

An Integrated Predictive Mobile-Oriented Bandwidth-Reservation Framework to Support Mobile Multimedia Streaming

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

Abstract—Bandwidth is an extremely valuable and scarce resource in wireless networks. Therefore, efficient bandwidth management is necessary to support service continuity, guarantee acceptable quality of service and ensure steady quality of experience for users of mobile multimedia streaming services. Indeed, the support of uniform streaming rate during the entire course of a streaming service, whereas the user is on the move is a challenging issue. In this paper, we propose a framework, together with schemes, which integrates user mobility prediction models with bandwidth availability prediction models to support the requirements of mobile multimedia services. More specifically, we propose schemes that predict paths to destinations, times when users will enter/exit cells along predicted paths, and available bandwidth in cells along predicted paths. With these predictions, a request for a mobile streaming service is accepted only when there is enough (predicted) available bandwidth, which is along the path to destination, to support the service. Simulation results show that the proposed approach outperforms existing bandwidth management schemes in better supporting mobile multimedia services.

Index Terms—QoS, QoE, mobility prediction, handoff time estimation, available bandwidth estimation, bandwidth reservation, handoff prioritization, admission control, and mobile networks.

I. INTRODUCTION

As wireless services become ever more ubiquitous, there is a growing demand for the provisioning of multimedia services with diverse quality-of-service (QoS) requirements; QoS is defined as the ability of the network to provide a service at an assured service level while QoE is how a user perceives the usability/quality of a service when in use; more specifically, how satisfied the user is with a service in terms of usability, accessibility, retainability and integrity of the service. Mobile calls/users may experience performance degradations due to handoffs (i.e., procedure that allows mobile users to change their points of attachment to the network [1]). Thus, to support QoS from source to destination, the dynamics of every mobile user, such as his path to destination and his entry/exit times

to/from each cell along the path, should be known in advance [2]–[4]. The main limitation of these schemes [2]–[4] is that they do not scale well with the number and frequency of user requests; indeed, they process requests individually.

In this paper, we intend extending our schemes [2]–[5], making them scale with the number of users, by developing (a) an Aggregate Path Prediction Model called APPM; (b) an Aggregate Handoff Times Estimation Scheme called AHTES; and (c) an Integrated Predictive Mobile-oriented Bandwidth Reservation Framework called IPMBRF. APPM estimates paths to destinations for groups of users (not only for a single user) and takes into account (a) road intersections (i.e., junctions of two or more roads) and road segments (i.e., a segment is a road portion between two adjacent road intersections or between a road intersection and a handoff location, i.e., a location at which the user exits/enters a cell); (b) preferences of users in terms of road characteristics (e.g., highway, multi-lane, one-way, without traffic light and without stop sign); (c) spatial conceptual maps; (d) current locations of users; and (e) estimated destinations of users; these destinations are determined by the Destination Prediction Model (DPM) proposed in [2]. AHTES estimates the handoff times for groups of users (not only for a single user), and takes into account (a) traffic flow condition (i.e., free flowing, under-saturated and congested) and (b) current driving behavior in terms of speeds and stopping times; we assume that a road segment has one or two traffic directions and each traffic direction is divided into sub-road segments (i.e., portions of a road segment which has a predefined length). IPMBRF integrates mobility and bandwidth availability prediction models to better support user calls (e.g., multimedia streaming sessions) from source to destination. It consists of two main components, namely User Equipment (UE) (e.g., mobile smart device) and the Controller (CTL) located in the network system (NS). To the best of our knowledge, IPMBRF is the first framework which takes into account traffic flow conditions to predict users' mobility; this allows reducing the amount of exchanged messages between UE and the network (CTL) and thus improves scalability and shortens response time. IPMBRF is also the first distributed and predictive mobile-oriented framework for bandwidth management and call admission control (CAC) that proposes an aggregate scheme to predict paths and handoff times.

In this paper, 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 smart phones on board vehicles, they can always consider charging

Manuscript received December 6, 2013; revised May 4, 2014; accepted June 21, 2014. Date of publication July 15, 2014; date of current version December 8, 2014. The associate editor coordinating the review of this paper and approving it for publication was S. Mao.

A. Nadembega is with the Network Research Laboratory, University of Montreal, Montréal, QC H3C 3J7, Canada (e-mail: nadembea@iro.umontreal.ca).

A. Hafid is with the Network Research Laboratory of the University of Montreal, Montréal, QC H3C 3J7, Canada (e-mail: ahafid@iro.umontreal.ca).

T. Taleb is with the School of Electrical Engineering, Aalto University, FI-00076 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/TWC.2014.2336809

them while being on the move. This is not to mention all recent findings about increasing battery lifetime (e.g., [46]–[48]). 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.

The remainder of this paper is structured as follows. Section II presents some related work. Section III describes the proposed framework along with the envisioned mobile network architecture. Section IV evaluates the performance of the proposed framework and showcases its potential in achieving its design objectives. Section V concludes this paper.

II. RELATED WORK

CAC and bandwidth management schemes can be classified into two categories: (a) non-predictive schemes [6]–[16] (based only on the source cell information) and (b) predictive schemes [17]–[26] (based on the mobility information of mobile users). CAC and bandwidth reservation schemes can be also classified based on (a) the number of cells where call admission is performed (e.g., a single cell [7], [14], [15], [17], [18], [21]–[24], [26]–[28] and two or more cells [6], [19]) and (b) the way handoff requests are handled (e.g., non-prioritized or prioritized handoff [6], [8], [9], [14]–[21], [23]–[29]). According to [1], prioritized handoff schemes which are distributed and predictive are those which better satisfy bandwidth requirements of users from source to destination. Many aggregate CAC and bandwidth reservation schemes have been proposed in the literature [6]–[8], [15], [16], [20]; however, they are not predictive mobile-oriented schemes; indeed, CAC and bandwidth reservation schemes, in mobile networks, that better satisfy bandwidth requirements of users from source to destination are those which are predictive and distributed, and support prioritized handoff [1]. The aggregation proposed in [6]–[8], [15], [16], [20] is used only to estimate available bandwidth. For example, Jun *et al.* [6] proposed a CAC and bandwidth reservation scheme which is cell-oriented, distributed and supports prioritized handoff based on the historical available bandwidth data while Wu *et al.* [15] proposed a CAC and bandwidth reservation scheme based on the load and the ratio of high speed users in the next cell as input variables of fuzzy inference system (FIS).

In the following, we briefly overview representative schemes [17]–[19] that are closely related to our proposed approach (i.e., predictive mobile-oriented schemes). In these three schemes [17]–[19], a handoff call is admitted if there is enough available bandwidth in the next cell; otherwise, it is dropped. Dias *et al.* [17] present a scheme for CAC and bandwidth reservation that avoids per-user reservation in order to meet scalability requirements. More specifically, based on GPS data traces (movements of users), they determine the next cell likely to be visited by a mobile user. Then, based on observations of changes of available bandwidth due to users' mobility, Dias *et al.* predict

available bandwidth of next cell. This scheme [17] suffers from three key limitations: (a) aggregation is used only to estimate available bandwidth which negatively impacts the scalability of the scheme; (b) available bandwidth estimation is not predictive (i.e., based on the behavior/state of network cells [1], [30]); in addition, the use of historical data of load, exchanged between neighboring cells, does not provide an accurate estimation of available bandwidth compared to schemes which use users' behaviors (e.g., handoff times estimation) as reported in [7], [18]–[20], [24]; and (c) the new call is accepted only when available bandwidth minus the virtual bandwidth reservation is enough in the source cell; in this case, the accepted call may be dropped in subsequent cells (if one is congested) to destination; this negatively impacts the handoff call dropping rate of the scheme. Rashad *et al.* [18] analyze previous movements of mobile users in order to generate mobility profiles; the profiles are based on transit times of the cells. More specifically, based on users' mobility history, the authors generate global mobility profile for a user which contains a set of 3-tuple $\langle (c_{i-1}, t_{i-1}); (c_i, t_i); (c_{i+1}, t_{i+1}) \rangle$ where t_i represents transit time of cell c_i (current cell, i.e., the cell where the user is located at the moment of prediction), t_{i-1} represents transit time of cell c_{i-1} (previous cell) and t_{i+1} represents transit time of cell c_{i+1} (next cell). Then, they compute probability $p()$ of each possible 3-tuple, taking into account the time-of-day. They select the 3-tuple with the largest value of $p()$ and define c_{i+1} as next cell to be visited and t_{i+1} as the estimated transit time of cell c_{i+1} . Making use of this prediction, Rashad *et al.* perform bandwidth reservation for handoff calls at next cell. A new call request is accepted only if the remaining/available bandwidth after the reservation, defined as $[BC_{c_{i+1}} - (BE_{c_{i+1}} + BR_{c_{i+1}})]$ where $BC_{c_{i+1}}$ (resp. $BE_{c_{i+1}}/BR_{c_{i+1}}$) is the bandwidth capacity of next cell c_{i+1} (resp. the estimated bandwidth used by users in next cell c_{i+1} /the bandwidth reserved by next cell c_{i+1} for handoff calls) is enough to accommodate the call. A strong point of this work is the fact that Rashad *et al.* perform available bandwidth estimation based on the aggregate behavior/profile of mobile users; however, a new call is accepted based only on the amount of available bandwidth of source/current cell. Wu *et al.* [19] propose a prediction system (based on the aggregate behavior/profile of mobile users), which predicts bandwidth utilization and call dropping probability in advance, and a distributed CAC scheme. More specifically, based on historical data of users' traces, Wu *et al.* determine the periodicity of patterns and handoff times of users; the users with the same periodicity of patterns are grouped in the same mobility profile. Then, they make use of periodic patterns to predict possible patterns and handoff times in future for the users with the same mobility profile. Indeed, if a segment of a repetitive pattern matches with the inputs (e.g., current pattern), it may happen that the following segment of that specific repetitive pattern has the possibility of reoccurrence; however, in case of several possible results, the authors select the prediction result with the highest probability. Thus, based on these prediction results and the bandwidth consumption of adjacent cells, the scheme [19] is able to decide to admit or not a new call. Also, the authors propose a throttle flag that can indicate the usage of current cell to prevent the newly admitted call request from being blocked

in next cell if handoff is needed. The main limitation of this scheme is the fact that the new call is accepted based only on the amount of available bandwidth of the source and next (adjacent) cells; subsequent cells in the path to destination are not considered; thus, even if a new call is accepted, it may be dropped in subsequent cells (if one is congested) to destination.

We can summarize the limitations of existing CAC and bandwidth reservation schemes in mobile networks as follows: (a) They rely on the current behavior/state of the network cells [6], [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 user/call is accepted to the time of his entry into cells towards destination; (b) The schemes that make use of mobility prediction techniques either do not take into account users' aggregation [14], [17], [20], [32]–[38], require additional equipment [14], [27], generate significant traffic overhead in terms of mobility data exchanges between users and network backbone [39], do not consider stopping times [14], [24], [27], [31], [39]–[41], make use of old road traffic data [24], [41] or rely only on historical data about previous users [24]; (c) admission control procedures are limited to source cell and possibly also next cell [14], [15], [17]–[19], [24], [27]; and/or (d) they rely only on historical network bandwidth observations or users transit times in cells [6], [16]–[18]. In this paper, we propose a scheme, to process call requests, that proposes solutions to these limitations.

III. INTEGRATED PREDICTIVE MOBILE-ORIENTED BANDWIDTH RESERVATION FRAMEWORK

The objective of the proposed IPMBRF is to satisfy the requirements, in terms of bandwidth, of each mobile user along his movement path across cells towards destination. For this purpose, the framework predicts (a) the mobile user path to destination; (b) the entry/exit times of the mobile user to/from cells along the path to destination; and (c) the available bandwidth in each cell that will be transited by the user to destination. It then accepts the user request, if there is sufficient available bandwidth along the path to accommodate the request; otherwise, it rejects the user request. In the following section, we present the architecture of IPMBRF and the IPMBRF operations in processing new call requests. Then, after stating our assumptions, we present the aggregate path prediction model (APPM) and the aggregate handoff times estimation scheme (AHTEs).

A. Architecture

Fig. 1 shows a network configuration that consists of two parts: wired and wireless networks which are inter-connected via gateways. The wired network represents, generally, Internet that connects a number of multimedia servers. The wireless network operator administrates a new entity, called Controller, that performs bandwidth management and call admission control. A number of multimedia calls data collectors are deployed over the entire network to collect data about calls and forward them to the Controller for processing. The backbone (see Fig. 1) allows cell towers and gateways to be inter-connected.

Fig. 2 shows the architecture of IPMBRF whose operation is performed by User Equipment (UE) and the Controller (CTL)

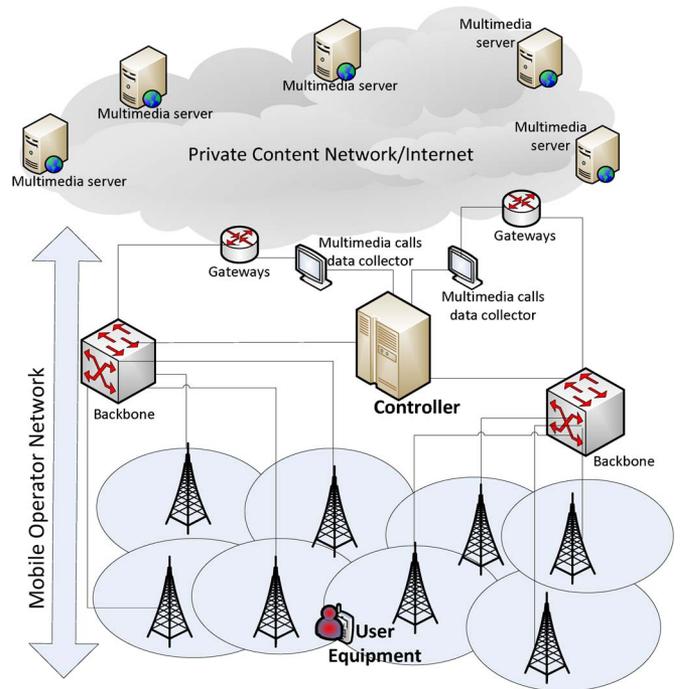


Fig. 1. Envisioned.

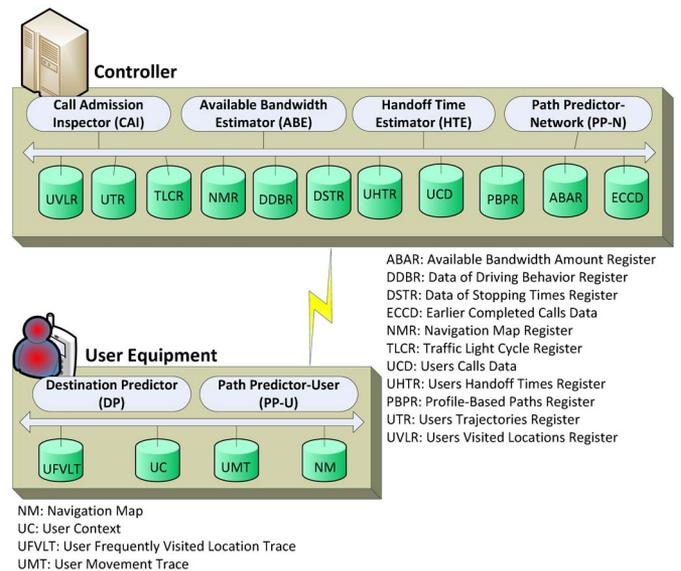


Fig. 2. IPMBRF Architecture.

which is located in the network system (NS). UE is responsible for predicting the mobile user path to destination when the navigation zone is lightly dense while CTL is responsible for predicting the path of the group of users when the navigation zone is highly dense. CTL is also responsible for predicting the entry/exit times of the mobile user to/from cells along the path to destination and available bandwidth in each cell along the path to destination. In the rest of the paper, current user and the person who uses the UE are used interchangeably.

UE consists of two modules, namely Destination Predictor (DP) and Path Predictor-User side (PP-U). DP (resp. PP-U) predicts the user's destination (resp. the user's path to the predicted destination). Fig. 2 shows CTL that consists of four main modules namely Path Predictor-Network side (PP-N), Handoff

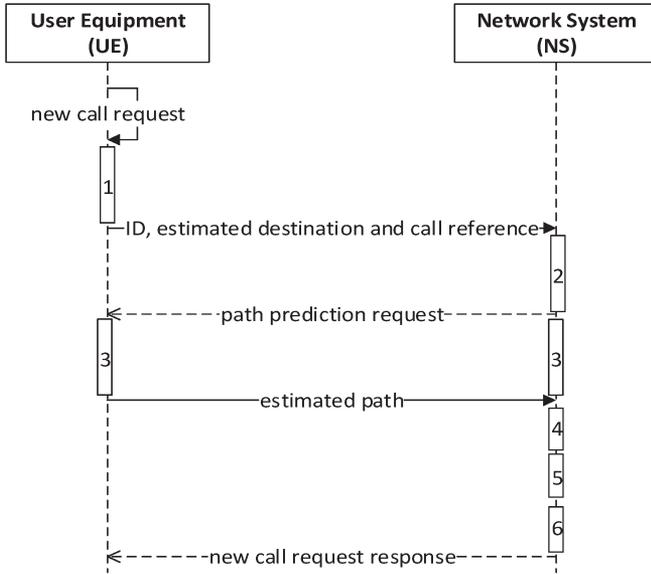


Fig. 3. IPMBRF process for new call acceptance/rejection.

Time Estimator (HTE), Available Bandwidth Estimator (ABE) and Call Admission Inspector (CAI). PP-N predicts the path of a group of mobile users from their source (e.g., source DSZ) to their destination (e.g., destination DSZ); HTE predicts the entry/exit times of a group of mobile users to/from cells along their path to destination; ABE predicts the available bandwidth in each cell along the path to destination; CAI makes decision on accepting or rejecting new calls. For reason of space, we do not describe in detail the databases which are maintained by UE and CTL; however, the descriptions may be found in our previous contributions [2]–[5]. For reason of space, we cannot describe in details how IPMBRF will be applied to nowadays technologies; however, these details can be found in our previous paper [32].

B. New Call Request Acceptance/Rejection Process

The role of IPMBRF is to decide on whether to accept or reject a new call request based on predicted available bandwidth in each cell along the path to destination. Fig. 3 illustrates the process of new call acceptance/rejection that consists of (1) *destination prediction*: indeed, upon receipt of a call request, IPMBRF determines the current user’s destination using DPM [2]; this operation is performed by destination prediction module (DP) of UE.

Then, UE creates a message, that contains the predicted destination, the reference of current user’s new call request (i.e., call ID or name and multimedia server where the call is located which allows identifying the multimedia application in order to get the required bandwidth and call time) and his ID, and sends it to CTL located in NS; (2) *navigation zone density estimation*: IPMBRF uses the user’s destination (forwarded by his UE) and information about users’ locations, to compute the density of the current user navigation zone (i.e., $E_{destination_DSZ}^{current_location}$) using (1); this operation is performed by CTL; (3) *path prediction according to the navigation zone density*: if the navigation zone is lightly dense, CTL sends a message to UE that determines, by using PP-U (implements PPM [3]), the predicted path and sends

TABLE I
SUMMARY OF NOTATIONS

Symbol	Description
D	Boundary density value of subzone
R_{H_i}	Preference rate of road segments characterized by H_i
R_{I_j}	Preference rate of road intersection characterized by I_j
r_H	Preference threshold of road segment
r_I	Preference threshold of road intersection
V_{H_i}	Preference of road segment characterized by H_i
V_{I_j}	Preference of road intersection characterized by I_j
$r(\theta_s)$	Deviation rate of segment S
Ψ	Pre-selected adjacent road segments
N_a	Cardinality of Ψ
L_R	Length of sub-road segment R
L	Length of road segment
E_B^A	Selection area from A towards B
$PBP(P_i, x, y)$	Profile-based path from $DSZ x$ to $DSZ y$ of road profile P_i
$p(s, P_i)$	Probability that adjacent road segment s is the next road segment towards destination DSZ , given road profile P_i
$D(Z, t)$	Density of subzone Z at time t
L_l	Length of portion l of sub-road segment already transited by a user
$n(t)$	Number of users on road segment at time instant t
$d(t)$	Density of road segment at time instant t
d_l	Lower boundary density value of road segment
d_u	Upper boundary density value of road segment
$u\Delta T^R$	Transit time on sub-road segment R by user u who has already transited R
ΔT_R^u	Transit time that will be required to transit R by user u who is currently on sub-road segment R
$u\Delta S^J$	Transit time of road connection zone J by user u who has already transited J
$F_{SiGkSa}()$	CDF of transit/travel times of link $SiGkSa$
ΔT_{SiGkSa}	Transit time of link $SiGkSa$

it back to CTL; otherwise, CTL determines, by using PP-N (implements APPM; see Section B) the predicted path. The predicted path is stored in database UTR of CTL with type t_y equal to “predicted”; (4) *handoff times’ estimation*: With the knowledge of the predicted path, IPMBRF determines the user’s handoff times using HTEMOD [4] (resp. AHTES, see Section C) when the navigation zone is lightly (resp. not lightly) dense; this operation is performed by handoff time estimation module (HTE of CTL); HTE output is stored in UHTR; (5) *available bandwidth estimation*: CTL (via ABE) determines available bandwidth in each cell along the user’s path to destination; the results of ABE are stored in ABAR; and (6) *new call admission control*: CTL (via CAI) checks whether there is sufficient available bandwidth along the path to accommodate the user’s new call; if the response is no, it rejects the user’s new call.

In the following sections, after stating our assumptions and definitions, we present the aggregate path prediction model (APPM) and the aggregate handoff times estimation scheme (AHTES); for better understanding, Table I shows the list of symbols/variables that are used to describe APPM and AHTES.

C. Assumptions

We assume that the road topology consists of several roads and intersections while the mobile network topology consists of several cells and handoff points. We refer to the intersection of a road and the border of a cell as a handoff point. We also

refer to a location frequently visited by a user (e.g., home, school, shop, and mall) as a frequently visited location (FVL). We assume that a road intersection, a FVL or a handoff point is represented by a node, and identify each node using its geographic coordinates (i.e., latitude and longitude). We refer to the road between two nodes a and b as road segment identified by (a, b) where the navigation direction a towards b ($a \rightarrow b$) is different from the navigation direction b towards a ($b \rightarrow a$). We also assume that a spatial conceptual map consists of several subzones; we compute the density of a subzone in terms of the number of users in that specific subzone. The expression of density of subzone Z at time t is defined as follows:

$$D(Z, t) = \frac{Num_u(Z, t)}{A_z} \quad (1)$$

where $Num_u(Z, t)$ denotes the number of users in Z at time t and A_z is the size of Z . We define D as the boundary density between (i) *lightly dense zones* (i.e., $D(Z, t) < D$) and (ii) *highly dense zones* (i.e., $D(Z, t) \geq D$). Throughout this paper, we refer to a highly dense zone as a dense subzone (DSZ) and a lightly dense zone as a non-dense subzone (non-DSZ).

D. Definitions

In this section, we present the definitions of concepts and terms we use to describe APPM and AHTES.

1) *Concepts and Terms—APPM: Road segment/intersection characteristic*: a road segment/intersection characteristic defines the type of a road segment/intersection; for example, highway, multi-lane, one-lane, one-way, two-way (resp. without/with traffic light, without/with stop sign) are characteristics of road segments (resp. intersections).

a) *Road segment/intersection preference*: A road segment/intersection preference is the frequent use of a road segment/intersection with a specific characteristic; for example, the frequent use of highway instead of two-way road is a road segment preference. We assign Boolean value (true or false) to a road segment/intersection preference.

b) *Road profile*: A road profile is the combination of road segment and intersection preferences of a user; for example, a user who prefers highways and one-way roads, and does not prefer multi-lane roads, roads with traffic lights and roads with stop signs has the following road profile: yes for highway; no for multi-lane; yes for one-way; no for traffic light; no for stop sign.

c) *Profile-based-path (PBP)*: A profile-based-path is a path which is determined according to a specific road profile.

d) *Selection area*: Let A and B be road intersections or DSZs; a selection area from A towards B, denoted by E_B^A , is the rectangular zone whose straight line from A to B is the diagonal of the rectangle; a selection area consists of several road segments and road intersections.

e) *Navigation Map Register (NMR)*: NMR is a database; it contains node ID, node type/characteristic (e.g., without traffic light, without stop sign), road segment ID, road segment length and road segment type/characteristic (e.g., highway, multi-lane, one-way) that represent static data about geographic areas (road and mobile network topologies); NMR is updated only when changes happen in road topology.

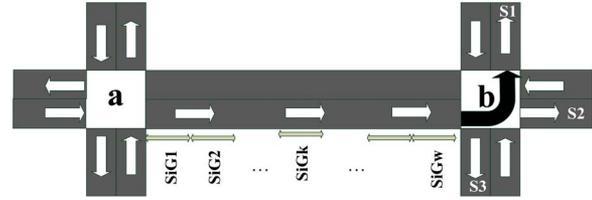


Fig. 4. Illustration of sub-road segments.

f) *Users Visited Location Register (UVLR)*: UVLR is a database of the users' current locations. A record in UVLR contains user ID, current cell ID and current road segment ID; UVLR is constantly updated according to users' movements; indeed, UVLR maintains only one record per user.

g) *Users Trajectories Register (UTR)*: UTR is a database; a record in UTR contains user ID, type ty , time t , date d and road segment or intersection ID that represents the location of the user on date d and time t ; the type ty assumes the value "predicted" if predicted trajectory and "real" if real/effective trajectory; to limit the size of UTR, predicted trajectories are deleted when users reach destination while real trajectories are deleted at the beginning of the day; indeed, PBPs are computed at the beginning of the day (e.g., midnight) based on the recorded real trajectories of the previous day (i.e., real trajectories of UTR); after the computation, the recorded real trajectories of UTR are deleted and the scheme starts a new real trajectories collection.

2) *Concepts and Terms—AHTES: Sub-road segment*: a sub-road segment is a portion of a road segment which has a predefined length q ; for example; a road segment of length l is divided in w (i.e., smallest integer value bigger than l/q) sub-road segments. In future work, we plan to propose a scheme that defines the best value, w , of sub-segments of a segment according to the average distance between cars; it is clear that in the case of congestion on the road segment (i.e., when average distance between cars is smaller than the safety distance of the navigation area), the impact of the value of w is negligible; this is due to the fact that the users have the same velocity on the road segment. We assume that each road segment S_i (e.g., $a \rightarrow b$ shown in Fig. 4) is divided into sub-road segments where $SiGk$ is the k th sub-road segment (see Fig. 4);

a) *Road segment portion*: A road segment portion is the road between two sub-road segments of the same road segment; for example, in Fig. 4, we refer to the road between sub-road segments $SiGk$ and $SiGw$ as the road segment portion $SiGkGw$.

b) *Road connection zone*: A road connection zone is the road between the last sub-road segment $SiGw$ and an adjacent road segment of the road segment of interest; for example, in Fig. 4, we refer to the road between the sub-road segment $SiGw$ and the road segment $S1$ as the road connection zone $SiS1$ (black arrow in Fig. 4).

c) *Link*: A link is the combination of a road segment portion and a road connection zone of an adjacent road segment to that specific road segment portion; for example, in Fig. 4, we refer to the road between sub-road segment $SiGk$ and adjacent road segment $S1$ as the link $SiGkS1$ (i.e., road segment portion $SiGkGw$ + road connection zone $SiS1$).

d) *Path portion*: A path portion is the road between two arbitrary sub-road segments located in two different road segments; a path portion consists of one or several links and one road segment portion; the road segment portion follows immediately the last link of the path portion; for example, in Fig. 4 the path portion that represents road between *SiGk* and *S1G4* consists of link *SiGkS1* and road segment portion *S1G1G4*.

e) *Data of Stopping Times Register (DSTR)*: DSTR is a database; whenever a user *UID* experiences a stop at time *t*, during a time period *t_s* at road intersection *ID*, the 4-tuple (*UID, t, t_s, ID*) is stored in the database DSTR; each entry/record in DSTR is deleted when the user reaches his trip destination (a FVL).

f) *Data of Driving Behaviors Register (DDBR)*: DDBR is a database; whenever user *UID* transits sub-road segment *SID* with acceleration *a* and velocity *v* at time *t*, the 5-tuple (*UID, t, a, v, SID*) is stored in DDBR; each entry/record in DDBR is deleted when the user reaches his trip destination (a FVL).

E. Aggregate Path Prediction Model

In this section, we present APPM that aims at determining paths, named profile-based paths (PBP), between two DSZs using spatial conceptual maps and road profiles; each PBP between two DSZs is determined according to a specific road profile; 2ⁿ PBPs are defined between two DSZs where *n* is the number of road profiles. Given the combination of preference in terms of road characteristics *CP* (i.e., road profile), we can estimate the proportion of users *X_i* with road profile *CP* that could select a specific road *i* making use of users' mobility history. Then, making use of *X_i*, we compute the probability of each possible road *i* and select the road *i* with the largest value of probability as the next road; this allows performing an aggregate prediction for all users who have road profile *CP*. It is worth noting that APPM is used only in highly dense navigation zones; in lightly dense navigation zones, we perform the prediction of path, per user, to destination using PPM as in the authors' previous work [3].

1) *Generation of User's Road Profile*: To generate users' road profiles, we first identify all possible road profiles according to a predefined number of road characteristics; it is worth noting that a road characteristic can be a road segment characteristic or road intersection characteristic. We use the data about road characteristics (maintained in NMR) to generate all possible road profiles that can be used to compute user's road profile and predict their movements. For example, based on the five road characteristics, namely highway, multi-lane, one-way, without traffic light and without stop sign, we obtain thirty two possible road profiles; the number of road profiles is equal to 2ⁿ where *n* denotes the number of road characteristics.

More specifically, based on databases UTR and NMR, we compute the ratio of utilization of each road characteristic. Whenever the user transits road segment *ID* that has characteristic *Hi* and length *l*, the triplet (*ID, Hi, l*) is stored in a list *L₁*. Making use of list *L₁*, we compute the pair (*Hi, d_{Hi}*) and store it in list *L₂*; *d_{Hi}* represents the total length of road segments

which are already transited by the user and have characteristic *Hi*. *L₂* can be easily computed using the following SQL query:

Select *Hi*, SUM(*l*) as *d_{Hi}* from *L₁* group by *Hi*;

The expression of the preference rate of road segments, characterized by *Hi* (e.g., highway, multi-lane or one-way), transited by the user is defined as follows:

$$R_{Hi} = \frac{d_{Hi}}{\sum_{i=1}^k d_{Hi}} \quad (2)$$

where *k* is the number of road segment characteristics.

Whenever the user transits road intersection *ID* that has characteristic *Ij*, the pair (*ID, Ij*) is stored in list *L₃*. Making use of list *L₃*, we compute the pair (*Ij, n_{Ij}*) and store it in list *L₄*; *n_{Ij}* represents the total number of times the user has transited road intersections characterized by *Ij*. *L₄* can be easily computed using the following SQL query:

Select *Ij*, COUNT(*) as *n_{Ij}* from *L₃* group by *Ij*;

The expression of the preference rate of road intersection, characterized by *Ij* (e.g., without traffic light or without stop sign), transited by the user is defined as follows:

$$R_{Ij} = \frac{n_{Ij}}{\sum_{i=1}^m n_{Ii}} \quad (3)$$

where *m* denotes the number of road intersection characteristics.

To determine the preference of the user in term of road segment (resp. intersection) characteristics, we define *r_H* (resp. *r_I*) as preference threshold of road segment (resp. intersection). The expression of *r_H* (resp. *r_I*) is derived from the number of road segment (resp. intersection) characteristics *k* (resp. *m*) and is defined as follows:

$$r_H = \frac{1}{k} \left(\text{resp. } r_I = \frac{1}{m} \right) \quad (4)$$

When *R_{Hi}* ≥ *r_H* (resp. *R_{Ij}* ≥ *r_I*), we assume that the user prefers road segments (resp. intersections) characterized by *Hi* (resp. *Ij*); thus, the (Boolean) expression of the preference of road segment (resp. intersection), characterized by *Hi* (resp. *Ij*) *V_{Hi}* (resp. *V_{Ij}*) is defined as follows:

$$V_{Hi} = R_{Hi} \geq r_H \quad (\text{resp. } V_{Ij} = R_{Ij} \geq r_I) \quad (5)$$

To obtain road profile *P* of the user, we combine the values *V_{Hi}* and *V_{Ij}* as follows:

$$P = \bigwedge_{i=1}^k V_{Hi} \bigwedge_{j=1}^m V_{Ij}. \quad (6)$$

2) *Generation of Profile-Based Path and Users' Path Prediction*: For each road profile *Pi*, APPM generates a profile based path, *PBP(Pi, x, y)*, from source DSZ *x* to destination DSZ *y* that best meets road profile *Pi*. More specifically, the operation of APPM consists of choosing a road segment (among one or more segments) at each road intersection towards the destination DSZ; notice that the selected road segment must be located in the current selection area $E_{\text{destination_DSZ}}^{\text{current_location}}$. The selection

process starts from the source DSZ and is repeated until the destination DSZ is reached. At each road intersection, APPM starts by a pre-selection process. The pre-selection process starts by identifying $E_{destination_DSZ}^{current_location}$. It then selects a set of road segments, among the adjacent road segments to the current road intersection, which are located in $E_{destination_DSZ}^{current_location}$. More specifically, the pre-selection process is performed making use of deviation function $r()$ to compute deviation rate $r(\theta_s)$ of each adjacent road segment s to the diagonal of $E_{destination_DSZ}^{current_location}$.

$$r : [0, 180] \rightarrow [0, 1]$$

$$r(\theta) = 1 - \frac{\theta}{180}$$

The pre-selected adjacent road segments are those that belong to the following set:

$$\Psi = \bigcup_s \left\{ s \mid r(\theta_s) \geq r(45^\circ) = \frac{3}{4} \right\} \quad (7)$$

The selection of an adjacent road segment, from within Ψ , as a next segment is performed using road profile P_i . Indeed, APPM computes the probability that an adjacent road segment s is the next road segment towards destination DSZ, given the road profile P_i . This probability is expressed as follows:

$$p(s, P_i) = \frac{Num(s, P_i)}{\sum_{a=1}^{N_a} Num(a, P_i)} \quad (8)$$

where $Num(s, P_i)$ is the number of times the transition from current intersection to road segment s is performed, in the past, by users, with road profile P_i , and N_a is the cardinality of Ψ . $Num(s, P_i)$ can be obtained using users' movement history. Indeed, whenever a user, with a road profile P_i , transits road segment ID, the pair (P_i, ID) is stored in the list L_5 (i.e., computed using database UTR). Making use of L_5 , we compute the pair (P_i, n_{ID}) and store it in the list L_6 . n_{ID} represents the total number of times that users, with road profile P_i , transit road segment ID. L_6 can be easily computed using the following SQL query:

```
Select  $P_i$ ,
COUNT(*) as  $Num(ID, P_i)$  from  $L_5$  group by  $ID, P_i$ ;
```

APPM chooses the adjacent road segment $s = (a \rightarrow b)$ with the largest value of $p(s, P_i)$ as a next road segment; the selected road segment is added to the list L of previous selected road segments. The selection process is repeated, making use of road intersection b as current road intersection (i.e., source of selection process) until destination DSZ is reached; when destination DSZ y is reached, the list becomes $PBP(P_i, x, y)$; i.e., the profile-based path from source DSZ x to destination DSZ y of the users who have road profile P_i . Making use of similar operations, we compute PBPs of road profile P_i between DSZs; the number of PBPs of road profile P_i is equal to $n(n-1)$ where n denotes the number of DSZs. The table of PBPs of road profile P_i is updated when DSZs' locations change. Indeed, when densities of subzones change, new DSZs may appear

while old DSZs may disappear (i.e., DSZ to non-DSZ or non-DSZ to DSZ); densities of subzones are computed periodically (e.g., morning, noon, afternoon, evening and night) to identify DSZs. APPM computes a table of profile-based paths for each road profile P_i ; each PBP, in such tables, is the predicted path for the set of users that share the same road profile. Upon receipt of a call request, with road profile P_i and located in DSZ x , APPM computes its predicted path by selecting $PBP(P_i, x, y)$ in the corresponding table of profile-based paths.

APPM calculates $g = N_a$ transition probabilities, at each road intersection, per road profile; thus, adding $O(g)$ complexity at each road intersection and $O(g^2)$ from source to destination; for e road profiles and n DSZs, the operation of APPM adds $O(n * e * g^2)$ complexity. The complexity of selection operation in a table of profile-based path is constant $O(1)$. Thus, the complexity of user path prediction is $O(n * e * g^2)$.

F. Aggregate Handoff Time Estimation Scheme

In this section, we present AHTES that aims at determining the times when a group of users would perform handoffs along their movement path to their destination using transit time tables of road segments of the spatial conceptual maps; the transit time table of a road segment is computed based on the driving behavior (i.e., velocity and stopping time at the road intersection) of previous users on the road segment; indeed, we divide the road segment into sub-road segments of predefined length q ; then, based on velocity and stopping time of previous users on the road segment, we derive the probability distributions (PDF) and cumulative distribution function (CDF) of transit times on each link (i.e., path portion from a sub-road segment to an adjacent road segment to the road segment of interest) of the road segment; finally, we compute the median of CDF of transit times on the link as the value of transit time on the link; i.e., the transit time to reach the adjacent road segment from the sub-road segment; this allows performing an aggregate transit time computation for all users who will transit on each link of this road segment.

1) Traffic Flow and Queuing Models:

a) *Traffic model*: In traffic flow theory, it is common to model vehicular flow and represent it with macroscopic variables of flow $f(t)$ (veh/s), density $d(t)$ (veh/m) and velocity $v(t)$ (m/s). Indeed, flow is defined as follows:

$$f(t) = d(t) \times v(t) \quad (9)$$

We make the assumption that the state of traffic flow is fully characterized by density d ; the expression of $d(t)$ is defined as follows:

$$d(t) = \frac{n(t)}{L} \quad (10)$$

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

b) *Multi-lane 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 implementation presented in this paper, we consider that all lanes have different queue lengths and model the different phases of traffic signals.

c) Model for differences in driving behavior: In this paper, driving behavior is based on the velocity model proposed in [42]; indeed, driving behavior is a cycle of acceleration, drive at a constant velocity, deceleration and finally stopping.

d) Stationarity of traffic: During each estimation interval, the parameters of traffic light cycles are constant; i.e., the time duration of color i is denoted q_i and the overall cycle time is denoted C . In case of absence of traffic lights, we apply the policy “first come, first serve”.

e) Road segment traffic dynamics: We define three discrete traffic conditions: free flowing, under-saturated and congested; they represent different dynamics of the road segments depending on the length of the queues at intersection. To determine these traffic conditions, we define d_1 and d_2 as boundary density values between (i) *free flowing conditions* $d(t) \leq d_1$; (ii) *under-saturated conditions* ($d_1 < d(t) < d_2$); and (iii) *congested conditions* $d(t) \geq d_2$.

The expression of transit time on sub-road segment R by user u (who has already transited R) is derived from his entry time uT_a^R to R and his exit time uT_e^R from R and is given by:

$$u\Delta T^R = uT_e^R - uT_a^R \quad (11)$$

However, for user u who is currently on sub-road segment R , we define the expression of transit time that will be required to transit R as follows:

$$\Delta T_R^u = L_R \times \frac{u\Delta T^l}{L_l} \quad (12)$$

where L_R is the length of the sub-road segment R , L_l is the length of portion l of R that is already transited by user u and $u\Delta T^l$ is the transit time on l by user u ; $u\Delta T^l$ is computed using (11). We also define the expression of transit time of road connection zone J by user u (who has already transited J) as follows:

$$u\Delta S^J = uS_e^J - uS_a^J \quad (13)$$

where uS_a^J is the entry time of user u to J and uS_e^J is his exit time from J . Throughout the remainder of this paper, current road segment and road segment of interest (i.e., road segment for which we compute the transit time table) are used interchangeably.

2) *Probability Distribution Function of Transit Times of Road Segment and Estimation of Handoff Times:* To estimate handoff times for a given user from source to destination, AHTES builds travel/transit time tables of road segments in the spatial conceptual maps. It is worth noting that AHTES is used only in under-saturated and congested conditions; in free flowing condition, the navigation zone is lightly dense and we perform the prediction/estimation of the entry/exist times to/from cells using HTEMOD [4].

To create or update travel/transit time tables, AHTES requires information about users driving behavior and the density of current road segment (e.g., average number of users on current road segment); the transit time table of a road segment is updated when the traffic flow condition of that specific road segment has changed (i.e., under-saturated to congested or congested to under-saturated); the number of times that the transit time table of a road segment is updated per day depends on its traffic condition dynamics. AHTES computes the density of current road segment, making use of (10). The final output of AHTES, for a given user request, is an n -tuple: $\Omega = \langle (t_1, c_1), \dots, (t_i, c_i), \dots, (t_n, c_n) \rangle$ where t_i is the value of the estimated time when the current user will reach cell C_i and C_1, \dots, C_n represent the cells the current user is predicted to cross towards destination. In the rest of the Section, *SiGw* represents the last sub-road segment (i.e., the sub-road segment immediately followed by the road connection zone to adjacent road segments) of any road segment Si ; the time unit of transit times is the minute and value of transit time is an integer; this helps regrouping users who have same transit time on a link.

We first propose estimating Probability Distribution Function (PDF) of the portion of current road segment *SiGkGw* transit times by users; it is worth noting that *SiGkGw* is the subsequent sub-road segments from *SiGk* to *SiGw*; thus, to estimate PDF of *SiGkGw* transit times, we need to estimate PDF of transit time of each sub-road segment *SiGk*. The probability population consists of the times to transit *SiGk* by (a) users who have already transited *SiGk* during the last 30 minutes; these times are computed based on (11), and (b) users who are currently on *SiGk*; these times are computed based on (12). Let n_{SiGk} be this population and $n_{SiGk}^{\Delta T_u}$ be the fraction of n_{SiGk} who transit *SiGk* within ΔT_u . Along *SiGk*, the transit time $v\Delta T$ is a random variable with distribution v . We derive the probability distribution v_{SiGk} of transit times of *SiGk* as follows:

$$v_{SiGk}(\Delta T_u) = \frac{n_{SiGk}^{\Delta T_u}}{n_{SiGk}} \quad (14)$$

Now, let us define PDF of road connection zone *SiSa* transit times by users. The probability population consists of the times to transit road connection zone *SiSa* by users who have already transited *SiSa* during the last 30 minutes; these times are computed based on (13). Let n_{SiSa} be this population and $n_{SiSa}^{\Delta S_u}$ be the fraction of n_{SiSa} who transit *SiSa* with ΔS_u as transit time. Along *SiSa*, the transit time $w\Delta T$ is a random variable with distribution w . We derive the probability distribution w_{SiSa} of transit times of *SiSa* as follows:

$$w_{SiSa}(\Delta S_u) = \frac{n_{SiSa}^{\Delta S_u}}{n_{SiSa}} \quad (15)$$

To derive PDF, ρ_{SiGkSa} , of link *SiGkSa* (i.e., combination of road segment portion *SiGkGw* and road connection zone *SiSa*) transit times by users, we use the following rule: If X and Y are two independent random variables with respective PDF f_X and f_Y , then PDF f_Z of the random variable $Z = X + Y$

TABLE II
TRANSIT TIME TABLE OF ROAD SEGMENT S_i

Sub-road segment	Adjacent road segment	Transit time of the link
SiG1	S1	Δt_{SiG1S1}
	S2	Δt_{SiG1S2}
	S3	Δt_{SiG1S3}
...
SiGw	S1	Δt_{SiGwS1}
	S2	Δt_{SiGwS2}
	S3	Δt_{SiGwS3}

is given by the convolution product of f_X and f_Y , $f_Z(Z) = (f_X * f_Y)(Z)$ defined as follows:

$$f_Z(Z) = \int_R f_X(t) f_Y(Z - t) dt$$

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 ρ_{SiGkSa} is given as follows:

$$\rho_{SiGkSa}(\Delta T) = (((v_{SiGk} * v_{SiGk+1}) * v_{SiGk+2}) * \dots * v_{SiGw}) * w_{SiSa}(\Delta T) \quad (16)$$

Using (16), we derive the cumulative distribution function, F_{SiGkSa} , of transit times of link $SiGkSa$ as follows:

$$F_{SiGkSa}(\Delta T') = \sum \rho_{SiGkSa}(\Delta T \leq \Delta T') \quad (17)$$

To estimate the times Δt_{SiGkSa} when a user will transit link $SiGkSa$, we compute the median of $F_{SiGkSa}()$; the expression of Δt_{SiGkSa} is defined as follows:

$$\Delta t_{SiGkSa} = F_{SiGkSa}^{-1}(0.5) \quad (18)$$

Thus, based on CDF of transit times of each link $SiGkSa$ of current road segment S_i , we obtain its transit time table of S_i . Using the example shown in Fig. 4, we compute the transit time table of S_i (see Table II).

AHTES computes a table (Table II) of transit times for each link on a road segment; a transit time, in the table, corresponds to the predicted transit time for all users who will transit the same link of the road segment. To compute the transit time table of a road segment S_i , AHTES calculates $f = w * h$ transit times where w is the number of sub-road segments and h is the number of adjacent road segments to S_i ; f is also the number of links of road segment S_i ; thus, adding $O(f)$ complexity for one road segment; for g road segments, AHTES adds $O(f * g)$ complexity. Finally, based on the transit time tables of all road segments of the navigation area, we estimate the time when a group of users will perform handoffs along their movements paths to their destinations; i.e., time when a group of users will transit the handoff points (i.e., the intersection of a road and the border of a cell); it is worth noting that the path from current location to a handoff point is a path portion from current sub-road segment to the sub-road segment where the handoff point

is located; thus, estimating the time when a group of users will perform a specific handoff consists of computing the time to transit the path portion to reach the handoff point of interest. To compute the transit time of a path portion, we sum the transit times of links and the road segment portion which compose that path portion. For better understanding, let S_1, S_3, S_4 , and S_5 be the road segments to reach the handoff point hp_1 , S_1G_3 the current sub-road segment (i.e., sub-road segment where the estimation starts), and S_4G_4 the sub-road segment where the handoff point hp_1 is located. We obtain the estimated time t_i when the group of users will reach the handoff point hp_1 as follows:

$$t_i = t_0 + \Delta t_{S1G3S3} + \Delta t_{S3G1S4} + (\Delta t_{S4G1S5} - \Delta t_{S4G4S5}) \quad (19)$$

where t_0 denotes the current time (i.e., time when the estimation starts); this allows aggregate computing of transit times for users who will transit on this same path portion (i.e., S_1, S_3, S_4 , and S_5) before the next update of transit time tables of S_1, S_3, S_4 , and S_5 .

IV. PERFORMANCE EVALUATION

In this Section, we evaluate, via simulations, the performance of APPM, AHTES and IPMBRF.

A. Mobility Schemes Performance Evaluation

We evaluate the performance of APPM and AHTES using two parameters: accuracy A_p and computational complexity C_t . A_p of APPM is defined as follows:

$$A_p(E_{act}, E_{pred}) = \frac{2 * |E_{act} \cap E_{pred}|}{|E_{act}| + |E_{pred}|} \quad (20)$$

where E_{act} is the actual set of transited road segments and E_{pred} is the set of the predicted set of road segments to be transited; E_{act} and E_{pred} are computed during simulation time d_t . A_p of AHTES is defined as follows:

$$A_p = 1 - \left(\frac{\sum_{i=1}^{N_U} E_i}{N_U} \right) \quad (21)$$

where N_U and E_i denote the number of users and the average handoff time prediction error gap (i.e., difference between real and predicted handoff time instants) of user i respectively; we assume that the average handoff time prediction error gap of 300 seconds (i.e., 5 minutes) represents an accuracy of 0%. C_t is defined as follows:

$$C_t = \frac{\sum_{i=1}^{N_U} T_i}{N_U} \quad (22)$$

where T_i denotes the prediction/estimation computation time of user i and the number of users respectively.

We compare the performance of APPM against PPM described in [3]. Indeed, APPM proposes an aggregate path prediction model while PPM proposes an individual path prediction model. Simulation results are averaged over multiple

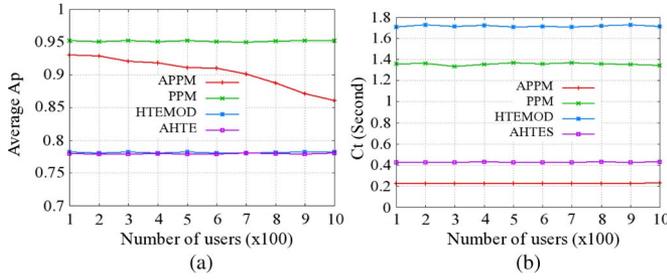


Fig. 5. The performance of mobility schemes.

runs; indeed, the simulation program is run *five hundred* times; one run of the simulation program provides *ten* prediction units; a prediction unit contains a destination and a path towards this destination. For each run, we compute A_p (resp. C_t) using (20) (resp. (21)); thus, to obtain the simulation results shown in Fig. 5 we compute the average of the *five hundred* runs.

1) *Simulation Setup*: To evaluate mobility prediction schemes (APPM versus PPM and AHTES versus HTEMOD), we use real mobile user traces (GPS trajectories), acquired from the Microsoft Research Asia laboratory's database available in the context of the GeoLife project [44]. A GPS trajectory of this dataset is represented by a sequence of time-stamped points, each of which contains the information of latitude, longitude, altitude, date and time. This dataset contains 17,621 trajectories with a total distance of about 1.2 million kilometers and a total duration of 48 000 hours; we use the 2,160 last hours (i.e., three months) of total duration for simulations; the two first months are used for the learning phase and the last month for prediction phase. This dataset recorded a broad range of users' outdoor movements, including not only life routines like going home or to work but also some entertainments and sports activities, such as shopping, sightseeing, dining, hiking, and cycling. By converting GPS coordinates to Cartesian coordinates, we identify the road segments in the map. The users' velocities are computed based on the GPS traces and the traffic light is not taking into account. We assume that the dense zones (DSZs) refer to the blocks of cells where more than 22 users/cell are located during specific times of the day. The values of the simulation parameters used by DPM [2] are: $f_{th} = 1/30$; $\theta_{th1} = 90^\circ$ and $\theta_{th2} = 45^\circ$. APPM may require a large number of mobility information to be collected and processed by a UE. However, the new generation UEs have sufficient storage space; for example, in our simulation the file (PLT file) to be maintained/used by a user is about 2.82 megabytes (for two months GPS trace collection). Recent mobile devices (e.g., Samsung galaxy) can use .XML or .TXT files (instead of database management system) that do not require large storage space. Indeed, for a mobile device of 16 gigabytes of storage space, APPM will use 0.002% of this storage space which is negligible. In this set of simulations, we use a Samsung galaxy note 2 with 1.6 GHz, 2 GB of RAM and 32 GB user memory.

2) *Results Analysis*: Fig. 5(a) shows that PPM slightly outperforms APPM while AHTES and HTEMOD provide the same average accuracy of 0.78 per user; indeed, PPM provides an average accuracy of 0.95 per user while APPM provides an average accuracy of 0.90 per user; thus, the average relative improvement (defined as [average A_p of PPM—average A_p

of APPM]) of PPM compared to APPM is about 5%. We observe that the average accuracy of APPM decreases when the number of users increases; this is expected since when the number of users increases, the number of road profiles increases and thus the number of assignation errors of road profiles to users increases; when a user is assigned the wrong road profile, his predicted path is also wrong and the prediction accuracy decreases. We also observe that the average accuracy of PPM remains constant even when the number of users increases; this is attributable to the fact that PPM performs an individual path prediction; thus, the accuracy is not impacted by the number of users. Fig. 5(b) shows that APPM (resp. AHTES) outperforms PPM (resp. HTEMOD). APPM (resp. AHTES) provides an average time complexity of 0.22 (resp. 0.42) second per user while PPM (resp. HTEMOD) provides an average time complexity of 1.35 (resp. 1.72) seconds per user; thus, the average relative improvement of APPM (resp. AHTES) compared to PPM (resp. HTEMOD) is about 1.13 (resp. 1.30) seconds per user. We conclude that, compared to PPM, APPM provides a reduction of 85.19% in time complexity and a slight decrease of 5% in accuracy; the 5% accuracy decrease is a very small price to pay for the small time complexity.

B. IPMBRF Performance Evaluation

In this sub-section, we evaluate the performance of IPMBRF in terms of new call blocking rate and handoff call dropping rate for different available bandwidth estimation errors. We define three parameters to evaluate the performance of IPMBRF: new call blocking rate, handoff call dropping rate and available bandwidth estimation error. The new call blocking rate, denoted by R_b , is computed as follows:

$$R_b = \frac{n_b}{m_b} \quad (23)$$

where n_b is the number of blocked new call requests and m_b is the total number of new call requests (i.e., accepted and blocked). The handoff call dropping rate, denoted by R_d , is computed as follows:

$$R_d = \frac{n_d}{m_d} \quad (24)$$

where n_d is the number of handoff calls dropped and $m_d = m_b - n_b$ is the number of accepted call requests. The available bandwidth estimation error, denoted by E_{bw} , is computed as follows:

$$E_{bw} = \frac{|BW_a - BW_e|}{BW_a} \quad (25)$$

where BW_a is the actual available bandwidth and BW_e is the estimated available bandwidth; BW_a and BW_e are measured whenever a call is blocked or dropped.

We compare the performance of IPMBRF against the schemes described in [19] and [18], referred to as AP1 and AP2, respectively. We selected AP1 and AP2 because they are aggregate and predictive mobile-oriented schemes; AP1 admission control procedures (resp. AP2) is limited to the source and next cells (resp. only to the source cell) while IPMBRF takes

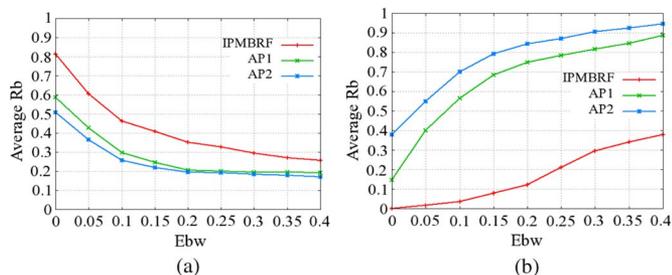


Fig. 6. Rb and Rd versus Ebw.

into account the cells along the path to destination. Simulation results are averaged over multiple runs; indeed, the simulation program is run *one thousand* times. For each run, we compute R_b (resp. R_d/E_{bw}) using Equation (22) (resp. 23/24); thus, to obtain the simulation results shown in Fig. 6 we compute the average of the *one thousand* runs.

1) *Simulation Setup*: To evaluate IPMBRF, we used mobile user traces acquired from the Generic Mobility Simulation Framework (GMSF) project [43]; 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). 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 specific time t , on a particular location (X and Y) of road segment $RSID$ in cell CID . The simulation environment is a two-dimensional environment; the roads are arranged in a mesh shape [26]; based on the results of our mobility prediction schemes (see Fig. 5(a)), each user has a predicted path with 95% of accuracy; indeed, at the intersection of two road segments, a user selects to continue straight, turn left or turn right according to his predefined path with 95% of accuracy; the cellular network is composed of 81 cells (i.e., a 9×9 mesh) and each cell's diameter is about 900 m [24]; it can typically be seen in a metropolitan downtown area.

Similar to [17], [20], [24], [26], new call requests are generated according to a Poisson distribution with rate λ (calls/second/user) and the minimum bandwidth granularity that may be allocated to any call is 1 *bandwidth unit* (BU) [8], [13], [17], [24]; in the simulations, we focus on the improvement of handoff call dropping rate; thus, similar to [6], [15], [17], [20], [24], [26], 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 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 [45]. The call time is assumed to be exponentially distributed with a mean of 300 sec. The values of the parameters used in the simulations are: $l = 400$ m, $d_1 = 0.025$ user/m; $d_2 = 0.075$ user/m; $D = 17.36$ users/km²; number of users = 1500, maximum speed = 14 m/s; BW_{req} is chosen from within the set $\{1, 2, 3, 4\}$ BUs with equal probability; the call arrival rate λ is 0.03 call/second/user; and the cell capacity is 100BUs.

2) *Results Analysis*: Fig. 6 shows (a) the average new call blocking rate and (b) the average handoff call dropping rate for different available bandwidth estimation errors. Fig. 6(a) shows that AP1 and AP2 outperform IPMBRF. Indeed, AP2 (slightly more efficient than AP1 in this scenario) provides an average call blocking rate of 0.25 per 5% of E_{bw} while IPMBRF provides an average call blocking rate of 0.42 per 5% of E_{bw} ; thus, the average relative improvement (defined as [average R_b of IPMBRF - average R_b of AP2]) of AP2 compared to IPMBRF is about 17% per 5% of E_{bw} . We observe that, for the three schemes, the average new call blocking rate decreases when E_{bw} increases. This is expected since when E_{bw} increases, the gap between the actual available bandwidth and estimated available bandwidth increases; thus, the number of successful/accepted new call requests increases and the new call blocking rate decreases. Fig. 6(b) shows that IPMBRF outperforms AP1 and AP2; IPMBRF provides an average handoff call dropping rate of 0.16 per 5% of E_{bw} while AP1 (slightly more efficient than AP2 in this scenario) provides an average handoff call dropping rate of 0.65 per 5% of E_{bw} ; overall, the average relative improvement (defined as [average R_d of AP1 - average R_d of IPMBRF]) of IPMBRF compared to AP1 is about 49% per 5% of E_{bw} . We observe that, for the three schemes, the average handoff call dropping rate increases with E_{bw} . This is expected since when E_{bw} increases, the available bandwidth seems to be enough and the number of successful/accepted new call requests increases; thus the number of handoff calls accommodated in a next cell decreases and thus the handoff call dropping rate increases. At 0% of E_{bw} , IPMBRF provides an average handoff call dropping rate of 0 while AP1 provides an average handoff call dropping rate of 0.15. This can be explained by the fact that IPMBRF makes passive reservation (in advance) along the user path to destination before the acceptance of the call. Even though AP1 uses mobility prediction, its prediction is limited to the next cell and the handoff calls can be dropped after the next cell; nonetheless, AP1 slightly outperforms AP2 in this scenario (Fig. 6(b)) because its new call admission control procedure takes into account the source and next cells while AP2 new call admission control procedure is limited to the source cell.

We conclude that, compared to AP1 and AP2, IPMBRF provides a considerable reduction of 49% per 5% of E_{bw} in handoff call dropping rate and an increase of 17% per 5% of E_{bw} in new call blocking rate. The 17% per 5% of E_{bw} new call blocking rate increase is a very small price to pay for the small handoff call dropping rate.

V. CONCLUSION

In this paper, a new aggregate and predictive mobile-oriented bandwidth reservation scheme for multimedia cellular networks is proposed. In order to strike the appropriate performance balance between handoff call dropping and new call blocking rates to ultimately support QoS-sensitive multimedia services, our proposed approach manages bandwidth by suitably combining various control techniques-bandwidth reservation and call admission; our call admission procedure makes its decision, on whether to accept or reject a new call request, based on predicted available bandwidth in each cell along the path

to destination. To make our approach scalable (with the number of users), we also proposed an aggregate path prediction model APPM (resp. aggregate handoff times estimation scheme AHTES) that estimates path to destination (resp. handoff times) for a group of users (not only for a single user). Therefore, it has low complexity, making our integrated framework practical for real mobile networks. We compared the performance of our scheme with two closely related schemes [18], [19]. Performance evaluation results did show that our scheme maintains a well-balanced network performance between bandwidth utilization, handoff dropping and new call blocking rates while other schemes cannot offer such an attractive performance balance; indeed, our scheme achieves considerably better handoff call dropping rate with slight new call blocking rate increase and efficient bandwidth utilization rate irrespective of cells capacities and call arrival rates.

As future work, we plan to work on a real-life implementation of IPMBRF in emerging mobile networks. More specifically, we envision integrating APPM and AHTES in: (a) orientation of the antennas of the base stations in cellular networks; indeed, based on the users mobility prediction thanks to APPM and AHTES, we can identify future dense zones according to the time of the day; then, we can program the directions of the antennas for different times of the day; (b) the discovery of a best gateway in vehicular networks; in vehicular networks, the gateway is a vehicle that forwards the data from vehicular network to a outside infrastructure (e.g., a cellular network) making use of a road side base station; this communication is called vehicle-to-infrastructure (V2I) communication; thanks to APPM and AHTES, we can obtain, in advance, the location of vehicles; then, we can identify a best gateway according to the time and the location of the road side base stations; and (c) small cell connection time estimation in LTE-Advanced; indeed, users mobility estimation provided by APPM and AHTES may help making decision about performing handoff or not into small cells.

REFERENCES

- [1] 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.
- [2] 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.
- [3] 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.
- [4] 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.
- [5] A. Nadembega, A. Hafid, and T. Taleb, "A framework for mobility prediction and high bandwidth utilization to support mobile multimedia streaming," in *Proc. IEEE Int. Conf. ANTS*, Chennai, Tamil Nadu, India, Dec. 2013, pp. 1–6.
- [6] Y. Jun, 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.
- [7] 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.
- [8] G. I. Tsiropoulos *et al.*, "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.
- [9] S. Al Khanjari, B. Arafeh, K. Day, and N. Alzeidi, "An adaptive bandwidth borrowing-based Call Admission Control scheme for multi-class service wireless cellular networks," in *Proc. Int. Conf. IIT*, Abu Dhabi, United Arab Emirates, Apr. 2011, pp. 375–380.
- [10] J. D. Mallapur, S. Abidhusain, S. S. Vastrad, and A. C. Katageri, *Fuzzy Based Bandwidth Management for Wireless Multimedia Networks Information Processing and Management*, vol. 70, V. V. Das, R. Vijayakumar, N. C. Debnath, J. Stephen, N. Meghanathan, S. Sankaranarayanan, P. M. Thankachan, F. L. Gaol, and N. Thankachan, Eds. Berlin, Germany: Springer-Verlag, Apr. 2010, pp. 81–90.
- [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] 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.
- [13] S. Kim, "Cellular network bandwidth management scheme by using nash bargaining solution," *IET Commun.*, vol. 5, no. 3, pp. 371–380, Feb. 2011.
- [14] 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.
- [15] 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. 109685–1–109685–10, Jan. 2012.
- [16] K. Sungwook and P. K. Varshney, "An integrated adaptive bandwidth-management framework for QoS-sensitive multimedia cellular networks," *IEEE Trans. Veh. Technol.*, vol. 53, no. 3, pp. 835–846, May 2004.
- [17] K. L. Dias, D. F. H. Sadok, S. F. L. 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.
- [18] S. Rashad, M. Kantardzic, and A. Kumar, "User mobility oriented predictive call admission control and resource reservation for next-generation mobile networks," *J. Parallel Distrib. Comput.*, vol. 66, no. 7, pp. 971–988, Jul. 2006.
- [19] C.-F. Wu, L.-T. Lee, H.-Y. Chang, and D.-F. Tao, "A novel call admission control policy using mobility prediction and throttle mechanism for supporting QoS in wireless cellular networks," *J. Control Sci. Eng.*, vol. 2011, pp. 21–31, Jan. 2011.
- [20] 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.
- [21] L. Shufeng, D. Grace, Y. Liu, J. Wei, and D. Ma, "Overlap area assisted call admission control scheme for communications system," *IEEE Trans. Aerosp. Electron. Syst.*, 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. SEISCON*, Chennai, Tamil Nadu, 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. 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.
- [25] S. N. Ahmed and B. Ferri, "Prediction based bandwidth reservation," in *Proc. IEEE CDC*, Atlanta, GA, USA, Dec. 2010, pp. 5302–5307.
- [26] 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.
- [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] 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.
- [30] A. Sgora and D. Vergados, "Handoff prioritization and decision schemes in wireless cellular networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 11, no. 4, pp. 57–77, 2009.
- [31] U. Javed *et al.*, "Predicting handoffs in 3G networks," *SIGOPS Oper. Syst. Rev.*, vol. 45, no. 3, pp. 65–70, Dec. 2011.

- [32] T. Taleb, A. Hafid, and A. Nadembega, "Mobility-aware streaming rate recommendation system," in *Proc. IEEE GLOBECOM*, Houston, TX, USA, Dec. 2011, pp. 1–5.
- [33] C.-Y. Wang, H.-Y. Huang, and R.-H. Hwang, "Mobility management in ubiquitous environments," *Pers. Ubiquitous Comput.*, vol. 15, no. 3, pp. 235–251, Mar. 2011.
- [34] U. Rathnayake, H. Petander, M. Ott, and A. Seneviratne, "EMUNE: Architecture for mobile data transfer scheduling with network availability predictions," *Mobile Netw. Appl.*, vol. 17, no. 2, pp. 216–233, Apr. 2012.
- [35] P. S. Prasad and P. Agrawal, "A generic framework for mobility prediction and resource utilization in wireless networks," in *Proc. 2nd Int. Conf. COMSNETS*, Bangalore, India, Jan. 2010, pp. 1–10.
- [36] H. Chenn-Jung *et al.*, "A probabilistic mobility prediction based resource management scheme for WiMAX femtocells," in *Proc. ICMTMA*, Changsha City, China, Mar. 2010, pp. 295–300.
- [37] F. Yu, V.W.S. Wong, and V.C.M. Leung, "Performance enhancement of combining QoS provisioning and location management in wireless cellular networks," *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, pp. 943–953, May 2005.
- [38] I. F. Akyildiz and W. Wang, "The predictive user mobility profile framework for wireless multimedia networks," *IEEE/ACM Trans. Netw.*, vol. 12, no. 6, pp. 1021–1035, Dec. 2004.
- [39] 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.
- [40] H. Jeung, M. L. 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.
- [41] K. Playtoni Meetei, G. R. Kadambi, B. N. Shobha, and A. George, "Design and development of a handoff management system in LTE networks using predictive modelling," *SASTECH*, vol. 8, no. 2, pp. 71–78, Sep. 2009.
- [42] 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.
- [43] R. Baumann, F. Legendre, and P. Sommer, "Generic mobility simulation framework (GMSF)," presented at the Proceedings 1st ACM SIGMOBILE Workshop Mobility Models, Hong Kong, Hong Kong, May 2008.
- [44] Y. Zheng, Y. Chen, Q. Li, X. Xie, and W.-Y. Ma, "Understanding transportation modes based on GPS data for web applications," *ACM Trans. Web*, vol. 4, no. 1, pp. 1–36, Jan. 2010.
- [45] J. Silvestre-Blanes, L. Almeida, R. Marau, and P. Pedreiras, "Online QoS management for multimedia real-time transmission in industrial networks," *IEEE Trans. Ind. Electron.*, vol. 58, no. 3, pp. 1061–1071, Mar. 2011.
- [46] N. Vallina-Rodriguez and J. Crowcroft, "Energy management techniques in modern mobile handsets," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 1, pp. 179–198, 2013.
- [47] 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.
- [48] C. Shi, V. Lakafosis, M. H. Ammar, and E. W. Zegura, "Serendipity: Enabling remote computing among intermittently connected mobile devices," in *Proc. 13th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, Hilton Head, SC, USA, Jun. 2012, pp. 145–154.



Abdelhakim Hafid is a Full Professor with the University of Montreal, where he founded the Network Research Laboratory (NRL), in 2005. He is also a Research Fellow with CIRRELT (Interuniversity Research Center on Enterprise Networks, Logistics and Transportation). He supervised to the graduation of more than two dozen graduate students. He published more than 170 journal and conference papers; he also holds three U.S. patents. Prior to joining the University of Montreal, he spent several years, as a Senior Research Scientist with Telcordia Technologies (formerly Bell Communications Research), NJ, USA, 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), Canada; a Research Director with Advance Communication Engineering Center (venture established by UWO, Bell Canada, and Bay Networks), Canada; a Researcher with CRIM, Canada, a Visiting Scientist with GMD-Fokus, Berlin, Germany; and a Visiting Professor with the University of Evry, France. He has extensive academic and industrial research experience in the area of the management of next-generation networks, including wireless and optical networks, QoS 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 sciences from the Graduate School of Information Sciences (GSIS), Tohoku University, Sendai, Japan, in 2001, 2003, and 2005, respectively.

He is a faculty member at the School of Engineering, Aalto University, Finland. He has been working as a Senior Researcher and 3GPP Standardization Expert with NEC Europe Ltd., Heidelberg, Germany. He was then leading the NEC Europe Labs Team

working on R&D projects on carrier cloud platforms. Prior to his work at NEC and until March 2009, he worked as an Assistant Professor with the Graduate School of Information Sciences, Tohoku University. From October 2005 until March 2006, he was working as a Research Fellow with the Intelligent Cosmos Research Institute, Sendai. His research interests lie in the field of architectural enhancements to mobile core networks (particularly 3GPP's), mobile cloud networking, mobile multimedia streaming, congestion control protocols, handoff and mobility management, inter-vehicular communications, and social media networking.

Dr. Taleb 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. He is a Board Member of the IEEE Communications Society Standardization Program Development Board. As an attempt to bridge the gap between academia and industry, he has founded and has been the General Chair of the "IEEE Workshop on Telecommunications Standards: from Research to Standards", a successful event that got awarded "best workshop award" by the IEEE Communication Society (ComSoC). He is on the Editorial Board of the IEEE Wireless Communications Magazine, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE Communications Surveys and Tutorials, and a number of Wiley journals. He is serving as Vice-Chair of the Wireless Communications Technical Committee, the largest in IEEE ComSoC. He also served as a Secretary and then as a Vice Chair of the Satellite and Space Communications Technical Committee of IEEE ComSoC (2006–2010). He has been on the Technical Program Committee of different IEEE conferences, including Globecom, ICC, and WCNC, and chaired some of their symposia. He is an IEEE Communications Society (ComSoc) Distinguished Lecturer. He is the recipient of the 2009 IEEE ComSoc Asia-Pacific Best Young Researcher award (June 2009), the 2008 TELECOM System Technology Award from the Telecommunications Advancement Foundation (March 2008), the 2007 Funai Foundation Science Promotion Award (April 2007), the 2006 IEEE Computer Society Japan Chapter Young Author Award (Dec. 2006), the Niwa Yasujiro Memorial Award (February 2005), and the Young Researcher's Encouragement Award from the Japan chapter of the IEEE Vehicular Technology Society (VTS) (October 2003). Some of his research work has been also awarded best paper awards at prestigious conferences.



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. thesis is 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 Burkina Faso Public Administration Staff Management Office. He is currently a member of the Network Research Laboratory, University of Montreal. His research interests lie in the field of handoff and mobility management, architectural enhancements to mobile core networks, mobile multimedia streaming, call admission control, bandwidth management, and mobile cloud computing.