Mobility-Prediction-Aware Bandwidth Reservation Scheme for Mobile Networks

Apollinaire Nadembega, Abdelhakim Hafid, and Tarik Taleb, *Senior Member, IEEE*

***Abstract*—Bandwidth is an extremely valuable and scarce resource in mobile networks; therefore, efficient mobility-aware bandwidth reservation is necessary to support multimedia applications (e.g., video streaming) that require quality of service (QoS). In this paper, we propose a distributed bandwidth reservation scheme called the mobility-prediction-aware bandwidth reservation (MPBR) scheme. The objective of MPBR is to reduce handoff call dropping rate and maintain acceptable new call blocking rate while providing efficient bandwidth utilization. MPBR consists of 1) a handoff time estimation (HTE) scheme that aims to estimate the time windows when a user will perform handoffs along the path to his destination, 2) an available bandwidth estimation (ABE) scheme that aims to estimate in advance available bandwidth, during the computed time windows in the cells to be traversed by the user to his destination, and 3) an efficient call admission control (ECaC) scheme that aims to control bandwidth allocation in the network cells. The simulation results show that MPBR outperforms existing schemes in terms of reducing handoff call dropping rate.**

***Index Terms*—Admission control, available bandwidth estimation (ABE), bandwidth reservation, handoff prioritization, handoff time estimation (HTE), mobile networks, quality of service (QoS).**

I. Introduction

APPLICATIONS, such as video streaming, Internet Protocol (IP) television, and voice over IP, are increasingly prevalent over telecommunication networks; thus, it becomes important to provide the QoS required by these applications to ensure an acceptable user satisfaction. The growth of these applications is due to the fact that new technologies, such as Worldwide Interoperability for Microwave Access and Third-Generation Partnership Project accesses, could offer anytime and anywhere access to mobile users [4]–[9]. However, these applications may experience performance degradation due to the intrinsic characteristics of users' mobility.

In mobile networks, QoS provisioning can be achieved by ensuring sufficient network resources (e.g., bandwidth) to mobile users during their movement and handoff operations [4]. Thus, at the start of a call, we need to be able to estimate/predict the times when handoffs will occur along the path to destination [5]. Furthermore, call admission control (CAC), at the level of each cell toward the destination, is needed to decide whether to accept a call into the corresponding cell [4], [6]–[8]. The objective is to accept as many calls as possible without degrading the QoS of ongoing calls; in particular, a new call request should be rejected if its acceptance, into a cell, will force the termination of an ongoing call handing off to this cell [8], [10], [11]. Therefore, a scheme capable of reducing handoff call dropping rate (ideally to zero) while maintaining an acceptable new call blocking rate and ensuring efficient bandwidth utilization is needed. In this paper, we propose an approach, called mobility-prediction-aware bandwidth reservation (MPBR), that provides QoS to mobile users while maintaining efficient bandwidth utilization. MPBR consists of three schemes: 1) a handoff time estimation (HTE) scheme, 2) an available bandwidth estimation (ABE) scheme, and 3) an efficient CAC (ECaC) scheme called ECaC.

HTE allows estimating the time windows when a user will perform handoffs along his movement path to his destination; it extends our proposed scheme, called HTEMOD [5], to improve estimation accuracy. Indeed, HTE estimates the time windows when the user arrives in each cell, along the path to the destination, and when he leaves the cell; we assume that the path of a user is known in advance (e.g., the schemes in [9] and [10] can be used to predict the path for a given user); in [9] and [10], user equipment (UE) embeds technology, such as GPS, that samples user coordinates of places visited by the user, along with the day and the time of the visits; however, the GPS network and user's call network are distinct. In this way, the GPS operation is not prior to the call setup. In this vein, it shall be noted that for mobile users with energy consumption constraints, some energy-aware settings can be envisioned in a way that the proposed solution is automatically disabled when the batteries of their devices go below a certain threshold. Furthermore, if the proposed solution is efficiently used for users without much constraint in energy consumption, the optimization and savings achieved in the network resources can be used to accommodate more mobile users with energy consumption constraints. More specifically, HTE uses the physics of traffic flows as a basis for designing probability distributions of traffic variables. HTE formulates specific assumptions on the physics of traffic flow to make the problem tractable while keeping it realistic. It derives analytical expressions for the probability distribution functions (pdfs) of travel times between two arbitrary locations l1 and l2 of the path to destination (i.e., a link/portion of path); this link may be the path portion from user's current location to his next handoff point (i.e., the location at which he enters his next cell).

Manuscript received December 12, 2013; accepted June 29, 2014. Date of publication August 1, 2014; date of current version June 16, 2015. The review of this paper was coordinated by Prof. H.-H. Chen.

A. Nadembega and A. Hafid are with the Network Research Labaratory, University of Montreal, Montreal, QC H3C 3J7, Canada (e-mail: nadembea@iro.umontreal.ca; ahafid@iro.umontreal.ca).

T. Taleb is with the School of Electrical Engineering, Aalto University, 02150 Espoo, Finland (e-mail: talebtarik@ieee.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2014.2345255

Indeed, during the HTE process, the coordinates of handoff points, along the path of the user under consideration (i.e., for whom handoff times are estimated), are estimated based on historical data of the coordinates of handoff points together with related information (e.g., time of the day and type of the day); thus, a probabilistic/learning-based scheme (e.g., [45], [46]) can be used to perform such an estimation. Notice that the travel time is the sum of the stopping times and the travel times on the road segments forming the link; a road segment refers to a road portion between two adjacent intersections or between an intersection and a handoff point. HTE first derives the pdf of travel times on the road segment, without considering the stopping times, making use of traffic flow conditions and current driving behavior on the road segment. To take into account the stopping times, HTE derives the stopping time function, making use of the stopping times of previous users or the previous stopping times of the user under consideration. Then, HTE sums the two functions to obtain the pdf of travel times, including stopping times, on the road segment. Finally, HTE derives the pdf of travel times on the link (i.e., between two locations), formed by all the road segments along this link, making use of the linearity of the convolution to convolve the pdf of travel times on each road segment forming the link. Thus, having the pdf of travel times of the link, HTE derives the cumulative distribution function (cdf) of travel times on the link. To set the desired level of accuracy, we use the inverse function of the cdf of travel times on the link to compute the lower and upper bound values of travel time on the link (i.e., travel time to reach the handoff point).

ABE allows estimating in advance (e.g., 30 min) the available bandwidth in a cluster of cells (e.g., a cluster of cells that will be visited by a set of users whose paths to destinations are known in advance); this cluster of cells represents the sequence of cells that were visited by the user in question throughout the trip. More specifically, taking into account the estimated handoff time windows of ongoing calls of mobile users (computed by HTE), ABE determines, at a given time in the future, the set of calls in each cell of interest (e.g., cells that will be traversed by a new call) and thus computes the available bandwidth in the cell.

ECaC allows controlling bandwidth allocation in the network cells. More specifically, taking into account the estimated/predicted available bandwidth in cells of interest (computed by ABE), ECaC accepts a new call request only if the estimated available bandwidth, in each cell that will be traversed by the new call, is sufficient to support the call when transiting the cell. Otherwise, the new call request is placed on hold if ECaC determines that the call can be accommodated soon in the future (e.g., in T seconds); if T, exceeds a predefined threshold (e.g., waiting time acceptable for this type of calls), the call is rejected.

To the best of the author's knowledge, the proposed approach (MPBR) is the first to consider estimating/predicting available bandwidth in cells to be traversed by new calls of mobile users to provide QoS support in mobile networks. MPBR considerably increases the probability of providing acceptable QoS to mobile users, in opposition to existing approaches that decide to accept/reject a new call based only on available bandwidth in the source cell; if one of the subsequent cells traversed by the call of a mobile user is congested, the call will be then simply dropped. In this paper, we do not take into account energy consumption of user equipment; indeed, we do believe that energy consumption is not an important constraint for vehicles and the impact on their batteries is expected to be negligible. For users using smartphones on board vehicles, they can always consider charging them while being on the move. This is not to mention all recent findings about increasing battery lifetime (e.g., in [12]–[14]).

The remainder of this paper is organized as follows. Section II presents related work. Section III presents a description of the HTE scheme, the ABE scheme, and the ECaC scheme. Section IV evaluates, via simulations, the proposed MPBR. Finally, Section V concludes this paper.

II. Related Work

The schemes proposed in [1], [2], [8], [11], and [15]–[20] decide to accept a new call or are not based on the behavior/state of the source cell and are usually simpler to implement but not efficient [4]. On the other hand, predictive mobile-oriented schemes [3], [6], [20]–[25] are based on the behavior/profile of mobile users and usually suffer from scalability issues, high computation and/or implementation complexity, signaling overhead, and unrealistic assumptions [4]. Vassilya and Isik [4] classified CAC and bandwidth reservation schemes based on various parameters, such as the number of cells where call admission is performed (e.g., a single cell, usually the source cell, for nondistributed schemes [2], [3], [15], [21]–[23], [25]–[28] and two or more cells for distributed schemes [1], [20]) and the way handoff requests are handled (e.g., nonprioritized or prioritized handoff). Nonprioritized handoff CAC schemes [29] do not differentiate between handoff calls and new calls; the main disadvantage of these schemes is that the forced termination probability of ongoing calls (i.e., a call moving to a congested cell is terminated/dropped) is relatively higher than it is normally anticipated. Prioritized handoff CAC schemes [1]–[3], [6], [8], [16], [20], [21], [23]–[28], [30] give handoff calls precedence over new calls (i.e., reject a new call to accommodate a handoff call); many attempts were made to address the issue of prioritized handoff CAC by making use of user mobility prediction (i.e., predictive mobile-oriented and prioritized handoff CAC schemes [3], [6], [20], [21], [23]–[25]). Thus, CAC and bandwidth reservation schemes, in mobile networks, that better satisfy bandwidth requirements of users from source to destination are those that are predictive and distributed, and support prioritized handoff [4]. They can be realized only if the dynamics of every user, such as the user's path to destination and his arrival/departure times in/from each cell in the path, are known in advance [31]. Having this knowledge in advance is not possible in realistic scenarios [4]; thus, a solution is to estimate/predict, as accurately as possible, the mobility of users and accordingly perform bandwidth allocation. More specifically, the solution should allow for 1) path prediction (list of cells to be traversed by the user from source to destination), 2) HTE (times of the user's entry/exit into/from each cell in the path, 3) bandwidth estimation (bandwidth available in each cell during the user presence at the cell along the movement path),

and (4) CAC (a call is accepted only if it can be accommodated by each cell along the entire path).

Many CAC and bandwidth reservation schemes have been proposed in the literature. In the following, we briefly overview some representative schemes [1]–[3] that are most related to our proposed approach. In these schemes, a handoff call is admitted if there is enough available bandwidth in the new cell; otherwise, it is dropped. However, the main question is how the available bandwidth is estimated and how handoff calls are prioritized relative to new calls. For example, Yao *et al.* [1] proposed a CAC and bandwidth reservation scheme that is cell oriented and distributed, and supports prioritized handoff. The current available bandwidth is estimated by making use of historical available bandwidth data. Indeed, by mapping the average value of historical bandwidth observations, the scheme estimates the available bandwidth in each cell of the network. Thus, a new call is accepted if the estimated available bandwidth is sufficient to accommodate the call, along the path to destination (it is assumed that the path is known in advance); otherwise, it is blocked. The main limitation of this scheme is that using historical network bandwidth observations (cell behavior/state) does not provide an accurate estimation of available bandwidth compared with using individual users' behaviors. Wee-Seng and Hyong [3] proposed a CAC and bandwidth reservation scheme that is predictive mobile oriented and nondistributed, and supports prioritized handoff. It reserves a certain amount of bandwidth, in the next cell, for handoff calls. The amount of reserved bandwidth is based on the estimation of the users' handoff times. The estimation procedure makes use of the pdf of the time taken by previous users to transit each road segment to the next cell. Thus, a new call will be accepted if the available bandwidth minus the reservation target is sufficient to accommodate the call in the current cell; otherwise, it is blocked. This scheme suffers from two key limitations: 1) The choice of the population to compute the probability may degrade the accuracy of predicted handoff times; indeed, the prediction error increases with the time period between the time the previous user left the next cell and the time of estimation, for the current user, is performed; and 2) the CAC is performed for only the source cell; even if a new call is accepted, it may be dropped in subsequent cells (if one is congested) to destination. Wu *et al.* [2] proposed a CAC and bandwidth reservation scheme which is cell oriented and distributed, and supports prioritized handoff. It is based on a threshold value computed by a fuzzy inference system (FIS) to prioritize handoff calls; it uses the load and the ratio of high-speed users in the next cell as input variables of FIS. By considering predefined FIS rules, they determine an output value $(0 \leq value \leq 1)$ called admission threshold parameter (TP). A new call is accepted if 1) the available bandwidth is sufficient to accommodate the call in the source cell, and 2) the TP value of the source cell is bigger than the rate of new calls in the cell. Otherwise, they apply an equal probability method to accept or block the new call request. Similar to the scheme in [3], the CAC proposed in [2] is performed for only the source cell. Furthermore, it uses current information in the next cell to determine TP; unfortunately, TP may not be valid when the user arrives in the next cell.

In conclusion, we summarize the limitations of existing bandwidth management schemes in mobile networks as follows: 1) They rely on the current behavior/state of the network cells [1], [31] to make their admission control decisions (this is not sufficient to support calls from source to destination since the state of a cell may change from the time the call of mobile user is accepted to his arrival time in the cell toward destination); 2) the schemes that make use of prediction techniques either require additional equipment [20], [27], generate a significant traffic overhead in terms of mobility data exchanges between users and network backbone [32], do not consider stop duration and traffic lights [3], [20], [27], [31]–[34], make use of old road traffic data [3], [34] or rely only on historical data about previous users [3]; 3) their admission control procedures are limited to the next cell [2], [3], [20], [26], [27] (the other cells in the path to destination are not considered); and/or 4) they rely only on historical network bandwidth observations [1].

In this paper, we propose a scheme, to process call requests, that incorporates solutions to the aforementioned limitations. In this paper, we use path prediction schemes proposed in [9] and [10]; indeed, our proposed scheme assumes the knowledge of the path from source to destination to process a call request. Thus, the key challenging issue we need to resolve is HTE; such estimation will allow to compute, with some accuracy, entry/exit times into/from cells along the path from source to destination, and this will help in computing available bandwidth and deciding to accept/reject calls.

III. Predictive Mobile-Oriented Bandwidth Reservation Scheme

Here, we present the details of the MPBR scheme. More specifically, we present the details of 1) the HTE scheme that estimates the time windows when a user will perform handoffs along his movement path to destination, 2) the ABE scheme that estimates the available bandwidth in advance in a cluster of cells using the estimated handoff times, and 3) the ECaC scheme that controls, given the estimated available bandwidth computed by ABE, bandwidth allocation. Table I shows the list of symbols/variables that are used to describe the proposed scheme.

A. Handoff Time Estimation

1) Traffic Flow and Queuing Models:

Assumptions: We assume that the road topology consists of several roads and intersections. We refer to the road portion between two road intersections or between a road intersection and a handoff point as a road segment, and identify each segment using a location pair (a, b) where $(a \rightarrow b \neq\!\!\rightarrow a)$. We refer to the intersection of a road and the border of a cell as a handoff point. We assume that a road intersection is represented by a node; each node is identified by a node ID that is related to its geographic coordinates (i.e., latitude and longitude). Based on the geographic coordinates, we build an oriented graph that represents the road topology for the proposed system; this graph may get downloaded by the entity that performs mobility prediction schemes.

TABLE I
SUMMARY OF NOTATIONS

Symbol	Description
l, ls	Length of road segment
n(t), d(t)	Number of users and density of road segment at time instant t
qi , C	Time of color I and Traffic light cycle time
d1 , d2	Lower and upper boundary density values
a, d, vm	Acceleration, deceleration and constant velocity
Δta la	Travel time and distance of acceleration phase
Δtd, ld	Travel time and distance of deceleration phase
Δtc, lc	Travel time and distance of constant phase
t_i^l, t_i^u	Lower and upper values of estimated handoff time of cell Ci
Δt	Travel time on road segment
Δd	Stopping time at road junction
ΔT	Travel time on road segment with stopping time
Fl1,l2 ()	CDF of transit/travel times of link (l1, l2)
[T0,Tz], z	Estimation time interval and number of time unit within this interval
$pbw_{alloc,l}^{i,T_k}$	Amount of passive allocated bandwidth to call l at Tk
pbw_{alloc}^{i,T_k}	Amount of passive allocated bandwidth to user ui at Tk
$PBW_{ava}^{c_j,T_k}$	Estimated available bandwidth in cell Cj
$\left[t_j^a, t_j^d\right]$	Time interval user u will spend in cell Cj

In the same way, the user's predicted path, which is an ordered list of road intersections that were visited by the user throughout the trip, can be downloaded by the entity that performs MPBR schemes. In traffic flow theory, it is common to model a vehicular flow as a continuum and represent it with macroscopic variables of flow $f(t)$ (veh/s), density $d(t)$ (veh/m), and velocity $v(t)$ (m/s). The definition of a flow gives the following relation between these three variables:

$$f(t) = d(t) \times v(t). \tag{1}$$

Thus, we assume that the state of traffic flow is fully characterized by the density d; the expression of $d(t)$ is as follows:

$$d(t) = \frac{n(t)}{l} \tag{2}$$

where $n(t)$ and l denote the number of users in the road segment at time instant t and the length of the road segment, respectively. We also make the following assumptions on the dynamics of traffic flow.

Multilane road segments: In this model, we do not take into account lane changes, passing or merging. For a road segment with several lanes, we assume that there is one queue per lane with its own dynamics. The parameters of the road network and the level of congestion may be different on each lane (e.g., to model turning movements) or equal (to limit the number of parameters of the model). In the numerical implementation presented in this paper, we consider that all lanes have different queue lengths and model the different phases of traffic signals.

Model for differences in driving behavior: In this paper, driving behavior is based on the velocity model proposed in [35]; indeed, driving behavior is a cycle of acceleration, maintaining of a constant velocity, deceleration and, finally, stopping. The free flow velocity is not the same for all users. Fig. 1 shows a simplified driving behavior cycle and the length of road segment portion associated to each phase of the cycle. In Fig. 1(a), the periods $[t_0 : t_1]$, $[t_1 : t_2]$, and $[t_2 : t_3]$ represent the acceleration phase, the constant velocity phase, and the deceleration (i.e., negative value of acceleration) phase, respectively.

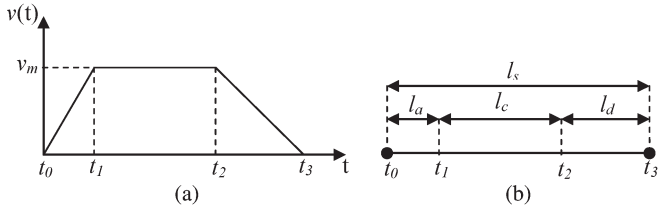


Fig. 1. (a) Simplified driving behavior cycle and (b) length of road segment portion associated to each phase of the driving behavior cycle.

TABLE II
TRAFFIC LIGHT CYCLE

time	$C = \sum_{i=1}^{n} q_i$						
color order	1	2	…	i	…	n-1	n
color time	q_1	q_2	⋯	q_i	⋯	q_{n-1}	q_n

Stationarity of traffic: During each estimation interval, the parameters of the traffic light cycles (i.e., the time of color i is denoted q_I, and the overall cycle time is denoted C) are constant (see Table II). In the case of lack of traffic lights, we apply the first-come–first-serve approach.

Road Segment Traffic Dynamics: In road networks, traffic is driven by the formation and the dissipation of queues at intersections. The dynamics of queues are characterized by shocks, which are formed at the interface of traffic flows with different densities. We define three discrete traffic conditions: free flowing, undersaturated, and congested; they represent different dynamics of the arterial link depending on the absence or the length of a queue at intersections. To determine these traffic conditions, we define d_1 and d_2 as the boundary density values between 1) *free-flowing conditions* $(d(t) \leq d_1)$ for which a user maintains more or less the same velocity and does not interact with other users; in this case, there is no queue; 2) *undersaturated conditions* $(d_1 < d(t) < d_2)$ for which users have the same velocity over a short queue; in this case, the queue fully dissipates with the end of stopping time (e.g., within the green time); and 3) *congested conditions* $(d(t) \geq d_2)$ for which the density of users forces them to slow down and thus have the same velocity over a long queue; in this case, there is a part of the queue that corresponds to vehicles which must stop multiple times before going through the intersection. Notice that our objective is to estimate the travel time on a link making use of pdf. Thus, depending on the traffic condition, we define the expression of the travel time on a road segment.

Free-flowing and undersaturated conditions: In this case, users do not stop multiple times before going through the intersection. Thus, for each road segment, each user performs only one driving behavior cycle. However, users do not experience the same stopping time, depending on the presence (respectively, the absence) of a queue at intersection. Indeed,

in the free-flowing condition, there is no queue, whereas in the undersaturated condition, there exists a short queue. For this reason, we first define the same travel time expression on a road segment for both conditions and then define a stopping time expression at the end of this road segment (i.e., intersection) for each condition.

Travel time on a road segment: The physical expression of velocity at time t_i is given by

$$v(t_i) = \sigma \times (t_i - t_{i-1}) + v(t_{i-1}) \tag{3}$$

where σ and $(t_i - t_{i-1})$ denote the acceleration/deceleration (depending on the movement) and the minimum time granularity, respectively. Based on (3) and the simplified driving behavior cycle shown in Fig. 1, we derive the expression of travel time of acceleration phase Δ_{t_a} and deceleration phase Δ_{t_d} as follows:

$$\Delta t_a = \frac{v_m}{a} \text{ and } \Delta t_d = \frac{v_m}{d} \tag{4}$$

where a, d, and v_m denote the acceleration, the deceleration, and the velocity during the constant velocity phase, respectively. It is possible that road segments do not have the same lengths. Thus, we need to know the travel distance of constant velocity phase l_c, which is given by

$$l_c = l_s - (l_a + l_a) \tag{5}$$

where l_s denotes the length of the road segment, and l_a and l_d denote the travel distance during the acceleration and deceleration phases, respectively. The expressions of l_a and l_d are derived from the integrand of the velocity function [see (3)] and is given by

$$\begin{aligned} l_a &= \left(a/2 \times (\Delta t_a)^2\right) \\ l_d &= \left(d/2 \times (\Delta t_d)^2\right) + (v_m \times \Delta t_d). \end{aligned} \tag{6}$$

Using (5), the expression of travel time of a constant velocity phase Δt_c is as follows:

$$\Delta t_c = \frac{l_c}{v_m}. \tag{7}$$

Finally, we sum the travel time of each phase of the driving behavior cycle to obtain the travel time on the road segment without stopping time. Its expression is given by

$$\Delta t = \Delta t_a + \Delta t_c + \Delta t_d. \tag{8}$$

Stopping time expression in the free-flowing condition: We define two stopping cases, i.e., traffic light case and stop sign case. In the case of a stop sign, we derive the expression of stopping time Δd as follows:

$$\Delta d = \frac{\sum_{\omega=1}^{i} \omega \Delta d_\omega}{\sum_{\omega=1}^{i} \omega} \tag{9}$$

where i and Δd_ω denote the number of stops already experienced by the user in the same condition during the current movement and its stopping time at the ωth stop sign, respectively. Notice that ω is used as a weight for Δd_ω; this mechanism allows giving more importance to the more recent stopping time. In the case of a traffic light, the stopping time depends on the time at which the user reaches the traffic light position; let t_s be this time. Using data shown in Table II, we derive the expression of the remaining time of color i when the user reaches the position of traffic light as follows:

$$r = q_i \left[(t_s \bmod C) - \sum_{f=1}^{i-1} q_f \right]. \tag{10}$$

If color i requires a stop, the stopping time $\Delta d = r$; otherwise, $\Delta d = 0$. Thus, Δd is defined as follows:

$$\Delta d = \begin{cases} 0, & \text{if non stop} \\ r, & \text{if stop is required.} \end{cases} \tag{11}$$

Stopping time expression in the undersaturated condition: We also define the expression of stopping time in the case of a traffic light or a stop sign due to the presence of queue as follows:

$$\Delta d = m \times \Delta D \tag{12}$$

where m and ΔD denote the length of the front queue and the average stopping time of the considered stopping position, respectively. Using (8), (9), (11), and (12) (depending on the condition and the case), we derive the travel time on road segment with stopping time as follows:

$$\Delta T = \Delta t + \Delta d. \tag{13}$$

Congested condition: In this condition, users stop multiple times before going through the intersection. Thus, for each road segment, each user performs several driving behavior cycles. Therefore, it is not possible to know the total number of cycles performed by a user. In addition, users have the same velocity over a long queue. Thus, for these two reasons, we do not estimate travel times based on the cycles. Due to the length of the queue, we do not estimate stopping times; they are included in the travel times of the road segment. Indeed, the expression of travel time on a road segment is derived from the arrival time t_a and the exit time t_e. It is simply given by

$$\Delta T = t_e - t_a. \tag{14}$$

2) Databases: To implement HTE, we assume that the UE maintains two main databases.

1) Data of driving behavior (DDB): It stores the essential information about the user's driving behavior required for making estimations; indeed, when the user's velocity exceeds a predefined value, HTE assumes that the user is in motion; in this case, the user's driving behavior characteristics are measured and stored in DDB every $\Delta T'$ (e.g., 1 s); an entry/record in DDB contains time t, acceleration a, velocity s, and road segment ID that

represents the user location at time t and moving at acceleration a and velocity s.

2) Data of stopping times (DST): This stores the required information about the stopping times experienced by a user; an entry/record in DST contains time t, stopping time d, and road segment ID that represents the user stop location at time t during stopping time d.

We also assume for each user request, the availability of the user's predicted path to his destination (i.e., sequence of road segments from current location to destination); a path to a destination is formed by a set of roads segments $\eta = (S_1, \ldots, S_n)$. The path prediction scheme reported in [9] and [10] can be used to compute, upon receiving a user request, the path from source to destination of the user.

To maintain DDB and DST, HTE requires information about the geographic areas covered by the network and equipment to record the user's driving behavior. Thus, we assume the following.

1) UE maintains a database, called navigation map (NM), that stores the road topology; an entry/record in NM consists of first intersection ID_1, second intersection ID_2, velocity v, road segment ID, and the length l of the road segment ID formed by the location pair (intersection ID_1, intersection ID_2) where the velocity limit is v.
2) UE embeds a GPS, a stopwatch, and an accelerometer; indeed, at regular epoch $\Delta T'$, the GPS samples the user's current location, the stopwatch samples his stopping times, and the accelerometer samples his acceleration and his velocity, together with the timestamp.

Due to the stochastic nature of the radio coverage environment, the network maintains a database called handoff point location (HPL); whenever the geographic coordinates (lat, lon) of location of a handoff point HID changes in the specific weather condition w at time t and day d, the six-tuple $(\mathrm{HID}, t, d, w, lat, lon)$ is stored in the database HPL. indeed, based on the data in HPL filtered using the current weather condition, the current time, and the current day, we compute the probability of each coordinate of the handoff point HID location and select the biggest as the coordinate of the handoff point HID location.

To limit the size of data of driving behavior (DDB) and DST, they are deleted when the user reaches the destination. MPBR may require some storage space and processing power for collecting and processing data at UE. While this would intuitively incur some additional cost, new generation UEs are manufactured with large storage space and sophisticated processing power. For example, in our simulation of mobility prediction, the PLT file requires only 2.82 MB to maintain two months of GPS trace collection. Recent mobile devices (e.g., Samsung Galaxy) can use XML or TXT files (instead of database management system); these types of files do not require large storage space. Indeed, for a mobile device of 16 GB of storage space, MPBR will use only 0.002% of this storage space that can be seen negligible.

3) Probability Distribution of Travel Times and Estimation of Time Windows: To estimate handoff time windows for a given user from source to destination, in addition to the user's predicted path, NM, DDB, and DST, HTE requires information about the density of navigation zones of interest (e.g., average number of users on a road segment). This information can be provided by a network component that has access to the database storing information about users and their locations at any time; the network component can compute, making use of (2), and transmit density information of navigation zones of interest, to HTE; this will consume a small amount of bandwidth (i.e., few bytes per transmission), which is generally negligible in the context of broadband wireless networks. The output of HTE, in return of a given user request, is an n-tuple

$$\Omega = \left\langle \left(t_1^l, t_1^u, c_1\right), \left(t_2^l, t_2^u, c_2\right), \ldots, \left(t_n^l, t_n^u, c_n\right) \right\rangle$$

where t_i^l and t_i^u denote the lower and upper bound values of the estimated time when the user will reach cell C_i, and $C_1, \ldots, C_n$ represent the cells the user is predicted to traverse toward the destination.

We first propose estimating the pdf of road segment transit/travel times by mobile users. The transit time, by a user traveling on road segment S toward the destination, is mainly impacted by traffic flow conditions (i.e., density) on S. Thus, we define three density-aware probability populations.

1) *Population in the free flow condition*: These are the times to transit S by the user under consideration; these times are computed based on the user's driving behavior (i.e., acceleration, deceleration, constant velocity and stopping times) on the road segments already transited, in the same traffic flow condition, just before entering S. HTE makes use of (8)–(11) and (13). Indeed, (8) is derived using (4)–(7). Based on the last values of acceleration, deceleration, and constant velocity of users stored in database DDB, we compute travel time during these three phases, using (4); then, using the constant velocity stored in database DDB and travel distance during the constant phase, we compute travel time during this phase using (7). Notice that travel distance during the constant phase is derived using (5), where travel distance during the acceleration and deceleration phases are computed using (6) and the last values of acceleration, deceleration, and constant velocity of users stored in database DDB. Equations (4) and (6) derive from the physical law of movement. To compute stopping times, (9) makes use of database DST, whereas (11) uses (10), where we assume that traffic light cycles are known *a priori*. Arrival time t_s is the median of the cdf of travel times $F_{l_1,l_2}(\cdot)$, which is defined in the following. The expression of t_s is as follows:

$$t_s = F_{l_1,l_2}^{-1}(0.5) \tag{15}$$

where l_1 and l_2 denote the current location and the traffic light location, respectively.

2) *Population in the undersaturated condition*: These are the times to transit S by users who are currently on S; these times are computed based on the driving behaviors, of these users, on S, or the last adjacent road segments just before entering S. HTE uses (8), (12), and (13). Equation (8) is used in the same way as the free flow

condition. For (12), AD is computed based on database DST, whereas m is assumed known *a priori*. m may be determined by a central location controller where all locations of users, on the adjacent road segment of the junction, are stored.

3) *Population in the congested condition*: These are the times to transit S by users who have already transited S and are currently located on adjacent road segments toward their destinations. Indeed, based on arrival times and exit times of users stored in database DDB, HTE computes travel times by making use of (14).

Thus, depending on the traffic flow condition, HTE determines the probability population. Let n_s denote this population and $n_s^{\Delta T_i}$ denote the fraction of n_s who transit S with ΔT_i as transit/travel time. Along the road segment S the transit/travel time ΔT is a random variable with distribution p. We derive the probability distribution p_S of transit/travel times of road segment S as follows:

$$p_s(\Delta T_i) = \frac{n_S^{\Delta T_i}}{n_S}. \tag{16}$$

To derive the pdf of travel times on a link (i.e., between two locations l_1 and l_2), p_{l_1,l_2}, we use the following fact: If X and Y are two independent random variables with respective pdf f_X and f_Y, then the pdf f_Z of the random variable $Z = X + Y$ is given by the convolution product of f_X and f_Y, which is denoted by $f_Z(Z) = (f_X * f_Y)(Z)$ and defined as

$$f_z(Z) = \int_R f_X(t) f_Y(Z-t)\, dt.$$

This classical result in probability is derived by computing the conditional pdf of *Z*, given *X*, and then integrating over the values of *X* according to the total probability law. Thus, the expression of the pdf of transit/travel times of link (l_1, l_2) p_{l_1,l_2} is given as

$$p_{l_1,l_2}(\Delta T) = \left(\prod_{i=1}^{m} p_{s_i}\right)(\Delta T) \tag{17}$$

where m denotes the number of road segments S_i forming the link (l_1, l_2). Using (17), we derive the cdf of transit/travel times of link (l_1, l_2) $F_{l_1,l_2}(\cdot)$ as follows:

$$F_{l_1,l_2}(\Delta T') = \sum p_{l_1,l_2}(\Delta T \le \Delta T') \tag{18}$$

Our goal is to estimate the time windows when a user will perform handoffs along his movement path to his destination. Therefore, for each handoff point h_k along the path to the destination of the user in question, we define the cdf on link (l_c, h_k) (i.e., between the current location l_c and the handoff point h_k) of transit/travel times by mobile users $F_{l_c,h_k}(\cdot)$. To set the desired level of accuracy, we select two values of probabilities δ_L and δ_U that determine the lower bound $\Delta t_{h_k}^l$ and upper bound $\Delta t_{h_k}^u$ of transit/travel time on link $((l_c, h_k)$. The expressions of $\Delta t_{h_k}^l$ and $\Delta t_{h_k}^u$ are derived from the inverse function of the cdf of travel times on the link $F_{l_c,h_k}(\cdot)$ and given by

$$\forall\, \delta_u, \delta_l \in [0;1] F_{l_c,h_k}^{-1}(1-\delta_l) = \Delta t_{h_k}^l; F_{l_c,h_k}^{-1}(\delta_u) = \Delta t_{h_k}^u. \tag{19}$$

We obtain the lower bound t_k^l and upper bound t_k^u of the estimated time when the user will reach the handoff point h_k as follows:

$$t_k^l = t_0 + \Delta t_{h_k}^l \text{ and } t_k^u = t_0 + \Delta t_{h_k}^u \tag{20}$$

where t_0 denotes the initial time of the estimation.

B. Available Bandwidth Estimation

The objective of ABE is to estimate available bandwidth in cells, at a given time in the future, assuming prior knowledge about all the incoming/outgoing handoffs that will occur within a limited time into the future in these cells.

Here, we describe the details of ABE and explain how the n-tuple Ω predictions, computed by HTE (see Section III-A) are used. We make use of both incoming and outgoing handoff predictions to achieve more efficient tradeoffs between handoff call dropping rate and new call blocking rate. We assume that a user may initiate several calls with different durations.

1) System Model: Similar to [25], we do not consider delay-insensitive calls that can tolerate long handoff delays and soft handoffs in code-division multiple-access systems, in which a mobile user can simultaneously connect with two or more cells. In our model, we only consider calls that require fixed bandwidth guarantees. We follow the common assumption of existing reservation schemes that each cell j has a fixed capacity of BW^{c_j} [3], [25]. Given the bandwidth demand of individual calls, the cell performs admission control to ensure that the total demand of all active calls does not exceed BW^{c_j}.

2) Databases: We make the following two assumptions.

a) The network maintains a database, called user call data (UCD), which records data about users' calls. An entry/record in UCD contains bandwidth b, time t, call duration d, call ID, and user that represents the user who makes the call ID at time t during call duration d and required bandwidth b. To limit the size of UCD, each entry/record in UCD is deleted when the call is completed.

b) Prior knowledge of the time windows when a user will perform handoffs along the predicted path to a destination $\Omega = \langle (t_1^l, t_1^u, c_1), \ldots, (t_j^l, t_j^u, c_j), \ldots, (t_n^l, t_n^u, c_n)\rangle$.

3) Description: To estimate available bandwidth in advance, ABE makes use of UCD and Ω. More specifically, taking into account the estimated handoff time windows of users Ω (computed by HTE), ABE determines, at a given time T_k in the future, the set of ongoing calls in each cell of interest (e.g., cells $C = \{c_1, \ldots, c_j, \ldots, c_n\}$ that will be traversed by the user while making the new call) and thus computes the available bandwidth in the cell. Indeed, we compute the available bandwidth in each cell in C at $T_k \in [T_0, T_z]$, where $[T_0, T_z]$ denotes the estimation time interval, and $T_k = T_0 + k\Delta t$, with $k \in N$, Δt denoting the time unit of estimation, and index z denoting the number of time units within time interval $[T_0, T_z]$.

Let $U = \{u_1, \ldots, u_i, \ldots, u_m\}$ denote the list of users who are expected to transit at least one of the cells in C. U is obtained using the n-tuple predictions Ω of all users in a predefined navigation zone. Thus, at T_k, based on the characteristics of calls (e.g., duration from UCD) of $u_i \in U$, we compute

TABLE III
MATRIX OF PASSIVE ALLOCATED BANDWIDTH

time \ user	u_1	⋯	u_i	⋯	u_m
T_0			pbw_{alloc}^{i,T_0}		
...			…		
T_k			pbw_{alloc}^{i,T_k}		
...			…		
T_z			pbw_{alloc}^{i,T_z}		

TABLE IV
MATRIX OF ESTIMATED AVAILABLE BANDWIDTH

time \ cells	c_1	⋯	c_j	⋯	c_n
T_0			$PBW_{ava}^{c_j,T_0}$		
...	…	…	…	…	…
T_k			$PBW_{ava}^{c_j,T_k}$		
...	…	…	…	…	…
T_z			$PBW_{ava}^{c_j,T_z}$		

the amount of passive allocated bandwidth $\text{pbw}_{\text{alloc}}^{i,T_k}$ (i.e., the amount of bandwidth to be reserved to user u_i at T_k to prevent his calls from being dropped). Its expression is given by

$$\text{pbw}_{\text{alloc}}^{i,T_k} = \sum_{l=1}^{q} \text{pbw}_{\text{alloc},\,l}^{i,\,T_k} \tag{21}$$

where q is the number of calls of u_i that are expected to be ongoing at T_k, and $\text{pbw}_{alloc,\,l}^{i,\,T_k}$ is the amount of passive allocated bandwidth to call l at T_k. Using (21), we compute the amount of passive allocated bandwidth of each element of U at each time T_k; then, we derive the matrix of passive allocated bandwidth at time T_k within time period $[T_0,\, T_z]$ (see Table III). Using Ω of each user $u_i \in U$, we compute, for each user u_i, his transit time in each cell in C that it is expected to traverse. Indeed, for cell $c_j \in C$, we consider t_j^l of $(t_j^l,\, t_j^u,\, c_j)$ as the arrival time t_j^a at c_j and t_{j+1}^u of $(t_{j+1}^l,\, t_{j+1}^u,\, c_{j+1})$ (i.e., the next cell to be visited after cell c_j) as the departure time t_j^d from c_j. With respect to the last cell, in C, to be visited by a user u_i, we compute only his t_n^a. Thus, we obtain the time interval each user $u_i \in U$ will spend in each cell $c_j \in C$: Soj $= \{\ldots, (t_j^a, t_j^d, c_j), \ldots, (t_n^a, c_n)\}$.

Using Soj and the matrix of passive allocated bandwidth (see Table III), we compute the estimated available bandwidth in each cell $c_j \in C$ at T_k as follows:

$$\text{PBW}_{\text{ava}}^{c_j,T_k} = \text{BW}^{c_j} - \sum_{i-1}^{g} \text{pbw}_{\text{alloc}}^{i,T_k} \tag{22}$$

where g denotes the number of users $u_i \in U$ who are expected to be located in cell c_j at T_k. Using (22), we compute the matrix of estimated available bandwidth (see Table IV).

C. Call Admission Control

To better understand the logic leading to the ECaC, we raise the following question: Suppose we have perfect knowledge about bandwidth available in a cluster of cells that will be traversed by a new call within a limited time in the future: What needs to be done in case of lack of bandwidth in certain cells of the cluster to accommodate this new call, and how much bandwidth should be reserved in each cell of the cluster to prevent any of the handoff calls from being dropped?

We assume that ECaC uses the same system model as ABE.

1) Databases: We assume first that the network maintains a database, called earlier completed call data (ECCD), which records data about earlier completed calls; an entry/record in ECCD contains call ID, bandwidth b, time t_1, time t_2, cell ID, and date d. The call is initiated with b as allocated bandwidth and completes in cell ID, at d, at t_1, whereas it was estimated/predicted to complete at t_2 $(t_2 > t_1)$. The entry/record in ECCD is extracted from UCD before its deletion; indeed, when an entry/record in UCD is completed before its expected time (defined as t of UCD $+\, d$ of UCD), it is extracted from UCD (before its deletion) and inserted into ECCD. To limit the size of ECCD, each entry/record in ECCD is deleted after one week. Second, we assume prior knowledge of the time windows when the user will perform handoffs along the predicted path to a destination $\Delta = \langle (t_1^l, t_1^u, c_1), \ldots, (t_j^l, t_j^u, c_j), \ldots, (t_n^l, t_n^u, c_n) \rangle$ (computed by HTE). Finally, we assume prior knowledge of the matrix of estimated available bandwidth (computed by ABE).

2) Description: ECaC uses ECCD, Ω, and the matrix of estimated available bandwidth (see Table IV) to manage the bandwidth allocation in the cells along the path to the destination of the user in question. To prioritize handoff calls over new calls, each cell reserves some bandwidth that can only be used by handoff calls. This reservation takes into account the users' transit and arrival time in each cell and his required bandwidth at this arrival time. Specifically, a new call request is accepted if the available bandwidth after its acceptance is sufficient to accommodate handoff calls when they enter into the cell. Let nc be a new call initiated by user u, $C_u = \{c_1, \ldots c_j, \ldots, c_w\}$ the list of cells to be traversed by the user u to destination, $\Omega_u = \langle (t_1^l, t_1^u, c_1), \ldots, (t_j^l, t_j^u, c_j), \ldots, (t_w^l, t_w^u, c_w) \rangle$ the handoff time windows of the user u along the path to destination, and $\text{Soj}_u = \langle \ldots, (t_j^a, t_j^d, c_j), \ldots, (t_w^a, c_w) \rangle$ the time intervals the user u will spend in each cell along the path to destination. The new call nc is accepted when available bandwidth $\text{PBW}_{\text{ava}}^{c_j,T_k}$ (i.e., the amount of bandwidth that should not be reserved or used in cell $c_j \in C_u$ at $T_k \in [t_j^a,\, t_j^b]$) is bigger than or equal to the bandwidth BW_{req} required by nc, during the time interval $[t_j^a, t_j^d]$; otherwise, it may be blocked. For the sake of better understanding, let us consider the example shown in Fig. 2. In this example, the new call should be blocked due to insufficient bandwidth in Cell 2. However, it may happen that bandwidth be sufficient in the cells, along the path, after the cell without sufficient bandwidth, called *critical cell* (e.g., cell 2 in Fig. 2). Indeed, if the new call starts in cell 3 it will be accepted. ECaC defines the concept of best instant to start (BIS) as the time the

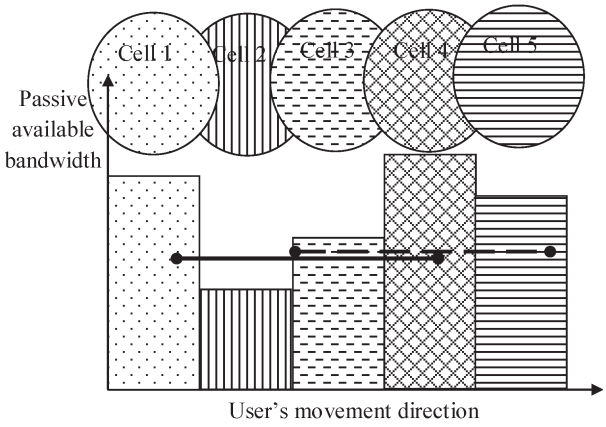


Fig. 2. User's required bandwidth and estimated available bandwidth along the user's path to destination.

user has to wait before successfully starting a new call. The expression of BIS is given by

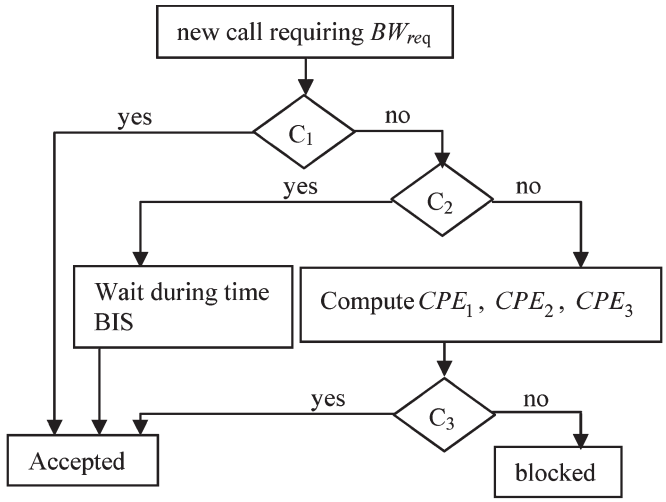
$$\text{BIS} = t_j^a - T_0 \tag{23}$$

where T_0 denotes the time of the call request, and t_j^a (e.g., t_3^a in Fig. 2) the time when the user is expected to enter cell c_j (e.g., cell 3 in Fig. 2). To avoid long delays, BIS should be shorter than a predefined delay threshold BISt according to the call type, e.g., if the new call requires an immediate connection, BISt is set to zero. In the following, "current user" and "user in question" (i.e., user who makes the new call request) are interchangeably used, referring to the same user.

To limit new call blocking rate that may be caused due to errors in estimation/prediction, ECaC introduces the concept of catching of prediction errors (CPE) in *critical cells*. CPE is applied when BIS exceeds BISt. CPE is computed as the sum of the following measurements.

1) *The average of released bandwidth per call for calls that complete earlier than expected* (CPE_1): The released bandwidth corresponds to the bandwidth passively reserved between the actual completion and the estimated completion.
2) *The sum of reserved bandwidth for handoff calls that are late according to their estimated arrival time and to the arrival time of current user* (CPE_2): The reserved bandwidth corresponds to the bandwidth reserved passively between the estimated arrival time and the actual arrival time.
3) *The sum of allocated bandwidth to ongoing calls that leave the cell earlier than their estimated exit time and the arrival time of current user* (CPE_3): The allocated bandwidth corresponds to the bandwidth allocated between the actual exit time and the estimated exit time.

Indeed, when a new call cannot be accepted due to a *critical cell*, ECaC checks whether its BIS is shorter than or equal to its BISt. If yes, the new call is accepted and the bandwidth reservation process starts after BIS time; otherwise, ECaC checks whether the sum of the estimated values of released bandwidth (i.e., $\text{CPE}_1 + \text{CPE}_2 + \text{CPE}_3$) is bigger than or equal to the required bandwidth of the new call BW_{req}. If yes, the new call is accepted; otherwise, the new call is blocked. Fig. 3 shows the operation of ECaC in processing a new call request.



C_1: $BWreq \le PBW_{ava}^{c_j,T_k}$ along the path of new call according to user's arrival time at each cell.
C_2: BIS $\le$ BISt.
C_3: $\text{CPE}_1+\text{CPE}_2+\text{CPE}_3 \ge BWreq$

Fig. 3. Operation of ECaC to accept or block a new call request.

To compute CPE_1, we use the database ECCD that records data about earlier completed calls. Let L_w be the list of entries/records in ECCD, where: 1) the cell ID is equal to the cell ID of the critical cell; 2) the average $(t_2 - t_1)$ per call is longer than or equal to the time interval the current user will spend in the critical cell; and 3) the earlier call has the same type of day (e.g., weekend or weekdays) as the new call. The expression of CPE_1 is defined as follows:

$$\text{CPE}_1 = \frac{\sum_{l=1}^{n_w} \text{bw}_{\text{alloc}}^l}{n_w} \tag{24}$$

where $\text{bw}_{\text{alloc}}^l$ is the amount of allocated bandwidth to call l in L_w, and n_w is the cardinality of L_w.

CPE_1 computation is based on one-week historical data; each entry in ECCD is deleted after one week from insertion. CPE_1can be easily computed using the following SQL query:

"**Select AVG**(b) **from** ECCD **where** cell ID $= critical\ cell\ ID$ **and** type_day(d) = type_day(current day) **and** [select AVG$(t_2 - t_1)$ from ECCD **where** $cellID = critical\ cell\ ID$ **and** type_day(d)=type_day(current day)] $\ge (t_j^d - t_j^a)$".

To compute CPE_2 and CPE_3, we use the database UCD that records data about ongoing calls. The list of users required to compute CPE_2 is obtained based on the maximum average velocity per road segment of users who are expected to enter the critical cell before the current user, whereas the list of users required to compute CPE_3 is obtained by making use of the minimum average velocity per road segment of users who are currently located in the *critical cell* and expected to exit after the current user reaches the *critical cell*. The maximum average and minimum average velocities of users are extracted from the database DDB. Thus, ECaC computes the minimum

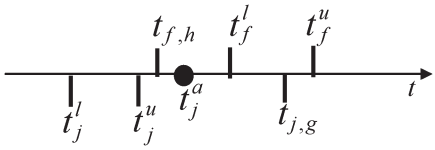


Fig. 4. "Arrive late" and "exit earlier" users.

(respectively, maximum) travel time, which is defined as [(distance to the *critical cell* border/maximum (respectively, minimum) average velocity) + current time], to reach the *critical cell* (respectively, cell to be visited after the *critical cell*) border. Let $c_j \in C_u$ be a *critical cell* with respect to the current user u, (t_j^a, t_j^d, c_j) be the user's estimated time window when the user will transit cell c_j, (t_j^l, t_j^u, c_j) (where $t_j^u < t_j^a$) be the estimated time window when user g will enter cell c_j, and (t_f^l, t_f^u, c_f) (where $t_j^a < t_f^l$) be the estimated time windows when user h will handoff from cell c_j to cell c_f. User g (respectively, h) is said to "arrive late" (respectively, "exit earlier") when his minimum (respectively, maximum) travel time $t_{j,g}$ (respectively, $t_{f,h}$) to reach the border of *critical cell* c_j is bigger (respectively, smaller) than t_j^a. In other words, user g (respectively, h) cannot reach (respectively, exit), with maximum (respectively, minimum) average velocity, the *critical cell* c_j before (respectively, after) the estimated arrival time of user u (see Fig. 4).

In the following, we refer to the reserved bandwidth for an incoming call or the allocated bandwidth to an outgoing call as required bandwidth.

Let L_g (respectively, L_h) be the list of "arrive late" (respectively, "exit earlier") users. These lists are extracted from the set of users who are expected to be located in the critical cell during the transit time, $[t_j^a, t_j^d]$ of current user in this critical cell. To compute these lists, we use Ω of all users in U (i.e., users who are expected to transit at least one of the cells in C); then, based on their maximum average or minimum average velocities, we identify the users of each list (L_g or L_h). The expression of CPE_2 and CPE_3 are defined as follows:

$$\text{CPE}_2 = \sum_{g=1}^{n_g} \text{bw}_{\text{req}}^g \text{ and } \text{CPE}_3 = \sum_{h=1}^{n_h} \text{bw}_{\text{req}}^h \tag{25}$$

where bw_{req}^g is the total amount of required bandwidth for calls of user g in list L_g, n_g is the cardinality of L_g, bw_{req}^h is the total required bandwidth for calls of user h in list L_h, and n_h is the cardinality of L_h. For each user x in L_g or L_h, the total amount of required bandwidth in the critical cell is recorded in the database UCD; let L be the list of entries/records in UCD where the user is equal to x, and $t + d$ is bigger than t_j^a. The expression of bw_{req}^x is defined as follows:

$$\text{bw}_{\text{req}}^x = \sum_{l=1}^{n} \text{bw}(l) \tag{26}$$

where $\text{bw}(l)$ is the amount of required bandwidth of call l (i.e., value of bandwidth b of the entry/record l in UCD), and n is the cardinality of L. bw_{req}^x can be easily computed using the following SQL query:

"**Select** $\text{SUM}(b)$ **from** UCD **where** $user = x$ **and** $t + d > t_j^a$".

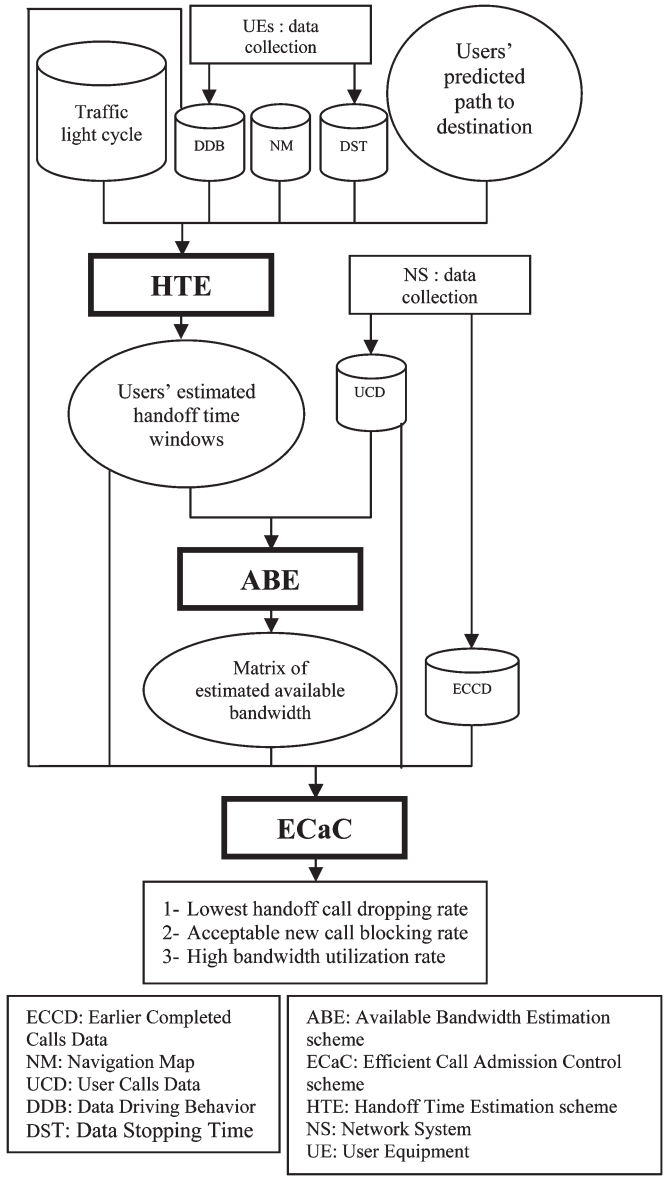


Fig. 5. MPBR processes.

Once a new call is accepted, it is assigned high priority over upcoming new calls. In each cell c_j along the path to destination, the required bandwidth is reserved during $[t_j^a, t_j^d]$. However, the call may lose its high-priority status in case of HTE errors (i.e., arrival outside the estimated handoff time windows); therefore, its passive bandwidth reservation is immediately released. Notice that bandwidth reservation is passive, i.e., the reserved bandwidth may be used by any handoff call; however, when a high-priority handoff call arrives and the available bandwidth, i.e., bandwidth that is not used, is not enough, ECaC drops some low-priority handoff calls to release bandwidth. When all low-priority handoff calls are dropped and the available bandwidth is still not enough, the high-priority handoff call may be dropped. These cases may happen due to mobility prediction errors.

Fig. 5 shows the architecture of MPBR, including all databases and components, together with their interactions.

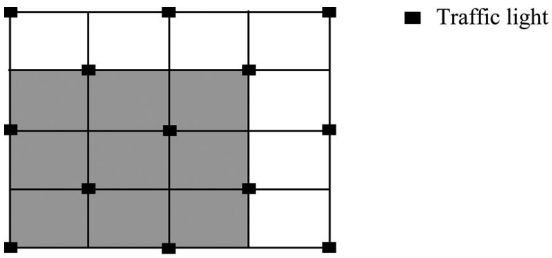


Fig. 6. Cell coverage and traffic light locations.

IV. Performance Evaluation

Here, we evaluate via simulations the performance of MPBR in terms of 1) handoff time prediction error, 2) new call blocking rate, 3) handoff call dropping rate; and 4) bandwidth utilization rate. We compare MPBR against the schemes described in [1]–[3], which are referred to as AP1, AP2, and AP3, respectively. We selected AP1 and AP2 because, to the best of our knowledge, they represent the most recent work related to CAC and bandwidth reservation in wireless mobile networking that outperforms existing approaches (e.g., [6], [16], [20], [21], and [27]). However, they do not use prediction techniques to perform call request. Thus, we also selected AP3, which is more related to MPBR in terms of prediction.

A. Simulation Setup

To evaluate MPBR, we used mobile user traces acquired from the Generic Mobility Simulation Framework (GMSF) project [36]. GMSF proposes new vehicular mobility models that are based on highly detailed road maps from a geographic information system (GIS) and realistic microscopic behaviors (car-following and traffic lights management). We developed programs to process GMSF traces to take into account handoff events and traffic light cycles. We also changed the selection process (random in GMSF models) of initial velocity, stopping time, maximum velocity, acceleration, and deceleration to obtain more realistic traces of users. An entry/record in user trace database contains user UID, time t, acceleration a, velocity v, road segment RSID, cell CID, Cartesian coordinates (X and Y), and event e that represents the user action (e.g., move, handoff, stop or change road segment) at a specific time t, on a particular location (X *and* Y) of road segment RSID in cell CID.

The simulation environment is a 2-D environment: the roads are arranged in a mesh shape, the cell coverage is formed by nine blocks (i.e., rectangular area formed by three road segments per side), and only one on the two ends of a road segment has a traffic light, as shown in Fig. 6.

The cellular structure can be typically seen in a metropolitan downtown area. We make the following assumptions for this 2-D environment: 1) Each user has a predefined (predicted) path; 2) at the beginning of the simulations, each user u randomly chooses acceleration A_u (in m/s^2) from within [0.1,0.2], deceleration D_u (in m/s^2) from within $[-0.2, -0.1]$, stopping time S_u (in seconds) from within [0,5], and maximum velocity V_{m_u} (in meters per second) from within $[V_{m_u} - 1, V_{m_u} + 1]$ [10], [14]; 3) after each stop, the initial velocity is 0; 4) at the intersection of two road segments, a user selects to continue straight, to turn left, or to turn right according to his predefined path; 5) on each road segment, a user u reaches a maximum velocity V_m chosen randomly from within $[V_{m_u} - 1, V_{m_u} + 1]$, and the user's acceleration A_u and deceleration D_u do not vary during the simulations; 6) the cellular network is composed of 81 cells (i.e., a 9∗9 mesh), and each cell's diameter is 1200 m; 7) at each stop sign, user u experiences stopping time S randomly chosen from within $[S_u - 1, S_u + 1]$; and 8) a traffic light signal switches from red (60 s) to orange (5 s) and then to green (60 s).

TABLE V
Simulation Parameters

Parameter	Value
GMSF-Mobility model - Manhattan Model (MN)	simulation area size=100km^2 (10000 m/dimension), number of blocks in one dimension=25, number of users=1500, maximum speed=14 m/s, simulation time =1200sec, $T_k - T_{k-1}$ =1sec
l_s, d_1 and d_2	400m, 8 users/cell and 22 users/cell
σ	uniformly distributed between -0.2 m/s^2 and 0.2 m/s^2
$\delta_L = \delta_U$	0.6
BW_{req}	chosen from the set {1, 2, 3, 4} Mbps with equal probability
T_{th} [2]	10 m/s

Similar to [3], [6], [25], and [26], new call requests are generated according to a Poisson distribution with rate λ (calls/second/user). In the simulations, we focus on the improvement of handoff call dropping rate; thus, similar to [1]–[3], [6], [25], [26], and [42]–[44], we do not consider call characteristics in terms of required bit rate; we simply assume that each call requires a constant amount of bandwidth and receives this amount of bandwidth when it is accepted. Even in the case of a variable bit rate (VBR) stream, the call can be simulated, in a simple manner, as a constant bit rate stream with its bit rate being set to the highest instantaneous rate of the VBR stream [42]. The call time is assumed exponentially distributed with a mean of 300 s. Table V shows the values of the parameters used in the simulations.

B. Results Analysis

Simulation results are averaged over multiple runs with different pseudorandom number generator seeds. We define four parameters to evaluate the performance of MPBR.

- Average handoff time prediction error gap (i.e., difference between real and predicted handoff time instants) per user denoted by average_error; it is computed as follows:

$$\text{Average\_error} = \frac{\sum_{u=1}^{q} \varepsilon_u}{q} \tag{27}$$

where q denotes the total number of users, and ε_u is the average handoff time prediction error gap per handoff point for each user u.

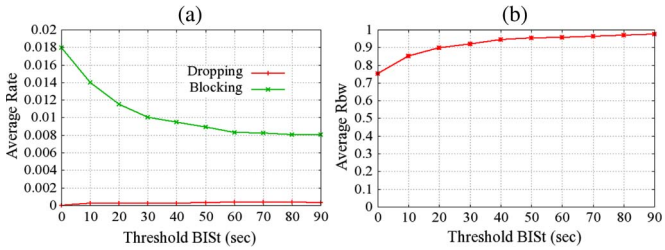


Fig. 7. MPBR performance metrics (Rb, Rd, and Rbw) versus BISt variation.

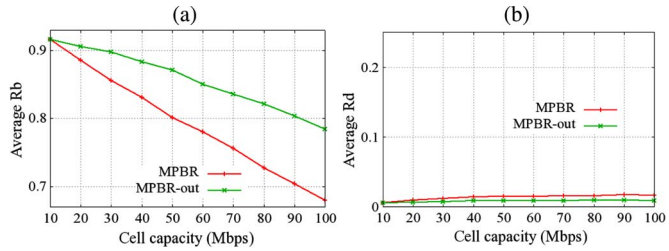


Fig. 8. Impact of CPE on MPBR performance metrics (Rb and Rd) versus cell capacity variation.

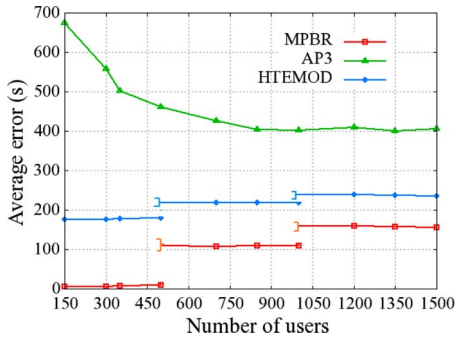


Fig. 9. Average prediction error gap versus number of users.

- New call blocking rate denoted by Rb; it is computed as follows:

$$\text{Rb} = \frac{n_b}{m_b} \tag{28}$$

where n_b denotes the number of new call requests blocked and m_b is the total number of new call requests (i.e., accepted and blocked).

- The handoff call dropping rate denoted by Rd is computed as follows:

$$\text{Rd} = \frac{n_d}{m_d} \tag{29}$$

where n_d denotes the number of handoff calls dropped, and $m_d = m_b - n_b$ is the number of call requests accepted.

- The bandwidth utilization rate denoted by Rbw is computed as follows:

$$\text{Rbw} = \frac{\text{bw}_{\text{alloc}}}{bw} \tag{30}$$

where bw_{alloc} denotes the average amount of allocated bandwidth per time unit $(T_k - T_{k-1})$, and bw is the overall cell capacity.

Fig. 7 shows the average rate of new call blocking, the handoff call dropping, and the bandwidth utilization of MPBR when varying delay threshold BISt. In this set of simulations, the call arrival rate λ is set to 0.001 call/s/user, the cell capacity is set to 300 Mb/s, and the number of users in the simulation area is 1500. Fig. 7(a) shows that the average rate of new call blocking is below 0.98%, whereas the average rate of handoff call dropping is almost null when the value of BISt is set to 0 s. In Fig. 7(a), we observe that, when BISt increases from 0 to 90 s, the average new call blocking rate decreases by 1% (i.e., [average Rb −at 0 s average Rb at 90 s]). This is expected since, when BISt increases, the number of successful/accepted new call requests increases, and thus, the new call blocking rate decreases. However, we observe that the average handoff call dropping rate remains constant even when BISt increases; this means that about 50% of the successful/accepted new call requests, due to BIS concept, have not been dropped. In Fig. 7(b), we also observe that the average bandwidth utilization rate increases with BISt. This is also expected since, when BISt increases, the amount of allocated bandwidth increases; thus, the bandwidth utilization rate increases. We conveniently conclude that the BIS concept improves the performance of MPBR.

Fig. 8 shows the average new call blocking rate and the average handoff call dropping rate of MPBR for varying cell capacities. In the figure, the MPBR version not integrating the concept of CPE is referred to as MPBR-out. In this set of simulations, the call arrival rate λ is set to 0.03 call/s/user, BISt is randomly chosen from {0, 30, 60, 90} s, and the number of users in the simulation area is 1500. Fig. 8(a) shows that MPBR outperforms MPBR-out; indeed, MPBR provides an average of 0.79 per 10 Mb/s, whereas MPBR-out provides an average of 0.86 per 10 Mb/s. The average relative improvement (defined as [average Rb of variant − average Rb of MPBR]) of MPBR compared with MPBR-out is about 7% per 10 Mb/s. Fig. 8(b) shows that MPBR is slightly less efficient than MPBR-out: it provides an average of 0.008 per 10 Mb/s, whereas MPBR provides an average of 0.01 per 10 Mb/s; the average relative improvement (defined as [average Rb of MPBR—average Rb of variant]) of MPBR-out compared with MPBR is about 0.2% 10 Mb/s, which is negligible. Thus, we conclude that MPBR provides a reduction of 7% 10 Mb/s of the new call blocking rate with negligible increase in handoff call dropping rate.

Fig. 9 shows the average prediction error gap [computed by (27)] of MPBR and AP3 for varying populations of users; AP1 and AP2 are not shown since they do not predict handoff times of users. We observe that MPBR handily outperforms AP3. Indeed, the relative improvement of MPBR compared with AP3 is about 77.1% per 150 users in the free flow condition (from 150 to 500 users), about 42.9% per 150 users in the undersaturated condition (from 501 to 1000 users), and about 35.3% per 150 users in the congested condition (from 1001 to 1500 users). Overall, the average relative improvement of MPBR compared with AP3 is about 54.3% per 150 users. This can be explained by the fact that MPBR selects the probability population according to the traffic flow condition; the selection allows for more accurate computation of the corresponding pdf. This is in opposition to AP3 that considers all previous users as the probable population in all cases. Furthermore, MPBR

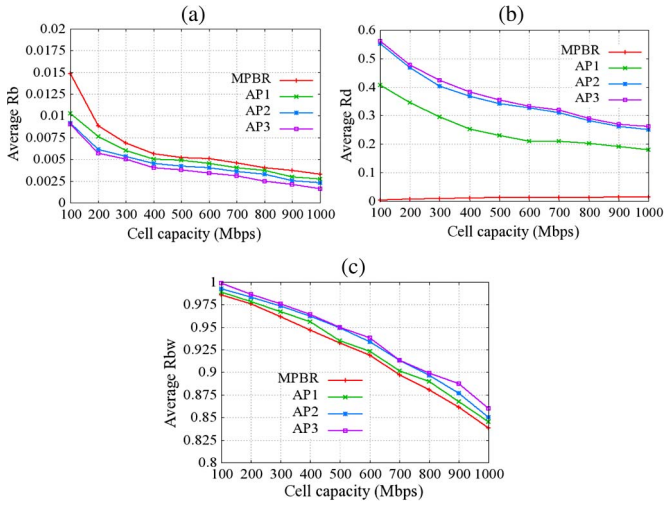


Fig. 10. Performance metrics (Rb, Rd, and Rbw) versus cell capacity variation.

uses the velocity function (in contrast to average velocity in case of AP3) and takes into account traffic light scheduling (not the case of AP3). Fig. 9 also shows that AP3 prediction error decreases when the number of users increases; this can be explained by the fact that, when the number of users increases, their speeds tend to be equal; thus, prediction based on the speeds of all previous users provides better performance (compared with the case of a small number of users). We also observe that MPBR prediction error increases with the number of users; this can be explained by the fact that, when the number of users increases, their stopping times estimation error increases, and thus, prediction based on the stopping times of previous users does not provide better performance (compared with the case of a small number of users).

We also observe that MPBR outperforms HTEMOD. The relative improvement of MPBR compared with HTEMOD is about 170 s per 150 users in the free flow condition, about 110 s per 150 users in the undersaturated condition, and about 80 s per 150 users in the congested condition. This can be explained by the fact that 1) HTEMOD does not take into account the road junctions with traffic light, whereas the HTE of MPBR, for this type of junctions, defines a function that estimates the stopping times that a user will spend at each road junction with traffic light along her/his path to destination; 2) HTEMOD makes use of predefined values of acceleration and deceleration to compute the travel time during these phases, whereas HTE of MPBR computes these travel times based on the last recorded values of acceleration, maximal speed, and deceleration; and 3) the HTE of MPBR considers the stopping times function in the pdf of travel time definition in contrast to HTEMOD, which considers the stopping time as a constant value that is added to the travel times.

Fig. 10 shows (a) the average new call blocking rate, (b) the average handoff call dropping rate, and (c) the average bandwidth utilization rate for different cell capacities. In this set of simulations, the call arrival rate λ is set to 0.001 call/s/user, BISt is randomly chosen from {0, 30, 60, 90} s and the number of users in the simulation area is 1500. Fig. 10(a) shows that AP3, AP1, and AP2 outperform MPBR. Indeed, AP3 (slightly more efficient than AP1 and AP2 in this scenario) provides an average call blocking rate of 0.0041 per 100 Mb/s, whereas MPBR provides an average call blocking rate of 0.0063 per 100 Mb/s; thus, the average relative improvement of AP3 compared with MPBR is about 0.22% per 100 Mb/s. We observe that, for the four schemes, the average new call blocking rate decreases when the cell capacity increases. This is expected since, when the cell capacity increases, the number of successful/accepted new call requests increases, and thus, the new call blocking rate decreases. Fig. 10(b) shows that MPBR outperforms AP1, AP2, and AP3. For example, MPBR provides an average of 0.009 per 100 Mb/s, whereas AP1 (slightly more efficient than AP2 and AP3 in this scenario) provides an average handoff call dropping rate of 0.25 per 100 Mb/s; overall, the average relative improvement of MPBR compared with AP1 is about 24% per 100 Mb/s. We observe that the average handoff call dropping rate of AP3, AP1, and AP2 decreases when the cell capacity increases; this is expected since. when the cell capacity increases, the number of handoff calls accommodated in a next cell increases, and thus, the handoff call dropping rate decreases. Although AP3 uses mobility prediction, it does not outperform AP1 and AP2 because its prediction is limited to the next cell while AP1 and AP2 estimate the available bandwidth, at the time of the call request, along the path to the destination (they assume *a priori* knowledge of the destination). We also observe that the average handoff call dropping rate of MPBR remains constant even when the cell capacity increases. This can be explained by the fact that MPBR makes passive reservation (in advance with good accuracy; see Fig. 9) along the user path to destination before the acceptance of the call. Although AP1 uses similar reservation mechanism (i.e., reservation along user's path to destination), it does not make use of the efficient available bandwidth estimation scheme; nonetheless, AP1 slightly outperforms AP2 and AP3 in this scenario [see Fig. 10(b)]. Fig. 10(c) shows that, for the four schemes, the average bandwidth utilization rate decreases when the cell capacity increases. This is expected since, when the cell capacity increases, the amount of available bandwidth increases, and thus, the bandwidth utilization rate decreases; indeed, when the cell capacity increases, the amount of accepted calls increases, and when these calls are completed, they release bandwidth that is not immediately used when the call arrival rate remains constant. In this case, the bandwidth utilization rate decreases. Fig. 10(c) also shows that AP1, AP2, and AP3 outperform MPBR. AP3 (slightly more efficient than AP1 and AP2 in this scenario) provides an average bandwidth utilization rate of 0.94 per 100 Mb/s, whereas MPBR provides an average bandwidth utilization rate of 0.92 per 100 Mb/s; the average relative improvement of AP3 compared with MPBR is about 2% per 100 Mb/s, which is negligible.

We conclude that, compared with AP1, AP2, and AP3, MPBR provides a considerable reduction of 24% per 100 Mb/s in handoff call dropping rate and a slight increase of 0.22% per 100 Mb/s in new call blocking rate with similar bandwidth utilization, irrespective of the network cell capacities. The 0.22% new call blocking rate increase is a small price to pay for the small handoff call dropping rate.

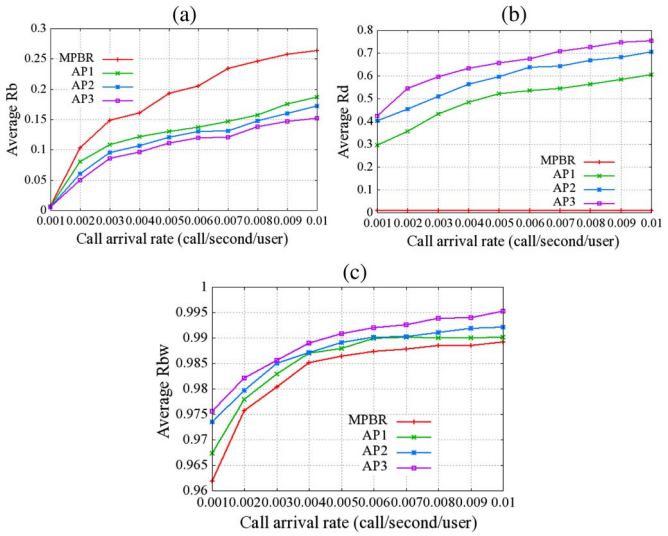


Fig. 11. Performance metrics (Rb, Rd, and Rbw) versus call arrival rate variation.

Fig. 11 shows (a) the average new call blocking rate, (b) the average handoff call dropping rate, and (c) the average bandwidth utilization rate for varying call arrival rates. In this set of simulations, the cell capacity is set to 300 Mb/s, BISt is chosen from within the set {0, 30, 60, 90} s with equal probability, and the number of users in the simulation area remains 1500. Fig. 11(a) shows that AP3, AP1, and AP2 outperform MPBR. Indeed, AP3 (slightly more efficient than AP1 and AP2 in this scenario) provides an average new call blocking rate of 0.10 per 0.001 call arrival rate, whereas MPBR provides an average new call blocking rate of 0.18 per 0.001 call arrival rate; the average relative improvement of AP3 compared with MPBR is about 8% per 0.001 call arrival rate. We observe that, for the four schemes, the average new call blocking rate increases along with the call arrival rate. This is expected since when the call arrival rate increases, the number of successful/accepted new call requests decreases, and thus, the new call blocking rate increases. Fig. 11(b) shows that MPBR outperforms AP1, AP2, and AP3; MPBR provides an average handoff call dropping rate of 0.009 per 0.001 call arrival rate, whereas AP1 (slightly more efficient than AP2 and AP3 in this scenario) provides an average handoff call dropping rate of 0.49 per 0.001 call arrival rate. Overall, the average relative improvement of MPBR compared with AP1 is about 48% per 0.001 call arrival rate. We observe that the average handoff call dropping rate of AP1, AP2, and AP3 increases along with call arrival rate. This is expected since, when the call arrival rate increases, the number of handoff calls accommodated in a next cell decreases, and thus, the handoff call dropping rate increases. Although AP3 uses mobility prediction, it does not outperform AP1 and AP2 because its prediction is limited to the next cell, whereas AP1 and AP2 estimate the available bandwidth, at the time of the call request, along the path to the destination. We also observe that the average handoff call dropping rate of MPBR remains constant even when the call arrival rate increases. This is attributable to the fact that MPBR makes passive reservation (in advance with good accuracy; see Fig. 9) along the user path to destination before the acceptance

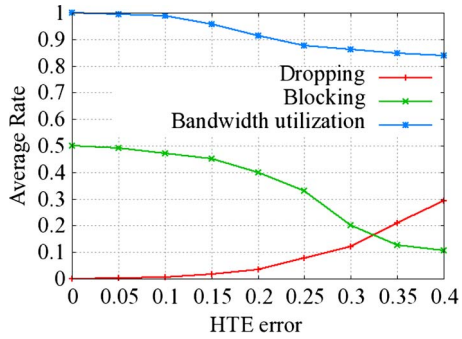


Fig. 12. MPBR performance metrics (Rb, Rd, and Rbw) versus HTE error variation.

of the call. Fig. 11(c) shows that, for the four schemes, the average bandwidth utilization rate increases along with the call arrival rate. The figure also shows that AP1, AP2, and AP3 outperform MPBR. AP3 (slightly more efficient than AP2 and AP1 in this scenario) provides an average of 0.99 per 0.001 call arrival rate, whereas MPBR provides an average of 0.98 per 0.001 call arrival rate; the average relative improvement of AP3 compared with MPBR is about 1% per 0.001 call arrival rate, which is still negligible. We conclude that, compared with AP1, AP2, and AP3, MPBR provides a considerable reduction of 48% per 0.001 call arrival rate in handoff call dropping rate and an increase of 8% per 0.001 call arrival rate in new call blocking rate with similar bandwidth utilization irrespective of call arrival rates. The 8% new call blocking rate increase is a small price to pay for the small handoff call dropping rate.

Fig. 12 shows the average rate of new call blocking, the handoff call dropping, and the bandwidth utilization of MPBR when varying HTE scheme error. In this set of simulations, the call arrival rate λ is set to 0.03 call/s/user, the cell capacity is set to 100 Mb/s, and the number of users in the simulation area is 1500. We observe that, when HTE error exceeds 0.15, the average handoff call dropping rate begins its increase, whereas the average new call blocking and bandwidth utilization rates begin their decrease. This is expected since, when HTE error increases, the number of successful/accepted new call requests (respectively, the amount of allocated bandwidth) increases, and thus, the new call blocking rate (respectively, bandwidth utilization rate) decreases; when HTE error increases, the number of handoff calls accommodated in a next cell decreases, and thus, the handoff call dropping rate increases.

V. Conclusion

In this paper, we have proposed a distributed bandwidth reservation scheme, called MPBR, that ensures QoS to mobile users while maintaining efficient bandwidth utilization. MPBR consists of three schemes: 1) an HTE scheme that aims to estimate the time windows when a user will perform handoffs along his movement path to destination; 2) an available bandwidth estimation scheme that aims to estimate in advance the available bandwidth, during the computed time windows, in the cells to be traversed by the user to destination; and 3) ECaC scheme that aims to control bandwidth allocation in cells to reduce handoff call dropping rate while maintaining acceptable new

call blocking rate. We evaluated, via simulations, MPBR and compared it against two recent related schemes [1], [2] and one closely related scheme [3]. The simulation results did show that MPBR exhibits considerably better handoff call dropping rate at the price of a slightly high new call blocking rate. MPBR also ensures efficient bandwidth utilization, irrespective of cell capacities and call arrival rates.

As future work, we plan to work on a real-life implementation of MPBR in 3GPP networks. More specifically, we envision integrating it into the Access Network Node Discovery Function (ANDSF) node [37]–[39], using a system design similar to that proposed in [40], and envisioning similar additional components to ANDSF. Indeed, ANDSF can be used to 1) collect mobility features from UE, using in turn HTE; 2) predict available bandwidth of mobile backhaul along the predicted trajectory of user using time series as in [41] and ABE proposed in this paper; and 3) then recommend to the UE the most suitable rate to be receiving an IP session at or simply reject the request for the IP session (i.e., enforcing ECaC).

REFERENCES

[1] J. Yao, S. S. Kanhere, and M. Hassan, "Improving QoS in high-speed mobility using bandwidth maps," *IEEE Trans. Mobile Comput.*, vol. 11, no. 4, pp. 603–617, Apr. 2012.

[2] J. S. Wu, S. F. Yang, and C. C. Huang, "Admission control for multiservices traffic in hierarchical mobile IPv6 networks by using fuzzy inference system," *J. Comput. Netw. Commun.*, vol. 2012, pp. 1–10, Jan. 2012.

[3] S. Wee-Seng and S. K. Hyong, "A predictive bandwidth reservation scheme using mobile positioning and road topology information," *IEEE/ACM Trans. Netw.*, vol. 14, no. 5, pp. 1078–1091, Oct. 2006.

[4] A. Vassilya and A. Isik, "Predictive mobile-oriented channel reservation schemes in wireless cellular networks," *Wireless Netw.*, vol. 17, no. 1, pp. 149–166, Jan. 2011.

[5] A. Nadembega, A. Hafid, and T. Taleb, "Handoff time estimation model for vehicular communications," in *Proc. IEEE ICC*, Budapest, Hungary, Jun. 2013, pp. 1715–1719.

[6] L. Mokdad, M. Sene, and A. Boukerche, "Call admission control performance analysis in mobile networks using stochastic well-formed petri nets," *IEEE Trans. Parallel Distrib. Syst.*, vol. 22, no. 8, pp. 1332–1341, Aug. 2011.

[7] K. Keshav and P. Venkataram, "A dynamic bandwidth allocation scheme for interactive multimedia applications over cellular networks," in *Proc. ICN*, St. Maarten, The Netherlands, Jan. 2011, p. 32.

[8] S. Al Khanjari *et al.*, "An adaptive bandwidth borrowing-based call admission control scheme for multi-class service wireless cellular networks," in *Proc. IIT*, Abu Dhabi, United Arab Emirates, Apr. 2011, pp. 375–380.

[9] A. Nadembega, T. Taleb, and A. Hafid, "A destination prediction model based on historical data, contextual knowledge and spatial conceptual maps," in *Proc. IEEE ICC*, Ottawa, ON, Canada, Jun. 2012, pp. 1416–1420.

[10] A. Nadembega, A. Hafid, and T. Taleb, "A path prediction model to support mobile multimedia streaming," in *Proc. IEEE ICC*, Ottawa, ON, Canada, Jun. 2012, pp. 2001–2005.

[11] T. Son Vo, H. Lan Le, and T. Hai Nguyen, "A fuzzy logic call admission control scheme in multi-class traffic cellular mobile networks," in *Proc. Int Symp. Comput. Commun. Control Autom.*, Tainan, Taiwan, May 2010, pp. 330–333.

[12] N. Vallina-Rodriguez and J. Crowcroft, "Energy management techniques in modern mobile handsets," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 1, pp. 179–198, First Quarter 2013.

[13] J. Sorber, A. Balasubramanian, M. D. Corner, J. R. Ennen, and C. Qualls, "Tula: Balancing energy for sensing and communication in a perpetual mobile system," *IEEE Trans. Mobile Comput.*, vol. 12, no. 4, pp. 804–816, Apr. 2013.

[14] C. Shi, V. Lakafosis, M. Ammar, and E. Zegura, "Serendipity: Enabling remote computing among intermittently connected mobile devices," in *Proc. 13th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, Hilton Head Island, SC, USA, Jun. 2012, pp. 145–154.

[15] A. Esmailpour and N. Nasser, "Dynamic QoS-based bandwidth allocation framework for broadband wireless networks," *IEEE Trans. Veh. Technol.*, vol. 60, no. 6, pp. 2690–2700, Jul. 2011.

[16] G. I. Tsiropoulos, D. G. Stratogiannis, J. D. Kanellopoulos, and P. G. Cottis, "Probabilistic framework and performance evaluation for prioritized call admission control in next generation networks," *Comput. Commun.*, vol. 34, no. 9, pp. 1045–1054, Jun. 2011.

[17] J. D. Mallapur, S. Abidhusain, S. Vastrad, and A. Katageri, *Fuzzy Based Bandwidth Management for Wireless Multimedia Networks Information Processing and Management*, vol. 70, V. V. Das *et al.*, Ed. Berlin, Germany: Springer-Verlag, Apr. 2010, pp. 81–90.

[18] M. Ravichandran, P. Sengottuvelan, and A. Shanmugam, "An approach for admission control and bandwidth allocation in mobile multimedia network using fuzzy logic," *Int. J. Recent Trends Eng.*, vol. 1, no. 1, pp. 289–293, May 2009.

[19] S. Kim, "Cellular network bandwidth management scheme by using nash bargaining solution," *IET Commun.*, vol. 5, no. 3, pp. 371–380, Feb. 2011.

[20] K. Madhavi, R. K. Sandhya, and R. P. Chandrasekhar, "Optimal channel allocation algorithm with efficient bandwidth reservation for cellular networks," *Int. J. Comput. Appl.*, vol. 25, no. 5, pp. 40–44, Jul. 2011.

[21] L. Shufeng *et al.*, "Overlap area assisted call admission control scheme for communications system," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 47, no. 4, pp. 2911–2920, Oct. 2011.

[22] K. S. S. Reddy and S. Varadarajan, "Increasing quality of service using swarm intelligence technique through bandwidth reservation scheme in 4G mobile communication systems," in *Proc. Int. Conf. SEISCON*, Chennai, India, Jul. 2011, pp. 616–621.

[23] C.-J. Huang, H.-Y. Shen, and Y.-T. Chuang, "An adaptive bandwidth reservation scheme for 4G cellular networks using flexible 2-tier cell structure," *Expert Syst. Appl.*, vol. 37, no. 9, pp. 6414–6420, Sep. 2010.

[24] S. N. Ahmed and B. Ferri, "Prediction based bandwidth reservation," in *Proc. IEEE Conf. CDC*, Atlanta, GA, USA, Dec. 2010, pp. 5302–5307.

[25] S. Choi and K. G. Shin, "Adaptive bandwidth reservation and admission control in QoS-sensitive cellular networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 13, no. 9, pp. 882–897, Sep. 2002.

[26] K. L. Dias, D. F. H. Sadok, S. F. H. Fernandes, and J. Kelner, "Approaches to resource reservation for migrating real-time sessions in future mobile wireless networks," *Wireless. Netw.*, vol. 16, no. 1, pp. 39–56, Jan. 2010.

[27] Q. Lu and P. Koutsakis, "Adaptive bandwidth reservation and scheduling for efficient wireless telemedicine traffic transmission," *IEEE Trans. Veh. Technol.*, vol. 60, no. 2, pp. 632–643, Feb. 2011.

[28] L. Duan-Shin and H. Yun-Hsiang, "Bandwidth-reservation scheme based on road information for next-generation cellular networks," *IEEE Trans. Veh. Technol.*, vol. 53, no. 1, pp. 243–252, Jan. 2004.

[29] E. Stevens-Navarro, L. Yuxia, and V. W. S. Wong, "An MDP-Based vertical handoff decision algorithm for heterogeneous wireless networks," *IEEE Trans. Veh. Technol.*, vol. 57, no. 2, pp. 1243–1254, Mar. 2008.

[30] Y. Xiaobo, P. Navaratnam, and K. Moessner, "Distributed resource reservation mechanism for IEEE 802.11e-based networks," in *Proc. IEEE VTC*, Ottawa, ON, Canada, Sep. 2010, pp. 1–5.

[31] U. Javed *et al.*, "Predicting handoffs in 3G networks," *SIGOPS Operating Syst. Rev.*, vol. 45, no. 3, pp. 65–70, Dec. 2011.

[32] W. Wanalertlak *et al.*, "Behavior-based mobility prediction for seamless handoffs in mobile wireless networks," *Wireless. Netw.*, vol. 17, no. 3, pp. 645–658, Apr. 2011.

[33] H. Jeung, M. J. Yiu, X. Zhou, and C. S. Jensen, "Path prediction and predictive range querying in road network databases," *VLDB J.*, vol. 19, no. 4, pp. 585–602, Aug. 2010.

[34] K. Playtoni Meetei, G. Kadambi, B. N. Shobha, and A. George, "Design and development of a handoff management system in LTE networks using predictive modeling," *SASTECH*, vol. 8, no. 2, pp. 71–78, Sep. 2009.

[35] J. Markoulidakis, G. L. Lyberopoulos, D. F. Tsirkas, and E. D. Sykas, "Mobility modeling in third-generation mobile telecommunications systems," *IEEE Pers. Commun.*, vol. 4, no. 4, pp. 41–56, Aug. 1997.

[36] R. Baumann, F. Legendre, and P. Sommer, "Generic mobility simulation framework (GMSF)," in *Proc. 1st ACM SIGMOBILE Workshop Mobility Models*, Hong Kong, China, May 2008, pp. 49–56.

[37] "Architecture Enhancements for Non-3GPP Accesses," Sophia-Antipolis Cedex, France, TS 23.402.

[38] "Operator Policies for IP Interface Selection (OPIIS)," Sophia-Antipolis Cedex, France, TR 23.853.

[39] "Data Identification in Access Network Discovery and Selection Function (ANDSF) (DIDA)," Sophia-Antipolis Cedex, France, TR 23.855.

[40] T. Taleb, A. Ksentini, and F. Filali, "Wireless connection steering for vehicles," in *Proc. IEEE GLOBECOM*, Anaheim, CA, USA, Dec. 2012, pp. 56–60.

[41] T. Taleb and A. Ksentini, "QoS/QoE predictions-based admission control for femto communications," in *Proc. IEEE ICC*, Ottawa, ON, Canada, Jun. 2012, pp. 5146–5150.
[42] J. Silvestre-Blanes *et al.*, "Online QoS management for multimedia real-time transmission in industrial networks," *IEEE Trans. Ind. Electron.*, vol. 58, no. 3, pp. 1061–1071, Mar. 2011.
[43] P. McGovern, P. Perru, P. Murphy, and L. Murphy, "Endpoint-based call admission control and resource management for VoWLAN," *IEEE Trans. Mobile Comput.*, vol. 10, no. 5, pp. 684–699, May 2011.
[44] L. Caviglione, "A simple neural framework for bandwidth reservation of VoIP communications in cost-effective devices," *IEEE Trans. Consum. Electron.*, vol. 56, no. 3, pp. 1252–1257, Aug. 2010.
[45] D. Gundlegård and J. M. Karlsson, "Handover location accuracy for travel time estimation in GSM and UMTS," *IET Intell. Transp. Syst.*, vol. 3, no. 1, pp. 87–94, Mar. 2009.
[46] Y. Liang, Y. SiXing, H. Weijun, and L. ShuFang, "Spectrum behavior learning in cognitive radio based on artificial neural network," in *Proc. MILCOM*, Baltimore, MD, USA, Nov. 2011, pp. 25–30.

Apollinaire Nadembega received the B.E. degree in information engineering from Computer Science High School, Bobo-Dioulasso, Burkina Faso, in 2003; the Master's degree in computer science from the Arts and Business Institute, Ouagadougou, Burkina Faso, in 2007; and the Ph.D. degree in mobile networks from the University of Montreal, Montreal, QC, Canada, in 2014. The primary focus of his Ph.D. dissertation was to propose a mobility model and bandwidth reservation scheme that supports quality-of-service management for wireless cellular networks.

From 2004 to 2008, he was a programming engineer with the Burkina Faso Public Administration Staff Management Office. He is currently a member of the Network Research Laboratory, University of Montreal. His research interests include handoff and mobility management, architectural enhancements to mobile core networks, mobile multimedia streaming, call admission control, bandwidth management, and mobile cloud computing.

Abdelhakim Hafid recieved the Master's and Ph.D. degrees from the University of Montreal, Montreal, QC, Canada.

He has spent several years as a Senior Research Scientist with Telcordia Technologies (formerly Bell Communications Research), working on major research projects on the management of next-generation networks, including wireless and optical networks. He was also an Assistant Professor with the University of Western Ontario (UWO), London, ON, Canada; a Research Director of the Advance Communication Engineering Center (venture established by UWO, Bell Canada, and Bay Networks), Canada; a Researcher at CRIM, Canada; a Visiting Scientist with GMD-Fokus, Berlin, Germany; and a Visiting Professor with the University of Evry, Evry, France. He is currently a Full Professor with the University of Montreal, Montreal, QC, Canada, where he founded the Network Research Laboratory in 2005. He is also a Research Fellow with the Interuniversity Research Center on Enterprise Networks, Logistics, and Transportation. He has supervised more than 24 graduate students. He is the author or coauthor of more than 170 journal and conference papers and the holder of three U.S. patents. His research interests include the management of next-generation networks, including wireless and optical networks, quality-of-service management, distributed multimedia systems, and communication protocols.

Tarik Taleb (S'04–M'05–SM'10) received the B.E. degree (with distinction) in information engineering and the M.Sc. and Ph.D. degrees in information science from Tohoku University, Sendai, Japan, in 2001, 2003, and 2005, respectively.

He is a Faculty Staff at the School of Engineering, Aalto University, Espoo, Finland. He has been a Senior Researcher and a Third-Generation Partnership Project Standardization Expert with NEC Europe Ltd., Heidelberg, Germany. He was then leading the NEC Europe Labs Team, working on research and development projects on carrier cloud platforms. Prior to his work at NEC and until March 2009, he was an Assistant Professor with the Graduate School of Information Sciences, Tohoku University. He has been also directly engaged in the development and standardization of the Evolved Packet System as a member of 3GPP's System Architecture working group. His research interests include architectural enhancements to mobile core networks (particularly 3GPP), mobile cloud networking, mobile multimedia streaming, congestion control protocols, handoff and mobility management, intervehicular communications, and social media networking.

Dr. Taleb is an IEEE Communications Society (ComSoc) Distinguished Lecturer. He is a Board Member of the IEEE ComSoc Standardization Program Development Board. He is serving as the Vice Chair of the Wireless Communications Technical Committee, the largest in the IEEE ComSoc. He also served as Secretary and then as Vice Chair of the Satellite and Space Communications Technical Committee of the IEEE ComSoc (2006–2010). As an attempt to bridge the gap between academia and industry, he founded and has been the General Chair of the "IEEE Workshop on Telecommunications Standards: from Research to Standards," which is a successful event that received the "Best Workshop Award" from the IEEE ComSoc. He has been on the Technical Program Committee of different IEEE conferences, including the IEEE Global Communications Conference, the IEEE International Conference on Communications, and the IEEE Wireless Communications and Networking Conference, and he has chaired some of their symposia. He is/was on the Editorial Board of the IEEE WIRELESS COMMUNICATIONS MAGAZINE, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, the IEEE COMMUNICATIONS SURVEYS & TUTORIALS, and a number of Wiley journals. He received the IEEE ComSoc Asia Pacific Best Young Researcher Award in June 2009, the TELECOM System Technology Award from the Telecommunications Advancement Foundation in March 2008, the Funai Foundation Science Promotion Award in April 2007, the 2006 IEEE Computer Society Japan Chapter Young Author Award in December 2006, the Niwa Yasujirou Memorial Award in February 2005, and the Young Researcher's Encouragement Award from the Japan Chapter of the IEEE Vehicular Technology Society in October 2003. Some of his research has received the Best Paper Award at prestigious conferences.