Dynamic Bandwidth Reservation in Cellular Networks Using Road Topology Based Mobility Predictions

Wee-Seng Soh and Hyong S. Kim
Department of Electrical and Computer Engineering
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213
Email: {wsoh,kim}@ece.cmu.edu

***Abstract*— In cellular networks, an important call-level Quality-of-Service (QoS) issue is how to limit the probability of forced termination during handoffs. One solution is to predict the trajectory of mobile terminals so as to perform bandwidth reservation in advance. With the vision that future mobile devices are likely equipped with reasonably accurate positioning capability, we propose a novel mobility prediction technique that incorporates both mobile positioning information and road topology knowledge. We then develop an adaptive bandwidth reservation scheme that dynamically adjusts the reservation at each base station according to both incoming and outgoing handoff predictions generated using our mobility prediction technique. We evaluate the performance of the scheme via simulations, along with six other schemes for comparison purposes. Results agree with intuition that schemes which incorporate more knowledge are able to achieve better reservation efficiency. Our scheme is shown to achieve the best efficiency among all realizable schemes simulated.**

I. INTRODUCTION

When a mobile terminal (MT) attempts to hand off from one cell to another, it may encounter forced termination due to bandwidth shortage at the target cell. From a user's point of view, the forced termination of an ongoing call is more objectionable than the blocking of a new call request. Therefore, handoff-requests are generally prioritized over new call requests. In the classic handoff prioritization problem, each base station (BS) prioritizes handoff-requests by setting aside some bandwidth that could only be utilized by incoming handoffs. Since any such reservation would inevitably increase the blocking probability of new calls (P_{CB}), and reduce the system's utilization, it is extremely important that these reservations are made as sparingly as possible while meeting the desired forced termination probability (P_{FT}).

Early work in handoff prioritization proposes the static reservation of bandwidth at each BS as a solution [1], in which a fixed portion of the radio capacity is permanently reserved for handoffs. However, such a static approach is unable to handle variable traffic load and mobility [2]. In order to meet the desired P_{FT} without over-reserving precious radio bandwidth, the amount of reservation at each BS should be dynamically adjusted according to the requirements of anticipated handoffs.

The best tradeoff between P_{CB} and P_{FT} can only be achieved if every MT's path as well as its arrival and departure times in each cell are known in advance. However, such an ideal scenario is very unlikely to occur. The next best option is to predict the mobility of MTs, and perform reservations based on these predictions. Many predictive schemes have been proposed in the literature. For example, Liu *et al.* [3] uses pattern matching techniques and a self-adaptive extended Kalman filter for next-cell prediction based on cell sequence observations, signal strength measurements, and cell geometry assumptions. In [4], Levine *et al.* propose the concept of a shadow cluster – a set of BSs to which a MT is likely to attach in the near future. The scheme estimates the probability of each MT being in any cell within the cluster for future time intervals, based on individual MT's dynamics and call holding patterns in the form of probability density functions (pdfs). Other examples of predictive reservation schemes can be found in [2], [5]–[8]. In the process of meeting the same P_{FT}, a more efficient scheme will be able to accomplish the task with a lower P_{CB} than a less efficient one. The efficiency of a scheme depends on whether the reservations are made at the right place and time, i.e., it is closely associated with the prediction accuracy. Since reservation efficiency has a direct impact on operators' revenues, there are strong incentives to design more accurate prediction schemes.

In the United States, the FCC mandates that cellular-service providers must be able to pinpoint a wireless emergency call's location to within 125 m. This spurs research in mobile-tracking techniques. One promising approach is the integration of a global positioning system (GPS) receiver in each MT. According to [9], assisted GPS positioning methods are expected to yield an accuracy of under 20 m during 67% of the time. During 2003-2009, a new batch of GPS satellites will be launched in the US that could potentially yield an accuracy within 1 m for civilian users [10]. The European Space Agency has also planned to launch their own global navigation satellite system known as GALILEO, which is also expected to deliver real-time positioning accuracy down to the meter range (95% of the time within 10 m) [11]. As more breakthroughs in positioning techniques take place, fuelled by the strong interest in location-based services from the industry, future MTs are likely equipped with reasonably accurate location-tracking capability. The time is thus ripe for active research into how such inherent capability may be harnessed for QoS provisioning in cellular networks. Specifically, we are interested in designing mobility prediction techniques that utilize real-time positioning information. This could potentially give rise to better accuracy and greater adaptability to time-varying conditions than previous methods.

While there has been previous work in the literature that attempts to perform mobility prediction based on positioning information [3], [5], [6], none of them has addressed the fact that the cell boundary is fuzzy and irregularly shaped due to terrain characteristics and obstacles that interfere with radio wave propagation. Instead, either hexagonal or circular cell boundaries have been assumed for simplicity. Another observation is that none of the previous work has integrated the road topology information into its prediction algorithm. Since MTs that are carried in vehicles are the ones that demonstrate high mobility, the integration of road information into the algorithm could potentially yield better accuracy, which is crucial for more timely and efficient reservations. With real-time location information of MTs, it is now possible to take advantage of knowledge about road layouts.

In [8], we propose a dynamic bandwidth reservation scheme that utilizes mobility predictions based on real-time mobile positioning information. It is the first such scheme that is capable of handling irregular cell boundaries. The scheme uses linear extrapolation from a MT's recent positions to predict its handoff cell and time, whereby the cell boundary is approximated as a series of points around the BS that are computed using previous handoff locations. In this paper, we introduce a novel predictive reservation scheme that utilizes knowledge of road topology, in addition to positioning information. It could potentially achieve more accurate predictions at the cost of increased complexity, but the resulting gain in reservation efficiency may justify this cost.

The remainder of this paper is organized as follows. In Section II, we present the proposed road topology based prediction scheme, while Section III describes the algorithm that utilizes these predictions for adjusting the reservation at each BS. Section IV describes the simulations that have been carried out to compare the performance of the proposed scheme with several other schemes. Finally, we give our conclusions in Section V.

II. Road Topology Based Mobility Prediction

In our proposed technique, we require the serving BS to receive regular updates about each active MT's position every ΔT, say 1 sec. This will consume a small amount of uplink wireless bandwidth (several bytes per update for each MT), which might be negligible for future broadband services. The output of each prediction has the form of a 4-tuple: [target cell, prediction weight, lower prediction limit, upper prediction limit]. The *target cell* is the MT's predicted handoff cell. The *prediction weight* is a real number between 0 and 1 that indicates how likely the prediction is correct. The *lower prediction limit* (LPL) gives a lower statistical bound for the actual remaining time from handoff, t_{remain}, with probability ζ_{L}, i.e., $P[t_{\text{remain}} \geq \text{LPL}] = \zeta_{\text{L}}$. The *upper prediction limit* (UPL) gives an upper statistical bound for t_{remain} with probability ζ_{U}, i.e., $P[t_{\text{remain}} \leq \text{UPL}] = \zeta_{\text{U}}$. Note that each MT may have more than a single 4-tuple; a 4-tuple is specified for each possible path from its current position that may lead to a handoff within a time $T_{\text{threshold}}$.

In the following, we first describe the database that is maintained at each BS to store essential information required for making the predictions. The prediction algorithm will then be described.

A. Prediction Database

The prediction tasks are assigned to individual BSs, which are expected to have sufficient computational and storage resources. In order to incorporate the road information into mobility predictions, each BS needs to keep a database of the roads within its coverage area. We shall treat the road between two neighboring junctions as a road segment, and identify each segment using a junction pair (j_1, j_2), where a junction can be interpreted as an intersection of roads (e.g., T-junction). The approximate coordinates of each junction are to be stored in the database. Since a road segment may contain bends, it can be broken down further into piecewise-linear line segments. The coordinates defining these line segments within each road segment are also recorded. All the above coordinates could be easily extracted from existing digital maps previously designed for GPS-based navigational devices. Infrequent updates to these maps are foreseen because new roads are not constructed very often, while existing road layouts are seldom modified.

The database also stores some important information about each road segment. Since two-way roads would probably have different characteristics for each direction, the database shall store information corresponding to opposite directions separately. The following summarizes the information that is stored in the database:

- Identity of neighboring segments at each junction.
- Probability that a MT traveling along a segment would select each neighboring segment. Note that this transition probability could be easily computed from the previously observed paths of other MTs.
- Statistical data of time taken to transit each segment.
- Statistical data about possible handoffs along each segment, such as probability of handoff, time in segment before handoff, and handoff positions.

With the exception of the first item listed above, the other database entries will be updated periodically every T_{database} since they are dependent on current traffic conditions.

In reality, the transition probabilities between road segments would probably vary with time and traffic conditions. For stochastic processes whose statistics vary slowly with time, it is often appropriate to treat the problem as a succession of stationary problems. We shall model the transition between road segments as a second-order Markov process, and we assume that it is stationary between database update instances so as to simplify the computations. Based on this model, the conditional distribution of a MT choosing a neighboring segment given all its past segments is assumed to be dependent

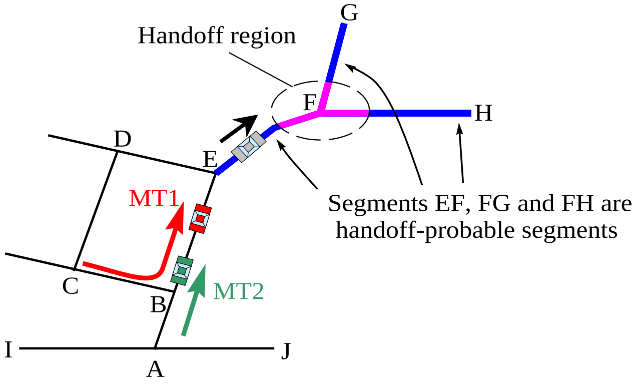


Fig. 1. Utilizing road topology information for mobility prediction.

TABLE I
NOTATIONS USED FOR ILLUSTRATING DATABASE MAINTENANCE.

Notation	Meaning
$T_{\text{thres\_max}}$	Maximum $T_{\text{threshold}}$ allowed.
$\mathcal{S}$	Set of road segments within BS's coverage area.
s_{ab}	Directional segment from junction j_a to j_b.
$\mathcal{N}(j_a)$	Set of neighboring junctions of junction j_a.
$\mathcal{N}_{\text{cells}}$	Set of neighboring cells of the cell of interest.
$\mathcal{S}_{\text{HPS}}$	Set of handoff-probable segments (HPSs) in $\mathcal{S}$.
$\mathcal{S}_{\text{RSV}}$	Set of segments in which MTs may be considered for reservations.
$P[s_{k+1} \vert s_k]$	1st order conditional transition probability, i.e., P[transit to s_{k+1}|currently s_k].
$P[s_{k+1} \vert s_k, s_{k-1}]$	2nd order conditional transition probability, i.e., P[transit to s_{k+1}|currently s_k, previously s_{k-1}].
$C_{\text{HO}}(s_{ab})$	Most probable target handoff cell if handoff occurs along s_{ab}, where $C_{\text{HO}}[s_{ab}] \in \mathcal{N}_{\text{cells}}$.
$P_{\text{HO}}[s_{ab}]$	P[handoff along s_{ab}|MT is currently on s_{ab}].
$f_{\text{transit},ab}(t)$	pdf of time taken to transit s_{ab}.
$g_{\text{HO},ab}(t)$	pdf of time spent in s_{ab} before handoff.
$h_{\text{HO},ab}(d)$	pdf of distance from j_b where handoff occurs.
X	Hop limit of routes that are considered.
$\mathcal{R}_X(s_{ab})$	Set of possible routes within X hops from s_{ab}. A route $\varphi \in \mathcal{R}_X(s_{ab})$ is a sequence of segments, starting with s_{ab}: $\{s_{ab}s_{bc}\ldots s_{yz}\}$.
$s_{\text{initial}}(\varphi)$	Initial segment of route φ.
$s_{\text{last}}(\varphi)$	Last segment of route φ.
φ'	Route φ without its initial and last segments, i.e., $\{\varphi\} = \{s_{\text{initial}}(\varphi)\} \cup \{\varphi'\} \cup \{s_{\text{last}}(\varphi)\}$.
$m_{\text{HO},ab \vert \varphi}(t)$	pdf of time taken to transit φ' and part of last segment $s_{\text{last}}(\varphi)$ before handoff.
$M^{-1}_{\text{HO},ab \vert \varphi}(q)$	q^{th} quantile of time taken to transit φ' and part of last segment $s_{\text{last}}(\varphi)$ before handoff.
$\mathcal{R}_{X,\text{HPS}}(s_{ab})$	A subset of routes from $\mathcal{R}_X(s_{ab})$, each of which terminates with a HPS, and, excluding the remaining time in current segment s_{ab}, has a median time to handoff that is within $T_{\text{thres\_max}}$.
$P_{\text{HO}}[\varphi \vert s_k]$	1st order conditional prob. that MTs in s_k would use φ and hand off at $s_{\text{last}}(\varphi)$, $\varphi \in \mathcal{R}_{X,\text{HPS}}(s_k)$.
$P_{\text{HO}}[\varphi \vert s_k, s_{k-1}]$	2nd order conditional prob. that MTs in s_k would use φ and hand off at $s_{\text{last}}(\varphi)$, $\varphi \in \mathcal{R}_{X,\text{HPS}}(s_k)$.

only on the current segment and the immediate prior segment. Using the road topology shown in Fig. 1 as an illustration, consider two MTs (MT1 and MT2) that are currently traveling from junction B towards junction E. MT1 came from segment CB previously, while MT2 came from segment AB. Based on the assumed model, the conditional probability of MT1 going to segment EF will be computed differently from that of MT2. The conditional probability of MT1 going to segment EF is

$$P[s_{k+1}{=}\text{EF} \vert s_k{=}\text{BE}, s_{k-1}{=}\text{CB}], \tag{1}$$

while that of MT2 is

$$P[s_{k+1}{=}\text{EF} \vert s_k{=}\text{BE}, s_{k-1}{=}\text{AB}], \tag{2}$$

where s_k is the current segment that the MT transits. Note that our stationarity assumption implies that the above conditional probabilities are independent of the value of k.

At the beginning of a new call, the previous segment of a MT is unknown, because it was not tracked previously. Therefore, we also need to store first-order conditional distribution in each segment, which are estimated from a subset of the data that are used to estimate the second-order conditional distribution. For instance, if we do not have any information about the previous segment of MT1 and MT2 in Fig. 1, their conditional probabilities of going to segment EF are both taken to be

$$P[s_{k+1}{=}\text{EF} \vert s_k{=}\text{BE}]. \tag{3}$$

We shall describe a road segment as a "handoff-probable segment" (HPS) if MTs have previously requested handoffs while traveling through it. For each HPS, we obtain the handoff probability as the ratio of MTs that made handoff-requests while on the segment. Also, for those MTs which made handoff-requests, we record their target handoff cell, and collect information about the time spent by them in the HPS before handoffs, as well as their handoff positions.

Using the model described above, we could determine via the chain rule the conditional probabilities of reaching and handing off at each of the HPSs from segments that are several hops away. We could also predict the remaining time before handoff for each of these possible paths, using previously collected statistical information from each segment along the path. Before we describe the prediction algorithm in Section II-B, we shall first explain how the prediction database is maintained. Table I shows the notations used. Since many of the database entries are dependent on current traffic conditions, a database update will be performed every T_{database} to ensure that the entries are current. Fig. 2 shows the procedure performed during each update.

We assume that in between the database updates, the BS shall collect all the relevant data required for the subsequent update. The procedure begins by emptying both $\mathcal{S}_{\text{HPS}}$ and $\mathcal{S}_{\text{RSV}}$ (Lines 1 and 2) so that they can be regenerated based on the newly collected data. From Lines 3 to 13, we sequentially examine every road segment within the BS's coverage area,

```
1   $\mathcal{S}_{\text{HPS}} \leftarrow \emptyset$
2   $\mathcal{S}_{\text{RSV}} \leftarrow \emptyset$
3   for each $s_{ab} \in \mathcal{S}$
4       evaluate $P[s_{k+1}{=}s_{bx} | s_k{=}s_{ab}]$
                $\forall j_x \in \mathcal{N}(j_b) - \{j_a\}$
5       evaluate $P[s_{k+1}{=}s_{bx} | s_k{=}s_{ab}, s_{k-1}{=}s_{ya}]$
                $\forall j_x \in \mathcal{N}(j_b) - \{j_a\}, \forall j_y \in \mathcal{N}(j_a) - \{j_b\}$
6       evaluate $f_{\text{transit},ab}(t)$
7       evaluate $P_{\text{HO}}[s_{ab}]$
8       if $P_{\text{HO}}[s_{ab}] > 0$
9           then $\mathcal{S}_{\text{HPS}} \leftarrow \mathcal{S}_{\text{HPS}} \cup \{s_{ab}\}$
10               $\mathcal{S}_{\text{RSV}} \leftarrow \mathcal{S}_{\text{RSV}} \cup \{s_{ab}\}$
11               evaluate $C_{\text{HO}}(s_{ab})$
12               evaluate $g_{\text{HO},ab}(t)$
13               evaluate $h_{\text{HO},ab}(d)$
14  for each $s_{ab} \in \mathcal{S}$
15      $\mathcal{R}_{X,\text{HPS}}(s_{ab}) \leftarrow \emptyset$
16      for each $\varphi \in \mathcal{R}_X(s_{ab})$
17          if $s_{\text{last}}(\varphi) \in \mathcal{S}_{\text{HPS}}$
18            then evaluate $m_{\text{HO},ab|\varphi}(t)$ and $M^{-1}_{\text{HO},ab|\varphi}(0.5)$
19                if $M^{-1}_{\text{HO},ab|\varphi}(0.5) \leq T_{\text{thres\_max}}$
20                    then $\mathcal{R}_{X,\text{HPS}}(s_{ab}) \leftarrow \mathcal{R}_{X,\text{HPS}}(s_{ab}) \cup \{\varphi\}$
21                         $\mathcal{S}_{\text{RSV}} \leftarrow \mathcal{S}_{\text{RSV}} \cup \{s_{ab}\}$
22                         evaluate $P_{\text{HO}}[\varphi | s_k{=}s_{ab}]$
23                         evaluate $P_{\text{HO}}[\varphi | s_k{=}s_{ab}, s_{k-1}{=}s_{ya}]$
                                   $\forall j_y \in \mathcal{N}(j_a) - \{j_b\}$
24                         evaluate $M^{-1}_{\text{HO},ab|\varphi}(1 - \zeta_{\text{L}})$,
                                   $M^{-1}_{\text{HO},ab|\varphi}(\zeta_{\text{U}})$
```

Fig. 2. Prediction database update procedure.

one at a time. Lines 4 and 5 evaluate the first and second order transition probabilities from the segment examined to its neighboring segments. They are calculated based on the paths of MTs previously served by the BS. Line 6 evaluates the pdf of the time spent by previous MTs in the segment. Note that the pdf may be estimated based on histograms with appropriate bin size. In Line 7, we compute the probability that a MT would request a handoff while transiting the segment. If handoffs have occurred along this segment previously, then the segment is identified as a HPS, and is entered into both $\mathcal{S}_{\text{HPS}}$ and $\mathcal{S}_{\text{RSV}}$ (Lines 9 and 10). Its membership in $\mathcal{S}_{\text{RSV}}$ signifies that MTs traveling in this segment are potential candidates for resource reservation. Lines 11 to 13 simply evaluate the database entries that describe the handoff behavior of MTs traveling in this segment.

From Lines 14 to 24, we make a second pass through all the road segments, again processing each segment sequentially. For each segment s_{ab}, we reset $\mathcal{R}_{X,\text{HPS}}(s_{ab})$ so that it will be regenerated using newly computed database entries (Line 15). For each hop-limited route that originates from segment s_{ab}, we test whether its last segment is a HPS (Lines 16 and 17). Note that a "route" must include the origin segment s_{ab}, and at least one other segment. A hop limit is specified so as to reduce the computational load required. Also, note that $\mathcal{R}_X(s_{ab})$ is pretty much static, and is modified only when there are changes to the road topology within the BS's coverage area. Therefore, it does not need to be recomputed during each database update. If the examined route is found to have a last segment that is a HPS, we estimate the pdf $m_{\text{HO},ab|\varphi}(t)$ of the time taken to transit φ' and part of the last segment $s_{\text{last}}(\varphi)$ before handoff (Line 18). It is obtained from the convolution of the pdfs $f_{\text{transit}}(t)$ of segments in the partial route φ', and also the pdf $g_{\text{HO}}(t)$ of the last segment $s_{\text{last}}(\varphi)$ of route φ. For example, if the segment we are currently processing is s_{ab}, and we consider one of its routes $\varphi = \{s_{ab}, s_{bc}, s_{cd}, s_{de}\}$. This route has three hops, with partial route $\varphi' = \{s_{bc}, s_{cd}\}$, and the last segment $s_{\text{last}}(\varphi)$ is s_{de}, which is assumed to be a HPS. The pdf $m_{\text{HO},ab|\varphi}(t)$ is then obtained as:

$$m_{\text{HO},ab|\varphi}(t) = f_{\text{transit},bc}(t) \otimes f_{\text{transit},cd}(t) \otimes g_{\text{HO},de}(t). \qquad (4)$$

Note that $m_{\text{HO},ab|\varphi}(t)$ does not include the time taken to complete the current segment, s_{ab}. The latter will be added during the prediction phase because we wish to utilize the dynamics of individual MT for its computation. Once the pdf $m_{\text{HO},ab|\varphi}(t)$ is obtained, we calculate the median time $M^{-1}_{\text{HO},ab|\varphi}(0.5)$. In Line 19, we compare the median time with the limit $T_{\text{thres\_max}}$. If it is found to be within $T_{\text{thres\_max}}$, we add the route φ to the set $\mathcal{R}_{X,\text{HPS}}(s_{ab})$, and include the segment s_{ab} in $\mathcal{S}_{\text{RSV}}$ (Lines 20 and 21). We then compute via the chain rule the conditional probabilities that MTs currently in segment s_{ab} would follow this route and hand off at its last segment (Lines 22 and 23). Finally, in Line 24, we compute the quantiles $M^{-1}_{\text{HO},ab|\varphi}(1-\zeta_{\text{L}})$ and $M^{-1}_{\text{HO},ab|\varphi}(\zeta_{\text{U}})$ for this route, which will be needed later to specify the prediction limits LPL and UPL.

One important point to emphasize for the above database update algorithm is that all the above database entries only need to be calculated once during each database update. Therefore, they should be well within the computational capability of a dedicated, average processor at the BS. Having seen the prediction database update procedure, we shall proceed to describe the mobility prediction algorithm in the following section.

B. Prediction Algorithm

In order to perform the predictions, the BS needs to map each MT's current position onto the correct road segment within the road topology database (a process known as map-matching [12]). In the prediction algorithm to be presented next, we do not describe how the map-matching is performed. Instead, we assume for simplicity that the MT's current road segment and estimated speed are already computed based on its recent positions. Interested readers can refer to relevant literature from Intelligent Transportation Systems (ITS) research for additional information, such as [12].

During the prediction phase, we need to specify two additional quantiles for every MT that is currently traveling within any HPS. These quantiles will be used to calculate the LPL and UPL of the predicted time from the handoff if the MT were to hand off within that segment. They are dependent on the MT's current position within the segment, therefore they have

TABLE II

ADDITIONAL NOTATIONS USED TO PRESENT ALGORITHM.

Notation	Meaning
v^i	Estimated speed of MT i.
s^i_{ab}	Current road segment in which MT i is traveling.
s^i_{prev}	Previous segment from which MT i came from (may or may not be known).
$d^i_{\text{EOS}}(s^i_{ab})$	MT i's estimated distance from end junction, j_b.
$t^i_{\text{EOS}}(s^i_{ab})$	MT i's estimated time from end junction, j_b.
$T_{\text{thres}}(C_j)$	$T_{\text{threshold}}$ of neighboring cell C_j.
$\hat{c}^i_{\text{target}}(\varphi)$	MT i's most probable target handoff cell if it follows route φ and hands off at $s_{\text{last}}(\varphi)$.
$w^i(\varphi)$	Prediction weight specifying the probability that MT i may follow route φ and hands off at $s_{\text{last}}(\varphi)$.
$\hat{t}^i_{\text{L}}(\varphi, \zeta_{\text{L}})$	LPL of MT i's remaining time from handoff (t^i_{remain}) if it follows route φ and hands off at $s_{\text{last}}(\varphi)$, s.t. $P[t^i_{\text{remain}} \geq \hat{t}^i_{\text{L}}(\varphi, \zeta_{\text{L}})] = \zeta_{\text{L}}$.
$\hat{t}^i_{\text{U}}(\varphi, \zeta_{\text{U}})$	UPL of MT i's remaining time from handoff (t^i_{remain}) if it follows route φ and hands off at $s_{\text{last}}(\varphi)$, s.t. $P[t^i_{\text{remain}} \leq \hat{t}^i_{\text{U}}(\varphi, \zeta_{\text{U}})] = \zeta_{\text{U}}$.
$\hat{t}^i_{\text{L}}(s^i_{ab}, \zeta_{\text{L}})$	LPL of t^i_{remain} if MT i hands off in s^i_{ab}.
$\hat{t}^i_{\text{U}}(s^i_{ab}, \zeta_{\text{U}})$	UPL of t^i_{remain} if MT i hands off in s^i_{ab}.
$\mathcal{Z}^i$	Set of predictions made for MT i. Each prediction is a 4-tuple with the following form: [target cell, prediction weight, LPL, UPL]. For a prediction that MT i may follow route φ and hands off at $s_{\text{last}}(\varphi)$, the corresponding 4-tuple is: $[\hat{c}^i_{\text{target}}(\varphi), w^i(\varphi), \hat{t}^i_{\text{L}}(\varphi, \zeta_{\text{L}}), \hat{t}^i_{\text{U}}(\varphi, \zeta_{\text{U}})]$. If s^i_{ab} is a HPS, then the 4-tuple for a prediction that a handoff may occur along s^i_{ab} itself is: $[C_{\text{HO}}(s^i_{ab}), P_{\text{HO}}[s^i_{ab}], \hat{t}^i_{\text{L}}(s^i_{ab}, \zeta_{\text{L}}), \hat{t}^i_{\text{U}}(s^i_{ab}, \zeta_{\text{U}})]$.

to be recomputed during each prediction. Let D_{ab} be a random variable representing the distance from junction j_b in segment s_{ab} where handoff occurs, with pdf $h_{\text{HO},ab}(d)$. Suppose the MT is currently at a distance D_t from junction j_b, and it has not yet made a handoff-request in s_{ab}. Using this information, we can derive a conditional pdf $h_{\text{HO},ab}(d|D_{ab} < D_t)$ for $d < D_t$:

$$h_{\text{HO},ab}(d|D_{ab} < D_t) = \frac{h_{\text{HO},ab}(d)}{P[D_{ab} < D_t]}. \qquad (5)$$

Note that $h_{\text{HO},ab}(d|D_{ab} < D_t) = 0$ for $d \geq D_t$. From the above conditional pdf shown in (5), its conditional cdf can be obtained as:

$$H_{\text{HO},ab}(d|D_{ab} < D_t) = \int_0^d h_{\text{HO},ab}(u|D_{ab} < D_t)\; du. \qquad (6)$$

With the above conditional cdf, it is straightforward to approximate any q^{th} conditional quantile $H^{-1}_{\text{HO},ab}(q|D_{ab} < D_t)$. By estimating the time that the MT would take to reach two specific quantile points, namely $H^{-1}_{\text{HO},ab}(\zeta_{\text{L}}|D_{ab} < D_t)$ and $H^{-1}_{\text{HO},ab}(1 - \zeta_{\text{U}}|D_{ab} < D_t)$, we are able to specify the LPL and UPL for a possible handoff that might occur along s_{ab}.

Table II shows the additional notations used to present the prediction algorithm. As mentioned earlier, predictions are only performed for MTs that are currently traveling in segments that belong to the set $\mathcal{S}_{\text{RSV}}$. The algorithm for a MT i that is currently traveling in segment $s^i_{ab} \in \mathcal{S}_{\text{RSV}}$ is shown in Fig. 3. In Line 1, we empty the prediction output set $\mathcal{Z}^i$, as new predictions will be made. Line 2 ensures that the MT is not stationary, otherwise the algorithm exits without making any prediction. Next, in Line 3, we estimate the MT's remaining distance from the end of its current segment. The time for the MT to reach this end is then estimated (Line 4). From Lines 5 to 12, we examine previously recorded candidate routes that might lead to handoffs. Note that each of these routes will generate a 4-tuple prediction, which may or may not be inserted into the set $\mathcal{Z}^i$. For each of these routes, we estimate its LPL(UPL) as the sum of two estimates, namely, the estimated time taken to finish the current segment, and the LPL(UPL) of the time taken to follow the remaining segment sequence on the route and handing off at the very last segment. If the LPL $\hat{t}^i_{\text{L}}(\varphi, \zeta_{\text{L}})$ is found to be within the threshold time of the most probable target cell (which is the most commonly chosen target handoff cell in the last segment of this route), the weight of the prediction is taken to be either the first or second order conditional probability of route φ, depending on whether we know the previous segment of the MT (Lines 8 to 11). Then, in Line 12, we insert the 4-tuple prediction generated for this route into the set $\mathcal{Z}^i$ if the test in Line 8 is satisfied.

```
1  $\mathcal{Z}^i \leftarrow \emptyset$
2  if $v^i > 0$
3   then compute $d^i_{\text{EOS}}(s^i_{ab})$
4        $t^i_{\text{EOS}}(s^i_{ab}) \leftarrow d^i_{\text{EOS}}(s^i_{ab})/v^i$
5        for each $\varphi \in \mathcal{R}_{X,\text{HPS}}(s^i_{ab})$
6            $\hat{t}^i_{\text{L}}(\varphi, \zeta_{\text{L}}) \leftarrow t^i_{\text{EOS}}(s^i_{ab}) + M^{-1}_{\text{HO},ab|\varphi}(1 - \zeta_{\text{L}})$
7            $\hat{t}^i_{\text{U}}(\varphi, \zeta_{\text{U}}) \leftarrow t^i_{\text{EOS}}(s^i_{ab}) + M^{-1}_{\text{HO},ab|\varphi}(\zeta_{\text{U}})$
8            if $\hat{t}^i_{\text{L}}(\varphi, \zeta_{\text{L}}) \leq T_{\text{thres}}(C_{\text{HO}}(s_{\text{last}}(\varphi)))$
9               then if $s^i_{\text{prev}}$ is known
10                       then $w^i(\varphi) \leftarrow P_{\text{HO}}[\varphi|s_k{=}s^i_{ab}, s_{k-1}{=}s^i_{\text{prev}}]$
11                       else $w^i(\varphi) \leftarrow P_{\text{HO}}[\varphi|s_k{=}s^i_{ab}]$
12                   $\mathcal{Z}^i \leftarrow \mathcal{Z}^i \cup \{[\hat{c}^i_{\text{target}}(\varphi), w^i(\varphi),$
                             $\hat{t}^i_{\text{L}}(\varphi, \zeta_{\text{L}}), \hat{t}^i_{\text{U}}(\varphi, \zeta_{\text{U}})]\}$
13       if $s^i_{ab} \in \mathcal{S}_{\text{HPS}}$
14         then $\hat{t}^i_{\text{L}}(s^i_{ab}, \zeta_{\text{L}})$
                  $\leftarrow [d^i_{\text{EOS}}(s^i_{ab}) - H^{-1}_{\text{HO},ab}(\zeta_{\text{L}}|D{<}d^i_{\text{EOS}}(s^i_{ab}))]/v^i$
15              $\hat{t}^i_{\text{U}}(s^i_{ab}, \zeta_{\text{U}})$
                  $\leftarrow [d^i_{\text{EOS}}(s^i_{ab}) - H^{-1}_{\text{HO},ab}(1{-}\zeta_{\text{U}}|D{<}d^i_{\text{EOS}}(s^i_{ab}))]/v^i$
16              if $\hat{t}^i_{\text{L}}(s^i_{ab}, \zeta_{\text{L}}) \leq T_{\text{thres}}(C_{\text{HO}}(s^i_{ab}))$
17                then $\mathcal{Z}^i \leftarrow \mathcal{Z}^i \cup \{[C_{\text{HO}}(s^i_{ab}), P_{\text{HO}}[s^i_{ab}],$
                             $\hat{t}^i_{\text{L}}(s^i_{ab}, \zeta_{\text{L}}), \hat{t}^i_{\text{U}}(s^i_{ab}, \zeta_{\text{U}})]\}$
```

Fig. 3. Prediction algorithm for a MT i traveling in segment s^i_{ab}.

If the MT is currently in a HPS (Line 13), then there is a chance that a handoff may occur while it is traveling along this segment. In Lines 14 and 15, we obtain the LPL and UPL as the estimated time taken to reach the two quantile points $H^{-1}_{\text{HO},ab}(\zeta_{\text{L}}|D < d^i_{\text{EOS}}(s^i_{ab}))$ and $H^{-1}_{\text{HO},ab}(1 - \zeta_{\text{U}}|D < d^i_{\text{EOS}}(s^i_{ab}))$ described earlier. Having determined both $\hat{t}^i_{\text{L}}(s^i_{ab}, \zeta_{\text{L}})$ and $\hat{t}^i_{\text{U}}(s^i_{ab}, \zeta_{\text{U}})$, if $\hat{t}^i_{\text{L}}(s^i_{ab}, \zeta_{\text{L}})$ is found to be within the threshold time of the cell $C_{\text{HO}}(s^i_{ab})$, we insert the newly generated 4-tuple into the prediction set $\mathcal{Z}^i$ (Lines 16 and 17).

Note that the above algorithm only performs predictions for a single MT i. In order to perform bandwidth reservations, predictions must be made for all active MTs that are currently traveling in segments that belong to the set S_{RSV}. In the next section, we shall present the reservation scheme that we have developed, and explain how these predictions will be used.

III. Dynamic Bandwidth Reservation Scheme

This section describes the reservation scheme that we have developed. Unlike some existing schemes that only utilize incoming handoff predictions to adjust their reservations [2], our scheme utilizes predictions about both incoming and outgoing handoffs to achieve even more efficient tradeoffs between P_{FT} and P_{CB}. In the following, we shall first describe the system model assumed. We then explain the logic behind the scheme, before presenting its detailed algorithms.

A. System Model

We consider a cellular network with 2-dimensional cell layout, in which each cell is adjacent to several other cells. The minimum granularity of bandwidth resources that could be allocated to any call is assumed to be 1 *bandwidth unit* (BU) [2], [4]. Each BS j has a capacity $C(j)$, which is assumed to be constant for simplicity, although the proposed scheme may be extended to include time-dependent $C(j)$. Given the bandwidth demand of individual connections, the BS performs admission control to ensure that the total demand of all active connections are below or equal to $C(j)$. Although it is suggested in [7] that some adaptive applications might be able to accept a lower bandwidth at the expense of lower call quality during congestion, we do not consider them here. Such an assumption is likely to reduce P_{FT}, but it may make it harder to visualize the advantages of using mobility predictions, which is the main aim of our work. Similar to [2], we shall also preclude delay-insensitive applications that can tolerate long handoff delays, as well as, soft handoffs found in CDMA systems. All these preclusions could possibly be added to our proposed scheme as future extensions, when the advantages of utilizing mobility predictions can be clearly demonstrated.

In order to prioritize handoffs over new calls, each cell must reserve some bandwidth that can only be utilized by incoming handoffs. Specifically, each BS j shall have a "reservation target" $R_{\text{target}}(j)$ that is being updated regularly based on mobility predictions. A new call request is accepted if the remaining bandwidth after its acceptance is at least $R_{\text{target}}(j)$, i.e.,

$$C(j) - (\sum_x b_{x,j} + b_{\text{new}}) \geq R_{\text{target}}(j), \tag{7}$$

where b_{new} is the bandwidth required by the new call request, and $b_{x,j}$ is the bandwidth currently being used by an existing connection x in cell j. Note that $R_{\text{target}}(j)$ is merely a target, not the actual amount of bandwidth that is reserved. The BS can only attempt to meet this target by rejecting new call requests, while waiting for some existing calls within the cell to release bandwidth when they end, or hand off to other cells. For a handoff request, the admission control rule is more lenient – it is admitted so long as there is sufficient remaining capacity for the handoff, regardless of the value of $R_{\text{target}}(j)$:

$$C(j) - \sum_x b_{x,j} \geq b_{\text{handoff}}, \tag{8}$$

where b_{handoff} is the bandwidth needed by the handoff.

When a new call request is rejected, we assume that it is cleared. Subsequent new call requests are assumed to be independent of previous requests. When the BS has insufficient bandwidth to accommodate an incoming handoff-request, we assume that it is forced to terminate. We do not consider handoff queuing here, although it would likely improve the performance of our scheme (as well as other schemes simulated for comparison). As mentioned earlier, such extensions may make it difficult to visualize the advantages of using mobility predictions.

B. Logic Behind the Proposed Scheme

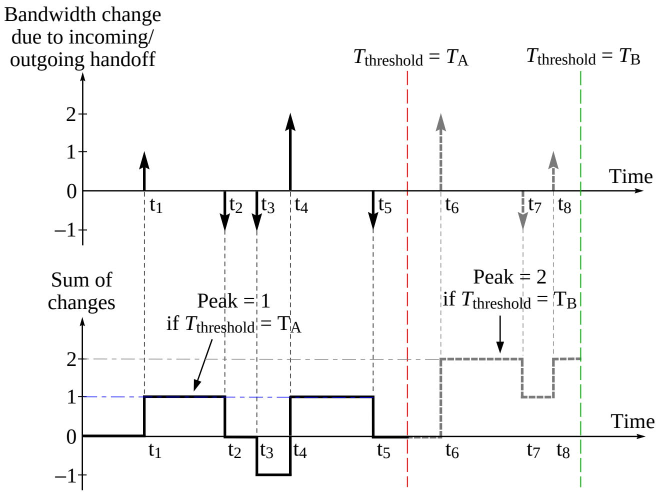
To understand the logic leading to the proposed scheme, we first ask ourselves the following question:

> *Suppose we have perfect knowledge about all the incoming/outgoing handoffs that will occur within a limited time into the future, how much bandwidth should be reserved to prevent any of these incoming handoffs from being dropped?*

Fig. 4 shows an example that we shall use to answer the above question. Here, we assume that we have perfect knowledge about future handoffs up to time $T_{\text{threshold}}$. Note that an incoming handoff into the current cell will lead to a positive change in the bandwidth used, while an outgoing handoff will lead to a negative change. Suppose $T_{\text{threshold}} = T_{\text{A}}$. By summing up all the bandwidth changes over the time interval $[0, T_{\text{A}}]$, we realize that the maximum peak bandwidth requirement within this interval is 1 BU. This implies that *if we succeed* in reserving 1 BU at the BS, we can ensure that *all* incoming handoffs within $[0, T_{\text{A}}]$ will not be dropped. Therefore, an appropriate $R_{\text{target}}(j)$ would be 1 BU.

For a reservation scheme that does not utilize outgoing handoff information (e.g., [2]), only the positive changes are summed up. As a result, the BS would set $R_{\text{target}}(j)$ to 3 BUs, which may lead to unnecessary blocking of new call requests that arrive within the interval $[0, T_{\text{A}}]$.

As mentioned earlier, $R_{\text{target}}(j)$ is merely a target. If there are insufficient existing calls that release bandwidth, $R_{\text{target}}(j)$ cannot be met. This will cause some of the incoming handoffs to be dropped, despite the fact that we have prior knowledge about them. However, the likelihood of this occurring will decrease if the BS is given more time to meet the target. The threshold $T_{\text{threshold}}$ can be viewed as the time given to the BS to set aside the required bandwidth to avoid a forced termination. Referring to Fig. 4 again, notice that handoffs beyond T_{A} are shown as gray dotted lines. This information is currently not available to the BS, therefore it will set $R_{\text{target}}(j) = 1$ BU. Suppose the BS has 2 BUs of spare capacity at time $t = 0$. If a new call from MT x requests 1 BU, the BS will accept the new call because it can still satisfy $R_{\text{target}}(j)$ after accepting



Bandwidth change due to incoming/ outgoing handoff
$T_{\text{threshold}} = T_A$
$T_{\text{threshold}} = T_B$
2
1
0
−1
Time
t1
t2
t3
t4
t5
t6
t7
t8
Sum of changes
Peak = 1 if $T_{\text{threshold}}$ = T_A
Peak = 2 if $T_{\text{threshold}}$ = T_B
2
1
0
−1
Time
t1
t2
t3
t4
t5
t6
t7
t8

Fig. 4. Perfect knowledge about handoffs up to $t = T_{\text{threshold}}$.

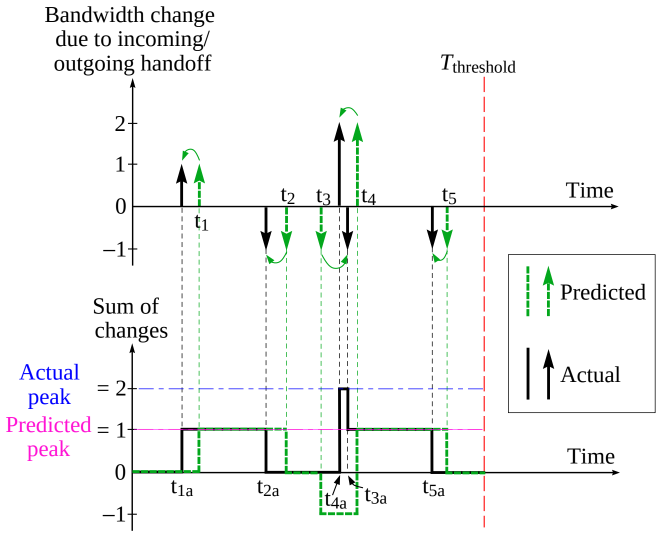


Fig. 5. Effects of prediction errors in handoff timings.

the new call. However, if no existing call ends before time t_6, then the spare bandwidth at time t_6 remains at 1 BU, thus causing the incoming handoff at time t_6 to be dropped. On the other hand, if we have set $T_{\text{threshold}}$ to T_B, then $R_{\text{target}}(j)$ would have been set to 2 BUs. The BS would then have rejected the new call request from MT x so as to maintain its spare capacity at 2 BUs. Consequently, the incoming handoff at time t_6 will not be dropped. This shows that it is possible to reduce P_{FT} by giving the BS earlier notice, which could be done by increasing $T_{\text{threshold}}$. Therefore, we could vary $T_{\text{threshold}}$ as an option to adjust P_{FT}.

The scenario examined thus far is for the ideal case of having perfect knowledge about handoffs within $[0, T_{\text{threshold}}]$, which is unlikely to happen in real-life. Now let us examine a more realistic scenario, whereby we only have handoff predictions. Fig. 5 gives an example of the possible effects of prediction errors in handoff timings. Here, handoffs are predicted at t_1, t_2, t_3, t_4 and t_5, but the actual handoffs occur at $t_{1a}, t_{2a}, t_{3a}, t_{4a}$ and t_{5a}. Using the predictions, the peak computed by the BS is 1 BU. However, the actual peak is 2 BUs. Therefore, the incoming handoff-request at time t_{4a} might be dropped. A closer look at Fig. 5 reveals that the error in predicted peak arises because the predicted sequence of a pair of incoming and outgoing handoffs is wrong. An outgoing handoff is predicted to occur (at t_3) before the incoming handoff does (at t_4). However, the incoming handoff actually occurs earlier (at t_{4a}) than the outgoing handoff (at t_{3a}). This reversal of predicted sequence and actual sequence causes the actual peak to become larger than the predicted peak. An interesting point to emphasize here is that, if, on the other hand, an incoming handoff is predicted to occur before an outgoing handoff, but the actual sequence is reversed, then the actual peak might be lower than the predicted peak. However, this type of prediction error is benign because it does not lead to a handoff being dropped. It may only result in over-reservation of spare bandwidth resources.

From the above, we observe that it is undesirable when an incoming handoff occurs earlier than its predicted time, and also, when an outgoing handoff occurs later than its predicted time. Either one or both of these scenarios may increase the chances of a forced termination. Therefore, we would like to reduce their likelihood. Recall that each prediction's 4-tuple consists of a LPL and a UPL. Suppose we use an incoming MT's LPL as its predicted arrival time, and use an outgoing MT's UPL as its predicted departure time. By specifying ζ_L and ζ_U to be larger than 0.5, we introduce some biases into the predicted times, such that the likelihood of the above scenarios may be reduced. If the injected biases are small, the predicted arrival and departure sequence for those handoffs that are sufficiently far apart would probably remain the same as though no biases have been injected. However, these biases could capture and correct those predictions that are close enough to result in under-reservation at the slightest prediction errors. Note that the parameters ζ_L and ζ_U are design parameters whose optimal values are best determined through experimentation in real cellular networks. A general rule of thumb is to set a value that is within the range of $0.5 \sim 0.7$. Any value that is under 0.5 will actually increase the likelihood of under-reservation, while a value that is too high may render the predictions too conservative and result in excessive over-reservation.

Having seen these key concepts, in the next section, we shall describe how each BS adjusts its $T_{\text{threshold}}$ to meet the desired P_{FT}. Section III-D will explain how R_{target} is adjusted at each BS.

C. Adjusting $T_{\text{threshold}}$ at each BS

In Section III-B, we have seen that the P_{FT} experienced by incoming handoff-requests may be indirectly controlled by adjusting $T_{\text{threshold}}$. Although there might exist an optimal value of $T_{\text{threshold}}$ for the desired P_{FT}, it would probably be different

TABLE III

NOTATIONS USED IN ALGORITHM THAT ADJUSTS $T_{\text{THRESHOLD}}$.

Notation	Meaning
$T_{\text{thres\_max}}$	The maximum $T_{\text{threshold}}$ value allowed.
$T_{\text{thres\_min}}$	The minimum $T_{\text{threshold}}$ value allowed.
$T_{\text{thres\_init}}$	The initial $T_{\text{threshold}}$ value.
n_{HO}	The number of handoffs counted.
n_{FT}	The number of forced terminations counted.
$P_{\text{FT,target}}$	The desired P_{FT} target.
w_{obs}	Observation window size.
μ	Scaling factor, an experimentally determined parameter.

```
$w_{\text{obs}} = \lceil \mu / P_{\text{FT,target}} \rceil$;
$T_{\text{threshold}} \leftarrow T_{\text{thres\_init}}$; $n_{\text{HO}} \leftarrow 0$; $n_{\text{FT}} \leftarrow 0$;
while (system running)
  if (incoming handoff-request occurs)
    then $n_{\text{HO}} \leftarrow n_{\text{HO}} + 1$;
         if (handoff accepted)
           then if ($n_{\text{HO}} \geq w_{\text{obs}}$)
                  then if (($n_{\text{FT}} = 0$) and ($T_{\text{threshold}} > T_{\text{thres\_min}}$))
                         then $T_{\text{threshold}} \leftarrow T_{\text{threshold}} - 1$;
                       $n_{\text{HO}} \leftarrow 0$; $n_{\text{FT}} \leftarrow 0$;
           else $n_{\text{FT}} \leftarrow n_{\text{FT}} + 1$;
                if ($n_{\text{FT}} > 1$)
                  then if ($T_{\text{threshold}} < T_{\text{thres\_max}}$)
                         then $T_{\text{threshold}} \leftarrow T_{\text{threshold}} + 1$;
                       $n_{\text{HO}} \leftarrow 0$; $n_{\text{FT}} \leftarrow 0$;
```

Fig. 6. Algorithm used by each BS to adjust its $T_{\text{threshold}}$.

in each cell, as it might be characteristic of the cell's coverage area, subscriber density, and so on. It might even fluctuate with user mobility and traffic load at different times of the day. Since there is no obvious way to compute the optimal $T_{\text{threshold}}$, we utilize an adaptive algorithm to approximate its value for any given P_{FT}. Table III shows the notations we have used in our algorithm, while the actual algorithm is shown in Fig. 6.

The basic idea of the algorithm is that it attempts to maintain approximately one forced termination out of every w_{obs} incoming handoffs that are observed. For this reason, w_{obs} is also referred to as the "observation window size". If there is no forced termination within w_{obs} handoffs, the value of $T_{\text{threshold}}$ is deemed to be too large, and will be decreased by 1 sec. A fresh observation window will be restarted when the current window is exhausted. If, at any time, more than one forced termination is observed within the observation window, the value of $T_{\text{threshold}}$ is immediately increased by 1 sec. When this happens, the observation window is also restarted.

For a desired P_{FT} target, the value of w_{obs} is chosen to be $\lceil \mu / P_{\text{FT,target}} \rceil$, where μ is a scaling factor close to 1. Ideally, if the algorithm were to succeed in achieving exactly one forced termination every w_{obs} handoffs, then w_{obs} should have simply been set to be the reciprocal of $P_{\text{FT,target}}$. However, through our simulations, we discover that the P_{FT} obtained by setting $w_{\text{obs}} = \lceil 1/P_{\text{FT,target}} \rceil$ is slightly different from the desired target $P_{\text{FT,target}}$ by an approximately constant factor (about $1.2 \sim 1.25$). A possible explanation for the above observation is that handoffs are bursty and the best that our adaptive algorithm could achieve is to allow the value of $T_{\text{threshold}}$ to fluctuate around its optimal value. This causes the average number of forced terminations per w_{obs} observations to deviate slightly from 1. To compensate for the above difference, the scaling factor μ is introduced for the calculation of w_{obs}. Note that the value of μ for an actual cellular system shall be determined experimentally.

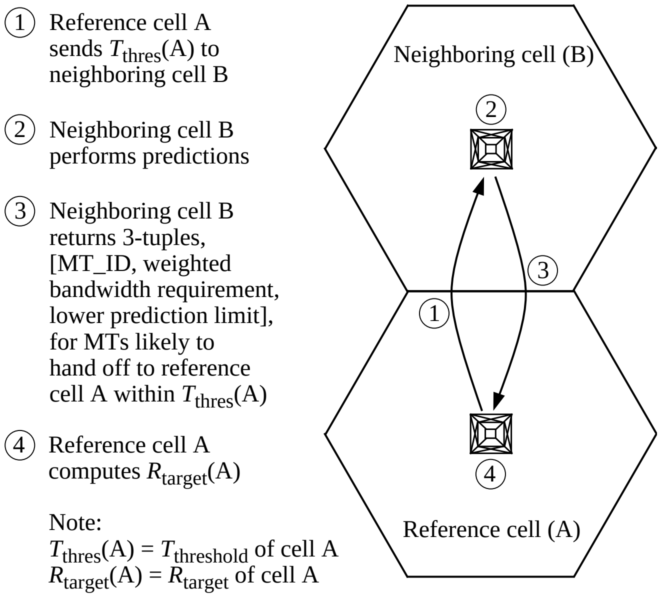

① Reference cell A sends T_{thres}(A) to neighboring cell B
② Neighboring cell B performs predictions
③ Neighboring cell B returns 3-tuples, [MT_ID, weighted bandwidth requirement, lower prediction limit], for MTs likely to hand off to reference cell A within T_{thres}(A)
④ Reference cell A computes R_{target}(A)
Note:
T_{thres}(A) = $T_{\text{threshold}}$ of cell A
R_{target}(A) = R_{target} of cell A
Neighboring cell (B)
Reference cell (A)
②
③
①
④

Fig. 7. Procedure performed every T_{predict}.

D. Adjusting R_{target} at each BS

The predictions used to compute $R_{\text{target}}(j)$ are made periodically every T_{predict}, which is a design parameter. If the predictions are performed very frequently, they are more accurate but a more powerful processor will be required at each BS. On the other hand, their accuracy may deteriorate if they are far apart, causing the tradeoff between P_{FT} and P_{CB} to become less efficient.

Fig. 7 depicts the procedure that is repeated every T_{predict}. For clarity, we only show two cells, A and B. Cell A is our reference cell for which we demonstrate the computation of its R_{target}(A), while cell B is one of A's neighboring cells. Note that in an actual cellular network, each cell is usually surrounded by several neighboring cells; Steps 1, 2 and 3 are simultaneously performed for every neighbor of cell A. Also, cell A concurrently serves as a neighboring cell for cell B; the procedure shown also applies when they interchange their roles.

An assumption made here is that inter-BS communications are possible and take place via wired links. The following describes each step of the procedure:

Step 1: Reference cell A transmits T_{thres}(A) to neighboring cell B. This will be used later by B to decide what prediction information needs to be sent to A.

Step 2: Neighboring cell B performs outgoing handoff predictions for the active MTs under its service. Each prediction is in the form of a 4-tuple described earlier. Note that cell A itself will also be performing outgoing handoff predictions at the same time for its role as some other cells' neighbor (not shown).

Step 3: For every active MT that is predicted to hand off into cell A with LPL $\leq T_{\text{thres}}$(A), the neighboring cell B transmits part of the predicted information to cell A in the form of a 3-tuple, with the format [MT_ID, weighted bandwidth requirement, predicted time]. The *weighted bandwidth requirement* is a real number calculated as the product of the prediction weight and the MT's bandwidth requirement, while the *predicted time* is its LPL.

Step 4: As cell A receives the 3-tuples from cell B, they are inserted into an ascending sorted list according to their predicted times. These represent the incoming handoff predictions. Cell A then examines its own outgoing handoff predictions. For those with UPL $\leq T_{\text{thres}}$(A), they are also inserted into the same list, but in the form of 3-tuples with format [MT_ID, $-$weighted bandwidth release, predicted time]. The *weighted bandwidth release* is the product of the prediction weight and the bandwidth that would be released when the MT leaves. The *predicted time* is its UPL. Finally, the completed list is used to calculate the value of R_{target}(A).

To calculate $R_{\text{target}}(j)$ for BS j, its sorted list is scanned and the bandwidth change from each entry are summed. Upon finishing the entire list, the overall peak discovered will be assigned to $R_{\text{target}}(j)$.

Although the predictions are performed every T_{predict}, $R_{\text{target}}(j)$ may be adjusted more than once between two successive predictions. This is because the BS may acquire updated information that renders some of the previous predictions invalid, before the next prediction takes place. Specifically, $R_{\text{target}}(j)$ of BS j may be updated when any of the following events occurs:

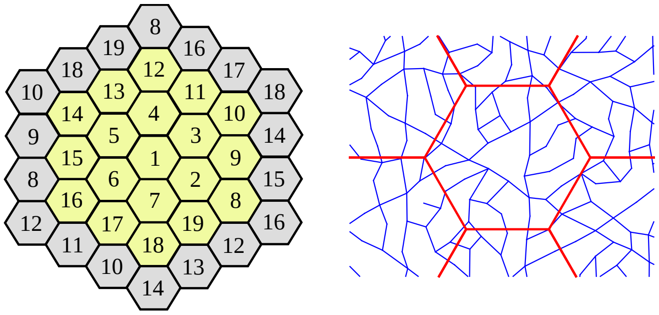
1) A previously predicted incoming handoff within the list has taken place.
2) A previously predicted outgoing handoff within the list has either handed off or ended its call.
3) A previously predicted incoming handoff within the list has either ended its call without handoff, or has handed off to another cell other than BS j. BS j needs to be informed by the neighboring BS that has previously sent the 3-tuple for that MT.

When an updated information is acquired due to any of the above conditions, the BS removes the affected entry from its sorted list, and recomputes $R_{\text{target}}(j)$.

IV. Simulations and Results

A. Simulation Model

To facilitate the evaluation of the schemes presented, a novel simulation model was designed. It incorporates road layouts that place constraints on MTs' paths, thus establishing a more realistic platform to evaluate the performance of any positioning-based prediction algorithm.



(a) (b)

Fig. 8. (a) Simulation network with wrap-around at network boundary, (b) a sample road layout generated using heuristic rules.

The simulation network consists of 19 wireless cells. In order to eliminate boundary effects that could make it very difficult to comprehend the results, a common approach found in the literature is used [2], [5]: cells at the boundary wrap around as shown in Fig. 8(a). In this way, whenever a MT travels out of the network boundary, it is re-injected into the network again via the appropriate wrap-around cell as though a handoff has occurred from outside the simulation network. This eliminates any traffic loss at the network boundary. The simulation model also consists of arbitrary road layouts that are randomly generated based on heuristic rules; real maps are not used because we require the roads to wrap around at the network boundary. The road layouts are designed to imitate those found in city areas. Fig. 8(b) shows an example of the road topology that was randomly generated.

Although the cell layout shown in Fig. 8(a) adopts the hexagonal cell model, we do not assume that handoffs occur at the hexagonal boundary. The hexagonal model is merely used to determine the relative positions of the cells. In contrast to previously mentioned work in which handoffs are assumed to occur at either circular or hexagonal cell boundaries, there are no well-defined cell boundaries. Suppose R is the designed *cell radius* (assumed to be 1000 m in the simulations), which is typically defined as the distance from the BS to the vertex of the hexagonal cell model. When a MT is between $1.1R$ and $1.2R$ from the BS, we assume that a handoff will occur during its transit through this region. The time at which the handoff shall occur is a random variable that is uniformly distributed over the total time spent in the region. The target BS is assumed to be the nearest neighboring BS at the time when the handoff occurs, although this may not be the case in real life.

To make the problem more interesting, traffic lights are introduced into the simulation model. Two sets of traffic lights are assumed. When one set is GREEN, the other set is RED. Each GREEN and RED signal shall last for 30 sec. A speed limit is also assigned to each road segment, chosen from the set 40 km/h, 50 km/h, and 60 km/h with equal probability. The speed of each MT is a random variable, drawn from truncated Gaussian distribution. The mean speed will be the speed limit of that particular road segment. The standard deviation is assumed to be 5 km/h, and the speed is truncated to a limit of three standard deviations from its mean.

We do not assume any particular positioning technology for the MTs, as new breakthroughs will continue to surface. The distribution and correlation of the possible positioning errors are thus unknown. For the sake of simplicity, we do not model the effects of positioning errors in the simulations. As mentioned earlier, our mobility prediction technique assumes that the positioning errors of future MTs are relatively small. Therefore, we do not foresee any drastic effect on the simulation results if positioning errors were to be introduced.

Each cell is assumed to have a fixed link capacity C of 100 BUs. For simplicity, the bandwidth requirement of each MT is assumed to be symmetric in both uplink and downlink, although it is straightforward to modify the scheme to handle asymmetric requirements. The traffic model used here is similar to the one used in [2]. Connection requests are generated according to Poisson distribution with rate λ (connections/sec/cell) in each cell. The initial position of a new call and its destination can be on any road segment with equal probability. The path chosen by the MT is assumed to follow the shortest path between its origin and its destination. Like in [2], we assume that each call request is either of type "voice" (requires 1 BU), or of type "video" (requires 4 BUs) with probabilities R_{vo} and $1-R_{\text{vo}}$ respectively, where R_{vo} is also called the *voice ratio*. In the simulations, R_{vo} is set to 0.5. All MTs are assumed to have the same P_{FT} requirement, regardless of their connection types. The lifetime for both types of connections are assumed to be exponentially distributed, with mean 180 sec. We define the *normalized offered load* per cell as

$$L_{\text{norm}} = \frac{[1 \cdot R_{\text{vo}} + 4 \cdot (1 - R_{\text{vo}})] \cdot \lambda \cdot 180}{C} \qquad (9)$$

In this paper, we shall only present the simulation results for $L_{\text{norm}} = 1$. The interval between predictions, T_{predict}, is set to 5 sec. The probabilities ζ_{L} and ζ_{U} that affect the prediction limits are both set to 0.65, as they are found to achieve the best performance for the simulation model used.

B. Other Schemes Simulated For Comparison

We shall call our proposed scheme the *road topology based* scheme (RTB). We have also simulated six other bandwidth reservation schemes for comparison purposes:

1) Benchmark Scheme: This is an idealized scheme that assumes perfect knowledge about every active MT's *next* cell and handoff time. It utilizes the same algorithms described in Sections III-C and III-D for adjusting $T_{\text{threshold}}$ and maintaining R_{target}. The only difference is that, instead of using prediction limits, it uses actual handoff times for the computation of R_{target} at each BS. The sorted list at each BS is created every T_{predict} as well, and only handoffs that are known to take place within the next $T_{\text{threshold}}$ of the BS are included in this list.

2) Reactive Scheme: This scheme is purely reactive with no prediction. It gives a lower bound for the efficiency of the schemes considered. The basic idea is to adapt the BS's R_{target} according to forced termination counts observed over w_{obs} handoff-requests. We utilize the same adaptive algorithm presented in Fig. 6 that was originally designed for adjusting $T_{\text{threshold}}$. Instead of adjusting $T_{\text{threshold}}$ (which does not exist in this scheme), the algorithm is used for adjusting R_{target} directly. If no forced termination occurs within w_{obs} handoff-requests, R_{target} is decreased by 1 BU. If more than one forced termination is observed, R_{target} is increased by 1 BU instead.

3) Static Scheme: This scheme utilizes a fixed R_{target} for each simulation run. The P_{CB} and P_{FT} obtained for different R_{target} values are plotted.

4) Choi's AC1 Scheme: This is one of the three schemes proposed in [2]. In their simulations based on 1-D cell layout, their AC3 method performed best among the three methods, namely AC1, AC2 and AC3. However, in our simulations based on our 2-D simulation network, AC1 has the best performance, whereas AC2 and AC3 are over-conservative and has much worse efficiency than the Reactive scheme. Therefore, we shall only present the results for AC1 here. This scheme works by estimating the probability that a MT would hand off into a neighboring cell within an estimation time window T_{est}, based upon its previous cell, and its extant sojourn time. The neighboring cell's R_{target} is then increased by the MT's bandwidth requirement, weighted by the estimated probability. The T_{est} of each cell is dynamically adjusted based on the measured forced termination ratio among a number of handoffs recently observed, so as to meet the desired P_{FT}.

5) Linear Extrapolation (LE) Scheme: This scheme utilizes the same algorithms described in Sections III-C and III-D for adjusting $T_{\text{threshold}}$ and maintaining R_{target}. The only difference from our RTB scheme is that, instead of using road topology based mobility prediction technique, a linear extrapolation based mobility prediction technique similar to the one we proposed in [8] is used.

6) RTB with Path Knowledge (RTB_PK) Scheme: This scheme is a variant of our RTB scheme. It assumes that there is a probability P_{known} that a MT's path may be known, either from the MT's past history, or via routes computed by an ITS navigation system. Note that even when the MT's path is known, we do not know the exact time and position that the handoff might occur, because it could happen anywhere within the handoff region. Also, note that we do not model errors in the presumed known path, although in an actual cellular system, there is a chance that the MT may deviate from its usual known path. We are only interested in understanding what is the *best* performance achievable if there is a probability P_{known} that we have prior knowledge about a MT's path. In this paper, we shall only assume that $P_{\text{known}} = 1$.

C. Simulation Results

All results shown here are the averages over 19 cells in the simulation network. When no handoff prioritization is performed, both P_{CB} and P_{FT} are 0.075. This is unacceptably high

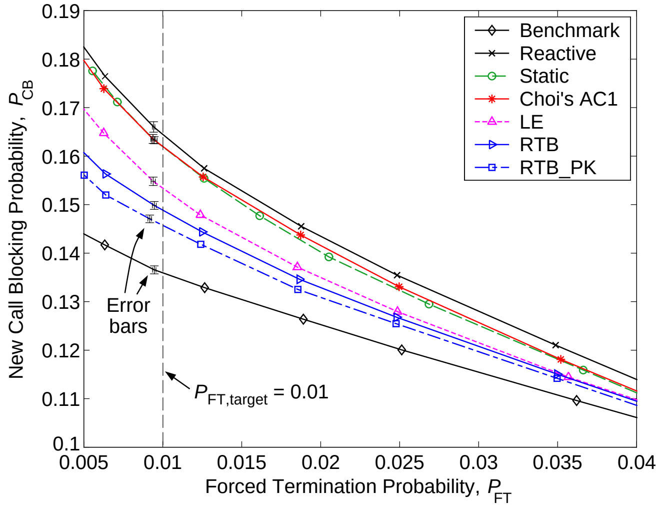


Fig. 9. P_{CB} versus P_{FT} for different schemes at $L_{\text{norm}} = 1.0$.

for P_{FT}, thus emphasizing the need for handoff prioritization. Fig. 9 shows the plots of P_{CB} versus P_{FT}. For each scheme, the target P_{FT} is varied so as to illustrate its tradeoff with P_{CB}. The relative positions of such plots demonstrate the relative efficiencies among the different schemes. A curve that is closer to the origin represents a more efficient scheme. It means that the scheme is able to achieve the same P_{FT} target while causing a smaller increase in P_{CB}.

The most efficient scheme among the seven schemes shown is the Benchmark scheme. It serves as a bound to the best efficiency that may be achieved by others. The Reactive scheme, on the other hand, has the worst efficiency. Recall that this scheme has little intelligence, as it merely adapts R_{target} according to forced termination counts over an observation window of past handoff-requests. Although L_{norm} is constant, new call and handoff call arrivals are random processes. Therefore, there might be times when many handoff-requests arrive together within a short period of time. Since the Reactive scheme has no predictive capability, it does not increase R_{target} even when there is a cluster of incoming handoff-requests in the near future, until forced terminations start to occur. The resulting large counts of forced terminations might cause the scheme to rapidly adapt its R_{target} to a much larger value, although there might be very few incoming handoff-requests after this busy period. This blocks new calls unnecessarily for extended periods of time, thus making the scheme the least efficient.

The Static scheme appears to be more efficient than the Reactive scheme. However, it is only useful if the average system load is constant all the time, which is unlikely to be the case. When load fluctuates with time, it may experience periods of over-reservation and under-reservation. While a static R_{target} may be sufficient to meet the desired target P_{FT} for a certain load, it may be too much or too little for some other loads. On the other hand, other adaptive schemes, including the Reactive scheme, can adapt to different loads.

Choi's AC1 scheme has slightly better efficiency than the Reactive scheme, because it is predictive and possesses some intelligence in where and when the bandwidth should be reserved. However, it only has about the same efficiency as the Static scheme, and has much lower efficiency than the remaining four schemes. This is probably because it might be insufficient to predict the mobility of a MT based on its previous cell information, and its extant sojourn time. In addition, calls that are newly generated in the cell do not have previous cell information. This hinders the scheme's prediction accuracy, thus lowering its efficiency. Moreover, the scheme does not utilize predictions about outgoing handoffs from each cell; it might over-reserve bandwidth resources, when sufficient resources would have been released by outgoing handoffs before the incoming ones arrive.

The LE scheme has better efficiency than Choi's AC1 scheme. The improvement is even more significant in the RTB scheme. These demonstrate that mobility prediction schemes based on mobile positioning information are more accurate, thus leading to more efficient reservations. Also, the LE and RTB schemes utilize both incoming and outgoing handoff predictions when determining the values of R_{target}, thus raising their potential to outperform other schemes that do not.

While the RTB_PK scheme performs better than the RTB scheme, it can be seen that the improvement is not very significant. In addition, P_{known} is unlikely to be 1 in real-life, therefore the actual improvement might be even lesser. It may not be worth the extra effort to implement the RTB_PK scheme in place of the RTB scheme, unless the additional information required by the RTB_PK scheme is readily available.

As we have seen from the simulation results, the plots agree with intuition that handoff prioritization efficiency improves as the amount of knowledge incorporated into the schemes increases. With the additional knowledge of real-time mobile positioning information, the LE scheme is able to outperform the Reactive scheme, the Static scheme, and Choi's AC1 scheme, even though it is based on a simple linear extrapolation approach. For the RTB scheme, the use of both real-time mobile positioning information and road topology knowledge allows it to perform better than the LE scheme. The RTB_PK scheme, which eliminates the uncertainty in predicting the MTs' future paths, further improves upon the RTB scheme, although the improvement is not dramatic. Finally, the Benchmark scheme sets a non-realizable bound for all the other schemes, using perfect knowledge about every MT's next handoff cell and time.

In Section III-B, we have explained the importance of utilizing both incoming and outgoing handoff predictions for adjusting the amount of reservation in each cell. Here, we shall demonstrate via simulations that the reservation efficiencies of such schemes are indeed better than those schemes that only utilize incoming handoff predictions.

We consider three additional schemes, which are variants of the Benchmark, LE, and RTB schemes. In these variant schemes, predictions about outgoing handoffs from each cell are purposely withheld from the computation of R_{target}. Fig. 10

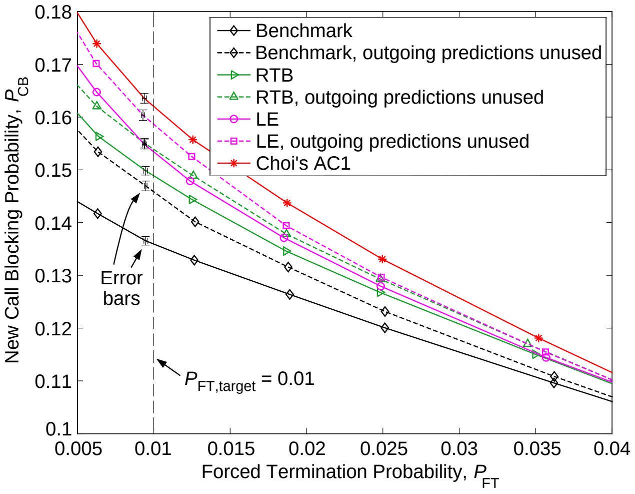


Fig. 10. Performance deteriorates when outgoing predictions not used.

shows the $P_{\rm CB}$ versus $P_{\rm FT}$ plots for these variant schemes and their original schemes. We also reproduce the plot for Choi's AC1 scheme, which does not utilize outgoing handoff predictions as well.

From the plots, we observe that the reservation efficiencies of the variant schemes are much worse than their original counterparts. This justifies the inclusion of outgoing handoff predictions for resource reservations. Another important observation is that even without using the outgoing handoff predictions, the variants of both our LE and RTB schemes outperform Choi's AC1 scheme. This reemphasizes the advantages of using predictive schemes that utilize real-time mobile positioning information, in contrast to the latter which only uses each MT's previous cell history and extant sojourn time.

V. Conclusion

We have presented a novel mobility prediction technique built upon the assumption that future MTs would likely be equipped with reasonably accurate positioning capability. Unlike previous attempts which have assumed either hexagonal or circular cell geometries, our technique caters for irregular handoff regions. We also incorporate road topology information into the prediction technique, which could potentially yield better prediction accuracy for MTs that are carried in vehicles.

Among the many possible applications for which mobility predictions could prove useful, we are interested in using it for handoff prioritization. We designed an adaptive bandwidth reservation scheme that dynamically adjusts the reservation at each BS according to both incoming and outgoing handoff predictions.

We have performed simulations to evaluate the performance of our scheme, and also simulated six other schemes for comparison. The results agree with intuition that schemes which incorporate more knowledge are able to achieve better reservation efficiency. The relative efficiencies of the different schemes can be summarized as: Reactive $<$ Static $\approx$ Choi's AC1 $<$ LE $<$ RTB $<$ RTB_PK $<$ Benchmark. Although the RTB_PK scheme is potentially realizable for $P_{\rm known} < 1$, its improvement over the RTB scheme is found to be small even when $P_{\rm known} = 1$. Therefore the RTB scheme is the preferred scheme for implementation, unless the extra knowledge required by the RTB_PK scheme is readily available, such as through dynamic route guidance in vehicular telematics systems.

In order to justify our claim that *both* incoming and outgoing handoff predictions should be used in order to maximize any reservation scheme's efficiency, we also examined variants of the Benchmark, LE, and RTB schemes that do not account for possible outgoing handoffs in their reservations. These variants suffer significant deterioration in performance compared to the original schemes. Nevertheless, both LE and RTB variant schemes still outperform Choi's AC1 scheme, demonstrating the improvement in prediction accuracy resulting from the use of real-time positioning information.

References

[1] D. Hong and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and non-prioritized handoff procedures," *IEEE Trans. Veh. Technol.*, vol. VT-35, no. 3, Aug. 1986, pp. 77–92.

[2] S. Choi and K. G. Shin, "Predictive and adaptive bandwidth reservation for handoffs in QoS-sensitive cellular networks," in *Proc. ACM SIGCOMM'98*, Vancouver, British Columbia, Sep. 1998, pp. 155–66.

[3] T. Liu, P. Bahl, and I. Chlamtac, "Mobility modeling, location tracking, and trajectory prediction in wireless ATM networks," *IEEE J. Select. Areas Commun.*, vol. 16, no. 6, Aug. 1998, pp. 922–936.

[4] 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. Networking*, vol. 5, no. 1, Feb. 1997, pp. 1–12.

[5] A. Aljadhai and T. Znati, "Predictive mobility support for QoS provisioning in mobile wireless environments," *IEEE J. Select. Areas Commun.*, vol. 19, no. 10, Oct. 2001, pp. 1915–1930.

[6] M.-H. Chiu and M. A. Bassiouni, "Predictive schemes for handoff prioritization in cellular networks based on mobile positioning," *IEEE J. Select. Areas Commun.*, vol. 18, no. 3, Mar. 2000, pp. 510–522.

[7] C. Oliveira, J. B. Kim, and T. Suda, "Adaptive bandwidth reservation scheme for high-speed multimedia wireless networks," *IEEE J. Select. Areas Commun.*, vol. 16, no. 6, Aug. 1998, pp. 858–874.

[8] W.-S. Soh and H. S. Kim, "Dynamic guard bandwidth scheme for wireless broadband networks", in *Proc. IEEE INFOCOM'01*, Anchorage, Alaska, USA, Apr. 2001, pp. 572–581.

[9] Y. Zhao, "Standardization of mobile phone positioning for 3G systems," *IEEE Commun. Mag.*, Jul. 2002, pp. 108–116.

[10] E. A. Bretz, "X marks the spot, maybe," *IEEE Spectrum*, Apr. 2000, pp. 26–36.

[11] J. Benedicto, S. E. Dinwiddy, G. Gatti, R. Lucas, and M. Lugert, "GALILEO: satellite system design and technology developments," European Space Agency, Nov. 2000.

[12] Y. Zhao, *Vehicle Location and Navigation Systems*, Chapter 4, Artech House, 1997.