^i_{r,0}$ for cell i) and will inform cell 0's BS of this value. Finally, cell 0's BS will calculate $B_{r,0}$ using equation (6).

4.2. Control of mobility estimation time window

Using the bandwidth reservation described above, the bandwidth for handoffs 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}. Hence, an adaptive algorithm is used to control the mobility estimation time window size based on the handoff dropping events in each cell so as to approximate the optimal T_{est} over time. Figure 5 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 handoffs, n_H, and handoff drops, n_{HD}, are reset to 0. As can be found in the pseudo-code of figure 5, w_{obs} is increased or decreased by the amount w, and the constraint $P_{HD} < P_{HD,target}$ can be translated into that to keep the counted number n_{HD} of handoff drops out of w_{obs} observed handoffs below w_{obs}/w. During the runtime, whenever there is a handoff drop after w_{obs}/w drops, the BS sets $T_{est} := T_{est} + 1$ and $w_{obs} := w_{obs} + w$. On the other hand, when there were less than, or equal to, w_{obs}/w handoff drops out of w_{obs} observed handoffs, $T_{est} := T_{est} - 1$ and $w_{obs} := w$. T_{est} is not greater than $T_{soj,max}$ in figure 5, which is the maxi-

mum T_{soj} derived from the handoff estimation functions in adjacent cells, because any value larger than that is meaningless. The minimum value of T_{est} is also set to 1 since if the value is too small, virtually no bandwidth will be reserved irrespective of the existing connections in adjacent cells.

4.3. Admission control

The basic idea of the admission decision is to check if there is enough bandwidth left unused after reserving the target reservation bandwidth. However, for the admission control of a newly-requested connection in a cell, sometimes it is required to check the reservation bandwidth in adjacent cells as well. Otherwise, the continuous connection admissions in a cell may result in continuous handoff drops in adjacent cells, thus violating the design goal, as discussed in [1].

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 recalculate $B_{r,i}$, and participate in the admission test. Now, for a new connection request, the admission test is performed as follows:

T1. For all $i \in \mathbf{A}_0$ such that $\sum_{j\in\mathbf{C}_i} b(C_{i,j}) + B_{r,i}^{curr} > C$, calculate $B_{r,i}$ newly, set $B_{r,i}^{curr} := B_{r,i}$, and check if $\sum_{j\in\mathbf{C}_i} b(C_{i,j}) \leqslant C - B_{r,i}$.

T2. Check if $\sum_{j\in\mathbf{C}_0} b(C_{0,j}) + b_{new} \leqslant C - B_{r,0}$.

T3. If all the tests are positive, the connection is admitted.

5. Admission control with estimation of incoming and outgoing handoffs

We now describe the distributed admission control scheme (referred to as NAG), originally proposed in [6], which utilizes the cell-specific history-based mobility estimation. Described here is more generalized than the original scheme in the sense that heterogeneous connections (in terms of connection bandwidths) are supported. The authors of [4] also presented another generalized version of the original scheme with a number of connection bandwidths. All of the previously-reported performance evaluations were based on exponentially-distributed sojourn times of mobiles in each cell and known connection handoff/termination rates.

5.1. Three state probabilities

The main difference between CHOI and NAG is that CHOI considers incoming handoffs only while NAG considers both incoming and outgoing handoffs in a cell. NAG is also based on the estimated mobility during $[t_o, t_o + T_{est}]$, in which t_o is the current time. Like in CHOI, we consider the behavior of a connection in the current cell. After T_{est} time units, connection $C_{0,j}$ can be in one of three different states with the corresponding probabilities shown in parentheses: (1) handoff into an adjacent cell i ($p_h(C_{0,j} \to i)$); (2) termination after completing the corresponding communication ($p_e(C_{0,j})$); and (3) staying in the current cell ($p_s(C_{0,j})$). We compute the probability of each event by utilizing the mobility estimation.

First, the handoff probabilities $p_h(C_{0,j} \to i)$ are defined in equation (4) for CHOI. Next, we consider how to estimate the probability that connection $C_{0,j}$ will terminate within time T_{est}, i.e. $p_e(C_{0,j})$. BSs utilize the average connection lifetime T_{ave_life} of each mobile, which is calculated over time by

$$T_{ave\_life} := (1-\alpha)T_{ave\_life} + \alpha T_{last\_life}, \tag{7}$$

where α (< 1) is a design parameter, and T_{last_life} is the connection lifetime obtained from the last connection of that mobile. We assume that the connection lifetime of $C_{0,j}$ follows an exponential distribution with mean $T_{ave\_life}(C_{0,j})$. In reality, the connection lifetime might not follow an exponential distribution, but this will be most likely dependent on each mobile, not on the cell in which it resides. Hence, this assumption does not have significant bearing on the results. Then, the probability is given by

$$p_e(C_{0,j}) = 1 - e^{-T_{est}/T_{ave\_life}(C_{0,j})}. \tag{8}$$

Finally, the probability that connection $C_{0,j}$ will stay in the cell for T_{est} time units is given by

$$p_s(C_{0,j}) = \big(1 - p_e(C_{0,j})\big)\left(1 - \sum_{i\in\mathbf{A}_0} p_h(C_{0,j} \to i)\right), \tag{9}$$

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

We assume that (1) the behavior of each connection is independent of others, and (2) the probability that a mobile hands off more than once during time T_{est} is negligible. Then, the required bandwidth $B_{T_{est},0}$ for handed-off and existing connections in cell 0 during T_{est} will be the sum of the bandwidths of (1) the connections which stay in cell 0 during T_{est} and (2) the connections which hand off into cell 0 from an adjacent cell during T_{est}. Using the Central Limit Theorem [7], this can be approximated to have a Gaussian distribution as

$$\Pr_{B_{T_{est},0}}(k) \approx G(m_{B,0}, \sigma_{B,0}), \tag{10}$$

where the mean

$$m_{B,0} = \sum_{i\in\mathbf{A}_0}\sum_{j\in\mathbf{C}_i} b(C_{i,j})p_h(C_{i,j} \to 0) + \sum_{j\in\mathbf{C}_0} b(C_{0,j})p_s(C_{0,j}), \tag{11}$$

and the variance

$$\sigma_{B,0}^2 = \sum_{i\in\mathbf{A}_0}\sum_{j\in\mathbf{C}_i} b^2(C_{i,j})p_h(C_{i,j} \to 0)\big(1 - p_h(C_{i,j} \to 0)\big) + \sum_{j\in\mathbf{C}_0} b^2(C_{0,j})p_s(C_{0,j})\big(1 - p_s(C_{0,j})\big). \tag{12}$$

Recall that $b(C_{i,j})$ is the connection $C_{i,j}$'s bandwidth, $\mathbf{C}_i$ is the set of connections' indices in cell i, and $\mathbf{A}_i$ is the set of cell i's neighbors' indices.

5.2. Admission control

To make an admission decision, we define the overload probability after T_{est} in cell i as follows:

$$P_{\text{O},i} = \Pr(B_{T_{\text{est}},i} > C) \approx Q\left(\frac{C - m_{B,i}}{\sigma_{B,i}}\right), \qquad (13)$$

where C is the link capacity. $m_{B,i}$ and $\sigma_{B,i}$ are obtained from equations (11) and (12), respectively, after replacing i with k, then replacing 0 with i in the equations. Now, for a new connection request, the admission test is performed as follows:

T1. For all $i \in \mathbf{A}_0 \cup \{0\}$, check if $P_{\text{O},i} \leqslant P_{\text{HD,target}}$.

T2. If all the tests are positive, the connection is admitted.

Note that for this scheme, the specific amount of bandwidth to be reserved is not defined. So, the relation between the value of T_{est} and the bandwidth reserved for handoffs is not clear. Basically, the larger T_{est}, the larger P_{h}'s and P_{e}'s, hence the smaller P_{s}'s. It is not clear whether m_B and σ_B^2 would increase or decrease as T_{est} increases. There may exist an optimal T_{est} which achieves the smallest P_{CB} while keeping P_{HD} under the target value, but it is not possible to adopt a similar scheme to the mobility estimation window control used for CHOI. We will later evaluate the effect of the value of T_{est} using simulations.

6. Per-connection bandwidth reservation

Now, we describe three admission-control schemes based on per-connection bandwidth reservation: AG, BHARG, and NCBF.

6.1. Control AG: No handoff drop

This subsection describes an admission-control scheme (referred to as AG, meaning "Absolute Guarantee") which guarantees no handoff drop. This is possible by checking the bandwidth in all cells which the mobile requesting a new connection will traverse, then reserving the required bandwidth in each of those cells. So, this admission scheme involves per-connection bandwidth reservation in each cell. This per-connection reservation and the corresponding admission control were proposed in the context of measurement-based admission control in [9].

For this scheme to work, each mobile should inform the wired network (or the corresponding BS) of the *mobility specification* that is composed of the cells the mobile will traverse during the lifetime of the requested connection. It is generally impossible to know a mobile's direction in advance. The navigation systems [10] of Intelligent Transportation Systems (ITS) might be used to predict the mobiles' path/direction accurately, and might be used to predict the mobility specification. The problem is that using a navigation system, it is possible to know the cell to which the corresponding mobile will move next, but we do not know if the mobile's connection will continue when the mobile enters the next cell. So, it is practically impossible to know the exact mobility specification at the time of admission control. But, we describe the admission-control scheme assuming the availability of the mobility specification as in [9].

For the mobility specification M_{sp} of a newly-requested connection, which consists of a set of cells, and its required bandwidth b_{new}, admission control and per-connection bandwidth reservation are as follows:

T1. For each cell i in the mobility specification M_{sp}, check if $\sum_{j\in\mathbf{C}_i} b(C_{i,j}) + b_{\text{new}} \leqslant C - B_{\text{r},i}$.

T2. If all the above tests are positive, for each cell i in the mobility specification M_{sp}, $B_{\text{r},i} := B_{\text{r},i} + b_{\text{new}}$, and the connection is admitted.

Here $B_{\text{r},i}$ is the sum of all per-connection bandwidths reserved in cell i. Whenever a mobile enters a cell, the cell's reserved bandwidth (for handoffs) will be decreased: upon handoff of connection $C_{i,j}$ into cell i, $B_{\text{r},i} := B_{\text{r},i} - b(C_{i,j})$.

Note that the cell index i used in this subsection is different from the relative index defined in section 2 and used for the previously-described two schemes. Cell i here should be considered as the ith cell in the entire cellular system. Through per-connection reservation in the cells within the mobility specification, it is possible to make the handoff drop probability zero, but we will show how inefficient this scheme is in terms of the bandwidth utilization and the connection-blocking probability.

6.2. BHARG: Per-connection reservation in next cell after admission

This subsection describes the second admission control scheme based on per-connection reservation referred to as BHARG. This scheme does not try to limit P_{HD} nor to eliminate handoff drops, but just reserves each connection's bandwidth in the predicted next cell of the mobile which has an on-going connection. The key aspect of this scheme is how to predict the next cell of a mobile, and it was proposed for indoor mobile computing environments [4,5]. We assume here that a perfect next-cell estimator, which informs the BS whether a mobile is terminating its connection in the current cell or moving into an adjacent cell with the connection, is available to evaluate the performance of per-connection bandwidth reservation. Admission control and per-connection bandwidth reservation work as follows:

T1. Check if $\sum_{j\in\mathbf{C}_0} b(C_{0,j}) + b_{\text{new}} \leqslant C - B_{\text{r},0}$.

T2. If the above test is positive, for the predicted next cell *next* of the connection, $B_{\text{r},next} := B_{\text{r},next} + b_{\text{new}}$, and the connection is admitted.

Here $B_{r,i}$ is the sum of all per-connection bandwidths reserved in cell i. Whenever a mobile enters a cell, the cell's reserved bandwidth (for handoffs) will be decreased: upon handoff of connection $C_{i,j}$ into cell i, $B_{r,i} := B_{r,i} - b(C_{i,j})$.

Note that the admission test checks for bandwidth availability in the mobile's current cell only. Then, the BS in the predicted next cell of the mobile will try to reserve the mobile's connection bandwidth. However, this is not always possible since bandwidth availability in this next cell was not a condition for admitting the connection. In that sense, $B_{r,i}$ in cell i is not a real reserved bandwidth, but a target reservation bandwidth. Even though this scheme was not aimed for no handoff drops, it will achieve virtually no handoff drops as will be shown later, but at a very high cost which is comparable to that of AG.

6.3. NCBF: Per-connection reservation in next cell before admission

This subsection describes the third per-connection reservation scheme referred to as NCBF (meaning "Next Cell Bandwidth reservation First"). This scheme lies in between the previous two schemes since it predicts the next cell of a mobile requesting a new connection during the admission control phase, and that mobile is admitted only when both the current cell and the predicted next cell of the mobile have enough bandwidth to support the requested connection. The difference between NCBF and AG is that the former reserves the bandwidth in the next cell only, and the difference between NCBF and BHARG is that the former will not admit the new connection if the predicted next cell does not have enough bandwidth to support the requested connection. This scheme does not try to limit P_{HD} nor to eliminate handoff drops. The next-cell prediction schemes [4,5] proposed for BHARG are expected to be used for this scheme as well since it is a slight twist of BHARG. We again assume that a perfect next-cell estimator, which informs the BS whether a mobile is terminating its connection in the current cell or moving into an adjacent cell with the connection, is available to evaluate the performance of this per-connection bandwidth reservation. Admission control and per-connection bandwidth reservation work as follows:

T1. For $i \in \{0, next\}$ where *next* is the index of the predicted next cell, check if $\sum_{j \in \mathbf{C}_i} b(C_{i,j}) + b_{new} \leqslant C - B_{r,i}$.

T2. If both of the above tests are positive, the connection is admitted, and $B_{r,next} := B_{r,next} + b_{new}$.

Here $B_{r,i}$ is the sum of all per-connection bandwidths reserved in cell i. Whenever a mobile enters a cell, the cell's reserved bandwidth (for handoffs) will be decreased: upon handoff of connection $C_{i,j}$ into cell i, $B_{r,i} := B_{r,i} - b(C_{i,j})$.

Note that $B_{r,i}$ in cell i for NCBF is the real reserved bandwidth (so it differs from that for BHARG). Even though this scheme was not aimed to avoid handoff drops, it will also achieve virtually no handoff drops, but at a very high cost comparable to that of AG. This scheme is claimed to be the best per-connection bandwidth reservation scheme.

7. Comparative performance evaluation

This section presents and discusses the comparison results of the five schemes discussed thus far. We first describe the assumptions and specifications used in our simulation study.

7.1. Simulation assumptions and specifications

In the system under consideration, cells are structured as a one- or two-dimensional array. For the one-dimensional case, 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 shown in figure 1(a). For the two-dimensional case, the roads are mapped into a mesh as shown in figure 6. A BS is located at each intersection of two crossing roads. The coverage of each cell is also shown in the figure. This cellular structure can typically be seen in a metropolitan downtown area. First, the following assumptions are made for our simulation study of one-dimensional case:

A1.1. The whole cellular system is composed of 10 linearly-arranged cells, and the diameter of each cell is 1 km. Cells are numbered from 1 to 10, i.e. cell $\langle i \rangle$ represents the ith cell.

A1.2. Connection requests are generated according to a Poisson process with rate λ (connections/s/cell) in each cell. A newly-generated connection can appear anywhere in the cell with an equal probability.

A1.3. A connection is either for voice (requiring 1 BU of bandwidth) or for video (requiring 4 BUs) with probabilities R_{vo} and $1 - R_{vo}$, respectively, where the *voice ratio* $R_{vo} \leqslant 1$.

A1.4. Mobiles can travel in either of two directions with an equal probability with a speed chosen randomly between SP_{min} and SP_{max} (km/h). Each mobile will run straight through the road with the chosen speed, i.e. mobiles will never turn around.

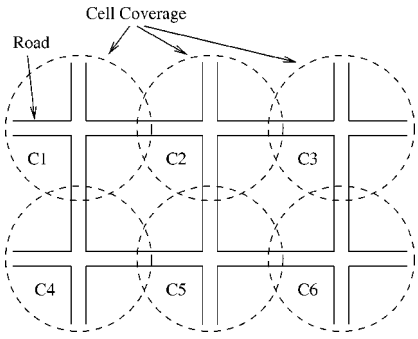


Figure 6. A two-dimensional cellular structure.

A1.5. Each connection's lifetime is exponentially distributed with mean 120 s.

A1.6. Connections are generated and behave in a stationary manner, i.e. there will be no fluctuations in the connection-generation rate and mobility.

A1.7. Each cell has a fixed link capacity 100 BUs (unless stated otherwise).

Next, for the two-dimensional case, we make the following additional assumptions:

A2.1. The whole cellular system is composed of 25 cells arranged as a 5×5 mesh, and the diameter of each cell is 300 m.

A2.2. At the intersection of two roads (i.e. at the center of each cell), a mobile might continue to go straight, or turn left, right, or around with probabilities 0.55, 0.2, 0.2, and 0.05, respectively.

A2.3. If a mobile chooses to go straight or turn right at the center of a cell, it might need to stop there with probability 0.5 for a random time from 0 to 30 s due to a red traffic light. The probability 0.5 roughly represents the fact that one of two crossing roads will see the green light at a time, and the time 30 s represents the duration of a traffic signal light (e.g., the time from the start to the end of a green light).

A2.4. If a mobile chooses to turn left or around, it needs to stop there for a random time from 0 to 60 s due to the traffic signal.

A2.5. Each cell has a fixed link capacity of 50 BUs.

The rationale behind the assumed mobile's delay at the intersection is that there are four traffic signals at the intersection for mobiles arriving from the four directions. A traffic signal will have the red (for stop), left-turn, green (for going straight and turning right) lights in order, then back to the red light. The period from a red to the next red is $60 + \varepsilon$ s in which the red light will last for 30 s, then the turn-left light will turn on for a very short time ε, then, finally, the green light will last for 30 s.

Each simulation run starts without any prememorized handoff event quadruplets. As simulations are run, quadruplets will be collected, and will affect the handoff estimation functions $F_{\text{HOE}}(t, \mathit{prev}, \mathit{next}, T_{\text{soj}})$. Two cases of user mobility are considered: *high* user mobility with $[SP_{\min}, SP_{\max}] = [80, 120]$, and *low* user mobility with $[40, 60]$. Both cases are considered for the one-dimensional structure, but only the low mobility case is considered for the two-dimensional structure since high user mobility is not likely in a downtown area, which is modeled as a two-dimensional structure. Under the above assumptions, the border cells (i.e. cells $\langle 1 \rangle$ and $\langle 10 \rangle$) for the one-dimensional structure 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 $\langle 5 \rangle$ and $\langle 6 \rangle$) 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 assess their operations correctly. So, we connected two border cells, i.e. cells $\langle 1 \rangle$ and $\langle 10 \rangle$, artificially so that the whole cellular system forms a ring architecture as was assumed in [1,6]. For the same reason, two end roads in the two-dimensional structure are also connected. For example, in figure 6, the left-most (upper-most) road in cell $C1$ is connected to the right-most (lower-most) road in cell $C3$ ($C4$).

The parameters used include: $P_{\text{HD,target}} = 0.01$; for the mobility estimation of CHOI and NAG, $N_{\text{quad}} = 100$ (unless stated otherwise), $T_{\text{int}} = \infty$, $N_{\text{win\_days}} = 0$, and $w_0 = 1$; for CHOI, $T_{\text{start}} = 1$ s. The choice of $T_{\text{int}} = \infty$ is reasonable since it was assumed that there is no time-variation in the user mobility and traffic. 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:

$$L = \big(1 \cdot R_{\text{vo}} + 4 \cdot (R_{\text{vo}} - 1)\big) \cdot \lambda \cdot 120. \tag{14}$$

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 0 to $2C$ (i.e. 200 for the one-dimensional structure, and 100 for the two-dimensional structure). Generally, the desirable range of the offered load is less than, or equal to, the link capacity (i.e. 100 BUs in the one-dimensional structure), of each cell. It is undesirable to keep a cell over-loaded (i.e. the offered load is >100) for a long time, and in such a case, the cell must be split into multiple cells to increase the total system capacity. However, cells can get overloaded temporarily. Suppose a mobile user's connection request is blocked once. Then, he/she is expected in most cases to continue to request the connection until it becomes successful or he/she gives up. This likely behavior of mobile users will affect the offered load. Near the offered load $= 100$, P_{CB} will be about, or larger than, 0.1 in most cases, due to some reserved bandwidth for handoffs. In such a situation, if each connection-blocked user attempts to make the connection about 5 times, then the offered load will increase by about 150 in a very short time. Likewise, there might be some cases with the offered load of 200. 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 200 for the one-dimensional structure and 100 for the two-dimensional structure, since even for these large offered loads, the design goal to keep P_{HD} below a target value should be achieved.

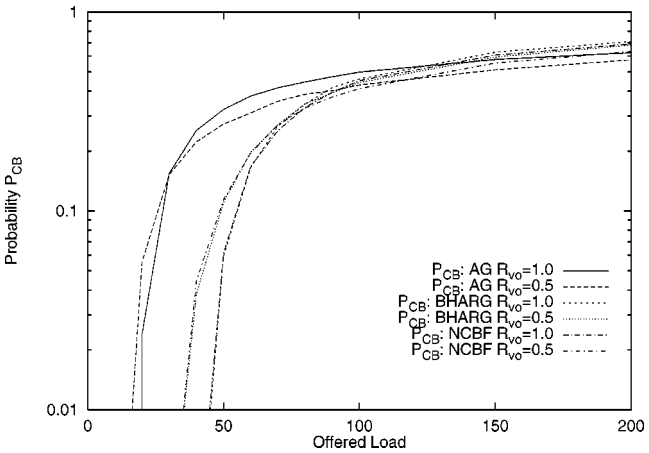
7.2. Simulation results and discussion

We first compare three admission-control schemes based on per-connection bandwidth reservation, and then compare CHOI with NCBF which is claimed to be the best per-connection bandwidth reservation scheme, and, finally,

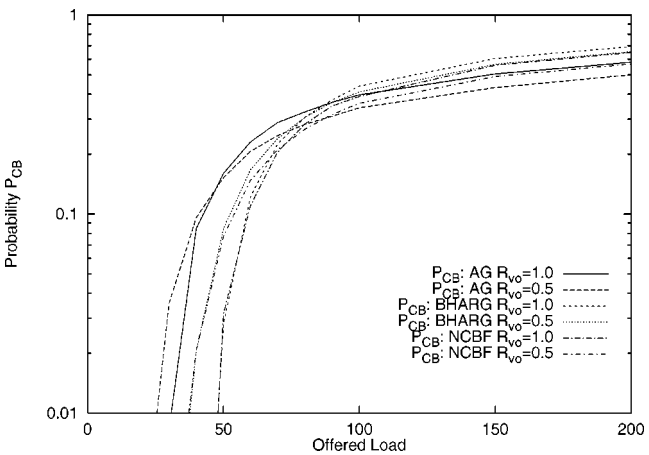
compare CHOI with NAG. The one-dimensional structure is considered for all comparisons. The two-dimensional structure is also considered for the comparison of CHOI and NAG, which needs a more careful comparison, in section 7.2.4. CHOI and NAG were claimed – in [1,6], respectively – to be superior to the conventional static bandwidth reservation scheme, while showing that the static reservation scheme cannot meet the design goal.

7.2.1. Comparison of AG, BHARG, and NCBF

Figure 7 shows P_{CB} of three per-connection bandwidth reservation schemes as the offered load increases for the voice ratio $R_{\mathrm{vo}} = 0.5$ and 1.0. It was observed that P_{HD} of all the three are (virtually) zero irrespective of the offered load, voice ratio, and user mobility, and thus omitted in the plots. In fact, we never observed any handoff drops with AG and NCBF as it should be with AG especially, and observed only two handoff drops throughout the whole simulations involving more than 100,000 handoffs for each simulation run. It should be noted that the handoff drops are (virtually) eliminated at the expense of blocking a large number of new connection requests even in lightly-loaded situations. The fact that P_{CB} is larger than



(a) high mobility



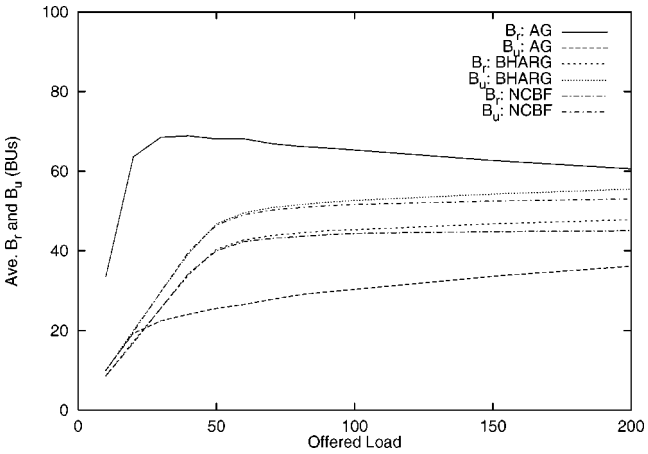
(b) low mobility

Figure 7. Comparison of AG, BHARG, and NCBF using P_{CB} versus offered load.

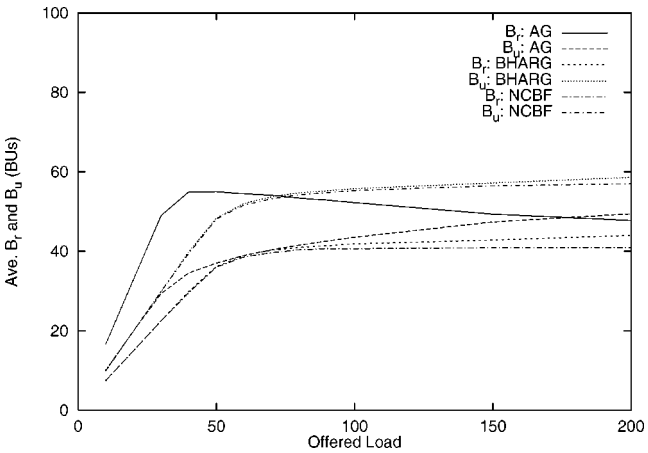
0.1 even for $L = 30$ with AG and for $L = 60$ with the other two schemes where $C = 100$ implies that these schemes severely under-utilize the link capacity. P_{CB}'s of BHARG and NCBF are observed to be less than that of AG. That is basically because AG reserves more bandwidth for handoffs. In fact, P_{CB}'s of BHARG and NCBF get closer to that of AG for the low mobility case since the average number of cells within the mobility specification used for AG is small in this case. Comparing BHARG and NCBF, we observe that P_{CB}'s of both schemes are almost the same in a lightly-loaded region, but that of NCBF is a little bit smaller than that of BHARG in a heavily-loaded region. For example, for $R_{\mathrm{vo}} = 1.0$ and high (low) mobility, P_{CB} of BHARG is about 0.80 (0.79) while that of NCBF is about 0.77 (0.74) at the offered load $L = 300$. This result is not easy to understand since this implies that NCBF reserves less bandwidth for handoffs. Intuitively, NCBF is expected to reserve more bandwidth for handoffs since it checks the bandwidth availability in the predicted next cell, and reserves the bandwidth before admitting a newly-requested connection. This will become clear when we examine the average reserved bandwidth with the next figure.

Figure 8 shows the average (target) reservation bandwidth B_{r} and utilized bandwidth B_{u} by the existing connections as the offered load increases for $R_{\mathrm{vo}} = 1.0$. Note that B_{r} is a target for BHARG while it is the real reserved bandwidth for AG and NCBF. First, BHARG and NCBF work desirably by reserving less bandwidth when the system is lightly-loaded, and increasing the reservation bandwidth as the offered load increases. B_{u} is observed to be larger than B_{r} throughout the whole offered loads examined. For the case of AG, when the system is lightly-loaded, B_{r} is larger than B_{u} because new connections will rarely be blocked in this case, and for each admitted connection, its bandwidth is reserved in all cells within the mobility specification, which includes, on average, more than two cells in our experiments. The number of cells within a connection's mobility specification is dependent on the connection's lifetime and the mobile's speed. Accordingly, B_{r} is found to be smaller for figure 10(b) with low user mobility. Contrary to intuition, B_{r} starts to decrease beyond a threshold offered load even though B_{u} continues to increase. This phenomenon can be explained as follows. After the threshold offered load, the degree of blocking new connection requests becomes severer, implying that a connection with a smaller mobility specification (i.e. a smaller number of cells in its mobility specification) will have a better chance to be admitted. As the offered load increases, connections with large mobility specifications will be more likely to be blocked, and hence, there will be more connections with small mobility specifications in the system. The smaller the mobility specification, the smaller total bandwidth will be reserved throughout the system. So, the bandwidth reservation will decrease with the increase in offered load.

Both B_{u} and B_{r} of NCBF are also observed to be smaller than those of BHARG in moderately-loaded and heavily-loaded regions while they are almost the same in lightly-

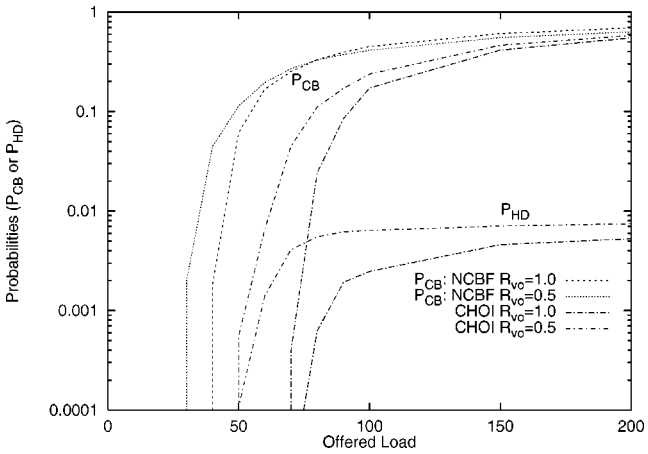


(a) high mobility

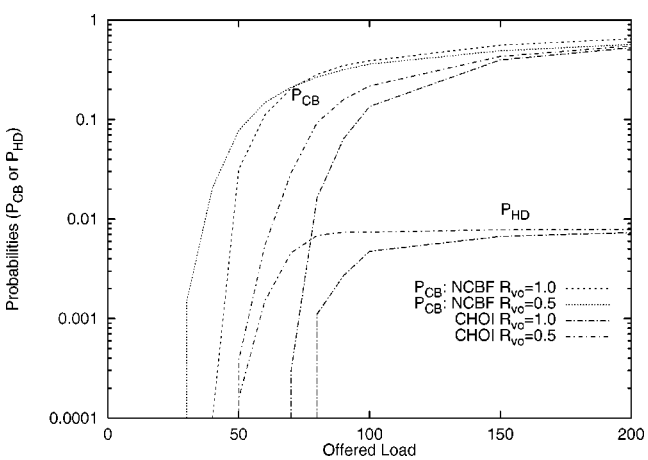


(b) low mobility

Figure 8. Comparison of AG, BHARG, and NCBF using the average B_r and B_u versus offered load for $R_{vo} = 1.0$.



(a) high mobility



(b) low mobility

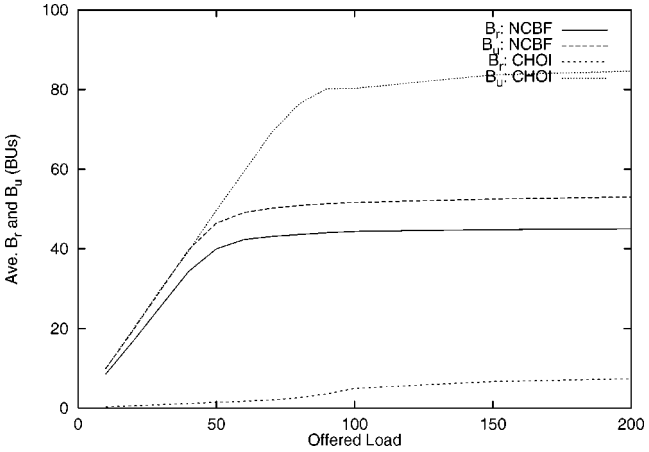
Figure 9. Comparison of NCBF and CHOI using P_{CB} and P_{HD} versus offered load.

loaded regions. This is due to a similar reason why B_u of AG decreases beyond a threshold, i.e. connections which will end in the current cell without incurring any handoff (or connections with relatively shorter connection lifetimes) will have more chance to be admitted since they do not require any bandwidth to be reserved in the predicted next cell while other connections will have less chance to be admitted. Note that BHARG does not differentiate these two types of connections during the admission control phase. Accordingly, BHARG will admit more connections requiring bandwidths to be reserved in the predicted next cells. This is why B_r of BHARG is larger than that of NCBF. Moreover, this explains why P_{CB} of BHARG was larger than that of NCBF in figure 7, since the larger B_r, the larger P_{CB} generally. The reason why B_u of BHARG is larger also can be explained similarly. That is, with NCBF, the admitted connections' lifetimes are shorter than those with BHARG, meaning more connections in the system with BHARG on average for a given number of admitted connections. Even though NCBF admits more connections on average, the effect of shorter connection lifetimes appears stronger. Comparing these three schemes, NCBF is more attractive since it results in virtually no handoff drops while achieving lower new connection blocks, i.e. less expensive in terms of bandwidth usage. In the next subsection, we compare NCBF with CHOI.

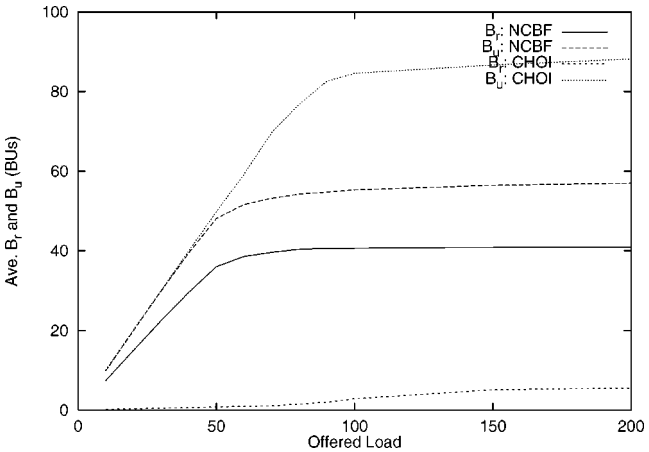
7.2.2. Comparison of CHOI and NCBF

Figure 9 shows P_{CB} and P_{HD} of CHOI and NCBF as the offered load increases for the voice ratio $R_{vo} = 0.5$ and 1.0. The P_{HD} of CHOI is observed to be upper-bounded by the target value $P_{HD,target} = 0.01$ irrespective of the voice ratio and user mobility over the entire offered loads examined, so CHOI attains the design goal. By comparing P_{CB} of two schemes, we can conclude that the handoff drops of NCBF (and hence, the other two per-connection bandwidth reservation schemes) are eliminated at the expense of blocking a large number of new connection requests even in lightly-loaded situations.

This is clearer in figure 10 which shows the average (target) reservation bandwidth B_r and utilized bandwidth B_u by the existing connections as the offered load increases for $R_{vo} = 1.0$. Note that B_r is a target for CHOI while it is a real reserved bandwidth for NCBF. CHOI works desirably by reserving less bandwidth when the system is lightly-loaded, and increasing the reservation bandwidth as

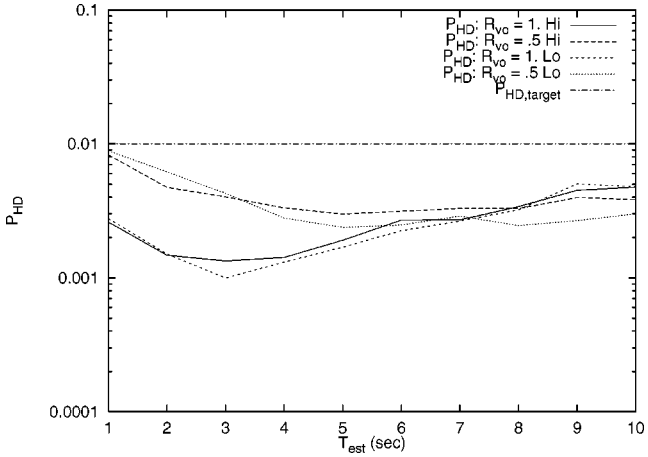


(a) high mobility

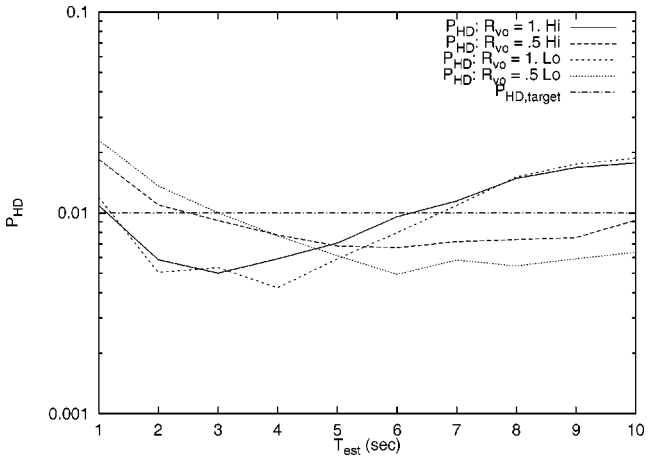


(b) low mobility

Figure 10. Comparison of NCBF and CHOI using the average B_{r} and B_{u} versus offered load for $R_{\mathrm{vo}} = 1.0$.



(a) $L = 100$



(b) $L = 200$

Figure 11. P_{HD} versus estimation time T_{est}: NAG.

the offered load increases. B_{u} is observed to be larger than B_{r} throughout the whole offered loads examined. NCBF has the same tendency, but the problem is that it reserves too much bandwidth compared to CHOI, so less bandwidth is utilized, i.e. B_{u} is smaller. This is why P_{CB} of NCBF is much larger than that of P_{HD}. From all the above observations, one can conclude that guaranteeing (virtually) no handoff drops through per-connection reservation is too expensive to be practically useful. Since wireless resources are scarce in general, per-connection bandwidth reservation schemes are practically unattractive. In practice, the service provider may support any of these three schemes (e.g., NCBF since it is the best among the three) as an option available to customers who are willing to pay the high price.

7.2.3. Comparison of CHOI and NAG: One-dimensional case

Now, we compare CHOI and NAG, both of which have the same design goal to keep P_{HD} below a given target value. First, we consider the performance of NAG to show the degree of its dependence on the choice of T_{est}. Figure 11 plots the P_{HD} of NAG with different values of T_{est} for the offered load (a) $L = 100$ and (b) $L = 200$, where "Hi" and "Lo" in the figures represent high and low user mobility, respectively. Four different (R_{vo}, mobility) pairs were considered. First, from figure 11(a), all ranges of T_{est} satisfy the design goal for $L = 100$. Next, from figure 11(b), NAG is observed to achieve the design goal only with certain values of T_{est} for $L = 200$. Especially, this plot of NAG shows the trade-off between large and small T_{est}'s, which was discussed at the end of section 5.2.

The smaller P_{CB} the better as long as $P_{\mathrm{HD}} \leqslant P_{\mathrm{HD,target}}$. The values of P_{CB} were observed to be almost constant for all the examined values of T_{est} even though the corresponding graphs are not included here due to lack of space. So, the smaller P_{HD} the better in this case. The problem is that the dependence of P_{HD} on T_{est} is a function of user mobility and R_{vo}. Especially, the optimal T_{est} which achieves the smallest P_{HD} depends greatly on R_{vo}. We also conducted the same experiment to obtain figure 11 for capacity $C = 20$, $L = 40$, and $R_{\mathrm{vo}} = 1.0$, and found that the optimal T_{est} depends also on the link capacity as shown in figure 12. Determination of the optimal T_{est} should involve a form of experiment similar to the above. However, the optimal T_{est} depends on user mobility, voice ratio, and link capacity. Moreover, user mobility and voice ratio are ac-

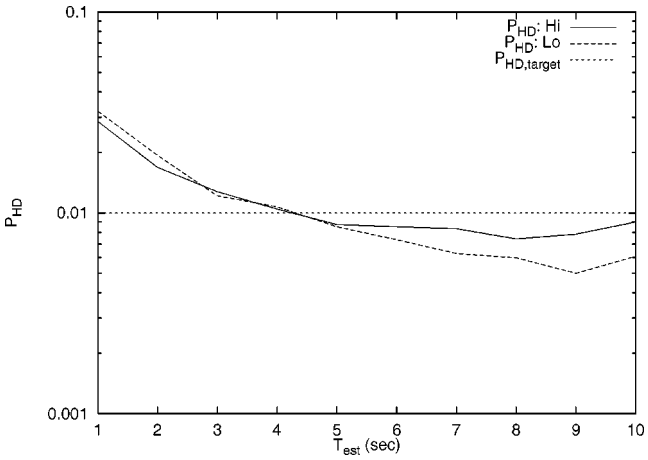
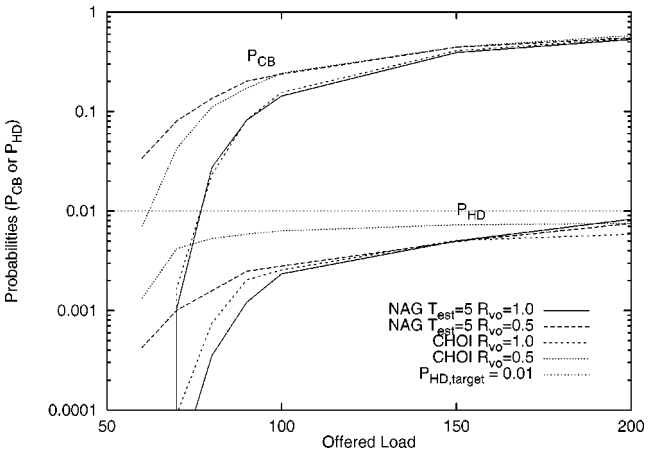
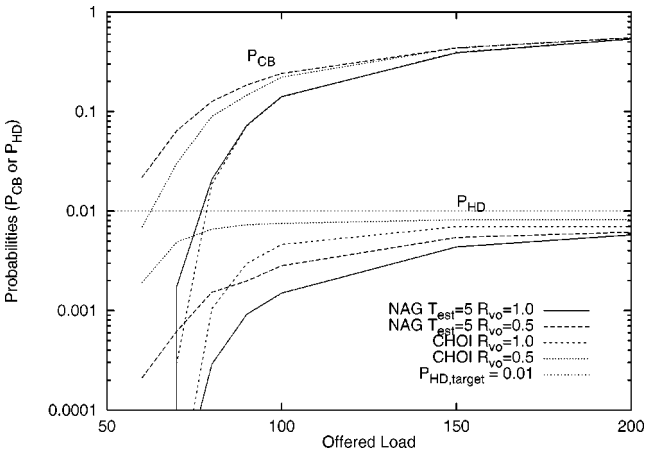


Figure 12. P_{HD} versus estimation time T_{est} for $C = 20$ and $L = 40$: NAG.



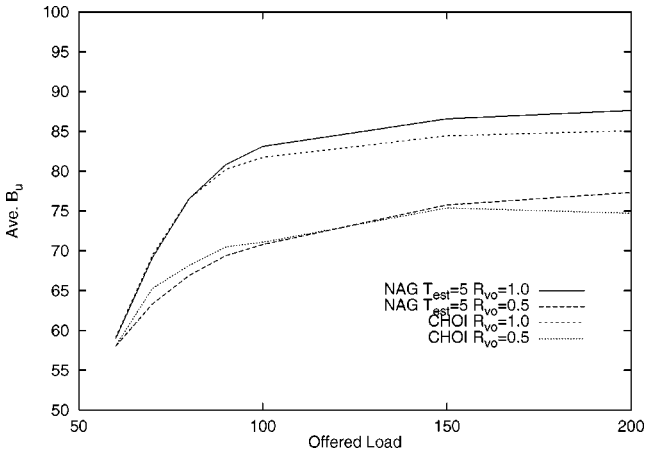
(a) high mobility



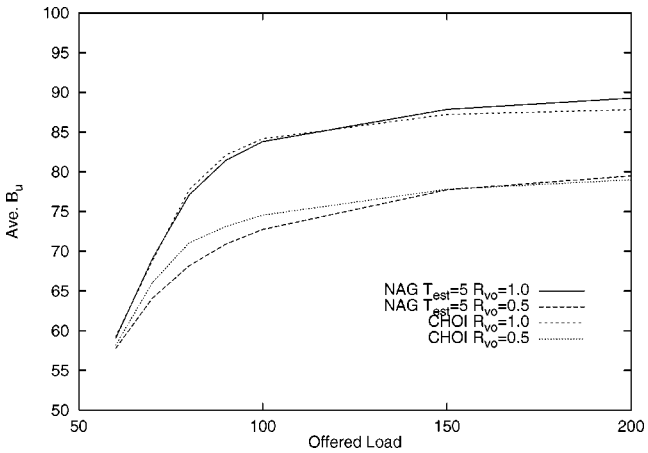
(b) low mobility

Figure 13. Comparison of NAG and CHOI using P_{CB} and P_{HD} versus offered load.

tually time-varying, so it is difficult to determine the best value of T_{est} for a system. For further experiments, we choose $T_{\mathrm{est}} = 5$ s, which is about the average of four different optimal T_{est}'s for four different cases in figure 11, for the one-dimensional environment.



(a) high mobility

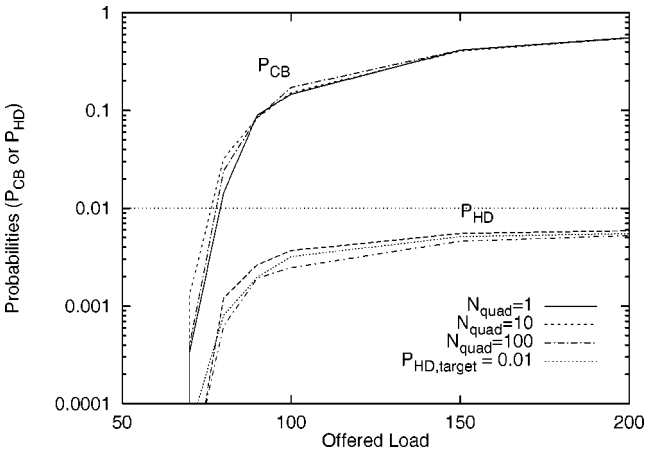


(b) low mobility

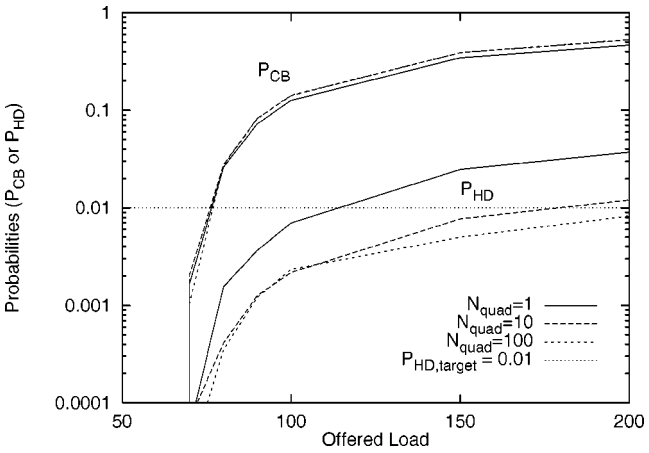
Figure 14. Comparison of NAG and CHOI using the average utilized bandwidth B_{u} versus offered load.

Figure 13 plots P_{CB} and P_{HD} as the offered load increases for NAG with $T_{\mathrm{est}} = 5$ and CHOI. Both schemes are found to achieve the design goal for the most of offered loads examined. As long as the design goal is met, which of the two achieves a smaller P_{HD} does not matter. In terms of P_{CB}, CHOI performs better than NAG for the lightly-loaded region, and worse for the heavily-loaded region. For a very heavily-loaded region, both schemes yield about the same P_{CB}, but NAG is slightly better; the rightmost points of the graphs for $R_{\mathrm{vo}} = 1.0$ are: (1) high-mobility: 0.695 (CHOI) and 0.672 (NAG); and (2) low-mobility: 0.682 (CHOI) and 0.676 (NAG).

Figure 14 shows the average utilized bandwidth B_{u} in a cell for both schemes. Note that in NAG, the bandwidth reservation is not explicitly defined, so the reserved bandwidths cannot be compared. This utilized bandwidth shows a similar comparison to that observed from figure 13 between the two schemes, i.e. CHOI is better for the lightly-loaded region, and worse for the heavily-loaded region. By examining the utilized bandwidth, CHOI might appear worse than NAG in the highly-loaded region of the high mobility case, but actually it is not, because P_{CB} is an important performance measure, and P_{CB}'s are almost



(a) CHOI



(b) NAG

Figure 15. Comparison of NAG and CHOI using the dependence on the value of N_{quad} for $R_{\text{vo}} = 1.0$.

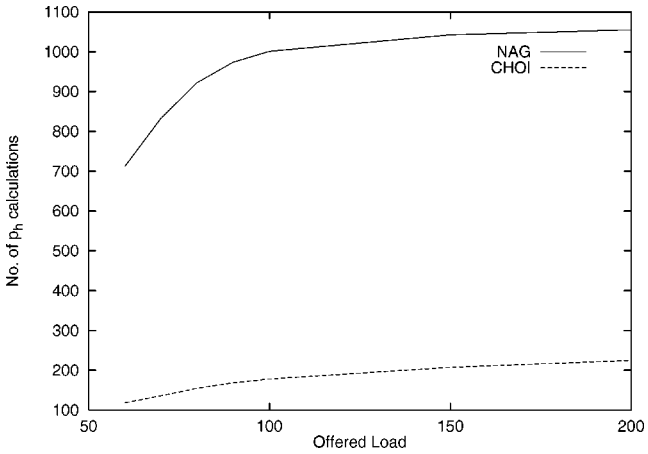


Figure 16. Complexity comparison of NAG and CHOI using the number of handoff probability p_{h} calculations for an admission test for $R_{\text{vo}} = 1.0$.

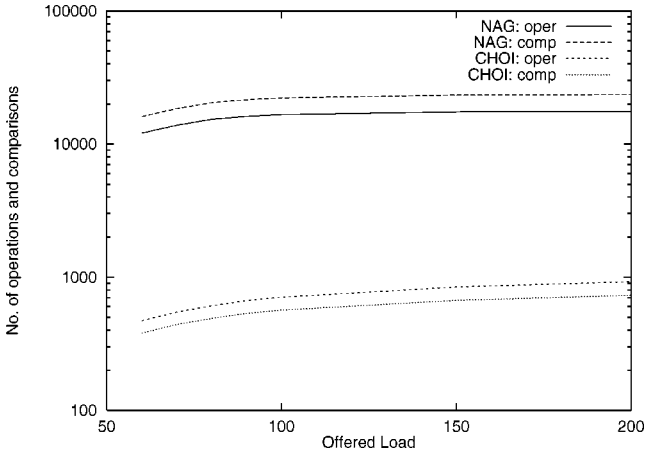


Figure 17. Complexity comparison of NAG (with $N_{\text{quad}} = 10$) and CHOI (with $N_{\text{quad}} = 1$) using the average numbers of numerical operations and comparisons for an admission decision for $R_{\text{vo}} = 1.0$.

the same for both schemes for the highly-loaded region. Note that usually the higher average utilized bandwidth, the lower P_{CB} in a system, but it is not always true for different systems or even in a system with different traffic conditions.

Next, we compare the complexity of the two schemes. We first examine their dependence on the accuracy of mobility estimation, which can be represented by the size of the cached history used for mobility estimation for each *prev*, i.e. the size of maximum handoff estimation function, N_{quad}. Figure 15 plots P_{CB} and P_{HD} as the offered load increases for (a) CHOI and (b) NAG with $R_{\text{vo}} = 1.0$ and $N_{\text{quad}} = 1$, 10, and 100. (Note that we have thus far used $N_{\text{quad}} = 100$.) From figure 15(a), we observe that CHOI does not depend much on N_{quad} as different values of N_{quad} yield almost the same performance. CHOI achieves the design goal even with $N_{\text{quad}} = 1$, implying that it uses only one cached history for mobility estimation. This indicates the robustness of CHOI to the inaccuracy of mobility estimation thanks to the mobility estimation time window control. On the other hand, figure 15(b) shows that NAG starts to violate the design goal in the over-loaded region with $N_{\text{quad}} = 10$. This implies that NAG requires very accurate mobility estimation. Note that this difference of dependence on the mobility estimation accuracy clearly separates the two in terms of memory and computation complexity. The memory required for cached history directly depends on N_{quad}, and the computation complexity of the handoff probability p_{h} in equation (4) is also affected greatly by N_{quad}.

Figure 16 shows the average number of calculations of the handoff probabilities p_{h} to decide admissibility upon request of a new connection. NAG considers both incoming and outgoing handoffs by calculating p_{h}'s and p_{s}'s. Calculation of p_{s} requires as many calculations of p_{h}'s as the number of adjacent cells. In addition, NAG requires the admissibility decision in both the current and all adjacent cells while CHOI determines it adaptively according to the condition of adjacent cells. We observe that NAG requires at least 4 times as many p_{h} calculations as CHOI does, where the lower the offered load, the more pronounced difference in the number of calculations between them.

Finally, we combine the dependence on N_{quad} and the number of p_{h} calculations. Figure 17 shows the average numbers of numerical operations (e.g., summations and

multiplications) and comparisons used to make an admission decision. Comparisons include the decisions such as if t_{soj} is larger than a value in the summations of equation (4). For NAG, $N_{quad} = 10$ is used even though the design goal is not always met with this value while $N_{quad} = 1$ is used for CHOI. The complexity to keep up with the average lifetime of each mobile's connections needed for NAG was not included. Moreover, the computation of the function $Q(\cdot)$ in equation (13) was also counted as one operation. Note that these are not fair to CHOI. For CHOI, the numbers of operations and comparisons used for the mobility time window control algorithm, normalized by the number of connection arrivals, are also added in the plots. From the graph, the complexity of NAG is found to be about 17.4–25.7 times of that of CHOI in terms of the number of numerical operations, and about 29.6–42.3 times in terms of the number of comparisons. The lower the offered load, the larger the difference between them. So, we can conclude that NAG is much more expensive than CHOI to attain a similar performance.

7.2.4. Comparison of CHOI and NAG: Two-dimensional case

In this subsubsection, we compare CHOI and NAG in the two-dimensional environment following a similar step in the previous subsubsection for the one-dimensional structure. Note that the link capacity $C = 50$, and we consider the low mobility case only. First, figure 18 shows P_{CB} and P_{HD} of NAG for different values of T_{est}. We observe that P_{HD}'s for both $R_{vo} = 1.0$ and 0.5 start to fall below the target value, 0.01, at around $T_{est} = 20$ s. On the other hand, we also observe that P_{CB}'s increase very slowly as T_{est} increases, i.e. the smaller T_{est}, the better in terms of P_{CB}. So, we select $T_{est} = 20$ s for the two-dimensional structure for further experiments.

Figure 19 plots the two probabilities P_{CB} and P_{HD} for NAG with $T_{est} = 20$ s and CHOI. Basically, the general tendency is the same as that from the one-dimensional structure. One difference we can observe is that P_{CB} of NAG is distinctly larger than that of CHOI. Note that P_{CB}'s were almost the same for both schemes in the one-dimensional structure. This difference seems more important in this case since P_{CB}'s are relatively large even in the lightly-loaded region. For example, the P_{CB}'s of both schemes are slightly larger than 0.1 (i.e. 10%) at the offered load $L = 40$ (30) and $R_{vo} = 1.0$ (0.5).

Now, we examine the sensitivity of the two schemes to the accuracy of mobility estimation. From figure 20, we observe that CHOI is still very robust to the inaccuracy of mobility estimation while NAG is also quite robust in terms of the design goal achievement. NAG attains the design goal even with $N_{quad} = 2$. However, P_{CB} of NAG seems to be affected by the accuracy of mobility estimation in this case since we observe that the smaller N_{quad}, the larger P_{CB}. Note that P_{CB} of NAG was almost independent of N_{quad} in the one-dimensional case as observed from figure 15(b). With NAG, either (or possibly both) of P_{CB} and P_{HD} appears to be affected by the value of N_{quad} depending on the cellular structure, user mobility pattern, and others. In any case, the performance of NAG is heavily dependent on the mobility estimation accuracy.

Finally, we compare the computation complexity of CHOI and NAG. Figure 21 shows the average number of numerical operations and comparisons used to make an admission decision for CHOI with $N_{quad} = 1$ and NAG with $N_{quad} = 2$ for $R_{vo} = 1.0$. We observed a similar tendency to that from the one-dimensional structure. From figure 21, one can find the complexity of NAG to be about 14.1–22.3 times of that of CHOI in terms of the number of numerical operations, and about 10.3–15.8 times in terms of the number of comparisons. The complexity difference between CHOI and NAG seems to be smaller here since NAG with $N_{quad} = 2$ (instead of $N_{quad} = 10$) was compared. The lower the offered load, the larger the difference between them. From the two-dimensional results, we reconfirm that NAG is much more expensive than CHOI to attain similar performance.

Table 1 summarizes the comparison results of the five different schemes considered thus far. Note that AG does

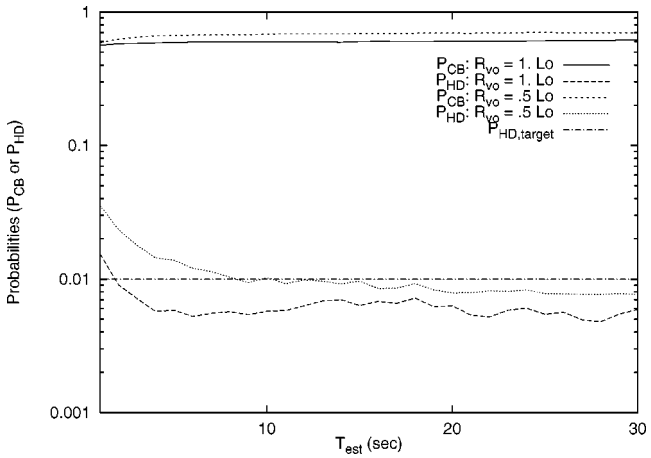


Figure 18. P_{CB} and P_{HD} versus estimation time T_{est} in the two-dimensional environment with $C = 50$ and $L = 100$: NAG.

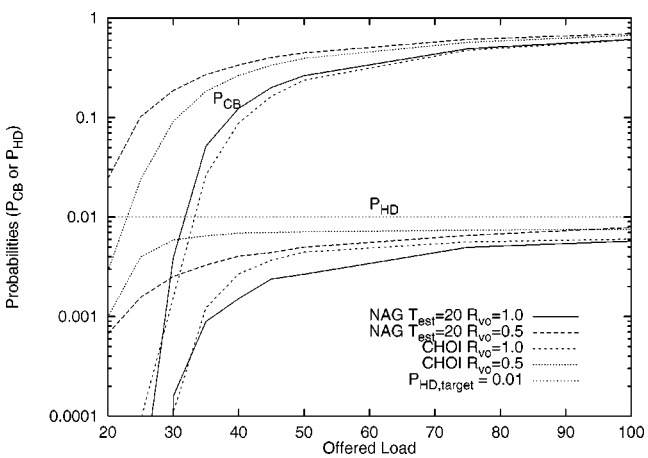
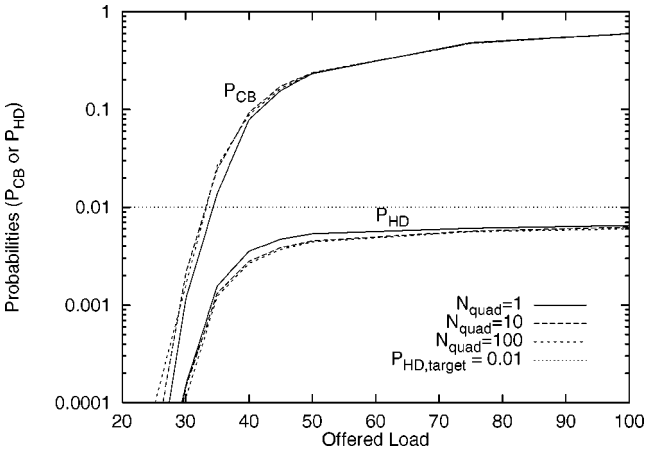


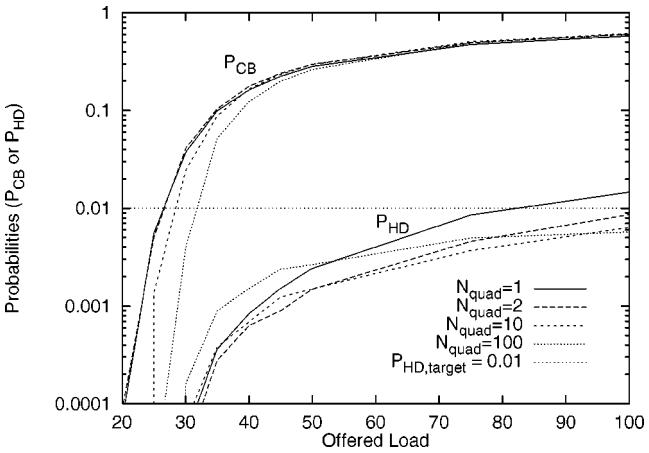
Figure 19. Comparison of NAG and CHOI using P_{CB} and P_{HD} versus offered load for the two-dimensional case.

Table 1
Summary of comparison among CHOI, NAG, AG, BHARG, and NCBF.

	CHOI	NAG	AG	BHARG	NCBF
P_{HD}	bounded	bounded with T_{est}	guaranteed zero	virtually zero	virtually zero
P_{CB}	about the same	about the same	worst	second worst	third worst
Complexity	1	at least 10 times	not based on history	N/A	N/A
T_{est}	adapted	shoud be assigned	N/A	N/A	N/A



(a) CHOI



(b) NAG

Figure 20. Comparison of NAG and CHOI using the dependence on the value of N_{quad} for $R_{\text{vo}} = 1.0$ and the two-dimensional case.

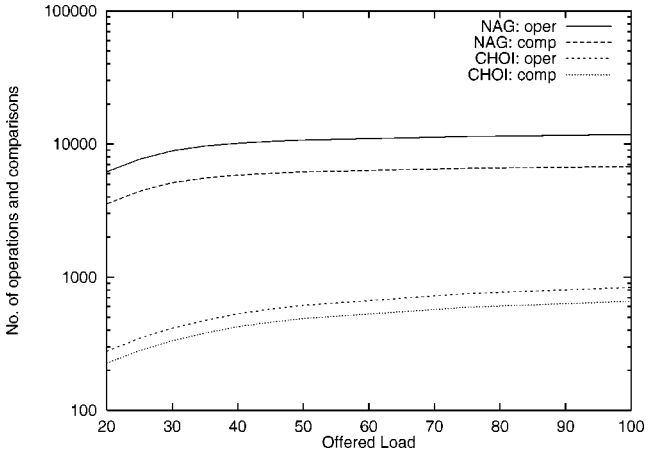


Figure 21. Complexity comparison of NAG (with $N_{\text{quad}} = 2$) and CHOI (with $N_{\text{quad}} = 1$) using the average numbers of numerical operations and comparisons for an admission decision for $R_{\text{vo}} = 1.0$ and the two-dimensional case.

not use the history-based mobility estimation, but relies on mobility specification, which is practically difficult to obtain. In fact, we did not account for how to predict the next cell of a mobile for BHARG and NCBF. So, their complexity cannot be compared fairly with the other two schemes.

8. Concluding remarks

In this paper, we compared five connection admission control schemes in QoS-sensitive cellular networks that either keep the handoff dropping probability below a prespecified target value or make it absolutely (or virtually) zero. NAG is made to utilize the mobility estimation scheme developed for CHOI since this mobility estimation is practically feasible. NAG was also generalized to accommodate heterogeneous connections. We showed how costly it is to make the handoff dropping probability zero even under an impractical assumption by evaluating the performance of AG. The other two per-connection bandwidth reservation schemes, BHARG and NCBF, achieved a lower P_{CB}, but were found to be very expensive (even though they are less expensive than AG). So, one can conclude that per-connection bandwidth reservation is too expensive to be practical.

NAG was shown to require much more memory and computation than CHOI in order to meet the design goal. NAG is also observed to depend greatly on the design parameter T_{est}, which is difficult to adjust in real world. By contrast, CHOI is robust to the inaccuracy of mobility estimation thanks to the mobility estimation window control while meeting the design goal over the entire range of the offered loads considered even with much less memory and computation. CHOI is, therefore, preferable to, and practically more attractive than, NAG.

References

[1] S. Choi and K.G. Shin, Predictive and adaptive bandwidth reservation for handoffs in QoS-sensitive cellular networks, in: *Proc. ACM SIGCOMM '98*, Vancouver, British Columbia (September 1998) pp. 155–166.

[2] D. Hong and S.S. Rappaport, Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and non-

prioritized procedures, IEEE Transactions on Vehicular Technology 35(3) (August 1986) 77–92.
[3] 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 Transactions on Networking 5(1) (February 1997) 1–12.
[4] S. Lu and V. Bharghavan, Adaptive resource management algorithms for indoor mobile computing environments, in: *Proc. ACM SIGCOMM '96* (August 1996) pp. 231–242.
[5] S. Lu, K.-W. Lee and V. Bharghavan, Adaptive service in mobile computing environments, in: *Proc. International Workshop on Quality of Service '97* (1997).
[6] M. Naghshineh and M. Schwartz, Distributed call admission control in mobile/wireless networks, IEEE Journal on Selected Areas in Communications 14(4) (May 1996) 711–717.
[7] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed. (McGraw-Hill, 1991).
[8] R. Ramjee, R. Nagarajan and D. Towsley, On optimal call admission control in cellular networks, in: *Proc. IEEE INFOCOM '96* (1996) pp. 43–50.
[9] A.K. Talukdar, B.R. Badrinath and A. Acharya, On accommodating mobile hosts in an integrated services packet network, in: *Proc. IEEE INFOCOM '97* (April 1997) pp. 1048–1055.
[10] Y. Zhao, *Vehicle Location and Navigation Systems* (Artech House, 1997).

Sunghyun Choi is a Senior Member of Research Staff at Philips Research, Briarcliff Manor, New York. He received his B.S. (summa cum laude) and M.S. degrees in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST) in 1992 and 1994, respectively, and received Ph.D. from the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, in 1999. His research interests are in the area of wireless/mobile networks with emphasis on the QoS guarantee and adaptation, in-home multimedia networks, connection and mobility management, multimedia CDMA, and link layer protocols. During 1997–1999, Dr. Choi was a recipient of the Korea Foundation for Advanced Studies Scholarship. During 1994–1997, he received the Korean Government Overseas Scholarship. He is also a winner of the Humantech Thesis Prize from Samsung Electronics in 1997.
E-mail: sunghyun.choi@philips.com.

Kang G. Shin is Professor and Director of the Real-Time Computing Laboratory, Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan. He has authored/coauthored about 600 technical papers and numerous book chapters in the areas of distributed real-time computing and control, computer networking, fault-tolerant computing, and intelligent manufacturing. He has coauthored (jointly with C.M. Krishna) a textbook *Real-Time Systems* (McGraw-Hill, 1997). In 1987, he received the Outstanding IEEE Transactions on Automatic Control Paper Award, and in 1989, Research Excellence Award from the University of Michigan. In 1985, he founded the Real-Time Computing Laboratory, where he and his colleagues are investigating various issues related to real-time and fault-tolerant computing. His current research focuses on Quality of Service (QoS) sensitive computing and networking with emphasis on timeliness and dependability. He has also been applying the basic research results to telecommunication and multimedia systems, intelligent transportation systems, embedded systems, and manufacturing applications. He received the B.S. degree in electronics engineering from Seoul National University, Seoul, Korea, in 1970, and both the M.S. and Ph.D. degrees in electrical engineering from Cornell University, Ithaca, New York, in 1976 and 1978, respectively. From 1978 to 1982 he was on the faculty of Rensselaer Polytechnic Institute, Troy, New York. He has held visiting positions at the U.S. Airforce Flight Dynamics Laboratory, AT&T Bell Laboratories, Computer Science Division within the Department of Electrical Engineering and Computer Science at UC Berkeley, and International Computer Science Institute, Berkeley, CA, IBM T.J. Watson Research Center, and Software Engineering Institute at Carnegie Mellon University. He also chaired the Computer Science and Engineering Division, EECS Department, University of Michigan, for three years beginning January 1991. He is an IEEE fellow, was the Program Chairman of the 1986 IEEE Real-Time Systems Symposium (RTSS), the General Chairman of the 1987 RTSS, the Guest Editor of the 1987 August special issue of IEEE Transactions on Computers on Real-Time Systems, a Program Co-Chair for the 1992 International Conference on Parallel Processing, and served numerous technical program committees. He also chaired the IEEE Technical Committee on Real-Time Systems during 1991–1993, was a Distinguished Visitor of the Computer Society of the IEEE, and Editor of IEEE Transactions on Parallel and Distributed Computing, and an Area Editor of International Journal of Time-Critical Computing Systems.
E-mail: kgshin@eecs.umich.edu