Comparison of Connection Admission-Control Schemes in the Presence of Hand-Offs in 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

In this paper, we compare four distributed connection admission control schemes in cellular networks where the first two schemes are to keep the hand-off dropping probability below a target value, the third scheme is to guarantee no hand-off drops through a per-connection bandwidth reservation, and the fourth scheme uses another type of per-connection bandwidth reservation. The first scheme predicts the bandwidth required to handle hand-offs by estimating possible hand-offs from adjacent cells, then performs the admission control for each newly-requested connection. On the other hand, the second scheme predicts the total required bandwidth in the current cell by estimating both incoming and outgoing hand-offs at each cell. The third scheme requires the set of cells which the mobile with a newly-requested connection will traverse, and reserves bandwidth for each connection in each of those cells. The fourth scheme reserves bandwidth for each connection in the predicted next cell of a mobile. We adopt the history-based mobility estimation for the first two schemes. Using extensive simulations, the four schemes are compared quantitatively in terms of (1) important performance measures such as the hand-off dropping probability, connection-blocking probability, and bandwidth utilization, (2) dependency on the design parameters, (3) dependency on the mobility estimation accuracy, and (4) complexity. The simulation results indicate that the first scheme is the most desirable in the sense that it achieves reasonably good performance while requiring much less memory and computation than the other three schemes.

1 Introduction

Connection-level QoS issues related to the establishment and management of connections are very important in QoS-sensitive cellular networks because users are expected to move during communication sessions causing hand-offs between cells. The current trend in cellular networks is shrinking cell size to accommodate more mobile users in a given geographical area. This results in more frequent hand-offs, 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 hand-offs 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 impractical to completely eliminate hand-off 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 hand-off dropping probability below a target value $P_{HD,target}$. Limiting P_{HD} below $P_{HD,target}$ is referred to as the *design goal* throughout this paper. Both schemes are based on the estimation of hand-offs that may occur in 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 hand-offs from adjacent cells upon arrival of a new connection request. The mobility (i.e., hand-off behavior) of each user is estimated using a history of hand-offs observed in each cell. Using this estimation, one can compute the bandwidth required to handle the hand-offs that are predicted to occur within a specific time window. It also adaptively controls the time window size depending on the observed hand-off dropping events.

In the second scheme proposed in [6] (referred to as **NAG**), the BS considers not only incoming hand-offs from adjacent cells, but also outgoing hand-offs into adjacent cells from the current cell. 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 spends in a cell; and (2) the perfect knowledge about mobility and lifetimes of user connections, i.e., known hand-off and connection termination rates. Under these assumptions, **NAG** was shown to achieve 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.

MOBICOM 98 Dallas Texas USA
Copyright ACM 1998 1-58113-035-x/98/10...$5.00

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 $P_{HD,target}$ 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 achieves the design goal with much less computation and memory complexity than NAG.

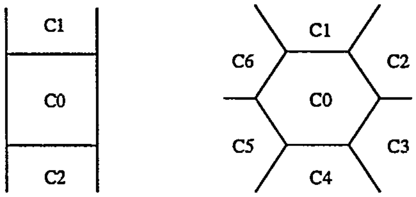
Also considered is an admission-control scheme (referred to as AG here) which guarantees no hand-off drop for any existing connection. Using the first two schemes, it is not possible to completely eliminate hand-off drops. No hand-off drops can be achieved only by checking the bandwidth availability and reserving each connection's bandwidth in *all* cells the mobile (which is requesting a new connection) is to traverse in future. Practically, it is impossible to know these cells in advance during the admission-control phase. The basic concept of this scheme was proposed in [8] assuming the availability of such information. We will show how costly it is to make the hand-off dropping probability zero even under this impractical assumption. The last admission-control scheme (referred to as BHARG) is based on per-connection bandwidth reservation [4, 5]. This scheme does not have any specific design goal, unlike the other three schemes described above. The next cell each mobile will move into is predicted, and its 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. It will be shown that BHARG is still costly compared to the first two schemes due to its per-connection bandwidth reservation requirement.

There is one more scheme that limits P_{HD} below a target [3]. The scheme 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 precise 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 didn't address clearly how to determine the shadow cluster either. More importantly, the scheme is computationally too expensive to be practical, as compared to the four schemes considered here.

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 two per-connection bandwidth reservation-based admission control schemes, AG and BHARG. Section 7 quantitatively compares these four 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,[1] 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 hand-off 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 hand-off could fail due to insufficient bandwidth available in the new cell, and in such a case, a *connection hand-off drop* occurs. Here, we preclude 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.



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

Figure 1: Indexing of cells.

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 won't 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:[2] A is labeled with 0, and the others with numbers beginning 1 as shown in Figure 1. Let $C_{i,j}$ be the j-th connection in cell i and $b(C_{i,j})$ be its required bandwidth. For simplicity, we assume that a mobile doesn't have multiple simultaneous connections, so that by an *active mobile*, we mean a mobile with one existing connection.[3]

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 re-routing upon hand-off 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 re-routing 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 describe here the mobility-estimation scheme [1] that is

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

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

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

based on a history of hand-offs 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 CHOI [1].

3.1 Hand-Off Estimation Functions

We now explain the 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 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 *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 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 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

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

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 (≥ 0) is an integer,

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

where $1 \geq w_n \geq w_{n+1}$, and $w_n = 0$ for all $n > N_{win\_days}$. 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_{win_days} is a design parameter so that the quadruplet observed more than $(N_{win\_days} \cdot T_{day} + T_{int})$ ago is determined to be out-of-date, and hence, 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 2. 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 2, in order to reduce the memory and computation complexity.[4] 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 *prev*. 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 that satisfies Eq. (1) with a smaller n gets higher priority. Second, among those satisfying Eq. (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 the hand-off estimation function 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 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 entries.

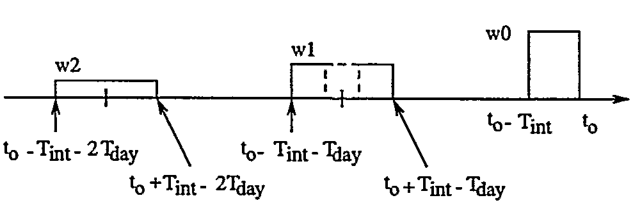

w2
w1
w0
$t_o - T_{int} - 2T_{day}$
$t_o + T_{int} - 2T_{day}$
$t_o - T_{int} - T_{day}$
$t_o + T_{int} - T_{day}$
$t_o - T_{int}$
t_o

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

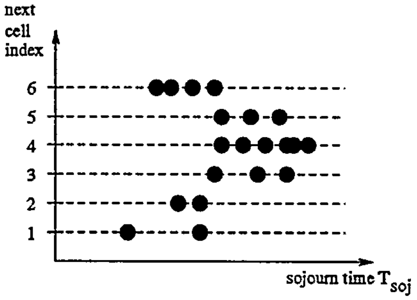


Figure 3: An example of the footprint of hand-off estimation function for $prev = 1$.

Figure 3 shows an example of footprint of the hand-off estimation function for $prev = 1$ without showing the values of w_n's. The hand-off estimation function in a 3-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 ajacent 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 given *prev* can generate a probability mass function for a two-dimensional random

[4] The calculations required for 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.

$$p_h(C_{0,j} \rightarrow next) := \begin{cases} \dfrac{\sum_{T_{ext\_soj}(C_{0,j})<t_{soj}\leq T_{ext\_soj}(C_{0,j})+T_{est}} F_{HOE}(t_o,prev(C_{0,j}),next,t_{soj})}{\sum_{next'\in A_0}\sum_{t_{soj}>T_{ext\_soj}(C_{0,j})} F_{HOE}(t_o,prev(C_{0,j}),next',t_{soj})}, \\ \quad \text{if } \sum_{next'\in A_0}\sum_{t_{soj}>T_{ext\_soj}(C_{0,j})} F_{HOE}(t_o,prev(C_{0,j}),next',t_{soj}) \neq 0, \\ 0, \quad \text{otherwise.} \end{cases} \tag{3}$$

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 $T_{min} < t_{soj} \leq T_{max}$, and depart to cell $next$ can be estimated by

$$Pr(T_{min} < t_{soj} \leq T_{max} \ \&\ \text{departure to cell } next) = \frac{\sum_{T_{min}<t_{soj}\leq T_{max}} F_{HOE}(t_o, prev, next, t_{soj})}{\sum_{next'\in A_0}\sum_{0<t_{soj}<\infty} F_{HOE}(t_o, prev, next', t_{soj})}, \tag{4}$$

where A_0 is the set of neighbor cells' indices of cell 0.

4 Admission Control CHOI with Estimation of Incoming Hand-offs Only

We first describe the admission control scheme CHOI in [1] to keep P_{HD} below $P_{HD,target}$ by utilizing the hand-off 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, $p_h(C_{0,j} \rightarrow i)$, 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*. 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 hand-off probability $p_h(C_{0,j} \rightarrow next)$ at time t_o is calculated by Eq. (3), where $prev(C_{0,j})$ is the cell in which $C_{0,j}$ resided before entering the current cell and 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 hand-off probability $p_h(C_{0,j} \rightarrow next)$.

Figure 4 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 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 light shaded regions) are summed to obtain the denominator in Eq. (3). 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. (3). 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. (3) is zero.

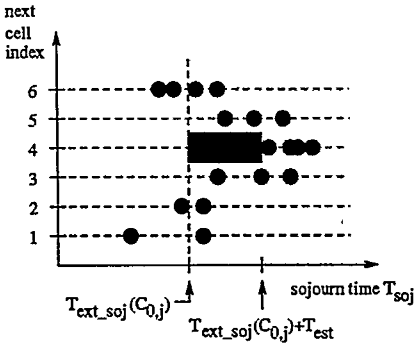


Figure 4: An example of calculating $p_h(C_{0,j} \rightarrow next)$ when $prev(C_{0,j}) = 1$ and $next = 4$ using the footprint of $F_{HOE}(t_o, 1, next', T_{soj})$.

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 given by:

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

where 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 hand-offs from adjacent cells within the estimation time T_{est}, is calculated as:

$$B_{r,0} = \sum_{i\in A_0} B^i_{r,0}. \tag{6}$$

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 hand-off dropping probability below the target. In this 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 subsection. Then, the estimation time T_{est} of cell $next$ (or $T_{est,next}$) will be used in Eq. (3). 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 use Eq. (5) to calculate the required bandwidth for the expected hand-offs from that cell, (i.e.,

```
if ($w = \lceil 1/P_{HD,target} \rceil$), 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 ($n_H \geq w_{obs}$) then {
            if ($n_{HD} \leq w_{obs}/w$ and $T_{est} > 1$) then
                $T_{est} := T_{est} - 1$;
            $w_{obs} := w$;  $n_H := 0$;  $n_{HD} := 0$;
        }
    }
}
```

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

$B'_{r,0}$ for cell i) and will inform cell 0's BS of this value. Finally, cell 0's BS will calculate $B_{r,0}$ using Eq. (6).

4.2 Control of Mobility Estimation Time Window

Using the bandwidth reservation described above, 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}. Hence, an adaptive algorithm is used to control the mobility estimation time window size based on the hand-off 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 hand-offs, n_H, and hand-off drops, n_{HD}, are reset to 0. As can be found in the pseudocode of Figure 5, w_{obs} is increased or decreased by w, and the constraint $P_{HD} < P_{HD,target}$ can be translated into 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 drops, the BS set $T_{est} := T_{est} + 1$ 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 5, which is the maximum T_{soj} derived from the hand-off 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 hand-off 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, $B^{curr}_{r,i}$, 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. Now, for a new connection request, the admission test is performed as follows:

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

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

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

5 Admission Control NAG with Estimation of Incoming and Outgoing Hand-Offs

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 hand-off/termination rates.

5.1 Three State Probabilities

The main difference between CHOI and NAG is that CHOI considers incoming hand-offs only while NAG considers both incoming and outgoing hand-offs in a cell. NAG is also based on the estimated mobility during $[t_0, t_0 + T_{est}]$, in which t_0 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) hand-off into an adjacent cell i ($p_h(C_{0,j} \rightarrow i)$); (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 hand-off probabilities $p_h(C_{0,j} \rightarrow i)$ are defined in Eq. (3) 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:

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

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 doesn't 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

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

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

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 from (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

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

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:

$$\begin{aligned} P_{O,i} &= Pr(B_{T_{est},i} > C), \\ &\approx Q\left(\frac{C - m_{B,i}}{\sigma_{B,i}}\right), \end{aligned} \tag{13}$$

where C is the link capacity. $m_{B,i}$ and $\sigma_{B,i}$ are obtained from Eqs. (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 $P_{O,i} \le P_{HD,target}$,

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

Note that for this scheme, a specific amount of bandwidth to be reserved is not defined. So, the relation between the value of T_{est} and the bandwidth reserved for the hand-offs 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^2_B would increase or decrease as T_{est} increases. There might 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 time window control used for **CHOI**. We will later evaluate the effect of the value of T_{est} using simulations.

6 Per-Connection Bandwidth Reservation-Based Schemes

Now, we describe two per-connection bandwidth reservation-based admission-control schemes, **AG** and **BHARG**.

6.1 Admission Control AG: No Hand-Off Drop

This subsection describes an admission-control scheme (referred to as **AG**, meaning "absolutely guaranteeing") which guarantees no hand-off 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 [8].

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. As described in [2], the route guidance system of Intelligent Transportation Systems (ITS) can be used to predict the mobiles' path/direction with a good accuracy, and might be used to predict the mobility specification. The problem is that using the route guidance 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 [8].

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 $\sum_{j \in \mathbf{C}_i} b(C_{i,j}) + b_{new} \le C - B_{r,i}$,

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

where $B_{r,i}$ is the sum of all per-connection bandwidths reserved in cell i. Whenever a mobile enters a cell, the cell's reserved bandwidth (for hand-offs) will be decreased,

> Upon hand-off of connection $C_{i,j}$ into cell i, $B_{r,i} := B_{r,i} - b(C_{i,j})$.

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 i-th cell in the entire cellular system. Through per-connection reservation in the cells within the mobility specification, it is possible to make the hand-off 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 Admission Control BHARG: Per-Connection Reservation in Next Cell

This subsection describes another per-connection reservation-based admission control scheme referred to as **BHARG.** This scheme doesn't try to limit P_{HD} nor to eliminate hand-off 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 $\sum_{j\in C_0} b(C_{0,j}) + b_{new} \le C - B_{r,0}$,

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

where $B_{r,i}$ is the sum of all per-connection bandwidths reserved in cell i. Whenever a mobile enters a cell, the cell's reserved bandwidth (for hand-offs) will be decreased:

> Upon hand-off 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 to make no hand-off drops, it will achieve virtually no hand-off drops as will be shown later, but at a very high cost which is comparable to that of **AG**.

7 Comparative Performance Evaluation

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

7.1 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 1 (a). We make the following assumptions for our simulation study:

A1. The 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* $R_{vo} \le 1$.

A4. Mobiles can travel in either of the two directions with an equal probability with a speed chosen randomly in $[SP_{min}, SP_{max}]$ (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. Connections are generated and behave in a stationary manner, i.e., there will be no fluctuations in terms of the connection-generation rate and mobility.

A7. The capacity C of each cell is 100 BUs (unless stated otherwise).

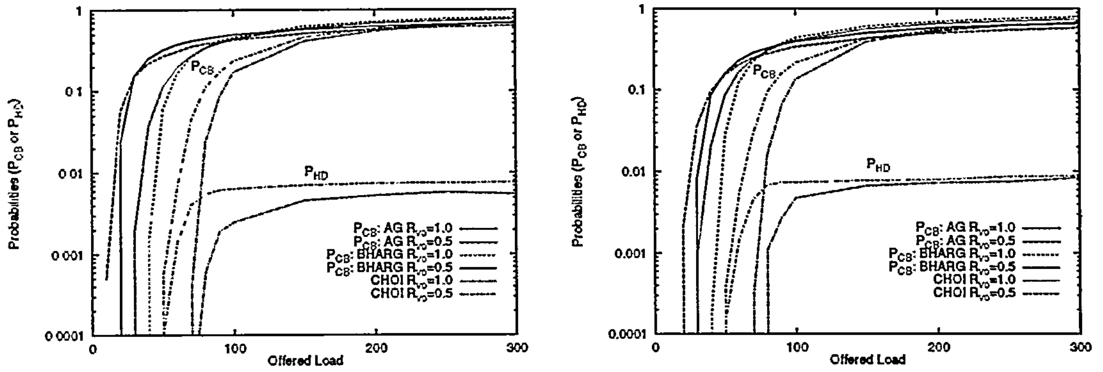
Each simulation run starts without any pre-memorized hand-off event quadruplets. As simulations are run, quadruplets will be observed, and will affect the hand-off estimation functions $F_{HOE}(t, prev, next, T_{soj})$. Two cases of user mobility are considered: high user mobility with $[SP_{min}, SP_{max}] = [80, 120]$, and low user mobility with $[40, 60]$. 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 with mobiles than those near the borders. This uneven traffic load can affect the performance evaluation of the four 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 cellular system forms a ring architecture as was assumed in [1, 6].

The parameters used include: $P_{HD,target} = 0.01$; for the mobility estimation of **CHOI** and **NAG**, $N_{quad} = 100$ (unless stated otherwise), $T_{int} = \infty$, $N_{win\_days} = 0$, and $w_0 = 1$; for **CHOI**, $T_{start} = 1$ (second). The choice of $T_{int} = \infty$ 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 × connections' bandwidth × average connection lifetime:

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

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 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 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 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 the connection until it becomes 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 about, or larger



(a) high mobility

(b) low mobility

Figure 6: Comparison of AG, BHARG, and CHOI using P_{CB} and P_{HD} vs. offered load.

than, 0.1 in most cases, due to some reserved bandwidth for hand-offs. 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 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, the design goal to keep P_{HD} below a target value should be achieved.

7.2 Simulation Results and Discussion

We first compare **CHOI** with **AG** and **BHARG**, then compare **CHOI** with **NAG**. **CHOI** and **NAG** were claimed to be superior to the conventional static bandwidth reservation scheme in [1] and [6], respectively, while showing that the static reservation scheme is unable to achieve the design goal.

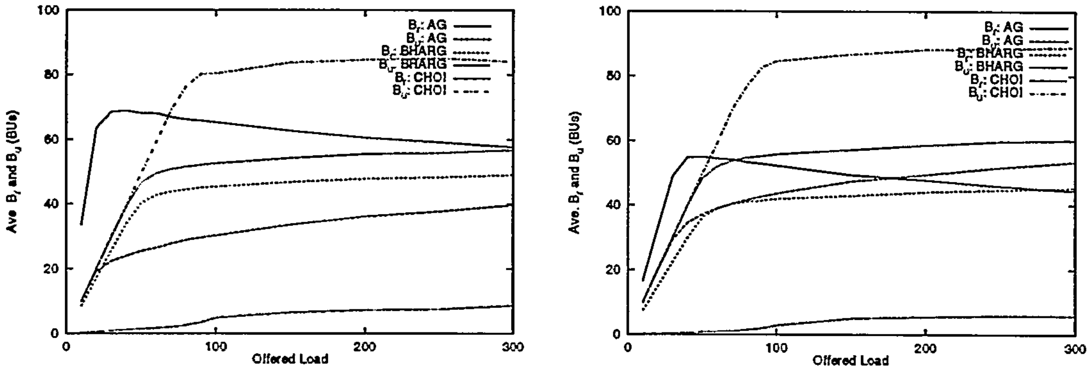
7.2.1 Comparison of CHOI, AG, and BHARG

Figure 6 shows P_{CB} and P_{HD} of three schemes as the offered load increases for the voice ratio $R_{vo} = 0.5$ and 1.0. First of all, it is observed that the P_{HD} of **AG** is zero irrespective of the offered load, voice ratio, and user mobility as it should be, and thus omitted in the plots. P_{HD} of **BHARG** is found to be almost zero, thus also omitted in the plots. On the other hand, 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. It should be noted that the hand-off drops of **AG** are eliminated at the expense of blocking a large number of new connection requests even in lightly-loaded situations, as the large P_{CB}'s shown. The fact that P_{CB} is larger than 0.1 even for $L = 30$ where $C = 100$ implies that **AG** severely under-utilizes the link capacity. P_{CB} of **BHARG** is observed to be less than that of **AG**, but still much larger than that of **CHOI**. In fact, P_{CB} of **BHARG** gets 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. We can see that **BHARG** also achieves virtually no hand-off drops at the expense of blocking a large number of new connection requests, implying that per-connection bandwidth reservation is basically too expensive to use.

This becomes clearer if we examine Figure 7, 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** and **BHARG** while it is a real reserved bandwidth for **AG**. First, **CHOI** works 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. The same tendency is also observed for **BHARG**. However, the problem is that it reserves too much bandwidth even in a lightly-loaded region. Note that B_r and B_u are comparable over the entire range of offered loads for the high mobility case.

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 7 (b) with low user mobility. Counter-intuitively, 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.

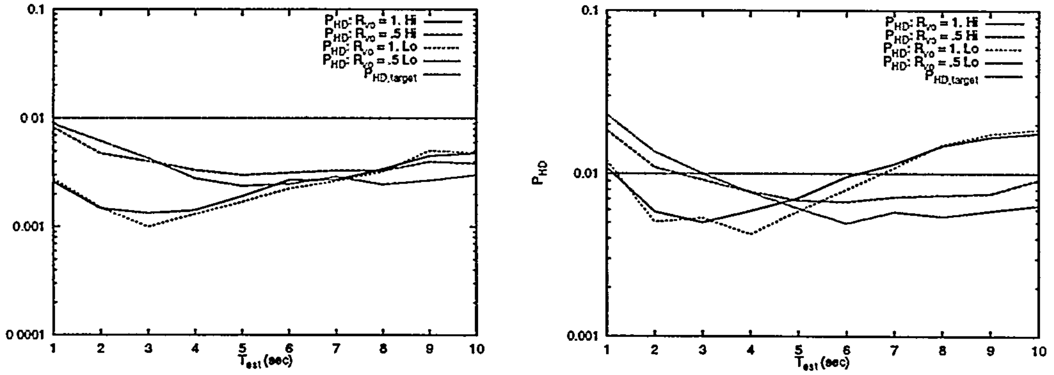
From the above observation, one can conclude that guaranteeing no hand-off drops through per-connection reservation is too expensive to be practically useful. Since wireless resources are very scarce and precious, **AG** is practically unattractive. **BHARG** is also observed to be too expensive even though it is less expensive than **AG**. Comparing these two schemes, **BHARG** seems more attractive since it results in virtually no hand-off drops while achieving lower new connection blocks. In practice, the service provider may support any of these two scheme as an option available to customers who are willing to pay the high price.



(a) high mobility

(b) low mobility

Figure 7: Comparison of **AG**, **BHARG**, and **CHOI** using the average B_r and B_u vs. offered load for $R_{vo} = 1.0$.



(a) $L = 100$

(b) $L = 200$

Figure 8: P_{HD} vs. estimation time T_{est}: **NAG**.

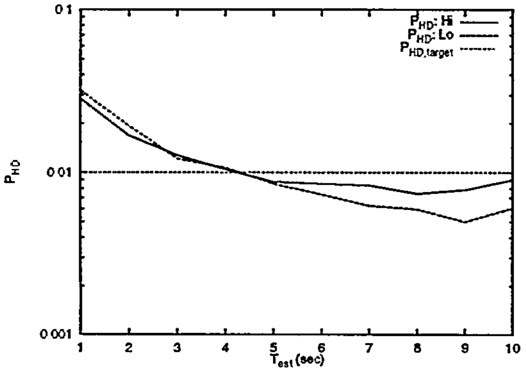


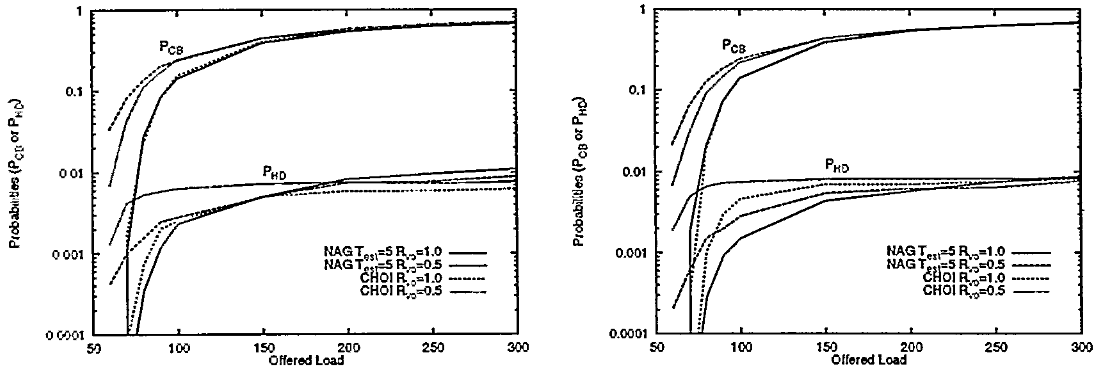
Figure 9: P_{HD} vs. estimation time T_{est} for $C = 20$: **NAG**.

7.2.2 Comparison of CHOI and NAG

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 dependency on the choice of T_{est}. Figure 8 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 8 (a), all ranges of T_{est} satisfy the design goal for $L = 100$. Next, from Figure 8 (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 that 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 $P_{HD} \leq P_{HD,target}$. 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 the space limit. So, the smaller P_{HD} the better in this case. The problem is that the dependency 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 8 for capacity $C = 20$, $L = 40$, and $R_{vo} = 1.0$, and found that the optimal T_{est} depends also on the link capacity as shown in Figure 9. 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 actually time-varying, so it is difficult to determine the best value of T_{est} for a system. For further experiments, we choose $T_{est} = 5$ (sec) which is about the average of four different optimal T_{est}'s for four different cases in Figure 8.

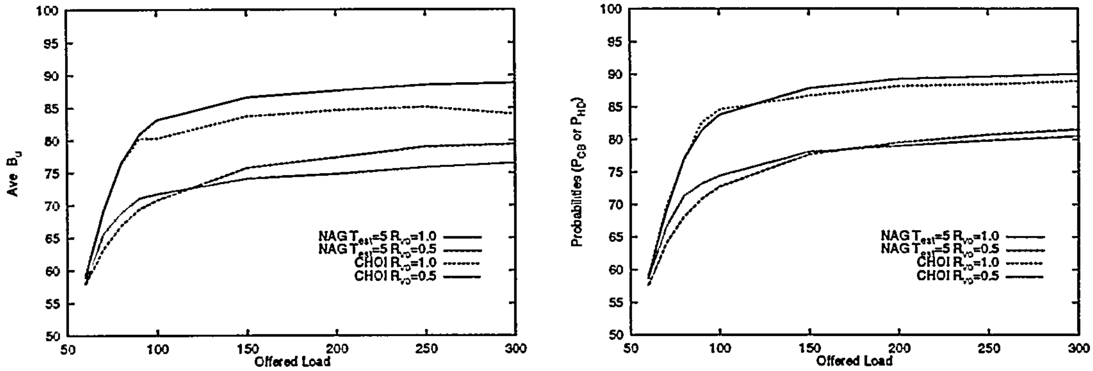
Figure 10 plots P_{CB} and P_{HD} as the offered load increases for **NAG** with $T_{est} = 5$ and **CHOI**. Both schemes are found to achieve the design goal for the most of offered loads examined. As long as the design goal is met, which of



(a) high mobility

(b) low mobility

Figure 10: Comparison of NAG and CHOI using P_{CB} and P_{HD} vs. offered load.



(a) high mobility

(b) low mobility

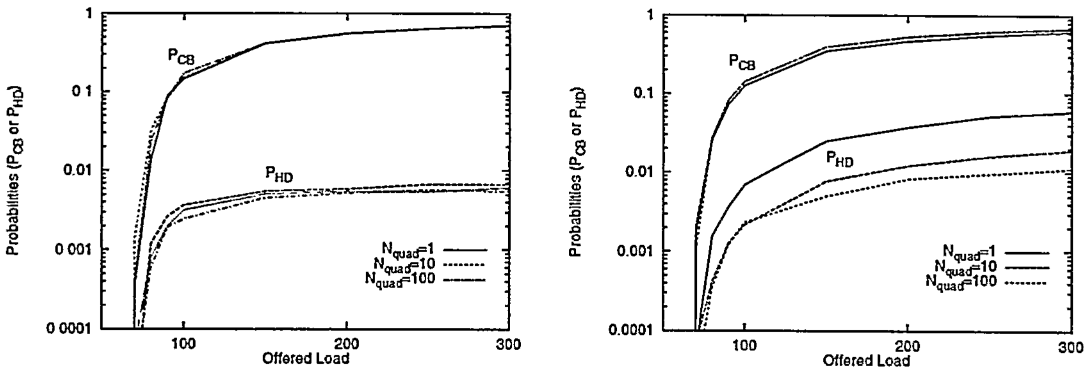
Figure 11: Comparison of NAG and CHOI using the average utilized bandwidth B_u vs. offered load.

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 $R_{vo} = 1.0$ are: (1) high-mobility: 0.695 (CHOI) and 0.672 (NAG); and (2) low-mobility: 0.682 (CHOI) and 0.676 (NAG).

Figure 11 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 10 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 much 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 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 two schemes. First, we examine their dependency on the mobility-estimation accuracy, which can be represented by the size of the cached history used for mobility estimation, i.e., the size of maximum hand-off estimation function, N_{quad}. Figure 12 plots P_{CB} and P_{HD} as the offered load increases for (a) CHOI and (b) NAG with $N_{quad} = 1$, 10, and 100. Note that we have thus far used $N_{quad} = 100$. From Figure 12 (a), we observe that CHOI doesn't depend much on N_{quad} as the performances for different values of N_{quad} are almost the same. It is remarkable that CHOI achieves the design goal even with $N_{quad} = 1$, 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 12 (b) shows that NAG starts to violate the design goal in the over-loaded region with $N_{quad} = 10$. This implies that NAG requires very accurate mobility estimation. Note that this difference of dependency 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 hand-off probability p_h in Eq. (3) is also affected greatly by N_{quad}.

Figure 13 shows the average number of calculations of the hand-off probabilities p_h to decide admissibility upon request of a new connection. NAG considers both incoming and outgoing hand-offs by calculating p_h's and p_s's. The calculation of p_s necessitates 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 depending on the condition of adjacent cells. We observe that NAG



(a) CHOI

(b) NAG

Figure 12: Comparison of NAG and CHOI using the dependency on the value of N_{quad}.

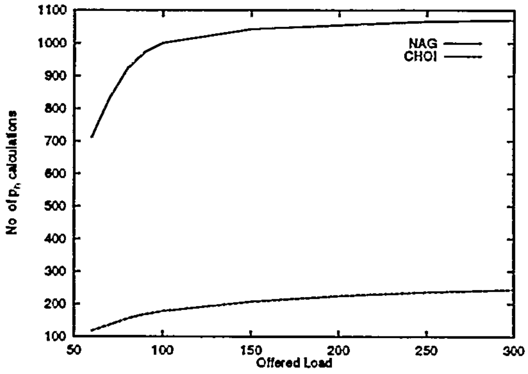


Figure 13: Complexity comparison of NAG and CHOI using the number of hand-off probability p_h calculations for an admission test.

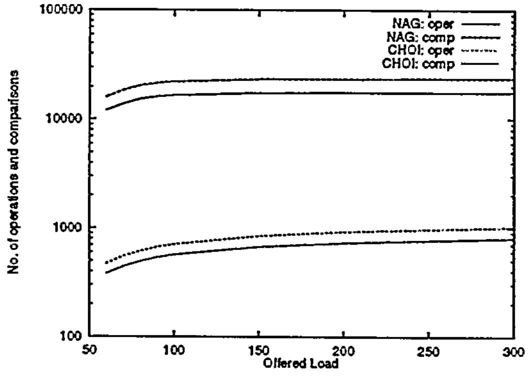


Figure 14: Complexity comparison of NAG (with $N_{quad} = 10$) and CHOI (with $N_{quad} = 1$) using the average numbers of numerical operations and comparisons for an admission decision.

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 dependency on N_{quad} and the number of p_h calculations. Figure 14 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 summations of Eq. (3). 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 Eq. (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 to 25.7 times of that of CHOI in terms of the number of numerical operations, and about 29.6 to 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.

Table 1 summarizes the comparison among the four different schemes considered in this paper. Note that AG is not based on the history-based mobility estimation, but based on the mobility specification, which is practically difficult to obtain. BHARG is not based on the history-based mobility estimation either. In fact, we didn't consider how to predict the next cell of a mobile for this scheme. So, their complexity can't be compared fairly with the other two schemes.

8 Concluding Remarks

In this paper, we compared four admission-control schemes for newly-requested connections in QoS-sensitive cellular networks in order to limit the hand-off dropping probability below a pre-specified target value or make it absolutely zero. We made the admission-control scheme NAG utilize the mobility-estimation scheme developed originally 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 hand-off dropping probability zero even under an impractical assumption by evaluating the performance of AG. Another per-connection bandwidth reservation scheme BHARG was also found to be very expensive even though it is less expensive than AG while achieving lower P_{CB}. We can conclude that per-connection bandwidth reservation is too expensive to be practical in general.

	CHOI	NAG	BHARG	AG
P_{HD}	bounded	bounded with T_{est}	virtually zero	guaranteed zero
P_{CB}	about the same	about the same	second worst	worst
Complexity	1	at least 17 times	N/A	not based on history
T_{est}	adapted	shoud be assigned	N/A	N/A

Table 1: Summary of the comparison among **CHOI**, **NAG**, **BHARG**, and **AG**.

NAG was shown to require much more memory and computation as compared to **CHOI** to meet the design goal. **NAG** is also observed to depend greatly on the design parameter T_{est}, which is difficult to adapt in real world. On the other hand, the admission-control scheme **CHOI** is robust to the inaccuracy of mobility estimation thanks to the mobility estimation time window control while meeting the design goal over the entire range of the examined offered loads even with much lower memory and computation complexity. It is concluded that **CHOI** is preferable to **NAG** and practically more attractive.

References

[1] S. Choi and K. G. Shin, "Predictive and adaptive bandwidth reservation for hand-offs in QoS-sensitive cellular networks," in *Proc. ACM SIGCOMM'98*, Vancouver, British Columbia, September 1998.

[2] S. Choi and K. G. Shin, "Exploiting mobility estimation for bandwidth reservation and admission control in cellular networks," submitted for publication, July 1998.

[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 Trans. on Networking*, vol. 5, no. 1, pp. 1–12, February 1997.

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

[5] S. Lu, K.-W. Lee, and V. Bharghavan, "Adaptive service in mobile computing environments," in *Proc. Intn'l Workshop on Quality of Service (IWQoS'97)*, 1997.

[6] M. Naghshineh and M. Schwartz, "Distributed call admission control in mobile/wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 14, no. 4, pp. 711–717, May 1996.

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

[8] 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.

TCP Source Activity and its Impact on Call Admission Control in CDMA Voice/Data Network

Sanjoy Sen. Jastinder Jawanda, and Kalyan Basu
Nortel Wireless Networks
2201 Lakeside Boulevard
Richardson, TX 75023
{sanjoy, jjawanda, kbasu}@nortel.com

Naveen K. Kakani and Sajal K. Das
Center for Research in Wireless Computing (CReW)
Department of Computer Science
University of North Texas
Denton, TX 76203-1366
{naveen, das}@cs.unt.edu

Abstract

A new call admission scheme is proposed for IS-99-like CDMA voice/data network. The scheme is based on controlling the activity of asynchronous data users at the base station controller (BSC) as well as controlling the round trip frame delay between the BSC and the mobile host (for each TCP segment) which effectively reduces the airlink activity, thus leading to an increase in the CDMA soft capacity. An expression for the average data activity is derived by modeling, the interaction between the TCP and the RLP (radio link layer protocol at the BSC providing limited reliability over the airlink), using a mean value analysis. Numerical results show the interactions between the various TCP/RLP parameters like TCP source activity and average round trip frame delay, and also capture the soft capacity enhancement. When the activities of data users are reduced, the soft capacity is enhanced by as high as 5 Erlangs.

MOBICOM 98 Dallas Texas USA
Copyright ACM 1998 1-58113-035-x/98/10...$5.00

1 Introduction

The TCP/IP has recently been recommended as a major protocol suite for the IS-99 standard for transmitting asynchronous data/fax over the CDMA network. Under such a scenario, the voice and data users will coexist in the network and vie for the available capacity. Hence, the network capacity for both types of users will degrade. With the growing popularity of the TCP, research was focused in improving its performance which resulted in the standard EIA/TIA/IS-99. The TCP is designed for wireline network where the link layer frame error rate (FER) is very small (below 10^{-8}). But due to random fading and shadowing effects of the radio link, the FER can be as high as 10^{-1}. To shield the TCP performance from the idiosyncracies of the wireless link, the link layer ARQ-based protocols (e.g., RLP in IS-99) are designed to provide mechanism to recover from large fading errors.

In this work, we evaluate the link error rate for the wireless link and use this as an upper bound for the wireline link performance. The *capacity* of a CDMA sector (cell) is essentially "soft" and depends on many parameters, such as soft hand-off, user activity, etc. Also, the majority of data transmissions are essentially non-real-time and can be given lesser priority compared to real-time voice users. This paper describes a simple technique which can be used to effectively increase the capacity in a CDMA data network, where multiple voice and data users are simultaneously active.

Upon receiving a call request, a TCP session is established between the mobile station (MS) and the corespondent host. The route taken by the TCP segments can be through many networks under varying network conditions. In our model, two important aspects of such internetworks have been considered, namely, the problem of packet losses due to congestion and the *round trip delay* (RTD) for the segment moving through intermediate nodes of the network. While the first aspect leads to a decrease in the *congestion window* (hence decreasing network throughput), the second one leads to an increase in the value of the *round trip timer* (RTT) and hence. modifies the retransmission rates. Both of these aspects have a significant impact on the TCP connection between the mobile host and the correspondent host.

After estimating the RTD and the average window size which follows from the protocol, the average data activity of the TCP source (MS or the remote host) can be computed. This data activity has a direct bearing on the activity of the radio frequency (RF) link-i.e.,IS-95 airlink-between the user and the base station. By controlling the *activity* of asynchronous data users, the system can decrease their RF capacity usage and hence admit new users into the system. Therefore, the fundamental contribution of this paper is to relate the TCP activity to the activity of the physical layer wireless link, and develop a call admission scheme to be implemented at the base station controller (BSC) of a network.

This paper is organized as follows. Section 2, starts with a brief summary of the salient features of the IS-99 protocol standard with primary focus on the TCP and RLP. This section also describes the motivation behind our work. and derives the main result dealing with the average activity of an asynchronous data user. Section 3 describes the new call admission algorithm to be implemented at the BSC. It also presents numerical results showing the effective capacity gain. Section 4 concludes the paper.

2 Performance Modeling of TCP/RLP

Let us consider a simple data flow path as shown in Figure 1 where the data from the application layer is broken down into TCP segments in the transport layer, and these segments are further divided into frames in the RLP layer before they are finally transmitted over the airlink. This is a simplistic model of the protocol stack recommended by the TIA/EIA/IS-99 standard. According to this standard, the maximum segment size (MSS) should not be smaller than 536 bytes, and the advertised window size should be no smaller than $2 \times MSS$ and no bigger than $4 \times MSS$. The radio link protocol (RLP) uses the *automatic repeat request* (ARQ) error control mechanism to reduce the FER.

When transmitting data, the RLP is a pure NAK-based finite selective repeat protocol with a bounded number of retransmission attempts. The essential scheme proposed in IS-99 is as follows. When the RLP at the receiving end finds a frame in error (or is missing), it sends back a NAK requesting for retransmission of this frame and sets a timer. When the timer expires for the first attempt, the RLP resets the timer and sends back NAK twice, each of which triggers a retransmission of the requested frame. Note that between two retransmission *trails*, either the base station receives a NAK to the RLP packets sent or RLP times out and the number of RLP packets sent during each retransmission trail depends upon the retransmission trail number. During each RLP trail, the (identical) RLP packets are sent one after the other without waiting for a NAK or RLP time out. Each RLP packet sent during each retransmission trail is called a retransmission attempt. The number of RLP packets sent during each retransmission trail is nothing but the retransmission trail number. Since the number of allowable retransmission trails is finite, it cannot completely eliminate detectable errors. In this way, the number of attempts per retransmission increases by one with every retransmission trail. The maximum number of retransmission trails recommended in the IS-99 standard is 3.

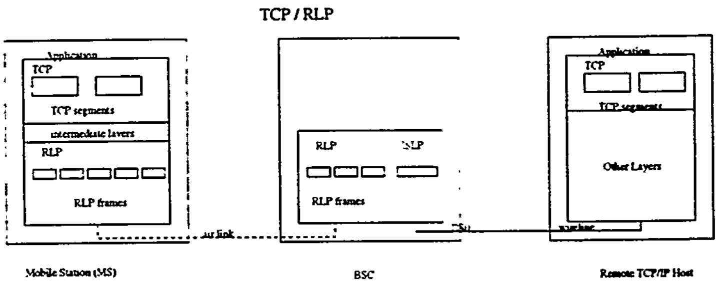


Figure 1: A simple Data flow path for the TCP/RLP

2.1 Performance Analysis at the RLP Layer

Recall that the RLP layer is introduced to reduce the FER so that the overall performance, especially *throughput*, at the TCP layer may not suffer significant degradation. The throughput of the TCP is largely dependent on the TCP packet error rate (PER) and the average packet round trip delay (RTD) which in turn is dependent on the FER of the RLP frame eventually passed on to its upper layer and the average round trip delay. Note that the RLP layer does not correct all the detectable errors.

Following the outline of the analysis by Bao [1], we proceed as follows and present our new results. Let p denote the probability of a frame being in error in the airlink, which is obtained from the frame error rate. Let us also define the following parameters for each frame, where $1 \leq i \leq j$ and $1 \leq j \leq 3$,

X_{ij} = ith retransmission frame at the jth retransmission trail received correctly at the destination

Y_{ij} = ith NAK frame at the jth retransmission trail received correctly at the source.

A_j = the missing frame not received correctly at the jth retransmission trail.

B_i = the missing frame not received correctly up to the end of the ith retransmission trail.

Assume the probability $P(X_{ij}) = P(Y_{ij}) = 1 - p$. Therefore, if a frame is not received correctly at the jth retransmission trail. we have

$$P(A_j) = \Pi_{i=1}^{i=j} (1 - P(X_{ij})P(Y_{ij})) = (p(2-p))^j.$$

Since the probability that the frame is aborted after the nth (maximum nuber of retransmissior trails) retransmission trail is equivalent to the probability that the frame is not received correctly upto the end of the nth retransmission trail. we obtain

$$P(B_n) = p\left[\Pi_{j=1}^{n} P(A_j)\right] = p(p(2-p))^{\frac{n+n^2}{2}}.$$

Let C_{ij} denote the first frame received correctly at the destination, which is the ith retransmitted frame at the jth retransmission trail. The probability of occurrance of this event is the probabilty that the packet was not sent successfully at the end of the $(j-1)$th retransmission trail times the probability of transmitting the frame successfully at the ith attempt during the jth retransmission trail and the probability of failure in sending the frame during the $i-1$ attempts during the jth retransmission trail. So

$$\begin{aligned} P(C_{ij}) &= P(B_{j-1})P(X_{ij})P(Y_{ij}) \times \\ &\quad \Pi_{k=1}^{k=i-1}(1 - P(X_{kj})P(Y_{kj})) \\ &= p(1-p)^2(p(2-p))^{\frac{j(j-1)}{2}+(i-1)}. \qquad (1) \end{aligned}$$

Therefore the probability, P_f, that a frame can be transmitted successfully over the RLP is given by the sum of the probabilities of sending the frame successfully during the first attempt (given by $P(C_{00})$) and the probabilties of successfull transmission during the retransmission trails.

$$\begin{aligned} P_f &= P(C_{00}) + \sum_{j=1}^{n}\sum_{i=1}^{j} P(C_{ij}) \qquad (2) \\ &= 1 - p + \frac{p(1-p)^2}{1-p(2-p)}\left[1 - (p(2-p))^{\frac{n(n+1)}{2}}\right] \end{aligned}$$

Assumed in this model is that a TCP segment can be divided into an integral number of frames. Let there be N_s frames per TCP segment. For the successful transmission of a segment, all the N_s frames have to be transmitted successfully over the RLP. Hence, the probability of a segment loss is

$$P_{segloss}^{airlink} = 1 - P_f^{N_s} \qquad (3)$$

Let us assume that the end-to-end frame transportation delay at the physical layer, denoted by T, is fixed. Denoting τ as the inter-frame time, the average time taken to transmit an RLP frame is given as

$$\begin{aligned} T_{avg}^{RLP} &= T(1-p) + \sum_{j=1}^{n}\sum_{i=1}^{j} P(C_{ij})(2jT + 2(i-1)\tau) \\ &= T(1-p) + p(1-p)^2 \times \\ &\quad \sum_{j=1}^{n} (p(2-p))^{\frac{j(j-1)}{2}} (2jT(\frac{1-(p(2-p))^j}{1-p(2-p)}) + \\ &\quad 2\tau(\frac{p(2-p)(1-(p(2-p))^{j-1})}{(1-p(2-p))^2} - \\ &\quad \frac{(j-1)(p(2-p))^j}{1-p(2-p)})) \qquad (4) \end{aligned}$$

2.2 Performance Analysis of the Wireline link

In this subsection, we consider the loss and delay conditions for the wireline network segment, assuming a simplistic network model. Let $P_{segloss}^{wireline}$ denote the segment loss probability in the wireline network due to congestion or other types of network specific error conditions.

The probability of a segment loss due to the airlink or wireline network failure can then be written as

$$P_{seg}^{loss} = 1 - (1 - P_{seg\,loss}^{air\,link})(1 - P_{seg\,loss}^{wireline}) \qquad (5)$$

From the segment loss probability, the average number of transmissions required to transmit a segment is obtained as follows. Let $P_{seg}(i)$ denote the probability that a segment is transmitted successfully at the ith attempt. Then

$$P_{seg}(i) = \left[P_{seg}^{loss}(2 - P_{seg}^{loss})\right]^{i-1} (1 - P_{seg}^{loss})^2 \qquad (6)$$

We shall use Equation (6) for estimating the average TCP window size.

Next, we consider two important aspects of a TCP/IP network, namely, variations of the *average congestion window*, in segments W_{avg} and the average *round trip delay* (RTD_{avg}) due to segment losses. We are interested in an expression for the average segment transmission rate, λ_{out}, of the TCP source, which we will refer to as the TCP *source activity*. In subsection 2.2.2, we will relate the source activity to the average activity of the RLP transmission over the airlink, and demonstrate how we can effectively enhance the reverse link capacity of a CDMA system by controling this activity.

The average segment transmission rate can be approximated as

$$\lambda_{out} = \frac{W_{avg}}{RTD_{avg}} \tag{7}$$

In the analysis that follows, we derive expressions for W_{avg} and RTD_{avg} using a TCP model similar to "TCP-OldTahoe". This is primarily used to keep the analytical model tractable. We also assume here that the delay-bandwidth product for the combined wireline-wireless network is high. This is particularly true for an IP network with wireless access in which a typical link bandwidth is of the order of Mbps and the radio link delay is of the order of 100 msec.

2.2.1 TCP Window Model

The TCP uses a *slow start* mechanism to avoid congestion [4], which means that for every (or a group of) successful transmissions, the window is increased by a certain amount. On the other hand when there is a segment loss which is notified to the source by a retransmission time-out or several duplicate acknowledgements, the window size is decreased. This increase or decrease of window size takes place in units of segments. There are two parameters, namely, the maximum window size (W_{max}) and the minimum window size (W_{min}), to restrict the variations of the TCP window.

In the following model, we have assumed a simplified general version of this scenario. According to the TCP standard for the IS-99 protocol, the congestion window is required to have a minimum size of $2 \times MSS$ and a maximum size of $4 \times MSS$, where MSS is the maximum segment size whose smallest value is 536 bytes. We assume a segment size equal to the MSS. It can be shown that the window increment process can be represented by a discrete Markov chain with state space $\{S_i = i\}$, where i is the window size. In our case, $i \in \{2, 3, 4\}$. This is a simplified model which captures the essence of modeling the congestion window and can be extended to more elaborate versions as demonstrated in [3]. If the initial window size is two or three (segments) and if there is a successful transmission, the window is incremented by one (see Figure 2). When the window has a size of four or three, on failing to transmit, the window is set to two segments (corresponding to state 2). When the system is in state 4 and there is a successful transmission, the window size remains unchanged.

Given these constraints, we develop a Markovian model as shown in Figure 2. The following transition probabilities are computed.

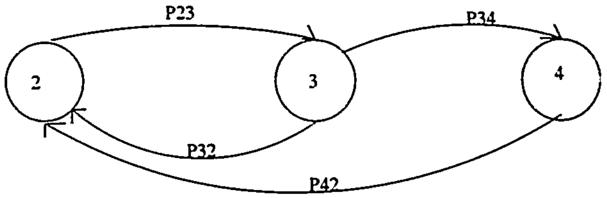


Figure 2: Modeling the TCP congestion window

Probability of a successful transmission is given by

$$P_{23} = \sum_{i=1}^{\infty} P_{seg}(i) \tag{8}$$

Where $P_{seg}(i)$ is the probability that the segment is transmitted successfully at the ith attempt.

Probability of a successful transmission in first attempt

$$P_{34} = P_{seg}(1) \tag{9}$$

Probability of an unsuccessful transmission

$$P_{32} = P_{42} = P_{seg}^{loss}(2 - P_{seg}^{loss}) \tag{10}$$

This Markov chain can be easily solved for the steady state probabilities Π_i (for state i) as follows :

$$\Pi_2 = \frac{1}{1 + \frac{P_{23}}{P_{32}+P_{34}} + \frac{P_{34}P_{23}}{P_{42}(P_{32}+P_{34})}} \tag{11}$$

$$\Pi_3 = \frac{P_{23}}{P_{32} + P_{34}}\Pi_2 \tag{12}$$

$$\Pi_4 = \frac{P_{34}P_{23}}{P_{42}(P_{32} + P_{34})}\Pi_2 \tag{13}$$

The average congestion window size is then given as

$$W_{avg} = \sum_{i=2}^{4} i\Pi_i. \tag{14}$$

2.2.2 Average Roundtrip Time Duration

We consider a simplified model of the network as shown in Figure 3. In the absence of any single model that can track the dynamics of the traffic build-up in the entire network, the idea here is to study the effects of congestion selectively in a few strategic nodes in the network. One such node is the gateway node between the wireless network and the wireline one, called the *interworking gateway* (IWG). The simplifying assumption here is that all the queues in the network are lumped together into an $M/G/1$ queue in the IWG. The IWG node is critical in the sense that there is a serious performance difference (e.g., link speed) between the wireless network with its wireline counterpart. Therefore, we estimate the effects of the delay at the IWG as well as the delay in the airlink, on the roundtrip time of a TCP segment. This will also determine their impact on the TCP source activity. All other links are assumed to have fixed amount of physical link transport delay. The fixed delays in the wireline links are assumed to be D_{IWG_BS} and D_{H_IWG} between the IWG and the base station (BS), and the host (H) and IWG, respectively. This model, albeit simplified, captures the dynamics of the system quite accurately. Then, the RTD can be computed as a function of the average frame delay T_{avg}^{RLP}, D_{IWG_BS}, D_{H_IWG} and the average delay in the IWG node of the M/G/1 queue.

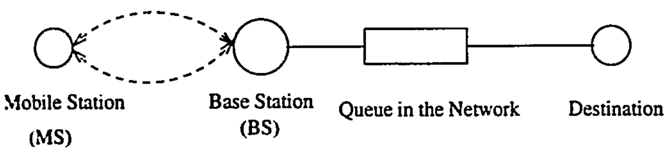


Figure 3: A typical model for the network

The M/G/1 model is chosen because on the input side of the queue, the packet arrival process may be approximated as a Poisson process (however, this is not always true given the nature of WWW sources) and with the arrival rate same as that of packets from the TCP source. The service rate is assumed as general because the network behavior is based upon the congestion avoidance algorithms, retransmission algorithms, varying data-rate networks and so on, and hence it is not easy to predict the service time distribution. Given this model, the average waiting time in the queue is given by the well known *Pollaczek-Khinchin* equation [5]:

$$T_{avg}^{queue} = \frac{\lambda \bar{X^2}}{2(1-\rho)} \qquad (15)$$

where λ is the Markovian packet arrival rate at the network queue,

$\bar{X}$ is the mean service time at the site of the queue,

$\bar{X^2}$ is the variance in the service time, and

$\rho = \lambda \bar{X}$

The average delay for transmitting a TCP segment over the airlink is given as

$$\begin{aligned} D_{seg}^{avg} &= [T + (N_s - 1)\tau](1-p)^{N_s} + \\ & \sum_{k=1}^{N_s} \binom{N_s}{k} p^k (1-p)^{N_s-k} (T + (N_s - 1)\tau \\ & + k \sum_{j=1}^{n} \sum_{i=1}^{j} P(C_{ij})\,(2jT + 2(i-1)\tau)) \\ &= T + (N_s - 1)\tau + \\ & pN_s \sum_{j=1}^{n} \sum_{i=1}^{j} P(C_{ij})\,[2jT + 2(i-1)\tau] \quad (16) \end{aligned}$$

The term T is the duration of transmission of the TCP acknowledgement packet over the airlink. Hence, the average value of the round trip delay is given as

$$RTD_{avg} = 2(D_{H\_IWG} + D_{IWG\_BS}) + D_{seg}^{avg} + T + T_{avg}^{queue} \, (17)$$

One way of estimating D_{H_IWG} by the IWG is by using the Unix *ping* tool or the ICMP *echo/echo reply* messages between IWG and the remote host. The delay in the part between the BS and IWG can be determined from the type of link (e.g. DS0's) between them.

Here, the assumption is that the TCP acknowledgement packet (assumed to be small compared to a segment) is never lost over the airlink. Also, we assume that the acknowledgement packet does not suffer any delay at the IWG which is true given that the link speed in the wireless network is considerably less than in the wireline network.

Using the average window size given in Equation (14) and the average value of the round trip delay as in Equation (17), we can estimate the *TCP source activity* as,

$$\lambda_{out} = \frac{W_{avg}}{RTD_{avg}} \qquad (18)$$

Note that, this mean value analysis can be easily extended to any other wireless data network such as GPRS or CDPD, leading to the computation of the TCP source activity. Since no variation of the packet transmission rate over the rest of the wireline network is assumed, the average packet rate at the input of the queue is the same as that at the output of the TCP source. Hence, $\lambda = \lambda_{out}$. Combining Equations (15), (17) and (18), we obtain

$$\frac{W_{avg}}{\lambda} = 2(D_{H\_IWG} + D_{IWG\_BS}) + D_{seg}^{avg} + T + \frac{\lambda \bar{X^2}}{1 - \lambda \bar{X}} \, (19)$$

which can be written as a quadratic equation of λ,

$$A\lambda^2 + B\lambda + C = 0 \qquad (20)$$

where,

$A = B'\bar{X} - \bar{X^2}$
$B = -(B' + W_{avg}\bar{X})$
$C = W_{avg}$
$B' = 2(D_{H\_IWG} + D_{IWG\_BS}) + D_{seg}^{avg} + T$

2.2.3 Numerical Results

Figure 4 shows the variation of the segment transmission rate (λ) computed by solving Equation (19) with various values of segment loss probability (P_{seg}^{loss}) due to congestion (and other reasons) in the wireline network. In this figure, $\bar{X} = 0.1$

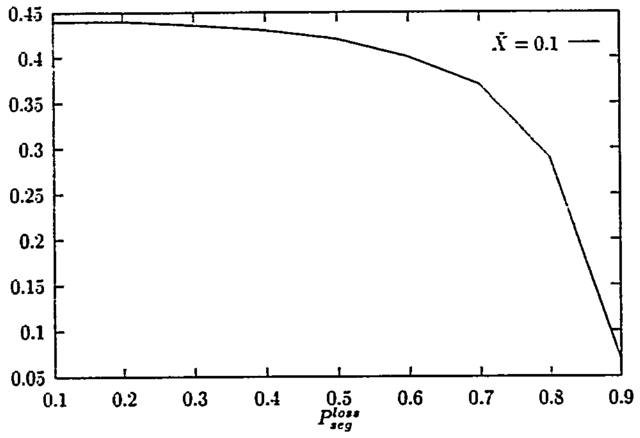


Figure 4: Variation of average segment rate with segment loss (congestion) probability

As the congestion increases, the TCP window shrinks which leads to a decrease in the average segment transmission rate of the source. Note that, as the segment loss probability increases, λ decreases slowly at first. but later plunges down drastically as the window decreases fast due to a large amount of congestion signals.

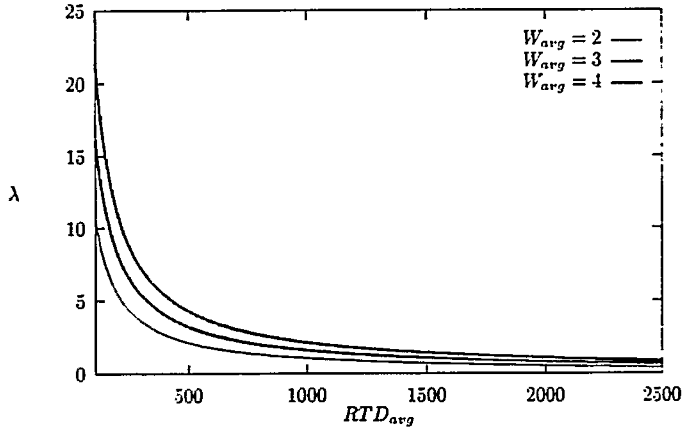


Figure 5: Relationship between average segment rate and average round trip delay (RTD_{avg}) for various TCP window sizes (W_{avg})

Figure 5 shows the variations of activity λ with RTD_{avg} and W_{avg} from Equation (18). The hyperbolic relationship between λ and RTD_{avg} is evident. Figure 6 shows the variation of RTD_{avg} with various frame error rates in the airlink. The variation in frame error probability leads to variations in the average frame transmission delay which in turn leads to a change in RTD_{avg}. It is seen that, although the average round trip time increases as expected with the frame error rate, it saturates to an almost fixed value for higher values of frame error rates.

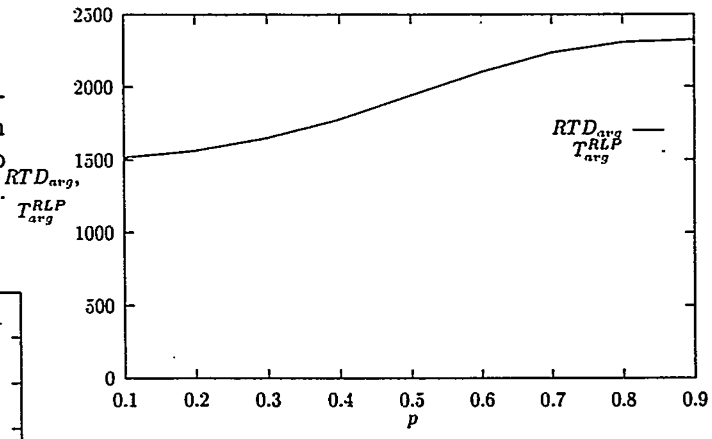


Figure 6: Variation of average RTD and airlink frame delay (T_{avg}^{RLP}) with airlink frame loss probability p

In the next section, we will show how we can provide an efficient call admission mechanism in an integrated voice/data CDMA network by controlling the source activity from the radio link layer protocol (RLP).

3 A Call Admission Scheme in IS-99

The user activity plays a crucial role in determining the CDMA soft capacity. If the average activity of one or more users decreases, additional calls can be admitted. Taking advantage of the period of inactivity of a typical voice user, data can be transmitted for the same user over the already established circuit. Thus, multiple frames can be transmitted for primary (voice) and secondary (data) traffic within the same frame-timeslot (20 msec) of the IS-95 CDMA physical layer.

The system does not have any control on the activity of the individual voice users. Due to the non-real-time nature of the average asynchronous data transmissions and also to a certain extent, fax (it does not matter if these transmissions are delayed by a couple of minutes or more), additional capacity can be gained by forcefully controlling the activity of these services at some intermediate node of the network, on top of the usual activity of the data source. The proposed simple scheme will affect the activity of the TCP/IP data source and help in controlling the activity of frame transmission over the airlink.

The average time the airlink is active in one direction for transmitting a segment successfully is given by

$$\begin{aligned} A_{seg}^{avg} &= [T + (N_s - 1)\tau](1 - p)^{N_s} + \\ &\sum_{k=1}^{N_s} \binom{N_s}{k} p^k (1 - p)^{N_s - k} (T + (N_s - 1)\tau \\ &+ k \sum_{j=1}^{n} \sum_{i=1}^{j} P(C_{ij})\,(jT + (i-1)\tau)) \\ &= T + (N_s - 1)\tau + \\ &pN_s \sum_{j=1}^{n} \sum_{i=1}^{j} P(C_{ij})(jT + (i - 1)\tau) \quad (21) \end{aligned}$$

The average time to transmit all the segments in an average congestion window is

$$A_w^{avg} = A_{seg}^{avg} \times W_{avg} \quad (22)$$

Hence, the data activity over the airlink is given by

$$A_D = \frac{A_w^{avg}}{RTD_{avg}} = A_{seg}^{avg} \cdot \lambda_{out} \quad (23)$$

where the expression for RTD_{avg} is given by Equation (17).

Our call admission scheme applies a fixed amount of delay to a fixed number of RLP frames in both the uplink and downlink directions at the BSC. This delay ultimately translates to increasing the RTD for the segments which leads to a decrease in the data transmission activity over the airlink, by virtue of Equation (23). Thus, the frames for a few particular transmissions are buffered for a fixed duration (determined by the call admission scheme) in the BSC, and this affects some TCP parameters like RTD and RTT estimates, forcing the reduction of the TCP source activity, λ_{out}. By providing the same amount of delay to the frames of a group of data transmissions, the average data activity over the airlink can be reduced to a significant extent. This provides additional soft capacity and allows the system to add new voice users to the sector.

Next we present numerical results to demonstrate the voice capacity gain for fixed amount of delay to data users.

3.1 Numerical Results

Considering the IS-95 reverse link, Viterbi's formula is used to estimate the gain in voice traffic capacity (in Erlang) when the data frames for a fixed number of users are delayed. The user capacity in Erlang is given as

$$K_u = \frac{\frac{W}{R} \cdot loading}{\frac{E_b}{I_0} \cdot A_D \cdot 1.55} \quad (24)$$

where,
$\frac{W}{R}$ is the spread gain. $\frac{E_b}{I_0}$ is the bit energy to interference noise density, A_D is the average user activity and 1.55 is the other-user-interference factor in the current and neighboring cells.

When the frames of K_u'' out of K_u users are delayed by an amount δ msec, the average data activity is given as

$$A_D' = \left(\frac{K_u - K_u''}{K_u}\right) \frac{A_w^a}{RTD} + \left(\frac{K_u''}{K_u}\right) \frac{A_w^a}{RTD + \delta} \quad (25)$$

The *loading* in the system due to data and voice users is assumed not to exceed 0.5 (i.e.,50%). The following equation determines the voice capacity K_u^{voice}, given an average voice activity of 0.45.

$$C(0.45\,K_u^{voice} + A_D' * K_u^{data}) = 0.5 \quad (26)$$

where $C = \left(\frac{E_b}{I_0} * 1.55\right) / \frac{W}{R}$
Let us evaluate the performance of the new algorithm using the following parameter values:

$$\begin{aligned} p &= 0.01 \\ P_{seg\ loss}^{wireline} &= 0.00001 \\ N_s &= 25 \\ T &= 100msec \\ \tau^{\cdot} &= 20msec \end{aligned}$$

The maximum number of data users in the sector is assumed to be 15.

Table 1: **Variation of airlink data activity with delay for a single user**

delay (msec)	data activity
20	0.57688542
100	0.56315985
120	0.5598299
200	0.54689482
500	0.50328739
750	0.47192912
1000	0.44424932
1200	0.42433852

The results are displayed in Tables 1 and 2. Table 1 and Figure 7 show the variation of users activity as we increase the delay. As the delay is increased for each user, the time taken to send a fixed amount of data increases and hence the activity of each user drops.
In Table 2, the column headings signify the amount of capacity enhancement obtained. For example, to increase the voice capacity by 1 Erlang, a 120 msec delay is to be applied to 7 data users. If the required number of users is not available in the system, then we can have the same capacity gain by applying a 500 msec delay to 2 data users. The blank entries in the table signify

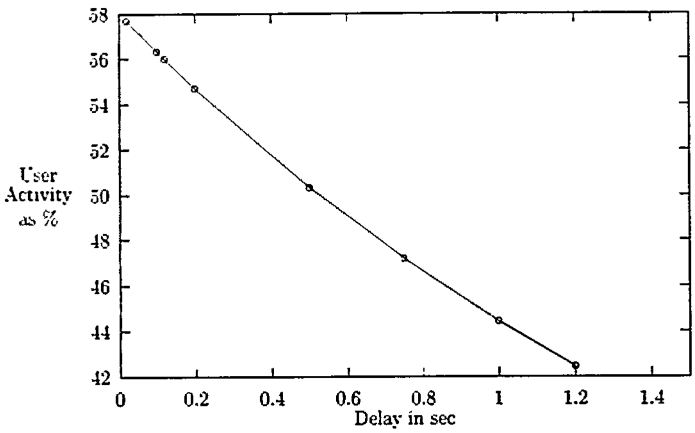


Figure 7: Data activity of each user vs. delay

Table 2: **Amount of delay and the # of users delayed for a given capacity enhancement**

delay (msec)	1 Erl	2 Erl	3 Erl	4 Erl	5 Erl
120	7				
200	4				
500	2	8	14		
750	1	6	10	14	
1000	1	4	8	11	14
1200	1	4	7	10	12

that the number of data users is insufficient to obtain the corresponding capacity increase with the specified amount of delay. Depending on the excess capacity required and the number of data users available in the sector, the operating point can be selected by the call admission algorithm from Table 2. This is the essence of our call admission scheme.

4 Conclusions

We have proposed a new call admission scheme for CDMA voice/data systems. The scheme is based on controlling the activity of asynchronous data users at the base station controller. In order to derive an expression for average data activity, the interaction between TCP and RLP (link layer protocol sitting in the base station controller) is modeled using a mean value analysis. By controlling the round trip frame delay between the base station and the mobile host, the airlink activity is effectively reduced, leading to an increase in the CDMA soft capacity. The decrease in the airlink activity with this round trip delay is formulated with the help of analytical model. Experimental results show that a soft capacity enhancement as high as 5 Erlang is possible with a maximum of 14 data users in the sector with a frame delay of 1200 msecs.

5 Acknowledgements

This work is supported by Texas Advanced Research Program (TARP-003594-013) and Nortel Wireless Networks, Richardson, Texas.

References

[1] G. Bao. "Performance evaluation of TCP/RLP protocol stack over CDMA wireless link", *ACM Wireless Networks* Journal, Vol. 2, 1996, pp 229-237.

[2] M. Mathis, J. Semke, J. Mahdavi, "The Macroscopic Behavior of the TCP Congestion Avoidance Algorithm", *Proceedings of ACM SIGCOMM*, 1996.

[3] A. Kumar, "Comparative Performance Analysis of Versions of TCP in a Local Network with Lossy Link". to appear in *IEEE/ACM Transactions on Networking*, 1998.

[4] *TCP/IP Illustrated*, W.R. Stevens, Addison-Wesley, 1994.

[5] *Probability, Stochastic Processes and Queueing Theory*, R. Nelson, Springer-Verlag, 1996.